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.

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.

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.

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.

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.

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.

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.

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.

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.

Minor issues:

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

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?

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.

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 ?
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?

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