Response (Referee #3 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:

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 Rn, 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 sol 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 (Rn) (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 (52-64 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 Rn.” 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 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.”

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 473-476 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.

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 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 fluxs are determined by the Rn. In fact, the intra-day distribution of these fluxs 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, 11, 31, 86, 103, and 462.

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. 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.*