Responses to RC3

Dear Reviewer #3:

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

Comments #1. Imbalance in Model Description: The description of the model in the methods section is unbalanced, with much more emphasis on Kriging compared to Random Forest (RF).

Responses #1: Thanks for your valuable comments on our manuscript. We have revised Sec. 3.2 to supplement some information about the advantages of RF, and more importantly restructured the section to more clearly describe how the RF models are integrated with two kriging techniques to meet the tasks of constructing hourly SAT estimation models in our study (Lines 234-237). In addition, there are two types of residual kriging techniques (OK and FRK) have been integrated into the SAT estimation models, which is described in Sec. 3.2. Therefore, the section contains more contents relevant to the two kriging methods than RF. But the content length for each kriging method is roughly the same as that for RF.

Comments #2. Handling Missing Data: The paper mentions removing records with poor quality in the ground station data but does not explain in detail how missing data is handled, especially missing data with different temporal and spatial resolutions. If there is a significant amount of missing data, the model's accuracy and generalizability may be impacted

Responses #2: Thanks for your important comments. Thanks for your important comments. There are complete and detailed official data documentation for the ISD observational dataset (see https://www.ncei.noaa.gov/data/global-hourly/doc/). For modeling SAT in our study, we used the station records from ISD without any quality-control issues, specifically the records passed all quality control checks (see page 10 of the isd-format-document.pdf), which has been pointed out in our revised manuscript (Line 206 and Lines 208-210). In fact, there is only a slight portion of records with missing SAT data or with QA issues. Even without any missing records, the current ground observational data is inadequate for use in the studies on SAT estimation and other applications. In general, there are high-coverage and high-density ground observational networks in developed countries and regions, such as America and

Europe, in contrast, the poor regions (for example, Africa) and polar areas have very limited coverage of ground stations. Estimation models for SAT will be more representative with higher accuracy when training using samples from high-coverage and high-quality networks of stations. However, there will be a long road ahead to establish such station networks, especially for undeveloped regions. The importance of ground station data for building SAT estimation models and the limitations of our models by ground station data have been discussed in Sec. 4.5 in our revised manuscript (Line 644-657).

Comments #3. Selection of Covariates: The paper uses multiple spatial covariates such as NDVI, elevation, latitude/longitude, and hour of the day, but it does not provide a detailed discussion of the rationale behind selecting these covariates or their applicability in different regions.

Responses #3: Thanks for your valuable comments. The primary fundamental of selecting the input variables for SAT estimation models is by considering whether incorporating the covariates 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 covariates for which datasets are available at the global scale** in the time period 2011-2023. There are some differences in the selection of covariates 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 availability**. For examples, previous studies have developed SAT estimation models considering covariates 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 covariates 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 covariates in other regions.

We should note that there are very limited input auxiliary covariates that available at the global scale for use in modeling hourly SAT. 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. In our exploratory experiments for building SAT models based on RF for different task regions using the selected covariates, we find that models with all the selected covariates exhibited better validation performance than models considering only subsets of the covariates. To more clearly state our consideration for selecting the auxiliary variables for modeling hourly SAT in our study, we have rewritten some parts of Sec. 2.3 (see Lines 155-166).

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.

Comments #4. Resampling of NDVI and Elevation Data: The resampling method for NDVI and elevation data is not clearly stated, which could affect data quality.

Responses #4: Thanks for your important suggestions. We agree with you that the resampling of various satellite-based and geoscientific data is very important and has impacts on data quality. We used the elliptical weighted averaging (EWA) method, which is stated Lines 201-202 of our revised manuscript. The EWA method is widely used for regridding MODIS data. More information about the method could be refereed to https://pyresample.readthedocs.io

Comments #5. Limited Model Performance Evaluation: The paper only uses RMSE and MAE as performance metrics, lacking analysis of systemic bias (Bias) or coefficient of determination (R^2), which makes it difficult to fully assess the model's performance.

Responses #5: Thanks for your important comments. We agree with you that the diverse metrics will help to fully assess the estimation models used for reconstructing the GHRSAT datasets. In addition to the RMSE for our validated models, we have provided the comparison of our models in terms of the performance metrics including Bias, MAE and coefficient of determination (see the revised supplement for our manuscript). We have additionally performed site-based cross-validation (CV) for all models developed in our study (see Lines 289-294, Sec. 3.3). 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 385-390).

Comments #6. Discussion on Practical Application: The discussion section could benefit from further elaboration on how the research results could be applied to real-world problems. Additionally, the limitations of the current study should be clearly stated, along with potential directions for future improvements.

Responses #6: Thanks for your valuable comments. In our revised manuscript, we have added Sec. 4.5 (Lines 643-657) to clearly discuss the limitations, practical applications and potential improvements of our study. We discussed the limitations of our study in modeling of hourly SAT from the aspects of methodology and data sources. The potential applications of our reconstructions for the field of remotely sensed estimation, and applications of our reconstructed GHRSAT dataset have been briefly discussed.