Author's response

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

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,

Thank you very much for your responses to review comments.

I have obtained further comments on the current revision and would like to invite you to study the comments carefully and submit a new revision.

The main issues to be addressed are:

1) What are the significantly added-values of your study compared to previous similar studies? You are invited to explicitly propose ideas and suggestions to the community on how to improve the algorithm and accuracy in G estimation at global scale.

2) Based on your interpretation and discussions can you identify the mechanisms behind the results, that may be investigated in future research to advance research in land surface energy balance.

I hope these comments are useful for improving the quality of your manuscript.

Best wishes

Bob Su

Editor HESS

****

Reply 1): The added-values of this study compared to previous similar studies were described in Line 90-106. It includes two points: (1), previous studies were largely limited to a single site scale, while this study focused on a global multi-site scales. (2), Purdy et al. (2016) evaluated six empirical methods at global scale. However, this

study evaluated G simulation against G' observations. As described in Line 45-54, there is a large difference between G' and G. We used the surface energy balance method to assess the G simulation methods. This method could avoid the inconsistent spatial scale of G with that of LE and H in field measurements, and makes full use of the surface energy term that can be accurately measured at present.

In Line 98-100, new sentences "Purdy et al. (2016) evaluated six empirical methods of G simulation against G' observations at 88 flux sites. This was a very meaningful study on a global scale. However, there is a large difference between G' and G which has been described above." have been added to make it more clear.

This study investigates temporal and spatial accuracy of five G simulation methods. The evaluation results of this study are expected to provide reference for RS ET model application and developers. For example, we find that the performance of these methods was good and poor at some sites and time periods and some land-cover types. RS ET modelers could check the advantage of the models at good performance regions, and find why the models are poor at some other areas, then revise the models to improve the accuracy at poor performance regions. Similar suggestions have been showed in Abstract, Discussion and Conclusion.

Several sentences have also been added in Abstract, Introduction, and Conclusion to make it more clearly, as follows,

Line 30-35, "Further improvement of G simulations at these sites and time periods is recommended for the RS ET modelers. In addition, variable parameters are recommended in empirical methods of G simulation to improve accuracy. Instead of the Rn, finding another variable that has a physical connection and strong correlation with the G might be a more efficient solution for the improvement, since the weak correlation between the G and Rn is the main reason for the poor performance at these regions."

Line 110-111, "The evaluation results of this study are expected to provide reference for RS ET model application and developers."

Line 538-543, "The performance was best in the Wetland and Other type sites, and was worst at the Savanna type sites. Improvement of G simulation at low-accuracy regions is recommended for the RS ET modelers, such as low-latitude regions and Savanna type sites. The weak correlation between the G and Rn is the physical reason for the poor accuracy of G simulation in these regions and sites. Instead of the Rn, finding another variable that has a physical connection and strong correlation with the

G might be a more efficient solution to improve the accuracy of the empirical estimation method for G."

Line 544-545, "Variable parameters are recommended in empirical methods of G simulation to improve accuracy."

Reply 2): The physical mechanism behind the evaluation results of the simulation accuracy of empirical methods is mainly affected by the correlation between G and Rn. Therefore, before the evaluation of empirical methods (Section 3.3), the Section 3.2 is "Temporal and spatial analysis of the empirical relationship between G and Rn and between G/Rn and NDVI".

Several new sentences have been added to make it more clear, as follows,

In Line 33-35, "Instead of the Rn, finding another variable that has a physical connection and strong correlation with the G might be a more efficient solution for the improvement, since the weak correlation between the G and Rn is the main reason for the poor performance at these regions."

In Line 498-504, "The accuracies of the LC_fc_SE and LC_fc_ST methods, which embed fractional vegetation coverage in the G simulation, were satisfied for monthly simulation, but were significantly lower than those of the previous three methods in simulating daily values. The weak correlation between G/Rn and NDVI might be the main reason for the poor performance of these methods. However, the coarse-resolution NDVI data used in this study are not sufficient to represent the scale of flux measurements, especially for sites with heterogeneous land surface. This might be the main reason for this weak correlation. The application of higher resolution and continuous vegetation index data series is expected to improve the simulation accuracy of these methods."

In Line 540-543, "The weak correlation between the G and Rn is the physical reason for the poor accuracy of G simulation in these regions and sites. Instead of the Rn, finding another variable that has a physical connection and strong correlation with the G might be a more efficient solution to improve the accuracy of the empirical estimation method for G."

In addition, the author also found that seasonal variations of G and Rn were also an important factor affecting the accuracy of empirical methods in simulating G. The new sentences have been added in Line 457-461, "The sites with satisfied model performance were generally characterized by seasonal variation in G and Rn due to

the climate in these regions. Conversely, the sites with poor model performance showed little seasonality. Whether G and Rn have seasonal variations was also an important factor affecting the accuracy of empirical methods in simulating G. This was consistent with the evaluation results of the LE simulation accuracy (Liu, 2021; 2022).”

Please find details in the revised manuscript with tracks.

Responses to Reviewer #4 Comments

This manuscript evaluates several empirical methods to estimate ground heat flux, which is interesting, but lack of creativity. Similar studies and conclusions could be found, e.g., Purdy et al. (2016) evaluated the soil heat flux at 88 sites globally based on FLUXNET2015 dataset. The large uncertainty in G estimation at global scale using empirical methods (including those evaluated in this manuscript) was clearly concluded by many previous studies. It's commonly agreed that these empirical equations should be carefully calibrated when applying them to new regions or sites. The site-to-site parameter optimization of empirical method in this study is helpful, but no significant added-values and explicit ideas/suggestions to the community on how to improve the algorithm and accuracy in G estimation at global scale. The author did quite a lot but yet superficial data analysis to show the relationship between G and Rn (and H) without insight interpretation or discussions on the mechanism behind the results, which did not bring sound scientific significance and inspirations to readers.

Reply: The added-values of this study compared to previous similar studies were described in Line 90-106. It includes two points: (1), previous studies were largely limited to a single site scale, while this study focused on a global multi-site scales. (2), Purdy et al. (2016) evaluated six empirical methods at global scale. However, this study evaluated G simulation against G' observations. As described in Line 45-54, there is a large difference between G' and G. We used the surface energy balance method to assess the G simulation methods. This method could avoid the inconsistent spatial scale of G with that of LE and H in field measurements, and makes full use of the surface energy term that can be accurately measured at present.

In Line 98-100, new sentences "Purdy et al. (2016) evaluated six empirical methods of G simulation against G' observations at 88 flux sites. This was a very meaningful study on a global scale. However, there is a large difference between G' and G which has been described above." have been added to make it more clear.

The added-values of this study and explicit suggestions for improving the accuracy in G estimation include several points:

(1) The evaluated flux sites are much more than previous studies.

(2) It investigates temporal and spatial accuracy of empirical methods of G simulation. The evaluation results are expected to provide reference for RS ET model application

and developers. We find that the performance of these methods was good and poor at some sites and time periods and some land-cover types. RS ET modelers could check the advantage of the models at good performance regions, and find why the models are poor at some other areas, then revise the models to improve the accuracy at poor performance regions.

(3) Empirical methods of estimating G from Rn were evaluated in this study. Results show that the simulation accuracy is mainly affected by the correlation between G and Rn. As added new sentences in Line 540-544, "the weak correlation between the G and Rn is the physical reason for the poor accuracy of G simulation in these regions and sites. Instead of the Rn, finding another variable that has a physical connection and strong correlation with the G might be a more efficient solution to improve the accuracy of the empirical estimation method for G."

(4) The optimal parameter value of each method varied significantly at different sites. Therefore, the fixed parameter values in the G simulation methods do not match the actual situation. Variable parameters are recommended in empirical methods of G simulation to improve accuracy.

The physical mechanism behind the evaluation results of the simulation accuracy of empirical methods is mainly affected by the correlation between G and Rn. Therefore, before the evaluation of empirical methods (Section 3.3), the Section 3.2 is "Temporal and spatial analysis of the empirical relationship between G and Rn and between G/Rn and NDVI".

The Abstract, Introduction, Discussion and Conclusion sections have been revised according to your valuable comments. Please find details in the revised manuscript with tracks.

Major issues:

Even though eddy covariance (EC) measurements are widely used, the uncertainty of EC measurements of turbulent fluxes should be carefully evaluated. Actually, the random uncertainty of the measured latent heat flux (LE) by EC could reach 16%, and the random uncertainty of sensible heat flux (H) could reach 18%. Considering the magnitude of G is much smaller than LE and H, the uncertainty of G estimated by SEBR method would be very large, even larger than the magnitude of G itself.

Reply: At present, there are uncertainties in turbulent fluxes observation, including EC measurements of flux variables. However, as described in Line 426-429, "The

eddy covariance measurements of H and LE are generally considered to be the most accurate observations available, and have been widely used as a reference to evaluate other simulation methods. 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 estimation of EC measurements uncertainty is out of the scope of this study.

The footprint of net radiation and soil heat flux observations are significantly different from that of H/LE observations. Large uncertainty is anticipated when estimating G directly using the surface energy balance residual (SEBR) method (Eq.1 in the manuscript). Energy imbalance has been an issue in the ground measurements for long time. The author does not seem to have assessed and corrected the energy imbalance at the flux tower sites before using the data, which could be the reason that the author has obtained very large G based on the SEBR method at site level and thus would bring unreliable relationships between G and Rn.

Reply: Yes. As described in Line 53-54, "The spatial scale of the G observation is also much smaller than that of the H and latent heat flux (LE) estimates (Shao et al., 2008; Verhoef et al., 2012)." Because the footprint of net radiation and soil heat flux observations are significantly different, the net radiation, H and LE observations with relatively similar footprint are used to calculate surface soil heat flux with corresponding spatial scale in this study. It makes full use of the advantages of relatively consistent footprint scale of the surface energy term (Rn, LE, and H).

Energy balance is a universal principle. The energy terms on the surface are also in balance. The surface energy imbalance, which is calculated from the observed data of the local surface energy term, might be mainly caused by the error of observation data. The uncertainty in measuring surface soil heat flux (G) includes: 1) the significant difference between the footprint of soil heat flux observation and other surface energy terms (e.g. Rn, LE, and H); 2) and a large difference between G and soil heat flux measured using heat flux plates near the surface, which is described in the second paragraph of the Introduction section (Line 45-54). These uncertainties of G measurements are important reasons for the imbalance of observed surface energy.

The mean value of G is definitely lower than the average of Rn. As described in the first paragraph of the section 3.1 (Line 175-178), based on the mean of 230

FLUXNET sites, LE, H, and G accounted for 34.5%, 46.3%, and 19.2% of Rn, respectively. G accounted for 28.8% of Rn when only daytime periods were considered. However, the daily or hourly value of G might be several times that of Rn. This is not rare in flux observations from 230 FLUXNET sites. For example, it is possible that G is larger than Rn on cloudy days or at night.

It seems some problems when the author processes the FLUXNET2015 data. According to my experience in the data-screening with FLUXNET2015 data, there are many sites cannot provide G observation (e.g., PA-SPn, and some others), and these sites should be eliminated from the analysis. But these sites were also listed in the Supplementary Table1. The author needs to check it more carefully.

Reply: Thanks very much for your valuable comments. There are 212 FLUXNET2015 sites and 81 FLUXNET-CH4 sites available from https://fluxnet.org/. However, there are many sites lack of measuring LE, H, Rn, and G'. The first three variables were used for simulating G in this study. Therefore, 230 sites with measurements of LE, H, Rn were used. These sites were listed in the Supplementary Table1. There were 167 sites observing G'. G' data series were only used in the Section 3.1 to identify intra-day distribution characteristics of observed surface energy balance items.

As described in Line 118-120, "All missing values were eliminated. For example, if there were missing values on a certain day, all data on that day were discarded. Therefore, only days with fully available half-hourly data were used in the analysis. Only sites with a data series longer than 360 days were used." In addition, there are also several problems in the raw dataset. For example, there are continuous identical values in Rn measurements at some sites. These continuous identical values were treated as missing values in this study. There are also many missing values in NDVI data covered flux sites. Overall, these data had been carefully preprocessed by the author. It cost a lot of time and energy.

According to your valuable comments, revisions are as follows,

Line 122, the new sentence "G' was not observed at 63 sites (**Table S1**)." was added.

Line 175-177, the sentence "**Figure 1** shows the intra-day distribution of half-hourly Rn, H, LE, G, and G', derived from the mean of 230 FLUXNET sites." has been revised to "**Figure 1** shows the intra-day distribution of half-hourly Rn, H, LE, G, and

G’. The first four variables and G’ were derived from the mean of 230 and 167 FLUXNET sites (**Table S1**), respectively.”

In Supplementary Materials, Table S1, the sites lack of G’ measurements have been marked by “*”. A new sentence “Note: * represents the soil heat flux (G’) is not observed.” has been added.

The diurnal variations in the averaged fluxes of the surface energy balance as shown in Fig. 1 seems too smooth to me. I cannot believe these curves come from actual measurements, they are just too perfect.

Reply: The Fig. 1 is the diurnal variations of surface energy balance in the averaged values from multi-sites. The Rn, H, LE, and G were derived from the mean of 230 sites, while G’ is calculated from the mean of 167 sites. The author could provide the raw data for this figure. Actually, the diurnal variations are not smooth at each single site.

Why the linear models with NDVI perform better than the model with Fc? The author should give explanation more clearly. Eq. 2 and Eq. 3 are almost the same, it does not make sense to take them as two different models.

Reply: The weak correlation between G/Rn and NDVI might be the main reason for the poor performance of the methods with Fc. In addition, the NDVI data was used for calculating the Fc in this study. The spatial resolution of the NDVI data was too coarse to represent the scale of flux measurements. This might explain the weak correlation between G/Rn and NDVI.

In Line 498-504, the sentence “The accuracies of the LC_fc_SE and LC_fc_ST methods, which embed fractional vegetation coverage in the G simulation, were significantly lower than those of the previous three methods.” has been revised to “The accuracies of the LC_fc_SE and LC_fc_ST methods, which embed fractional vegetation coverage in the G simulation, were satisfied for monthly simulation, but were significantly lower than those of the previous three methods in simulating daily values. The weak correlation between G/Rn and NDVI might be the main reason for the poor performance of these methods. However, the coarse-resolution NDVI data used in this study are not sufficient to represent the scale of flux measurements, especially for sites with heterogeneous land surface. This might be the main reason for this weak correlation. The application of higher resolution and continuous

vegetation index data series is expected to improve the simulation accuracy of these methods."

The five methods (Eq. 2 to 6) are empirical models based on the correlation between the G and Rn. The term containing the NDVI or Fc in Eq. 3 to 6 could be taken as a whole, which is similar to the coefficient in the Eq. 2. However, these methods were commonly used in different remote sensing evapotranspiration models. For example, the Eq. 2 is applied in the TSEB, ALEXI, DisALEXI, GLEAM, and other RS ET models to simulate G, but different models use different linear coefficient values. The Eq. 3 is applied in the SEBAL model. The Eq. 4 is applied in modified SEBAL (Singh et al., 2008) and SEBS (Chen et al., 2019) evapotranspiration models. Eq. 5 and 6 are applied in the S-SEBI, NTSG, BESS, METRIC, MOD16A2, and SEB-4S models. Therefore, Eq 2 to 6 are taken as different methods and evaluated in this study.

Indeed, the different performance of G estimation in different time (hours of a day) are closely related to the time-lag between G and Rn, which is important but not well explained.

Reply: Yes, the time-lag between G and Rn could be found in Figure 1. The temporal performance of G estimation in each half-hour of a day was shown in Figure 6. Figure 3 shows intra-day distribution of the linear fitted relationship between G and Rn. It is obvious from the Figure 3 and 6 that the performance of G estimation in each half-hour is closely related to the correlation between G and Rn. However, the author has not found the obvious relationship between the performance and the time-lag from the Figure 1 and 6.

At large scale application using satellite data (NDVI), the author has used the NDVI from AVHRR product with the spatial resolution of 0.05×0.05 °which is too coarse to compare with the footprint of ground measurements. This is particularly important for sites where the land surface is heterogeneous around the 0.05×0.05 °spatial domain.

Reply: Thanks very much for your valuable comments. Yes, the spatial resolution of the NDVI data was too coarse to represent the scale of flux measurements. Several new sentences have been added as follows,

In Line 501-504, "However, the coarse-resolution NDVI data used in this study are not sufficient to represent the scale of flux measurements, especially for sites with

heterogeneous land surface. This might be the main reason for this weak correlation. The application of higher resolution and continuous vegetation index data series is expected to improve the simulation accuracy of these methods."

In Line 518-519, "The poor performance might be mainly caused by the coarse-resolution vegetation index data, which could not represent the scale of flux measurements."

In the Discussion section the author state that "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", this is not true! Geostationary satellites can provide LST observations at 15min – 30min intervals.

Reply: This 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.

Minor issues:

In Figure 1, explanations for sub-figures are missing in the caption.

Reply: Explanations for sub-figures had been added in the caption of Figure 1, as follows, "a is raw Rn, H, LE, G, and G'; b-f are normalized Rn, H, LE, G, and G', respectively"

Figure 8: I do not understand how to get the median NSE value of each site. Shouldn't it be a single NSE value per site?

Reply: It is the median NSE of 48 half-hour values in diurnal period. The caption of Figure 8 has been revised to "…… a is the median NSE of 48 half-hour values; b-e represent the NSE values at 6:30, 10:30, 13:30 and 18:00, respectively……"

It's inappropriate to define 6:00-7:00 as sunrise periods and 17:00-18:00 as sunset periods globally, since the sunrise and sunset time vary with locations and seasons.

Reply: According to your valuable comments, the definitions of sunrise and sunset periods have been deleted in the revised manuscript.

Line 190: "During data processing, data points with absolute values greater than 10 in the G/Rn or G/H daily series of each period were deleted." Threshold 10 seems too large, can G be larger than Rn ?

Reply: There were several data points with absolute values greater than 10 in the observed G/Rn or G/H daily series. For example, in the daily series of observed soil heat flux (G') and net radiation (Rn) at the AR-SLu site, the value of G'/Rn is greater than 10 in 2010-6-18 and 2010-9-3, and is lower than -10 in 2010-5-15, 2010-6-26, 2010-8-3 and 2010-9-27. The value of G'/Rn is larger than 10 in 11 days, and is lower than -10 in 17 days at the AT-Neu site. Similar values could also found in other flux sites. The author could provide the raw data if it is possible.

Similar problem is in the range of coefficient α in Eq.1, the maximum value of 1.5 will lead to G is 1.5 times larger than Rn, does it have physical meaning?

Reply: The ratio of G and Rn is generally lower than 1.0 for the mean value at observed sites. However, as described above, the daily values of G/Rn might be larger than 1.0. This can happen. For example, the observed Rn and G' is 0.15 W/m^2 and 5.0 W/m^2 in 2002-10-14. The simulated G on the same day is -15.7 W/m^2. The absolute values of G/Rn and G'/Rn are larger than 1.5.

In Figure 4, analysis is also done for monthly temporal scale without explaining why it is needed.

Reply: Figure 4-d and **4-e** show the linear and exponential fitted R^2 values between the monthly series of G/Rn and NDVI at each observed site, respectively. The empirical correlation between G/Rn and NDVI was also analyzed in Figure 4. The results showed that the correlation between G/Rn and NDVI was weak in the daily series but strong in the monthly series. The last paragraph of the section 3.2 (Line 283-301) is the description of these sub-figures. In addition, the revision was also made in Line 498-500.