Responses to RC1

Dear Reviewer #1:

Thanks very much for your time on reviewing our manuscript. We sincerely thank the reviewer for your efforts on the reviewing of our manuscript. We deeply appreciate your valuable comments on our manuscript, and we have carefully revised the manuscript according to the comments. The point-by-point responses to your comments are provided in this document.

Best regards,
Zhenwei Zhang
Nanjing University of Information Science & Technology

The manuscript "GHRSAT: the first global hourly dataset of all-sky remotely sensed estimates of surface air temperature" developed a hybrid method integrate random forest models and kriging techniques to estimate all-weather air temperature, and generate global all-weather air temperature products from 2011-2023. The proposed method improved the model accuracy compared to traditional RF algorithm. The content structure of the preprint is clear, the method is innovative, and the topic is meaningful. My comments and questions for clarification can be found below.

Comments:

1. Line 120: The author uses zoning modeling and mentions that the zoning basis also refers to station density. I am confused that the model building effect depends largely on the representativeness of the sample. Is it reasonable to use low-density sites for modeling? In addition, will zoning modeling lead to boundary effects between regions? Why not build a global unified model ?

Responses #1: Thanks for your valuable comments. As you mentioned, representativeness of samples severely influences model building processes. Additionally, quantity and quality of the samples also impact model building. As stated in Lines 190-191 of our original manuscript, we obtained about 0.9 billion matched samples at ground stations for the global land study areas in the 2011–2023 period, and the average number of samples for each calendar month is more than 5.9 million. **It is computationally infeasible to build a global unified model using all matched samples.** Given the large numbers of samples across the global land areas in the long-term period, we chose to develop estimation models separately for eight different task regions **not only considering the computation issue, but also based on the following concerns**: locally developed models have high adaptability to specific local land areas. As an analogy, the reason why GWR (geographically weighted regression) has been widely adopted

for various modeling tasks in the broad fields of geosciences is that GWR exhibits high modeling capacity than global linear regression models. In essence, the GWR method is implemented by building a linear model for each sample point using the nearby samples within the local neighborhood of the point. The study for estimating long-term SAT in China (Yao R. et al., 2020) conducted comparative experiments indicating the advantage of locally modelling strategies in terms of prediction performance. For the modeling tasks involved in our study, as the global land areas cover a wide variety of geographical settings and backgrounds, a set of locally developed models targeted for each specific region will have high adaptability. **Therefore, it could be a good choice to adopt a locally (both time and geographic space) modeling strategy when building models for large-scale areas with huge number of samples.** Similarly, the study by Zhang T. et al. (2022) on estimating daily SAT for global land areas also developed models separately for different areas.

In fact, the locally or zoning modeling strategy try to build region-specific estimating models with high adaptability to each region. The boundary effects of the modelling strategy are inevitable as all modelling tasks involved in not only our work, but also other research fields have to be targeted for a predefined or manually selected study area.

We really appreciate your concern about the reasonability of modeling for regions with low-density sites. As you mentioned, the samples from low-density sites will severely impact the sampling representativeness. Generally, models developed using samples from study areas with low-density sites will have large errors and estimation uncertainties (see the reported performance metrics for regions of TR-1 and TR-6 in Figure 4), which is primarily due to severely inadequate sampling representativeness. Despite the issue for the areas with limited sites, such as polar areas and high-altitude areas, it is worth exploring to develop models for these regions, although the models have large errors and uncertainties. There are studies have been conducted for building SAT models for the areas with very limited stations (Nielsen E. B., et al, 2023; Meyer H., et al, 2016). **We admit that at present, building models for these areas with satisfactory performance is very challenging**, considering the limited stations installed in these areas. With more autonomous ground stations installed in the harsh polar and high-altitude areas, the estimation of SAT for these areas will be gradually improved. **Lastly, high-quality observations from ground stations are the basis for various research fields, and the scarcity of ground stations is the common issue for these fields**. The studies conducted for these fields can only be based on the available ground sites. For example, the studies on the estimation of surface radiation for large-scale areas all depend on very limited number of ground flux sites. Recently, the study on estimating surface long-wave downward radiation (Zeng Q. et al, 2024) published in the ISPRS journal utilized all available 51 ground flux sites across the central Asia. Thus, to advance the studies in the broad geoscientific community, it is very important for governments and research agencies to further improve ground observation networks, especially for polar and high-altitude areas.

Finally, we thank the valuable comments again. We have stated the reasons for adopting the locally model building strategy (Lines 122-133) and discussed the issue of sampling representativeness for building SAT estimation models for regions with limited stations (Lines 360-371).

Yao, R., Wang, L., Huang, X., Li, L., Sun, J., Wu, X., and Jiang, W.: Developing a temporally accurate air temperature dataset for Mainland China, Sci. Total Environ., 706, 136037, https://doi.org/10.1016/j.scitotenv.2019.136037, 2020.

Zhang, T., Zhou, Y., Zhao, K., Zhu, Z., Chen, G., Hu, J., and Wang, L.: A global dataset of daily maximum and minimum near-surface air temperature at 1 km resolution over land (2003–2020), Earth Syst. Sci. Data, 14, 5637–5649, https://doi.org/10.5194/essd-14-5637-2022, 2022.

Nielsen, E. B., Katurji, M., Zawar-Reza, P., and Meyer, H.: Antarctic daily mesoscale air temperature dataset derived from MODIS land and ice surface temperature, Sci Data, 10, 833, https://doi.org/10.1038/s41597-023-02720-z, 2023.

Meyer, H., Katurji, M., Appelhans, T., Müller, M., Nauss, T., Roudier, P., and Zawar-Reza, P.: Mapping Daily Air Temperature for Antarctica Based on MODIS LST, Remote Sens., 8, 732, https://doi.org/10.3390/rs8090732, 2016.

Zeng, Q., Cheng, J., Sun, H., and Dong, S.: An integrated framework for estimating the hourly all-time cloudy-sky surface long-wave downward radiation for Fengyun-4A/AGRI, Remote Sensing of Environment, 312, 114319, https://doi.org/10.1016/j.rse.2024.114319, 2024.

2. Line 158: Air temperature is related to many factors. The variables input into the model in this article are only LST, NDVI, latitude and longitude, elevation and hour of a day. What is the basis for selecting these variables? Among them, only LST and hour of a day change over hours. Is the result mostly dependent on LST? Please show the feature importance of the models.

Responses #2: Thanks for your valuable comments. The primary fundamental of selecting the input variables for modeling SAT is based on their connection to SAT, and specifically considers whether incorporating the variables into SAT estimation models will contribute the predictive performance of the models. As our study aimed at building estimation models for global land areas, it is inevitable to only consider the variables for which datasets are available at the global scale in the time period 2011-2023.

There are some differences in the selection of variables for SAT estimation among previous studies, which **is primary due to the localized consideration of modeling SAT for specific study areas and the constraints of data availabilit**y. For examples, previous studies have developed SAT estimation models considering variables for satellite-based snow cover (Wang W. et al., 2025) and surface structural properties derived from lidar data (Venter Z. S., et al., 2020). However, the models utilizing these variables are only restricted to the study areas that the studies focused on, and cannot be generalized to other regions due to data unavailability for these variables in other regions.

The auxiliary variables used in our study have been widely used in previous studies for building SAT estimation models. More importantly, data for the auxiliary variables used in our study are available at the global scale, and can be easily and publicly accessed online. The

hourly LST is the core input for our SAT estimation models, and as you pointed out, these models are primarily dependent on LST. **We have provided a figure for illustrating the feature importance of the input variables for the SAT models** (see Figure S3 in the revised supplement file), and we have rewritten some parts of Sec. 2.3 to more clearly state our consideration for selecting the auxiliary variables for modeling hourly SAT in our study (see Lines 155-166).

With more earth observation satellites operating at high temporal revisit cycles for large-scale areas in the future, we expect that the hourly estimation of SAT will be significantly improved by more available high-temporal satellite-based data for land surface properties relevant for modeling SAT.

Wang, W., Brönnimann, S., Zhou, J., Li, S., and Wang, Z.: Near-surface air temperature estimation for areas with sparse observations based on transfer learning, ISPRS Journal of Photogrammetry and Remote Sensing, 220, 712–727, https://doi.org/10.1016/j.isprsjprs.2025.01.021, 2025.

Venter, Z. S., Brousse, O., Esau, I., and Meier, F.: Hyperlocal mapping of urban air temperature using remote sensing and crowdsourced weather data, Remote Sens. Environ., 242, 111791, https://doi.org/10.1016/j.rse.2020.111791, 2020.

3. Line 255： In the validation part, the samples were randomly divided into ten parts, one of which was used to validate the model, which means that the training samples may include all sites, and there is no completely independent site for validation. What is the prediction accuracy of this method in non-site areas?

Responses #3: Thanks for your important questions. According to your comments, we have **additionally performed site-based cross-validation (CV) for all models developed in our study**. In the site-based cross-validation, the sites were first randomly divided into ten sets, and samples from the sites in each set are treated as one fold of samples (see Lines 289-294, in Sec. 3.3). For site-based CV, models are validated by completely independent sites. Overall, the averaged performance of the hybrid models for different regions ranges from 1.87 °C to 2.62 °C under site-based CV. The validation results for sited-based CV have been discussed in our revised manuscript (Lines 26-29 in the Abstract, Lines 332-359 in Sec. 4.1, and Lines 694-699 in Sec. 6). Fig. 4 has been revised to contain the overall validation results for our models under both sample-based and site-based cross-validation (Lines 380-385), and some additional figures have been added in our revised supplement file (Fig. S4, S5, S6).

4. Line 289: The author developed 156 models for each region. The temporal variation of air temperature has certain regularities, and data from the same period in different years may provide effective information. Why does the author establish a separate model for each month in 2011-2023?

Responses #4: Thanks for your important comments. The comments are related to the locally

(both time and geographical areas) modelling strategy adopted in our study, which has been discussed in our Responses #1. We obtained about 0.9 billion matched samples at ground stations for the global land study areas in the 2011–2023 period. In addition to the consideration of computational efficiency for model training, developing models built for each month will have high adaptability to the month.

5. Section 4.2: The spatial validation in the preprint is based on the station scale, which cannot reflect the continuity of the generated product and the estimated effect of non-site areas. Please further prove it at the spatial scale.

Responses #5: Thanks for your valuable comments. We really appreciate the comments. SAT is estimated under the assumption that the model fitted using in situ samples generalizes to other areas (pixels) without ground samples, which is the basis for all studies in the field of NSAT estimation. It is hard to reflect the continuity of the generated product in term of estimation errors. In other words, as there are no abundant and very high-density ground stations available at the global scale, it is the rare case that each ground pixel contains several ground stations. Therefore, the estimation errors for non-site areas can only be represented by the cross-validation results for the SAT models. We do think that spatially quantification of estimation errors across the areas is important. The cross-validation results for SAT models can be analyzed at the site-level to indirectly exhibit how the models will perform across different areas. For example, see Fig. 7 from Kilibarda et al. (2014) and Fig. 6 from Yao et al. (2023).

Kilibarda, M., Hengl, T., Heuvelink, G. B. M., Gräler, B., Pebesma, E., Perčec Tadić, M., and Bajat, B.: Spatio-temporal interpolation of daily temperatures for global land areas at 1 km resolution, J. Geophys. Res.: Atmos., 119, 2294–2313, https://doi.org/10.1002/2013JD020803, 2014.

Yao, R., Wang, L., Huang, X., Cao, Q., Wei, J., He, P., Wang, S., and Wang, L.: Global seamless and high-resolution temperature dataset (GSHTD), 2001–2020, Remote Sens. Environ., 286, 113422, https://doi.org/10.1016/j.rse.2022.113422, 2023.

6. The preprint lacks data cross-validation. For example, the air temperature estimated from geostationary satellites or reanalysis data all have hourly air temperatures. Please compare the with the published air temperature data or methods.

Responses #6: We very appreciate your comments. To offer a further validation of our models, we have additionally performed site-based cross-validation (CV) for all models developed in our study. For site-based CV, the models are assessed by completely independent sites. The validation results for sited-based CV have been discussed in our revised manuscript (Lines 26-29 in the Abstract, Lines 332-359 in Sec. 4.1, Lines 380-385, and Lines 694-699 in Sec. 6).

We agree with you that the comparison of our dataset with other SAT datasets (reanalysis or SAT from geostationary) is very important. However, there are no common, independent and

adequate station data held out for validating the estimated SAT developed by different researchers using different methodological approaches. That is, if we hold out some parts of ground stations for validating our estimated SAT data, **the validation of the estimated SAT data developed by previous studies using the same hold-out stations will not be objective and independent**, because it is unknow if the hold-out stations had been used to train the models for the estimated SAT data in the previous studies.

Reanalysis data have high temporal resolutions, but have relatively coarse spatial resolutions. More importantly, reanalysis data (such as ERA5, GLDAS) are generated by numerical models with assimilation of various observational data sources, such as ground-based meteorological observation, satellite data and sounding data from radiosondes. The organizations (for example, ECMWF, NOAA, NASA GMAO) for developing reanalysis data and corresponding assimilation systems have not disclosed the specific information on the ground station data assimilated into their systems. Thus, validation of reanalysis using ground stations **will not be independent, and has the risk of over-estimation of the accuracy of reanalysis data**. Although it is unknown as to the specific information on the assimilated ground station by reanalysis data, **it could be inferred that publicly available observational datasets for ground stations are very likely be assimilated, because it is easy and free to access these public observational ground datasets**.