General comments:

Liang et al. describe the development of a global 1km data product for land surface – air temperature differences and sensible heat fluxes. In principle such data-driven products are very valuable for the community. However, I have several methodological concerns, primarily related to the validation approach and with respect to variable selection as input to the models.

Re: Thank you very much for your comments. All comments have been well addressed one by one and provided in the follows.

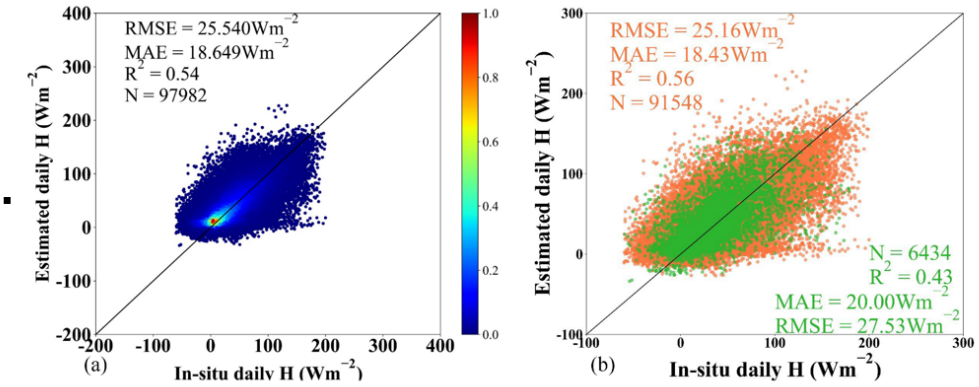
Major points

Point 1:

The cross-validation strategy chosen by the authors is not adequate and yields overoptimistic results. It is absolutely compulsory to stratify train and test data by sites and not (only) by time. This is simply because data from one site are not independent and the objective of the study is to estimate at unmeasured locations. This needs to be done correctly.

Response1: Thank you for this valuable comment. There are seventeen sites in the independent validation set that were not included in the training data. The model's performance on these sites is acceptable, with an RMSE of 27.53 Wm^{-2}, a MAE of 20 Wm^{-2} and an R^2 of 0.43. For clarify, we have revised Fig. 8 and the corresponding text have been already added in lines 516-522 in Section 4.2 in the revised manuscript:

> Fig.8a. The overall validation accuracy was deemed satisfactory, with an RMSE of 25.54 Wm^{-2}, MAE of 18.649 Wm^{-2} and R^2 of 0.54. To evaluate the model's ability to predict data from sites not included in the training set, we split all independent validation sample into those originating from sites used in training and those from "unseen" sites. The corresponding accuracies are presented in Fig. 8b, where the green scatter points represent samples from sites not included in the training process. The model performed reasonably well on the "unseen" sites, with an RMSE of 27.53 Wm^{-2}, a MAE of 20 Wm^{-2} and an R^2 of 0.43. These results are only slightly lower than those obtained for the sites included in the training set, which yielded an RMSE of 25.16 Wm^{-2}, MAE of 18.43 Wm^{-2} and R^2 of 0.56. Furthermore, the spatial



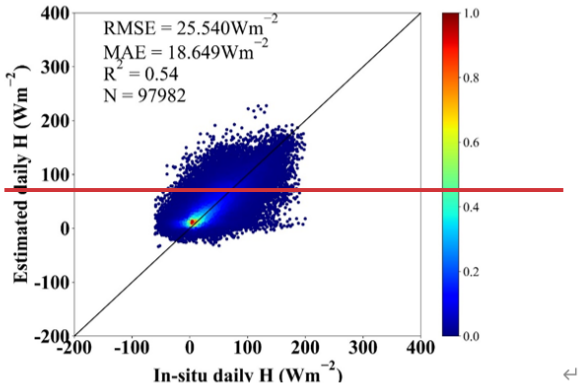


Figure 8. The ~~overall validation~~ accuracy of the estimated H based on (a) all ~~all~~ independent in-situ validation samples, (b) samples from sites used in training (orange scatter points) and from "unseen" sites (green scatter points). The values were obtained by replacing the results for areas with missing ABD in mod1 with those from mod2.

Moreover, H exhibits clear seasonal variations throughout the year. After stringent quality control, the daily in-situ measurements of H show substantial data gaps. To ensure the temporal continuity and completeness of the training data used in the LSTM model, we selected monthly datasets with less than 10% missing values for the training set. The remaining data were allocated to an independent validation set. The corresponding text was in lines 194-198 as "Due to significant gaps in the daily in-situ measurements of H after stringent quality control, a distinct strategy was implemented to segregate the samples for H and Tsa. For H, the methodology involved selecting monthly datasets with fewer than 10% missing values for the training set, while the rest were allocated to an independent validation set for evaluating model performance.". To maintain the representativeness of the independent validation samples, all subsequent analyses and conclusions are based on the entire independent validation set.

Point 2:

The authors chose Rn and ET as input to the model to predict H. In my opinion this is hard to justify as H=Rn-LE-G and predicting ET is a similar problem as predicting H. I would find it

conceptually more appealing if input variables are close to observations and not already derived products with additional layers of uncertainty.

Response2: Thank you for your thoughtful comment. We understand the concern regarding the use of derived variables such as Rn and ET as predictors for H, especially given the physical relationship H = Rn − LE − G.

Previous studies have shown that the variability of H is also influence by aerodynamic factors such as aerodynamic resistance which derived by wind speed. The corresponding text in lines 82-83 as "H estimation has traditionally relied on temperature-derived one-source and two-source models, incorporating ground-based observations of temperature and wind fields." and lines 89-90 as "Both models face common challenges in calculating aerodynamic resistance (r_{ah}) due to the complexities of Monin-Obukhov similarity theory (Monin and Obukhov, 1954; Brutsaert, 2013)". However, such variables are often unavailable at the spatial resolution required for this study. In preliminary experiments, we tested the inclusion of wind speed from the MERRA-2 reanalysis dataset. However, due to its coarse spatial resolution and relatively large uncertainties, incorporating wind speed resulted in reduced model performance. In contrast, using Rn and ET—although both are derived products—provided more spatially consistent information and led to better model performance in our experiment. These variables effectively integrate various surface and atmospheric processes, thus offering informative signals for estimating H at regional to global scales. The corresponding text have been added for clarify in lines 332-336 in Section 3.2.1 in the revised manuscript:

> FVC) and six radiation-related parameters (Tsa, DLW, Rn, ABD, DSR, and ET). Note that aerodynamic factors such as aerodynamic resistance, which is primarily derived from wind speed, were not included in this study. Preliminary experiments using wind speed from reanalysis datasets showed that its inclusion
> 335 decreased model accuracy, likely due to the coarse spatial resolution and associated uncertainties of the data. The significant correlations among these parameters are well-established; for instance, DSR is

In future work, we plan to incorporate higher-resolution observational datasets or improved reanalysis products to further enhance the physical interpretability and robustness of the model.

Point 3:

The authors chose slope and aspect as predictors. While it is clear that these variables are very relevant in principle, the footprint of flux towers is supposed to be restricted to reasonably flat terrain. Therefore, I cannot imagine that robust patterns wrt these terrain variables can be learned.

Response3: Thank you for your insightful comment. We agree that most flux towers are installed on relatively flat terrain to ensure the validity of flux measurements. However, due to the 1 km spatial resolution of our input data, the complex surrounding terrain within and beyond the flux footprint may still affect the surface energy and temperature dynamics in the target area, even when the tower itself is situated on relatively flat ground. Moreover, as shown in Fig. 20, slope and aspect demonstrate a strong contribution to the model, highlighting their potential relevance even for towers located in predominantly flat areas. The corresponding text have been already added in lines 762-764 in Section 5 for clarity in the revised manuscript:

760 terrain and radiation-related variables are integral to accurately estimating Tsa. Notably, slope and elevation were more critical than the other terrain-related variable, aspect, which accounted for 7.37%. Although flux towers are generally installed on relatively flat terrain to ensure measurement accuracy, the surrounding complex terrain within and beyond the flux footprint can still influence local surface energy and temperature dynamics near the flux towers. Similarly, LST and DSR proved to be more

Point 4:

The authors also chose day of year as predictor, which has no direct environmental meaning. I suggest to drop this or replace by e.g. potential radiation or sun angle.

Response4: Thank you for your valuable suggestion. We agree that *doy* does not represent a direct physical quantity; however, Tsa exhibits a seasonal cycle, and *doy* serves as an effective temporal indicator that helps the RF model capture this variation. To evaluate its contribution, we conducted experiment and found that removing *doy* from the RF model resulted in a noticeable decline in performance (RMSE increased from 1.459 K to 1.51 K, MAE from 1.071 K to 1.115 K, and R^2 dropped from 0.53 to 0.50). We also tested replacing *doy* with the solar height angle, which yielded comparable accuracy (RMSE = 1.454 K, MAE = 1.072 K, R^2 = 0.53), indicating that *doy* and solar geometry-related variables provide similar predictive value. Regarding your suggestion to use potential radiation, we appreciate the insight. In our current model, we have already included radiation-related parameters such as downward shortwave radiation (DSR) and downward longwave radiation (DLW), which provide more direct and dynamic representations of surface energy input. These results support the inclusion of *doy* as a simple but informative feature in the Tsa estimation. Additionally, the feature importance ranking in Fig. 20 shows that while *doy* ranks relatively low, it still contributes to the model's performance. The corresponding text has been added to the revised manuscript (lines 765–768) for clarification.

impactful than DLW, which held a contribution of 4.98%. In addition, the *doy* ranked as the second least important variable in our analysis. Although it does not represent a direct physical environmental variable, our experiments demonstrated that it serves as a simple yet informative seasonal indicator that helps the RF model capture temporal variations effectively.

Point 5:

The authors mentioned a 'circular' approach between training and testing for hyper-parameter tuning (line 300). It is absolutely forbidden to use test data for any kind of model tuning. Perhaps this is a misunderstanding. Please clarify.

Response5: Thank you for pointing this out. We apologize for the misleading wording. In our workflow, the training dataset was internally partitioned into subsets for model training and hyperparameter tuning. An independent validation set was reserved exclusively for evaluating model performance and was never involved in the training or tuning process. To avoid any misunderstanding, we have revised the term "test phase" to "parameter tuning" (line 317). In addition, we carefully reviewed the entire manuscript to eliminate similar ambiguities, and the description in Section 2.1 (lines 197–207) has been revised accordingly as follows:

10% missing values for the training set, while the rest were allocated to an independent validation set for evaluating model performance. Linear interpolation was employed to impute missing values within the training set, ensuring the integrity of the monthly datasets. A five-fold cross-validation was then applied~~,~~ by partitioning the data such that 80% of the months were used ~~designated~~ for training and the remaining 20% for ~~testing~~ tuning the model parameters ~~during~~ in each iteration. This process yielded a training set encompassing 121,542 daily H samples and an independent validation set containing 97,982 samples. In contrast, the Tsa analysis designated measurements from 2018 to 2019 as the independent validation set for model evaluation, with data from preceding years allocated to the training set. Specifically, for each site, 70% of the samples from 2000 to 2017 ~~samples~~ were randomly selected for ~~the~~ training ~~set~~, and the remaining 30% were used for ~~testing~~ tuning the model parameters. As a result, the Tsa training set included 564,918 daily samples, and the independent validation set comprised 84,977 daily Tsa samples.

Point 6:

The authors use the Twine et al approach to correct flux tower based sensible heat fluxes by forcing energy balance closure. This is a critical assumption, which needs through discussion because the uncertainty related to energy balance correction is very large, esp. for H (see

Mauder et al 2024, AFM)

Response6: Thank you for your valuable comment. We agree that correcting for energy balance closure (EBC) using the method of Twine et al. (2000) introduces uncertainty, especially for sensible heat flux. We have added a clarifying sentence in lines 188-191 in the Section 2.1 to acknowledge the assumptions and potential uncertainties associated with this correction:

> Where H_{cor} is corrected H; LE_{uncor} and H_{uncor} are uncorrected LE and H, respectively. It should be noted that this correction method relies on assumptions about the distribution of residual energy, which may still have uncertainties into the corrected flux values. These uncertainties and their potential impacts are further discussed in the Discussion section of this paper.↵

Additionally, we have expanded the Discussion section in lines 823-835 to include recent insights from Mauder et al. (2024) to discuss the uncertainties of the correct method.

> Furthermore, as the H in-situ measurements used as ground truth values in this study have undergone energy balance closure (EBC) correction, their reliability warrants thorough discussion. In this study, we adopted the widely used method proposed by Twine et al. (2000), which redistributes the residual energy between sensible and latent heat fluxes in proportion to their original magnitudes. Although this approach has been implemented in many large-scale studies and provides a practical solution when additional constraints are lacking, recent research has underscored its limitations. Notably, Mauder et al. (2024) highlighted that EBC remains a persistent issue in FLUXNET data, with a global average energy balance ratio of approximately 0.82, and identified unresolved processes such as mesoscale secondary circulations and unmeasured energy storage terms as major contributors to the energy gap. These uncertainties are particularly relevant for H, and their effects can propagate into downstream analyses and model training. Although the Twine method does not resolve these underlying physical mechanisms, it remains a necessary and pragmatic compromise for enabling the use of flux tower data in surface energy balance studies. ↵

Minor points:

Point 7:

I find the uncertainty estimates listed in Table 1 and referenced in the text a bit misleading as they are not comparable among the products because they were not calculated consistently

Response7: Thank you for this comment. We have added the corresponding caption in Table1 for clarify as:

▪ **Table 1.** The mainstream global product information. Note that the uncertainty estimates, coming from different sources (e.g., documentation and publications), serve only as general references and should not be directly compared between products.↵

and text in lines70-73 as:

generally provide long temporal coverage but tend to have coarse spatial resolution and exhibit varying levels of~~significant~~ uncertainty, as illustrated in Table 1. Notably, the uncertainty estimates were derived through different sources (including original documentation and associated publications), and should therefore be considered as approximate references rather than being directly comparable across products. ~~Even~~For instance, FLUXCOM_RS, ~~–~~ as the most recent and only satellite product ~~boasting~~ with the highest spatial resolution of 0.0833°, exhibits ~~encounters~~ a reported global uncertainty of 11.61% over

Additionally, the corresponding context has been modified accordingly to ensure better coherence between statements.

Point 8:

Model tree ensembles for FLUXCOM mentioned in table 1 is likely wrong as I suppose the authors used the ensemble product

Response8: Thank you for your helpful comment. The FLUXCOM RS and RS+METEO products were generated using nine and three machine learning methods, respectively. Therefore, we have revised the term “model tree ensembles” to “multi-model ensemble” in the manuscript to more accurately reflect the methodology used in Table1.

Point 9:

Line 113: sentence starting with “Therefore” seems incomplete

Response9: Thank you for pointing this out. We have revised the sentences to improve its grammatical accuracy and clarity in lines 116-120. The corresponding text have been revised as “Traditional physically-based models for estimating H are typically developed for specific areas and land surface conditions, and often require parameters that are not easily accessible (e.g. aerodynamic resistance to heat transfer, rah). As a result, these models tend to produce large uncertainties when applied to other areas. Therefore, a convenient and widely applicable method for estimating global H values is still lacking.”

Point 10:

The choice of LSTM for estimating H is unclear – have not seen a clear comparison to RF

and the other methods (did I miss this?)

Response10: Thank you for your comment. The choice of LSTM for estimating H was motivated by the limited availability of observations for H and the need to capture its temporal dependencies. This rationale is now clearly stated in lines 125–129 of the revised manuscript:

> improving the accuracy and spatial resolution of Tsa and H on a global scale. Considering the different characteristics of the target variables, we adopted two ML models tailored. Specifically, RF was used for Tsa estimation due to the availability of dense in-situ measurements and its robust performance in such scenarios, whereas LSTM was applied for H estimation to better handle the limited data samples and capture temporal dependencies. ~~Given the intricate interactions between Tsa and other land-atmosphere parameters, along with the significant temporal variations of H identified through our analysis, we utilized two machine learning methods, Random Forest (RF) and long short-term memory (LSTM), to predict Tsa and H, respectively.~~ Initially, we employed the RF method, utilizing pertinent parameters

Moreover, a detailed comparison of LSTM with other methods, including RF, DBN, and Transformer, is presented in Section 4.2 (lines 536–566).

Point 11:

Are the comparisons of global H values in section 5 based on exactly the same spatial domain. This matters as e.g. FLUXCOM does not cover deserts where H is particularly large.

Response11: Thanks for this valuable comment. Indeed, the spatial domains and temporal periods vary across the cited studies. While these differences preclude direct quantitative comparison, we have revised the text to explicitly specify each study's spatial coverage and time periods. The corresponding modifications in lines 740-747 are as follows:

> land surface H to be 35.29±0.71 Wm^{-2} over the 2000–2020 period~~.~~. This value is higher than the 27 W m^{-2} based on global land data from 2000 to 2004 reported by Trenberth et al. ~~surpassing the previously reported estimates of 27 Wm^{-2} by Trenberth et al.~~ (2009), and also exceeds the 32 Wm^{-2} estimated by Jung et al. (2019)~~.~~, which excluded barren regions, deserts, permanent snow or ice, and water bodies for the period 2000–2013. It is more consistent with the ~~ and aligning closely with the~~ 36-40 Wm^{-2} range reported by Siemann et al.(2018) for global land areas between 1984 and 2007. These figures are provided for general context, as differences in spatial coverage and temporal periods across studies limit direct comparability. Despite these advancements, certain aspects still require discussion, particularly