

Author's response

Ms. Ref. No.: hess-2022-125

Revised title: Accuracy of five ground heat flux empirical simulation methods in the surface energy balance-based remote sensing evapotranspiration models

Author(s): *Zhaofei Liu*

It would be greatly appreciated for your kind reviewing to this paper. Thanks very much for your valuable comments and suggestion. For your convenience to re-review the paper, the response corresponding to your comments are described in detail as follows:

Responses to Editor' Comments

Comments to the author:

Dear author,

On the basis of the review comments and your responses, you are invited to submit a revision for further consideration.

Best wishes

Bob Su

Reply:

The manuscript has been revised according to reviewer's comments. Please find the point-by-point reply to the comments as follows.

Responses to Reviewer #1 Comments

Remote sensing-based surface energy balance typically requires G simulation to close the surface energy balance, which is often a challenge given that G could not be easily sensed from the surface. Hence, most remote sensing-based ET models use an empirical approach to scale G between the two extreme limits of % or fraction of G/R_n within the open surface and full canopy. This % or fraction G/R_n is characterized by vegetation and remotely sensed indices like NDVI, LAI, albedo, LST, etc using simple empirically derived values. This paper aims to study the spatiotemporal variations of this empirical relationship (G , R_n , H) and evaluate some of the remote sensing-based empirical methods using half-hourly global flux observations data. While, I think it is important to improve remote sensing approaches to simulate G , as it will also improve remote sensing and surface energy balance-based ET models, the results and discussion, as presented in the paper, are a little challenging to follow with not many insights into how G simulation in remote sensing-based ET models could be improved. So, I have some major issues (and some minor issues) that the author needs to consider before the paper can be reevaluated.

Major Comments

The paper is more focused on the assessment of R_n and G relationships than the evaluation of simulated G within the existing remote sensing-based ET models. So I wonder if this should be reflected in the title of the paper, which suggests that the paper is focused on the evaluation of the existing methods. Note that the empirical nature of G simulations and their uncertainty in remote sensing-based ET models is a well-known issue. So while the optimization of regression coefficients (e.g., those in the LC methods) is nice, the finding that the coefficients differ across different parts of the world is obvious. What is more important is to present an idea about how this empiricism and uncertainty can be reduced and a globally applicable model could be developed. The paper falls short on this part.

Reply: Yes. This study is focused on the evaluation of the existing methods. The title has been revised to “Accuracy of five ground heat flux empirical simulation methods in the surface energy balance-based remote sensing evapotranspiration models” to make it more clear. As mentioned in Line 103-106 (in the revised manuscript), “This study addresses four key objectives: (1) investigating the temporal and spatial

variations and common characteristics of the empirical relationship between G and R_n ; (2) evaluating the accuracy of five empirical methods in simulating half-hourly G from R_n ; and (3) investigating the performance of five methods at different times during the intra-day and the spatial distribution of simulation accuracy at global flux observation sites.” However, it is out of the scope of this study to present an idea about how this empiricism and uncertainty can be reduced and a globally applicable model could be developed. These are issues that model users and developers need to consider more. Results of this paper can provide some references for RS ET data users and the remote sensing evapotranspiration modelers. For example, the applications of RS ET data sets need more caution in tropical regions, and further improvement of G simulations at low-latitude areas and noon periods are recommended for RS ET modelers.

The paper acknowledges the limitation of existing G derivation methods in remote sensing-based ET models but does not consider some of the widely used approaches, such as the one used in the SEBAL model, which would require albedo and LST. The author acknowledged that SEBAL based G method was found to be working better than other approaches in another study (Saadi et al. 2018). The G models evaluated in this paper are very similar in nature. Hence, it is important to incorporate G models with different structures/inputs. Note that obtaining albedo and LST for these sites is as easy as obtaining NDVI. The author should have incorporated some additional G derivation methods used in the common remote sensing-based ET model.

Reply: The author had used LST data at regional scales, while were not familiar with albedo data. The used LST data is the Terra Moderate Resolution Imaging Spectroradiometer (MODIS) MOD11A1 product, which is produced daily LST at a spatial resolution of 1 km. The evaluation in this study was based on daily data series. For daily series of the global LST dataset, the author only found that the MODIS MOD11 dataset was available. The MOD11B product provides daily per pixel Land Surface Temperature and Emissivity (LST&E) in a 1,200 by 1,200 kilometer (km) tile with a pixel size of 5,600 meters (m). There are hundreds of files for each day in the dataset (MOD11A and MOD11B) covering the global land. As for 230 flux sites used in this study, each site is needed to be corresponded to these hundreds of files. Huge amounts of data need to be downloaded, i.e. hundreds of files per day multiplied by number of days in the observed daily series of flux sites. In fact, Saadi et al. (2018)

only evaluated the methods at a single observed site. The NDVI dataset used in this study is a single file per day with global coverage, and the workload is relatively acceptable. In addition, the authors had tried several methods to download MODIS product but without success at the beginning of this study. It was also failed to download these products in the past few days. This work is beyond the author's capacity. Therefore, the methods embedded with LST data were not evaluated in this study.

In Line 485-486 (the revised manuscript), a new sentence “Evaluation of such methods embedded with the LST data is recommended for further research where data is available.” was added to complement the discussion of this issue.

The author mentioned that observed G is taken as the residual of the energy balance to evaluate different G models, assuming that all other components are perfectly derived. While the author acknowledges this in section 4.1 (Line 364-365), I think still problematic because no attempt has been made to address this issue. Here, there is no information on how the energy balance was closed (or was not unclosed) or corrected. The observed G used in this paper and all error metrics presented hence could be highly biased and uncertain.

Reply: It assumes that the measurements of R_n , H and LE are accurate in this study. These measurements might have some errors. However, it is not considered in this study. To the author's knowledge, the eddy covariance measurements of H and LE are generally considered to be the most accurate observations available.

As described in the second paragraph (Line 39-44), ground heat flux (G) is the soil heat flux at the surface. It is difficult to observe directly, due to technical limitations (Wang and Bou-Zeid, 2012; Gao et al., 2017). Soil heat flux (referred to as G') is generally measured using heat flux plates near the surface (within a few millimeters of the surface) in the flux tower observation sites. There were numerous studies investigated on the surface energy balance closure issue at flux sites. The observed G' instead of G was generally used to investigate the energy balance ratio (Wilson et al., 2002). However, the difference between G' and G could be 50% because of the soil heat storage within the layer from the surface to the flux plate (Heusinkveld, 2004; Yue et al., 2011; Wu et al., 2020). A large error is produced if the soil heat storage is ignored in the G calculation (Meyers and Hollinger, 2004; Lu et al., 2018). The energy balance closure problem might be largely caused by the soil heat storage

(Foken, 2008).

Theoretically, surface energy is balanced. The energy unclosure might be mainly caused by the error of the observed data. Compared with G, other energy terms can be observed more accurately. Therefore, the surface energy balance method was used as references in this study. As mentioned in the section of Discussion (Line 362-365), “The eddy covariance measurements of H and LE are generally considered to be the most accurate observations available. The **Eq. (1)** makes full use of the surface energy term that can be accurately measured at present. In other words, it assumes that the measurements of R_n, H and LE are accurate in this study. The uncertainties of measurements are not considered in this study.”

The residual of the surface energy balance method has been validated by an experimental site in the West of Spain (van der Tol, 2012).

References

- Foken, T.: The energy balance closure problem: An overview, *Ecol. Appl.*, 18, 1351–1367, <https://doi.org/10.1890/06-0922.1>, 2008.
- Gao, Z., Russell, E. S., Missik, J. E. C., Huang, M., Chen, X., Strickland, C. E., Clayton, R., Arntzen, E., Ma, Y., and Liu, H.: A novel approach to evaluate soil heat flux calculation: An analytical review of nine methods, *J. Geophys. Res. Atmos.*, 122, 6934–6949, doi:10.1002/2017JD027160, 2017.
- Heusinkveld, B. G., Jacobs, A. F. G., Holtslag, A. A. M., and Berkowicz, S. M.: Surface energy balance closure in an arid region: role of soil heat flux, *Agric. For. Meteorol.*, 122, 21–37, doi:10.1016/j.agrformet.2003.09.005, 2004.
- Lu, S., Wang, H., Meng, P., Zhang, J., and Zhang, X.: Determination of soil ground heat flux through heat pulse and plate methods: Effects of subsurface latent heat on surface energy balance closure, *Agric. For. Meteorol.*, 260–261, 176–182, doi:10.1016/j.agrformet.2018.06.008, 2018.
- Meyers, T. P., and Hollinger, S. E.: An assessment of storage terms in the surface energy balance of maize and soybean, *Agric. For. Meteorol.*, 125, 105–115, doi:10.1016/j.agrformet.2004.03.001, 2004.
- van der Tol, C.: Validation of remote sensing of bare soil ground heat flux, *Remote Sens. Environ.*, 121, 275–286, doi:10.1016/j.rse.2012.02.009, 2012
- Wang, Z. H., and Bou-Zeid E.: A novel approach for the estimation of soil ground heat flux, *Agric. For. Meteorol.*, 154-155, 214–221, doi:10.1016/j.agrformet.2011.12.001, 2012.

Wilson, K., Goldstein, A., Falge, E., Aubinet, M., Baldocchi, D., Berbigier, P., Bernhofer, C., Ceulemans, R., Dolman, H., Field, C., Grelle, A., Ibrom, A., Law, B. E., Kowalski, A., Meyers, T., Moncrieff, J., Monson, R., Oechel, W., Tenhunen, J., Valentini, R., and Verma, S.: Energy balance closure at FLUXNET sites, *Agr. Forest Meteorol.*, 113, 223–243, [https://doi.org/10.1016/S0168-1923\(02\)00109-0](https://doi.org/10.1016/S0168-1923(02)00109-0), 2002.

Wu, B., Oncley, S. P., Yuan, H., and Chen, F.: Ground heat flux determination based on near-surface soil hydro-thermodynamics, *J. Hydrol.*, 591, 125578, doi:10.1016/j.jhydrol.2020.125578, 2020.

Yue, P., Zhang, Q., Niu, S., Cheng, H., and Wang, X.: Effects of the soil heat flux estimates on surface energy balance closure over a semi-arid grassland, *Acta Meteorol. Sin.*, 25, 774–782. doi:10.1007/s13351-011-0608-4, 2011.

Note that typically in a remote sensing-based ET model, R_n is calculated using radiation balance using remote sensing and meteorological inputs, and G is estimated as a fraction of R_n . Hence, the uncertainty in R_n calculation is also a source of error in G . In some cases when G may be biased available energy ($R_n - G$, where R_n is coming from remote sensing-based radiation balance) may be reasonable. In this paper, the author used observed R_n in calibrating G , so when you compare coefficients, the uncertainty in R_n (even better when remote sensing-based R_n is used) needs to be mentioned too. Given that R_n is the key input used in all G methods considered in this study, additional assumptions (assuming that R_n is perfectly simulated by the remote sensing-based ET model) and uncertainties need to be discussed.

Reply: Thank you very much for your valuable comments. Yes, observed R_n was used for calibrating G in this study. It was assumed that R_n is perfectly simulated by the remote sensing-based ET models. Several sentences have been added in the Discussion section (Line 497-500 in the revised manuscript) to describe this issue, as follows, “In RS ET models, R_n is generally calculated using radiation balance with RS images and meteorological inputs. However, observed R_n was used for simulating G in this study. In other words, it was assumed that R_n is accurately simulated by the RS ET models. Therefore, it should be noted that the uncertainty in R_n calculation was also a source of error in G simulations in ET models.”

It is not clear how the coefficients of the LC methods are calibrated in this study. Are these just the regression coefficients or other optimization methods used? Was any calibration/validation approach used (using independent sets of data)?

Reply: The coefficients of the LC methods are calibrated by the NSE. The author realizes that there are many multi-objective parametric calibration methods. But these methods are too time-cost to be achieved for hundreds of sites. A new sentence “The parameters of these methods were calibrated by the Nash-Sutcliffe efficiency (NSE) at each observation site.” is added in Line 156-157 (the revised manuscript).

A new sentence “At each site, daily series of each half-hour were divided into two parts: the first 80% of the data were used for parameter calibration and the rest were used for validation.” is added in Line 155-156 (the revised manuscript) to make calibration/validation clear. In addition, the author tried to test robustness of the methods at some sites. Daily series were randomly assigned to one of two datasets: 80% were assigned to the calibration dataset and 20% to the validation dataset. The process of random assignment was repeated to generate 100 independent datasets. Results showed that these methods are robust. The author has added these results in Supplementary Materials (Figure S2).

In Line 320-324 (the revised manuscript), several new sentences “Daily series were randomly assigned to one of two datasets: 80% were assigned to the calibration dataset and 20% to the validation dataset. The process of random assignment was repeated to generate 100 independent datasets. Results showed that there was little difference between the performance of the models in calibration and validation datasets. It indicated that these methods are robust. The performance of the LC_NDVI_P in calibration and validation datasets at some sites could be found in **Fig. S2.**” have also been added in revised manuscript.

I am surprised why the author did not test the actual LC methods (i.e., the original coefficients) used in different ET models considered in this study. In addition, it is important to mention how these different ET models come up with different empirical coefficients.

Reply: This issue had been discussed in the first and second paragraphs of the Section 4.2. As mentioned in Line 394-398, “The LC method is most commonly used in the RS ET models. The coefficients applied to each model were different. The coefficients of the LC method in the TSEB (Norman et al., 1995), ALEXI (Anderson

et al., 1997), DisALEXI (Norman et al., 2003), MOD16A2 (Mu et al., 2011), and modified TSEB (Ait Hssaine et al., 2020) ET models were 0.35, 0.31, 0.30, 0.39, and 0.37, respectively. The coefficient of the method in the GLEAM model was 0.05, 0.2 and 0.25 for the tall canopy, short vegetation and bare soil, respectively (Miralles et al., 2011).” In this study, the parameters of the LC methods were calibrated for each half-hour periods at each site. Results showed that the optimal parameter values varied significantly in different sites and half-hour periods. The author had tested some actual LC methods, and found that the original parameter values could accurately simulate G at some sites, but induced large errors in the G simulations in other regions. Therefore, it is recommended that model developers consider the spatial variations of G simulation parameters in RS ET modeling on a global scale (Line 405-407).

According to your valuable comments, the title has been revised to “Accuracy of five ground heat flux empirical simulation methods in the surface energy balance-based remote sensing evapotranspiration models”. In addition, “empirical based” has also been added in the main text. In this study, the parameters of each empirical method were calibrated for each half-hour periods at each site. According to your valuable comments, descriptions of calibration/validation have been added in Line 155-157 (the revised manuscript), as follows “At each site, daily series of each half-hour were divided into two parts: the first 80% of the data were used for parameter calibration and the rest were used for validation. The parameters of these methods were calibrated by the Nash-Sutcliffe efficiency (NSE) at each observation site.”

I find no difference between the contents in the abstract and the conclusion. Both summarize key results with no discussion on the key reasons for differences in model performances and insights into how future remote sensing-based G models can be improved. I couldn't find the main objective of the paper in the abstract.

Reply: The abstract and the conclusion have been revised to avoid repeat problem. In the third paragraph of the section 4.1, it was found that “the accuracy of the G simulation is affected by the correlation between R_n and G.” However, the other (physical) reasons for differences in model performances have not been found in this study. It might be caused by the differences in climate, soil and land cover. According to your valuable comments, evaluations of seven land cover types have been added in revised manuscript. Because the observation sites used in this study has a land cover

classification. The sites were divided into seven land cover types: Forest, Grassland, Cropland, Wetland, Shrubland, Savanna, and Other types. Figure 3, 4 and 7 (Figure 8 in the revised version) have been revised according to your valuable comments. A new Figure 7 has been added. Descriptions of these figures have also been added as follows,

Line 229-237, “In terms of seven land cover types, the intra-day performance of each land type was similar to that of all sites except the Other type (Fig. 3-c and 3-d). The correlation between G and Rn was relatively high in the sunrise and sunset periods. The correlation in Other and Wetland types is generally higher than that of other land cover types. In each period, the median R2 of all sites in the two types generally exceeded 0.60, and the highest value even exceeded 0.80. Except Other type, the difference of correlation between G and Rn in different land types is mainly reflected in the daytime period except Other type. The correlation in the Forest and Savanna types was significantly lower than that of other types during daytime, especially for Savanna sites, most of which had R2 lower than 0.5 during daytime. In Other type sites, the correlation between G and Rn in the daytime is stronger than that in the night periods. The slope value of each land cover type in the daytime is lower than that in the night. This intra-day distribution of slope was consistent with that of all sites.”

Line 267-273, “For different land cover types, the correlation between G and Rn was the strongest at Other and Wetland sites. The mean value of the median R2 of statistical sites was 0.71 and 0.67 for these two types, respectively. There was also a strong correlation between G and Rn at Cropland, Shrubland and Grassland sites. The mean value of corresponding R2 is about 0.60. There was only one site with a median R2 lower than 0.4 in each of the three land cover types. The mean value of corresponding R2 was 0.55 for Forest sites. However, the correlation was relatively weak in the Savanna type sites, and the mean value of corresponding R2 is 0.49. The weak correlation between G and Rn ($R2 < 0.4$) were mainly distributed in the Forest (6/97), Grassland (7/42) and Savanna (3/15) sites.”

Line 351-361, “Figure 7 shows the NSE simulated by each method in seven land cover types. The intra-day performance of each land cover type was similar to that of all sites except for the Other type, with the highest simulation accuracy at sunrise and sunset periods. The intra-day accuracy varied greatest at the Forest and Savanna sites. The median NSE of all sites simulated by the LC_NDVI_E method was close to 0.8 at

the sunrise periods, while the corresponding NSE was only approximately 0.4. It varied little at other land cover types, especially for Wetland and Shrubland types. The greatest and lowest values of median NSE for all sites simulated by the LC_NDVI_E method were approximately 0.7 and 0.6, respectively. The NSE of the LC, LC_NDVI_P and LC_NDVI_E methods showed a unimodal distribution in the Other type sites. The NSE was significantly higher in the daytime than at night periods. The highest value was in the morning and noon periods, with the median NSE of all sites exceeding 0.8. The model performance was significantly better than other land cover types. In the Other type sites, the LC_NDVI_E method performed better than other methods, with the median NSE higher than 0.6 in each time period.”

Line 387-399, “For different land cover types, the LC method performed better in the Cropland, Wetland and Other type sites. The mean value of median NSE of Wetland and Other sites was 0.66 and 0.69, respectively. The method was also able to accurately simulate G in the Forest, Grassland and Shrubland type sites, with the corresponding mean NSE of 0.57 or 0.56. It performed the worst at the Savanna sites, with the corresponding mean NSE was only 0.47. Since the Savanna sites are mainly distributed in tropical regions, this is consistent with the relatively poor performance of tropical region site as mentioned above. The performance of the method varied significantly in each land cover types except for the Other type sites. In the Wetland type sites, there were 3 sites in the United States with the NSE value lower than 0.3. The NSE of other 35 sites was higher than 0.50, with the highest value was close to 0.90. The Grassland sites were distributed in Asia, Europe, North America and Oceania. The NSE value was greater than 0.5 at each Grassland site in Europe. Cropland sites were distributed in Asia, Europe, and the United States. The NSE value was lower than 0.60 at 8 sites in the United States, with the mean NSE value of only 0.45. The method was able to accurately simulate G at 11 sites in Europe except for one site in Mediterranean region, with the mean NSE value of 0.74. The NSE for the two Asian sites was 0.54 and 0.71, respectively.”

This study is focused on evaluation of five ground heat flux empirical simulation methods in ET models. It only provides some references for ET modelers. For example, consider the spatial variations of G simulation parameters in RS ET modeling on a global scale, and further improvement of G simulations at low-latitude areas and noon periods are recommended.

In Line 10-11, a new sentence “The G simulation methods had been evaluated at

many individual sites, while there were relatively few multi-site evaluation studies.”
has been added to make it clear.

Minor comments:

Line 7: Instead of saying “According to 230 flux site observations” better say Based on the assessment from 230....

Reply: Thanks for your valuable comments. “According to 230 flux site observations” has been revised to “Based on the assessment from 230 flux site observations”.

Line 8-9: Based on the previous statement, it shows that G accounts for a significant proportion of the daily surface energy balance.

Reply: Yes. It used “important role” to describe this issue.

Line 19: It’s not the accuracy of the sites. It’s rather the accuracy of the models in these sites.

Reply: According to your valuable comments, this sentence has been revised to “The accuracy of the model was generally higher in Northern Hemisphere sites than in Southern Hemisphere sites.”

Line 31-42: It’s better to differentiate “ground heat flux” or “soil heat flux” by providing their physical meanings and with more detailed descriptions. The author defines soil heat flux as the heat flux measured by the flux plates near the surface.

Reply: A new sentence “Soil heat flux is the heat flux occurring in a layer of soil.” has been added in Line 43 to make it clear. Ground heat flux is the soil heat flux at the surface.

Line 74-75: Suggest citing Roerink et al., 2000 and Merlin et al., 2014 right after the corresponding model names

Reply: Thanks very much for your valuable comments. This sentence has been revised to “The solutions of G in the first two models were also applied to the Simplified Surface Energy Balance Index (S-SEBI) (Roerink et al., 2000) and Four-source Surface Energy Balance (SEB-4S) (Merlin et al., 2014) models, respectively.”

Lines 101-110: Given the numbers of towers from different networks, could you please indicate how you came up with the number “230” (i.e., 230 sites used in this study).

Reply: There were 189 FLUXNET2015 sites and 60 FLUXNET-CH4 sites were used in the analysis. There were 19 sites belonging to both FLUXNET2015 and FLUXNET-CH4. Four sites obtained from the TERN OzFlux dataset were also included in FLUXNET products. Therefore, 230 sites used in this study.

The sentence “There were 19 sites belonging to both FLUXNET2015 and FLUXNET-CH4, and flux observation data from four sites in Australia were obtained from the TERN OzFlux dataset, which was a long and continuous series up to 2019 (Beringer et al., 2016).” has been revised to “There were 19 sites belonging to both FLUXNET2015 and FLUXNET-CH4. Flux observation data from four sites in Australia were obtained from the TERN OzFlux dataset. These four sites were included in FLUXNET products, but were with a longer and continuous series up to 2019 (Beringer et al., 2016).” to avoid misunderstanding.

Line 270-272: I do not think you can say NSE is suitable but RE and KGE for evaluation. Yet, you are using RE, RMSE, and KGE for model evaluation. Maybe you need to rephrase the sentence. It is better to justify the choice of model evaluation metrics in the methods section.

Reply: Line 298 in the revised manuscript, the sentence “The evaluation of the model in this study included four criteria.” has been revised to “In this study, four criteria were tried to evaluate the model.”. In addition, in Line 160 (the revised manuscript), the sentence “The criteria used to evaluate these simulations included...” has been revised to “The criteria tried to evaluate these simulations included”.

Line 340: Please mention the optimization process in the Methods section

Reply: Descriptions of the parameter calibration have been added in Line 155-157 (the revised manuscript), as follows, “At each site, daily series of each half-hour were divided into two parts: the first 80% of the data were used for parameter calibration and the rest were used for validation. The parameters of these methods were calibrated by the Nash-Sutcliffe efficiency (NSE) at each observation site.”

Line 329: How can daily G be simulated at 6:30? Shouldn't this be G only or

half-hourly G?

Reply: The sentence “The LC method accurately simulated daily G of most sites at 6:30” has been revised to “The LC method accurately simulated G at 6:30 in most sites” (Line 375 in the revised manuscript) to make it clear. In addition, similar revisions have also been made in Line 245, 247, 256, and 378 in the revised manuscript.

Line 383: MODIS is not used at 10:30 and 13:30. MODIS data represents conditions around these times.

Reply: Yes. Thanks for your valuable comments, this sentence has been revised to “For example, MODIS data represents conditions around 10:30 and 13:30”.

Line 393-398: redundant information in the paper

Reply: Yes. These sentences “The LC method is most commonly used in the RS ET models. The coefficients applied to each model were different. The coefficients of the LC method in the TSEB (Norman et al., 1995), ALEXI (Anderson et al., 1997), DisALEXI (Norman et al., 2003), MOD16A2 (Mu et al., 2011), and modified TSEB (Ait Hssaine et al., 2020) ET models were 0.35, 0.31, 0.30, 0.39, and 0.37, respectively. The coefficient of the method in the GLEAM model was 0.05, 0.2 and 0.25 for the tall canopy, short vegetation and bare soil, respectively (Miralles et al., 2011).” have been deleted.

Line 416-417: These data are easy to get. It may not be a good idea to ignore Bastiaanssen (1995) Method when it was found to be working better than other approaches in another study (Saadi et al. 2018).

Reply: This has been explained in the reply of the second Major Comment. The author has tried hard to download LST data, but failed. In Line 485-486 (the revised manuscript), a new sentence “Evaluation of such methods embedded with the LST data is recommended for further research where data is available.” was added to discuss this issue.

Line 420: The difference among different methods was not significant because NDVI and fc are highly correlated (in fact NDVI is likely used to derive fc) and they are calibrated similarly.

Reply: Yes. The author agrees with that. But the performance of the different methods varied at some sites.

Line 430: there may be a case when a large error in G may be canceled by a large error in R_n leading to reasonable estimates of available energy ($R_n - G$), which is further partitioned into sensible and latent heat fluxes.

Reply: Yes. The author agrees with that. This sentence has been revised to “A large error in the G simulation might be induced in the ET modelling process, thereby reducing the accuracy of the ET estimates.”

Responses to Reviewer #2 Comments

This paper analyzes the relationship between G and R_n at a continental scale with hundreds of flux site measurements. This work is interesting to RS energy balance ET model users. It concluded that the linear coefficient (LC) method and the methods embedded with the normalized difference vegetation index (NDVI) were able to accurately simulate a half-hourly G series at most sites. The methods using fractional vegetation coverage showed poor performance. The highest accuracy was exhibited during sunrise periods (6:00-7:00), followed by sunset periods (17:00-18:00). The lowest accuracy was observed at noon periods (10:00-15:30). These conclusions are important for RS ET simulation. From this point, this work deserves a publication on HESS. Meanwhile, it also has some shortages which needs more clarification. The following are some comments.

Two major comments:

G was taken as the residual of R_n-H-LE in this paper, without considering the energy balance issue. This method might work for some low canopies which has a relative homogeneous land surface. The measurement of H and LE might have problem for forest site, since H and LE sensor are not high enough to be out of the sub-roughness layer on the canopy top. Hereby, this paper needs some discussion on why the energy unbalance item can be all partitioned to G , or what kind of data quality controlling process can make him/her believe that H and LE measurement at the selected sites are accurate and they don't need energy balance correction.

Reply: Thanks for your valuable comments. The observation sites used in this study has a land cover classification. The sites were divided into seven land cover types: Forest, Grassland, Cropland, Wetland, Shrubland, Savanna, and Other types. Evaluations of seven land cover types have been added in revised manuscript. The low performance in some Forest sites might be due to the fact that the H and LE sensor are not high enough to be out of the sub-roughness layer on the canopy top as you mentioned.

As described in the second paragraph of the Introduction section, “ G , which is the soil heat flux at the surface, is difficult to observe directly, due to technical limitations (Wang and Bou-Zeid, 2012; Gao et al., 2017), and direct estimation of G using RS data is not possible (Kalma et al., 2008; Allen et al., 2011; Saadi et al., 2018).”

Therefore, as discussed in the first paragraph of the Discussion section, “The eddy covariance measurements of H and LE are generally considered to be the most accurate observations available. The Eq. (1) makes full use of the surface energy term that can be accurately measured at present. In other words, it assumes that the measurements of R_n , H and LE are accurate in this study.”

Eq.2-6, the author has optimized a, a1, a2, and b. However, they did not analyze the values of these optimized variable. Figure 8 only show optimized values for three methods, without show other two methods. a1 and a2 in eq. 5 has their definition or physical meaning in the original publication. Whether the optimized values for these two parameters still follow the range of their physical meaning? I suggest to do some statistical analysis of these optimized parameter values. This can help other users when using equation 2-5. Chen et al. 2019 AFM has optimized fc based G/Rn equation. Please make a comparison with this study. They have optimized a1, a2 with a classification of land covers and canopy types. Since these parameter values could varies due to canopy covers, I suggest this paper also use canopy classification to analyze the NSE values in figure 6, KGE, RMSE, RE in figure 5, R^2 and slope in figure 3. Figure 1 can be also divided into different land covers. And, please also conclude which of the five methods is the best for which land covers or canopy classification. This result will be more useful for the RS ET model users. Figure 4, it would be interesting to analyze the linear fitting R^2 between G/Rn and NDVI for different canopy. The same problem with figure 7. Figure5, please also add Re, RMSE and KGE for other methods, not only show the LC method.

Reply: According to your valuable comments, the evaluations of seven land cover types have been added in revised manuscript. Figure 3, 4 and 7 (Figure 8 in the revised version) have been revised according to your valuable comments. A new Figure 7 has been added. Descriptions of these figures have also been added as follows,

Line 229-237, “In terms of seven land cover types, the intra-day performance of each land type was similar to that of all sites except the Other type (Fig. 3-c and 3-d). The correlation between G and R_n was relatively high in the sunrise and sunset periods. The correlation in Other and Wetland types is generally higher than that of other land cover types. In each period, the median R^2 of all sites in the two types generally exceeded 0.60, and the highest value even exceeded 0.80. Except Other type, the

difference of correlation between G and Rn in different land types is mainly reflected in the daytime period except Other type. The correlation in the Forest and Savanna types was significantly lower than that of other types during daytime, especially for Savanna sites, most of which had R2 lower than 0.5 during daytime. In Other type sites, the correlation between G and Rn in the daytime is stronger than that in the night periods. The slope value of each land cover type in the daytime is lower than that in the night. This intra-day distribution of slope was consistent with that of all sites.”

Line 267-273, “For different land cover types, the correlation between G and Rn was the strongest at Other and Wetland sites. The mean value of the median R2 of statistical sites was 0.71 and 0.67 for these two types, respectively. There was also a strong correlation between G and Rn at Cropland, Shrubland and Grassland sites. The mean value of corresponding R2 is about 0.60. There was only one site with a median R2 lower than 0.4 in each of the three land cover types. The mean value of corresponding R2 was 0.55 for Forest sites. However, the correlation was relatively weak in the Savanna type sites, and the mean value of corresponding R2 is 0.49. The weak correlation between G and Rn ($R2 < 0.4$) were mainly distributed in the Forest (6/97), Grassland (7/42) and Savanna (3/15) sites.”

Line 351-361, “Figure 7 shows the NSE simulated by each method in seven land cover types. The intra-day performance of each land cover type was similar to that of all sites except for the Other type, with the highest simulation accuracy at sunrise and sunset periods. The intra-day accuracy varied greatest at the Forest and Savanna sites. The median NSE of all sites simulated by the LC_NDVI_E method was close to 0.8 at the sunrise periods, while the corresponding NSE was only approximately 0.4. It varied little at other land cover types, especially for Wetland and Shrubland types. The greatest and lowest values of median NSE for all sites simulated by the LC_NDVI_E method were approximately 0.7 and 0.6, respectively. The NSE of the LC, LC_NDVI_P and LC_NDVI_E methods showed a unimodal distribution in the Other type sites. The NSE was significantly higher in the daytime than at night periods. The highest value was in the morning and noon periods, with the median NSE of all sites exceeding 0.8. The model performance was significantly better than other land cover types. In the Other type sites, the LC_NDVI_E method performed better than other methods, with the median NSE higher than 0.6 in each time period.”

Line 387-399, “For different land cover types, the LC method performed better in the Cropland, Wetland and Other type sites. The mean value of median NSE of Wetland and Other sites was 0.66 and 0.69, respectively. The method was also able to accurately simulate G in the Forest, Grassland and Shrubland type sites, with the corresponding mean NSE of 0.57 or 0.56. It performed the worst at the Savanna sites, with the corresponding mean NSE was only 0.47. Since the Savanna sites are mainly distributed in tropical regions, this is consistent with the relatively poor performance of tropical region site as mentioned above. The performance of the method varied significantly in each land cover types except for the Other type sites. In the Wetland type sites, there were 3 sites in the United States with the NSE value lower than 0.3. The NSE of other 35 sites was higher than 0.50, with the highest value was close to 0.90. The Grassland sites were distributed in Asia, Europe, North America and Oceania. The NSE value was greater than 0.5 at each Grassland site in Europe. Cropland sites were distributed in Asia, Europe, and the United States. The NSE value was lower than 0.60 at 8 sites in the United States, with the mean NSE value of only 0.45. The method was able to accurately simulate G at 11 sites in Europe except for one site in Mediterranean region, with the mean NSE value of 0.74. The NSE for the two Asian sites was 0.54 and 0.71, respectively.”

Line 480-481, a new sentence “This has also verified by Chen et al. (2019).” was added to make it clear.

In Figure 5, it was focus on some problems about the KGE, RMSE and RE in evaluating the model performance at different sites and time periods. The land cover results of the KGE, RMSE and RE has been added in in the Supplementary Materials (Figure S1).

Some minor comments:

Figure 6. The NSE value is calculated after or before a, a1, a2, b were optimized? The figure description should include this information.

Reply: Yes. The NSE value is calculated after the parameters were optimized. The figure title has been revised to “Figure 6: The NSE simulated by the (a) LC, (b) LC_NDVI_P, (c) LC_NDVI_E, (d) LC_fc_SE and (e) LC_fc_ST methods based on optimized parameters in each site and half-hour intervals.” to make it clear.

Figure 8, the label for y-axis is not accurate, please revise it.

Reply: Yes. The label for y-axis in Figure 8 (Figure 9 in revised version) has been revised.

Figure 1a shows that G and Rn has a time phase difference in their diurnal variation. However, this paper does not consider this effect. Please explain why not consider this effect in their using G/Rn equations.

Reply: Yes. There is a time phase difference in the diurnal variation of G and Rn. The time phase difference varied at different sites. This effect has been reduced by parameter optimization at each site and half-hour period.

These ET datasets include, but are not limited to, the Breathing Earth System Simulator (BESS) (Jiang and Ryu, 2016), Moderate Resolution Imaging Spectroradiometer (MODIS; MOD16A2) (Mu et al., 2011), GLEAM (Miralles et al., 2011), and Numerical Terradynamic Simulation Group (NTSG) (Zhang et al., 2010) products. There are more global ET products which is based on energy balance method, such as EB-ET (Chen et al. 2021), <http://data.tpdc.ac.cn/zh-hans/data/df4005fb-9449-4760-8e8a-09727df9fe36/?q=energy%20balance>. This ET product is based on energy balance method. The author may think that this study is more useful for energy balance based ET models.

Reply: This sentence has been revised to “These ET datasets include, but are not limited to, the Breathing Earth System Simulator (BESS) (Jiang and Ryu, 2016), Moderate Resolution Imaging Spectroradiometer (MODIS; MOD16A2) (Mu et al., 2011), GLEAM (Miralles et al., 2011), Numerical Terradynamic Simulation Group (NTSG) (Zhang et al., 2010) and Thermal Energy Balance (Chen et al., 2021) products.”

The surface energy balance method provides an alternative solution for assessing the G simulation schemes (van der Tol et al., 2012). This method could avoid the inconsistent spatial scale of G with that of LE and H in field measurements. I don't understand what's the meaning of these two sentences, please rephrase them.

Reply: As mentioned in Line 85-93, the gradient and calorimetry approaches had been used for evaluations of G simulations. These evaluations were limited to a single site scale because field observations of soil thermal properties were available only at a few sites. Therefore, the surface energy balance method provides an alternative

solution for assessing the G simulation schemes (van der Tol et al., 2012). And this method could avoid the inconsistent spatial scale of G with that of LE and H in field measurements.

The slope and R2 of the linear fitting curve were -0.012 and 0.92, respectively. Are you sure the slope is negative value?

Reply: Yes. As shown in Figure 2-c, the slope of the linear fitting curve for mean G/Rn of all sites in the daytime periods is -0.012.

Change “use Rn to calculate G in the RS inversion of ET” to use Rn to calculate G in RS based energy balance ET models (Chen et al. 2019 AFM; Chen et al. 2021 JGR).

Reply: “use Rn to calculate G in the RS inversion of ET” has been revised to “use Rn to calculate G in the RS based energy balance ET models”.

Some references about energy balance ET models should be cited:

Chen, X., et al. (2019). "Optimization of a remote sensing energy balance method over different canopy applied at global scale." Agricultural and Forest Meteorology 279: 107633.

Chen, X., et al. (2021). "Remote Sensing of Global Daily Evapotranspiration based on a Surface Energy Balance Method and Reanalysis Data." Journal of Geophysical Research: Atmospheres 126(16): e2020JD032873.

Chen, X., et al. (2014). "Development of a 10-year (2001–2010) 0.1° data set of land-surface energy balance for mainland China." Atmos. Chem. Phys. 14(23): 13097-13117

Reply: These references have been cited in revised manuscript.

Responses to Reviewer #3 Comments

The main target of this paper is to test several empirical formulations of the ratio between the soil heat flux G and the net radiation R_n , which is a key issue for estimating evapotranspiration through surface energy budget models forced by instantaneous remote sensing surface temperature data.

Main issues with the paper are:

The evaluation dataset is based on the sole estimate of G as a residual of the energy budget from flux tower measurements; G being usually small compared to the turbulent fluxes, the total uncertainty is high, and a more robust method would have been to do, as classically done, a correction of the subsurface soil heat flux plates measurements, with potentially a further correction with the residual G estimate, bearing in mind that turbulent fluxes are generally underestimated. Furthermore, the FLUXNET dataset is not representative of the agro-eco-types where remotely sensed ET estimates are required; especially, crops in Mediterranean and semi-arid climates are largely underrepresented. This limits the study's impact.

Reply: As described in Line 36-38 (40-42 in revised manuscript), “Over bare soils or sparsely vegetated surfaces, G can reach half of the net radiation (R_n) (Heusinkveld et al., 2004). Even under full vegetation cover, G is significant, especially when turbulent processes are less active (Gentine et al., 2012).”

As described in Line 48-60 (53-65 in revised manuscript), “There are numerous schemes for estimating G (Wang and Bou-Zeid, 2012; Gao et al., 2017; Wu et al., 2020)...”. “However, applications of these physical mechanism-based approaches are restricted to only a few sites, due to the limitations of field observations of soil thermal properties (Mayocchi and Bristowa, 1995; Kustas et al., 2000). Soil thermal properties are affected by soil texture, mineralogical composition, bulk density, and the surrounding environment (e.g., soil moisture and temperature) (Peng et al., 2017; Ju and Hu, 2018). In other words, soil thermal properties vary with time and space.”

“To estimate ET in RS models, G is usually obtained from empirical relations with R_n .” In this study, accuracy of five ground heat flux empirical simulation methods in the surface energy balance-based remote sensing evapotranspiration models was evaluated by flux site observations.

The observation sites used in this study has a land cover classification. The sites were divided into seven land cover types: Forest, Grassland, Cropland, Wetland, Shrubland, Savanna, and Other types. It represents different agro-eco-types. According to your valuable comments, evaluations of seven land cover types have been added in revised manuscript. Figure 3, 4 and 7 (Figure 8 in the revised version) have been revised according to your valuable comments. A new Figure 7 has been added. Descriptions of these figures have also been added as follows,

Line 229-237, “In terms of seven land cover types, the intra-day performance of each land type was similar to that of all sites except the Other type (Fig. 3-c and 3-d). The correlation between G and Rn was relatively high in the sunrise and sunset periods. The correlation in Other and Wetland types is generally higher than that of other land cover types. In each period, the median R2 of all sites in the two types generally exceeded 0.60, and the highest value even exceeded 0.80. Except Other type, the difference of correlation between G and Rn in different land types is mainly reflected in the daytime period except Other type. The correlation in the Forest and Savanna types was significantly lower than that of other types during daytime, especially for Savanna sites, most of which had R2 lower than 0.5 during daytime. In Other type sites, the correlation between G and Rn in the daytime is stronger than that in the night periods. The slope value of each land cover type in the daytime is lower than that in the night. This intra-day distribution of slope was consistent with that of all sites.”

Line 267-273, “For different land cover types, the correlation between G and Rn was the strongest at Other and Wetland sites. The mean value of the median R2 of statistical sites was 0.71 and 0.67 for these two types, respectively. There was also a strong correlation between G and Rn at Cropland, Shrubland and Grassland sites. The mean value of corresponding R2 is about 0.60. There was only one site with a median R2 lower than 0.4 in each of the three land cover types. The mean value of corresponding R2 was 0.55 for Forest sites. However, the correlation was relatively weak in the Savanna type sites, and the mean value of corresponding R2 is 0.49. The weak correlation between G and Rn ($R2 < 0.4$) were mainly distributed in the Forest (6/97), Grassland (7/42) and Savanna (3/15) sites.”

Line 351-361, “Figure 7 shows the NSE simulated by each method in seven land cover types. The intra-day performance of each land cover type was similar to that of all sites except for the Other type, with the highest simulation accuracy at sunrise and

sunset periods. The intra-day accuracy varied greatest at the Forest and Savanna sites. The median NSE of all sites simulated by the LC_NDVI_E method was close to 0.8 at the sunrise periods, while the corresponding NSE was only approximately 0.4. It varied little at other land cover types, especially for Wetland and Shrubland types. The greatest and lowest values of median NSE for all sites simulated by the LC_NDVI_E method were approximately 0.7 and 0.6, respectively. The NSE of the LC, LC_NDVI_P and LC_NDVI_E methods showed a unimodal distribution in the Other type sites. The NSE was significantly higher in the daytime than at night periods. The highest value was in the morning and noon periods, with the median NSE of all sites exceeding 0.8. The model performance was significantly better than other land cover types. In the Other type sites, the LC_NDVI_E method performed better than other methods, with the median NSE higher than 0.6 in each time period.”

Line 387-399, “For different land cover types, the LC method performed better in the Cropland, Wetland and Other type sites. The mean value of median NSE of Wetland and Other sites was 0.66 and 0.69, respectively. The method was also able to accurately simulate G in the Forest, Grassland and Shrubland type sites, with the corresponding mean NSE of 0.57 or 0.56. It performed the worst at the Savanna sites, with the corresponding mean NSE was only 0.47. Since the Savanna sites are mainly distributed in tropical regions, this is consistent with the relatively poor performance of tropical region site as mentioned above. The performance of the method varied significantly in each land cover types except for the Other type sites. In the Wetland type sites, there were 3 sites in the United States with the NSE value lower than 0.3. The NSE of other 35 sites was higher than 0.50, with the highest value was close to 0.90. The Grassland sites were distributed in Asia, Europe, North America and Oceania. The NSE value was greater than 0.5 at each Grassland site in Europe. Cropland sites were distributed in Asia, Europe, and the United States. The NSE value was lower than 0.60 at 8 sites in the United States, with the mean NSE value of only 0.45. The method was able to accurately simulate G at 11 sites in Europe except for one site in Mediterranean region, with the mean NSE value of 0.74. The NSE for the two Asian sites was 0.54 and 0.71, respectively.”

The number of empirical equations under study is limited, esp. regarding previous works (Sun et al., 2013, Bonsoms and Boulet 2022**)*

Reply: These works have been cited in revised manuscript. The author has reviewed these references carefully, and found that the empirical equations missed in this study are some methods required albedo and LST data. As described in Line 479-482 in revised manuscript, these equations were not evaluated in this study due to data limitations.

The author had used LST data at regional scales, while were not familiar with albedo data. The used LST data is the Terra Moderate Resolution Imaging Spectroradiometer (MODIS) MOD11A1 product, which is produced daily LST at a spatial resolution of 1 km. The evaluation in this study was based on daily data series. For daily series of the global LST dataset, the author only found that the MODIS MOD11 dataset was available. The MOD11B product provides daily per pixel Land Surface Temperature and Emissivity (LST&E) in a 1,200 by 1,200 kilometer (km) tile with a pixel size of 5,600 meters (m). There are hundreds of files for each day in the dataset (MOD11A and MOD11B) covering the global land. As for 230 flux sites used in this study, each site is needed to be corresponded to these hundreds of files. Huge amounts of data need to be downloaded, i.e. hundreds of files per day multiplied by number of days in the observed daily series of flux sites. In fact, Saadi et al. (2018) only evaluated the methods at a single observed site. The NDVI dataset used in this study is a single file per day with global coverage, and the workload is relatively acceptable. In addition, the authors had tried several methods to download MODIS product but without success at the beginning of this study. It was also failed to download these products in the past few days. This work is beyond the author's capacity. Therefore, the methods embedded with LST data were not evaluated in this study.

In Line 485-486, a new sentence “Evaluation of such methods embedded with the LST data is recommended for further research where data is available.” was added to complement the discussion of this issue.

I am concerned with Figure 1a: H and Rn are equal ! Also, why are the flux values so low for half hourly flux estimates ? Some explanation is required here; if G is the residual, the energy budget is closed, the SEB average of all sites should also be closed for each half hourly value, i.e $Rn-G=H+LE$. Also, G' seems to be an uncorrected G measurement at a few cm depth (please confirm, G' is actually not defined properly in the paper), the corrected G' at the surface should be shown and analysed for all sites compared to G, esp. since the normalized (G) and (G') looks

similar (1e versus 1f).

Reply: Figure 1a includes the primary and secondary y-axes. In Figure 1a, the primary and secondary y-axes represent the H and Rn, respectively. The H is very different from Rn. For example, the greatest H value ($<150 \text{ W/m}^2$) accounts for only 40% of the highest Rn value (Figure 1a). The half hourly flux values shown in Figure 1 are calculated from the FLUXNET observations. G is the residual in this study.

As described in the second paragraph of the Introduction section, G' is soil heat flux measurement at a few cm depth. G is the soil heat flux at the surface, which is difficult to observe directly due to technical limitations (Wang and Bou-Zeid, 2012; Gao et al., 2017), and direct estimation of G using RS data is not possible (Kalma et al., 2008; Allen et al., 2011; Saadi et al., 2018). There are too many sites (230) used in this study, it is impossible to show intra-day distribution of flux values for each site. Therefore, the mean flux values of all sites were shown in Figure 1. Yes, the intra-day distribution characteristics of normalized G and G' are similar (1e and 1f). It could also be found that the normalized H and LE are also similar (1c and 1d). All of these intra-day distributions of fluxes are determined by the Rn. In fact, the intra-day distribution of these fluxes is also similar to the Rn (1b).

Detailed comments:

Line 7: what is the difference between “intra-day” and “diurnal” ?

Reply: “diurnal” was expected to describe of or belonging to or active during the daytime. It has been revised to “daytime” to avoid misunderstanding.

Line 9: add that G is required for RD ET models based on the SEB forced by radiative surface temperature (it is of no importance for other models).

Reply: This sentence has been revised to “This indicates that G plays an important role in remote sensing (RS) energy balance based evapotranspiration (ET) models.” according to your valuable comments.

Line 9: add “empirical”, i.e. “G empirical estimation methods”

Reply: “empirical” has been add in revised manuscript, including the Title, Line 10, 12, 32, 87, 104, and 468.

Line 13: “the two methods ... “: revise the sentence ; I find a bit contradictory that calibrated G/Rn based on NDVI and fractional cover have contrasted performances.

Reply: This sentence has been revised to “The linear coefficient (LC) method and the two methods embedded with the normalized difference vegetation index (NDVI) were able to accurately simulate a half-hourly G series at most sites.”

L65 to 77: all models based on forcing SEB with land surface temperature need an estimate of G/Rn, no need to review them all, better provide an updated review of all G/Rn equations

Reply: Four new references have been added in this paragraph. A new sentence has been added at the end of this paragraph, “More G empirical estimation methods could be found in Sun et al. (2013) and Bonsoms and Boulet (2020).”

Line 140: we can't use only calibrated parameters for operational applications (i.e. satellite products) so it is important to also test the default (published) parameter values (comment also made by other reviewers).

Reply: The default parameter values had been evaluated in this study. As described in the second paragraph of the section 4.2, “The fixed parameters might be suitable for some regions, but not on a global scale. This study confirmed that the optimal parameter values vary significantly from site to site.” In addition, as revised in the last paragraph of this section, “...the optimal values of the model parameters differed among the different sites. This has also verified by Chen et al. (2019).” The author found that the default parameter values published in the references were also optimized by some observations sites data. Therefore, it is recommended that model developers consider the spatial variations of G simulation parameters in RS ET modeling on a global scale.

Line 370: NO, Santanello and Friedl (2003) do NOT need LST

Reply: Thanks for your valuable comments. It includes a variable “t” in the equation (4) of this reference. The author made a mistake on that. The sentence “However, it requires intra-day land surface temperature (LST) data series, which cannot be

obtained by RS. Because RS can only monitor instantaneous LST when a satellite overpasses, it cannot obtain intra-day LST data series.” has been deleted.

Line 420: I don't understand this sentence

Reply: Line 420, “The results of this study indicate that the performance of the different methods varied at some site scales.” As described in Line 417-420, Saadi et al. (2018) found the accuracy of three methods was different at an observation site. The sentence in Line 420 means that the performance of methods evaluated in this study is also different at some sites. In Line 486 in the revised manuscript, this sentence has been revised to “The results of this study indicate that the performance of the different methods varied at some sites.” to make it clear.

* *Sun, Z., Gebremichael, M., and Wang, Q.: Evaluation of Empirical Remote Sensing-Based Equations for Estimating Soil Heat Flux, Journal of the Meteorological Society of Japan, 91, 627-638, 10.2151/jmsj.2013-505, 2013.*

** *Bonsoms, J., and Boulet, G.: Ensemble Machine Learning Outperforms Empirical Equations for the Ground Heat Flux Estimation with Remote Sensing Data, Remote Sensing, 14, 1788, 10.3390/rs14081788, 2022.*