Response (Referee #1 comment)

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:

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/Rn within the open surface and full canopy. This % or fraction G/Rn 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, Rn, 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 Rn 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 96-99, "This study addresses four key objectives: (1) investigating the temporal and spatial variations and common characteristics of the empirical relationship between G and Rn; (2) evaluating the accuracy of five empirical methods in simulating half-hourly G from Rn; 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.

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 Rn, 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 Rn, 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 evaluatesoil heat ux calculation: An analyticalreview 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, 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, Rn is calculated using radiation balance using remote sensing and meteorological inputs, and G is estimated as a fraction of Rn. Hence, the uncertainty in Rn calculation is also a source of error in G. In some cases when G may be biased available energy (Rn-G, where Rn is coming from remote sensing-based radiation balance) may be reasonable. In this paper, the author used observed Rn in calibrating G, so when you compare coefficients, the uncertainty in Rn (even better when remote sensing-based Rn is used) needs to be mentioned too. Given that Rn is the key input used in all G methods considered in this study, additional assumptions (assuming that Rn 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 Rn was used for calibrating G in this study. It was assumed that Rn is perfectly simulated by the remote sensing-based ET models. Several sentences have been added in the Discussion section (Line 431-434, Page 14) to describe this issue, as follows, “In RS ET models, Rn is generally calculated using radiation balance with RS images and meteorological inputs. However, observed Rn was used for simulating G in this study.

In other words, it was assumed that Rn is accurately simulated by the RS ET models. Therefore, it should be noted that the uncertainty in Rn 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 148-149.

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 147-148 to make calibration/validation more 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 would like to add these results in Supplementary Materials if possible.

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 148-150, 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 Rn 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 225-234, “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 263-269, “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 342-352, “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 378-389, "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: Yes.

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: 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 150, 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 152-154, 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" to make it clear. In addition, similar revisions have also been made in Line 232, 234, 243, and 339.

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.

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 Rn leading to reasonable estimates of available energy (Rn-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.”