Response to comments by Anonymous Referee #1

General Comments

This is a landmark paper. The authors introduce a wealth of hard-won, empirical data into a longstanding debate over how best to parameterize potential evapotranspiration (PET). Their analysis is meticulous, thorough, and well documented. Their main finding (robustness of Milly-Dunne and Priestley-Taylor methods, relative to Penman-Monteith method) is convincing and will surprise many investigators.

Specific Comments

1) *My most significant concern with this analysis was the use of evaporative fraction, LE/(LE+H), in the characterization of stress. I wondered if this might somehow bias the analysis in favor of the Milly-Dunne method, since MD posits a constant value of LE/(LE+H). For this reason, the sensitivity analysis using soil moisture as a stress criterion, and reaching similar conclusions, is a valuable part of the paper.*

    Response: The reviewer is indeed raising a legitimate concern. For this reason we included the soil moisture criterion as well, which nicely confirms the outcome of the analysis based on the evaporative fraction. This will be further stressed in the revised manuscript by adding to the end of Section 4.1.: "*Still, by using the evaporative fraction as a criterion for selecting unstressed days, we might bias the findings in favour of the PT and MD methods, as they are more sensitive to the available energy. However, the soil moisture criterion taken here provides an independent check of the results and confirms the robust and superior performance of the PT_b and MD_b methods*".

2) *One other concern that might be allayed by a little more information is the use of "data corrected by energy balance closure." For one who is not familiar with FLUXNET and might hesitate to dig into the Michel et al 2016 reference, could the authors say just a little more about how this method works, how big the typical adjustments are, and to what extent the correction method could potentially influence the findings?*

    Response: We will add a section explaining this in the text: "*In this approach, the Bowen ratio derived from the tower measurements is assumed to be correct, and the measured λE_a and H are multiplied by a correction factor derived from a moving window method; see http://fluxnet.fluxdata.org/data/fluxnet2015-dataset/data-processing/ for a detailed description.*" We tested both the use of the 'raw' and energy-balance-corrected values, and found very little differences in the results of both products.

3) *If the authors mean to suggest (this is not entirely clear, and should be clarified) that the simple radiation-based methods should now be incorporated into climate models, then I disagree. Climate models use the same physics upon which Penman-Monteith is based, but it needs to be recognized that in such models the stomatal conductance is calculated dynamically in response to controlling environmental variables. (There is nothing wrong with the Penman-Monteith approach in principle; it's just that it's hard to apply observationally, since it is sensitive to variables that are hard to know with sufficient precision.) Furthermore, the value of stomatal conductance is crucial for the computation of land-atmosphere carbon exchange. I would agree, on the other hand, that the MD method beats the PM method hands-down for application in global "offline" analyses in which atmospheric feedbacks are not present. However, it can also be argued that analysis of climate-model outputs themselves is a better way to spend one's time than doing offline analyses, which are sometimes amount to nothing better than attempts make a silk purse from a sow's ear.*

    Response:
    We fully agree with the reviewer on this matter. We do not want to suggest the incorporation in online climate models, which require solving for stomatal conductance to compute carbon fluxes,

and which do not rely on the calculation of potential evaporation to derive actual rates. We will make sure that in the revised version that that interpretation is not inferred. We do indeed recommend the inclusion of simple radiation-based methods in offline computations, such as drought monitoring systems of rainfall-runoff models. Also, we are in communication with the FLUXNET community to include the estimates of E_p as part of the FLUXNET synthesis dataset.

4) *To address the question of whether or not PET can be calculated correctly from actual Rn-G when the system is stressed takes this otherwise solid empirical paper into the metaphysical realm. What is the meaning of PET in a stressed system? And how does one empirically test that meaning? If one considers the feedback to surface temperature and albedo, why not also the feedback to lower atmospheric conditions, such as humidity and temperature, leading to changes in downward longwave radiation? This passage, for me, detracts from the paper and might better be presented as a technical note elsewhere. Highlighting it in the Conclusion, at the expense of more concrete and surprising findings, seems not to be an ideal choice.*

Response: We understand the reviewer's concern that this section deviates from the solid empirical evidence into a 'metaphysical realm'. However, the question on how to best calculate E_p – also in unstressed conditions – is an important one, and this section discusses the best ways to do it. Therefore, we decided to leave it into the Discussion section, yet to exclude it from the Conclusions. While it is clear that PET is also relevant in a stressed system – even more than in an unstressed system –it should also be clear that "*it is nearly impossible to define a correct and universally accepted definition of E_p, and the most appropriate definition should remain tied to the specific interest and application*", as mentioned in the introduction section. We understand that the correction of surface variables but not considering the feedback to near-surface atmospheric conditions is somewhat arbitrary. We wanted to draw the line at the surface and consider only the actual forcing variables as 'input' and not aim to correct for feedbacks into the radiation, which is extremely difficult and impractical to estimate reliably – as mentioned in a new section in the introduction discussing feedbacks "*Moreover, extensive reference surfaces can be expected to not only exert a feedback on the aerodynamic forcing, but also on the radiative forcing. Indeed, by altering the temperature, humidity and through cloud formation, extensive reference systems are likely to also affect incoming shortwave and longwave radiation. Yet, as this feedback is almost impossible to calculate, it is ignored in all methods considering extensive reference surfaces*". After all, that is the only way to still relay on real tower measurements of atmospheric forcing during stress times. We can certainly see the need to correct for the more direct effect of soil moisture on G, SW_{out} and LW_{out}, because they are more immediately influenced and thus different between stressed and unstressed ecosystems. Therefore, we prefer not to attempt to correct for near-surface atmospheric variables.

We also want to highlight that we also show that: (a) the simple alternative of ignoring this issue, and taking the actual (R_n–G) as forcing variable for calculating E_p, leads to a severe underestimation of E_p; (b) results suggest that this underestimation is largely caused by differences in LW_{out} between the stressed and the unstressed ecosystem; (c) there is a practical solution to overcome this underestimation, which only requires T_a as additional input, and which results in an almost unbiased estimate of E_p.

Finally, we will also acknowledge the complementary approach in our revised paper along these lines, after the suggestion by Reviewer #2.

5) *Could the authors compose a more fitting title? It is nice that they have provided estimates of PET all over the world, but the scientific value of the paper lies in its use of these data to test conceptual frameworks and related equations for quantifying PET.*

Response: We understand the rationale behind the comment, however, we do consider the title to be well suited for the article. In this publication, we are "targeting" different scientific communities, and probably, the scientific value for each community will be different. The current title highlights the end product (dataset of E_p), which can be of great interest for e.g. the FLUXNET community. A title that is more fitted to the findings, such as “Radiation-based methods are more optimal way to estimate potential evaporation at ecosystem scale”, would indeed be more appealing to the hydrological modelling community, but we believe that will narrow down the potential audience of the manuscript.

Technical Corrections/Comments

P1L6 (i.e., Page 1, Line 6). “forecastING”

Thanks. We will correct it.

P1,L14. “calibrated BY biome” (here and many places elsewhere through the paper)

Thanks, this will be changed throughout the paper.

P2L39. “compared”

True.

P3L16. “atmospheric demand” seems an inappropriate phrase, given the dependence on surface properties.

‘atmospheric’ was omitted.

P5L14. “will be used, IN ADDITION TO a biome-specific”

We will change it.

P7L5. “and WHERE u”*

Will be changed.

P8L17. “if FEWER than”.

True, thanks.

P8L18. “criteriON”.

It will be corrected.

P8L26-27. I don’t think the authors mean “actual crop” here but rather “actual vegetation”

Correct, and corrected accordingly

P9L34. Seems more significant than “marginal” to me.

The statement “*although these differences are only slightly significant in the case of* g_{c_ref} *(p=0.017 – see Table 2)”* will be removed.

P10L7. alpha_RB typo?

Yes, corrected to α_{MD}.

P10L17-18. Fig 3d rather than Fig 3c?

Indeed, adjusted.

P13L8. “SMOOTHS”.

True.

P13L13. “relating to whether leaves”.

It will be changed.

P13L17. “issues and would”.

It will be corrected.

P21. Thornthwaite is misspelled.

Thanks, corrected

Table 2. Use of color is a little distracting/ unnecessary, and dark shades obscure text in first column. Why not use horizontal and vertical lines to serve same purpose?

Response: The colours were used to group biomes into forest/savannah /grassland/crop/wetland ecosystems. This will be specified. The darker colours will be adjusted.

Table 2. I am not familiar with the a/b notation in the SUPERscripts (not subscripts). Is there a simple explanation so the reader doesn’t need to search through a statistics book?

Response: We explained it in the new version as “*Different alphabetic superscripts indicate significantly differing means (Tukey post-hoc test; p<0.05).*”

Table 4,5,6. Again, reconsider use of color.

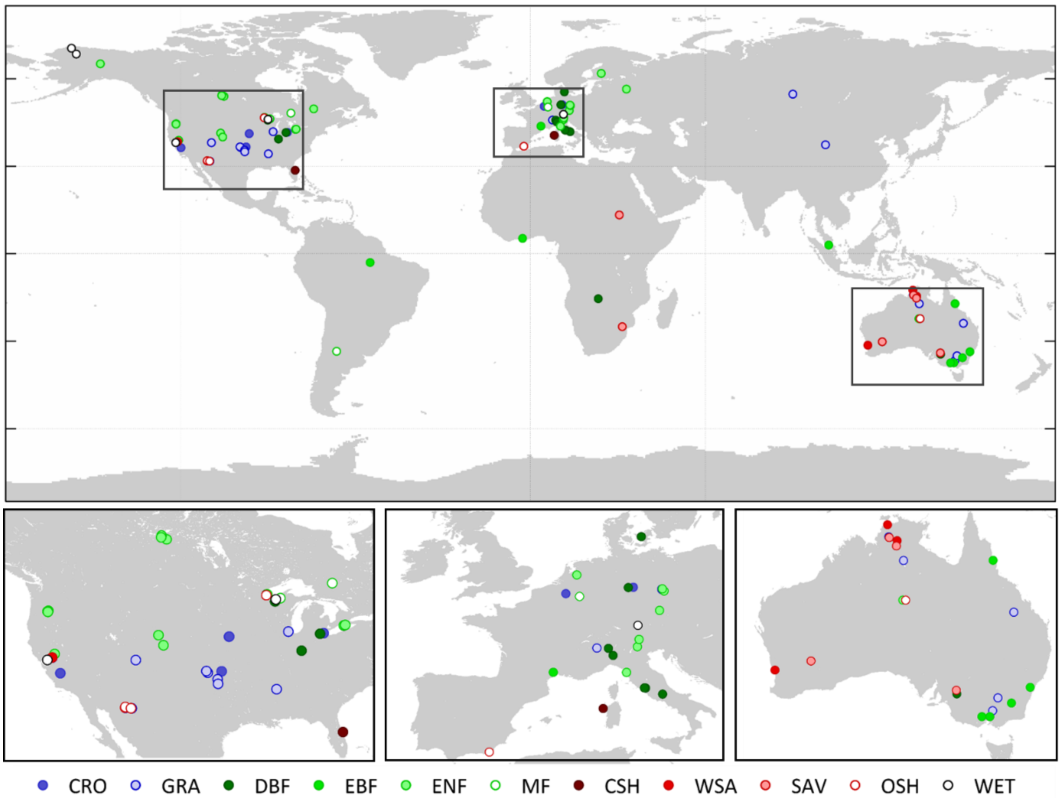
Response: *See response on Table 2.*

Table 5,6. Could it be informative/helpful also to highlight the values that give the best results within the limits of the “standard” approach?

Response: We tried this, but it results in a messy figure. However, since the best biome-specific approach can be calculated for nearly the entire world based on the datasets, there is arguably no need to revert to the standard approaches.

Figure 1. The grey background is so dark that it reduces contrast with colored symbols. Could be lighter, or just put in coastlines. The symbols are very difficult to differentiate within a given color set.

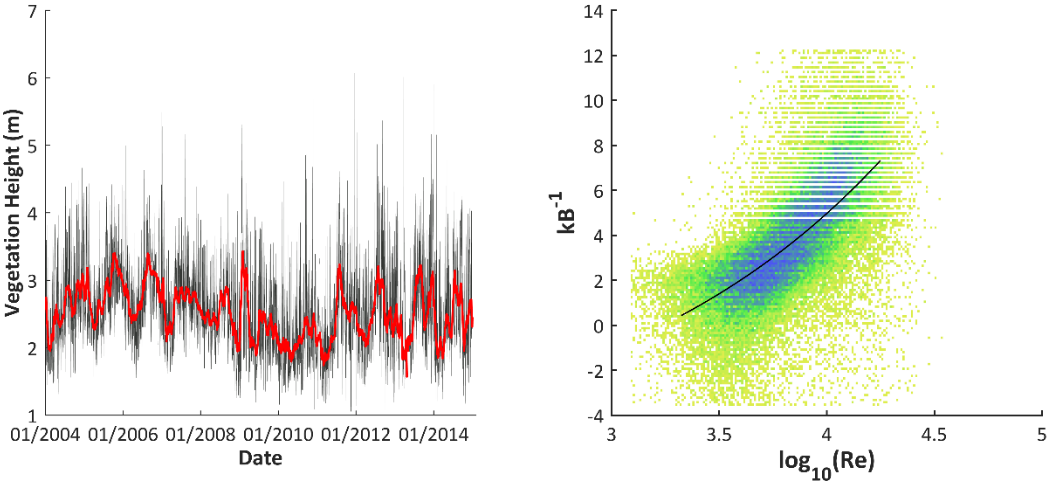
Response: Figure 1 has been redone – the grey background is now lighter, and symbols were changed to add clarity to the figure:



Updated version of **Figure 1. Location of the flux sites used in this study per biome. CRO=cropland; DBF=Deciduous Broadleaf Forest; EBF=Evergreen Broadleaf Forest; ENF=Evergreen Needle Forest; MF=Mixed Forest; CSH=Closed Shrubland; WSA=Woody Savannah; SAV=Savannah; OSH=Open Shrubland; GRA=Grasslands; WET=Wetlands.**

Figure 2. Label dates with full year, e.g., 2010.

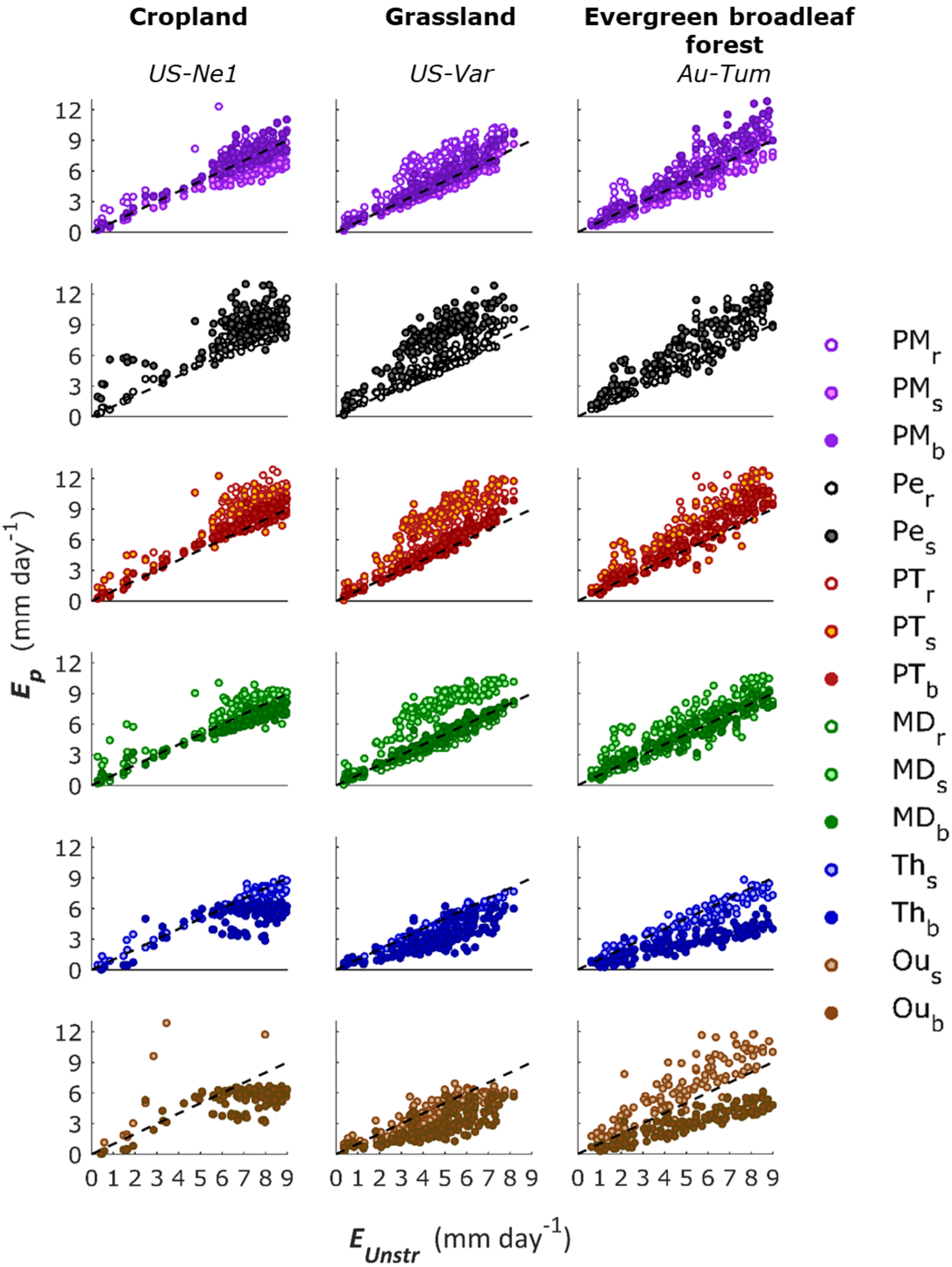
Response: Will be done as requested:



Updated version of **Figure 2 (a) Vegetation height dynamics in time (grey dots: half-hourly measurements; dark grey lines: daily mean vegetation height; red line: 30-day moving average (i.e. the final vegetation height dataset). (b) Relation between the Stanton number (kB^{-1}) and the Reynolds number (Re). Both plots correspond to the woody savannah site of Santa Rita Mesquite (Arizona, USA).**

Figure 4. The symbols are very difficult to differentiate within a given color set. And really there is too much information on these plots, making it difficult to see the authors' main point. What about a 3x6 matrix of panels (some r-s-b spaces empty) with a single panel showing data from the six (or other number of) example sites?

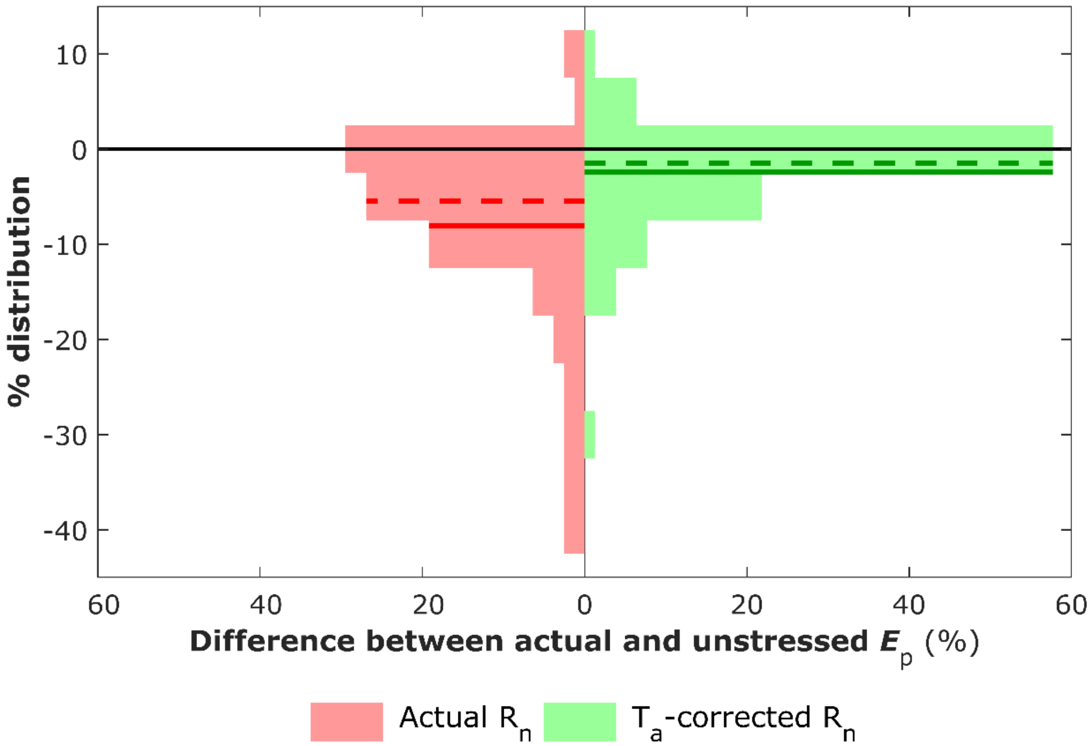
<u>Response</u>: We understand the comment, we will improve the figures as requested, using 6 different panels per site and showing three sites (showing 18 panels would only allow showing 1 site):



***Updated version of* Figure 4. Scatterplot of the measured E_{unstr} versus E_p calculated with the different methods for three selected sites. The discontinuous line is the 1:1 line. Based on unstressed days only defined using the energy balance criterion.**

Figure 6. I don't understand the dotted, dashed, and solid lines. This plot overall is hard to read and might possibly be improved upon.

<u>Response</u>: We understand the reviewer's comments and will improve the figure as follows:



Updated version of **Figure 6. Distribution of the mean error per fluxtower in the estimate of E_p of two empirical methods to calculate unstressed (R_n-G). The first empirical method simply take the actual (R_n-G) as input, the second method corrects the actual (R_n-G) with T_a (Eq. (16). Negative Y-values indicate an underestimation by the empirical methods. For each distribution, the mean and median are indicated with a full and dashed line, respectively.**