General Comments: The authors compare the spatial patterns, seasonality, interannual variations, and trends of evapotranspiration (ET) estimates from several satellite products, reanalysis, and climate models with catchment scale mass balance ET and ET estimates from flux towers. ET is the largest flux from the land to the atmosphere. Despite its importance, estimates of ET are often highly uncertain. This uncertainty in ET estimates poses a real challenge in hydrological and land atmospheric studies at several scales. In this regard the authors are trying to address an important open question by analyzing the consistency and reliability of a few remotely sensed products and modeled ET fluxes. To do so, they compared these ET estimates with long-term catchment mass balance ET.

Answer: We thank the referee for taking the time to read and critique our paper. We have responded to each of your comments in the text below.

Major comments

My major concern is that the uncertainty in catchment mass balance ET that the authors correctly report (Fig 6b and 6c) encapsulates almost all model estimates of ET and remotely sensed ET products. Despite being well written and well structured, I doubt that the basic method of using catchment mass balance ET with such large uncertainty is suitable to evaluate the performance of ET products. Presenting correlations of modeled/estimated ET with catchment mass balance values as performance measures is questionable given the large uncertainties involved in catchment mass balance estimates. Therefore, it is not clear to me in what ways the current paper adds up to the currently reported literature on uncertainties in ET products (e.g. Mueller et al 2014) or in other words what is new here that we didn't know before? For example in Figure 6 b and c, almost all model estimates of ET falling within the uncertainty band of catchment-balance ET. Given such a large uncertainty, one cannot judge the suitability of any ET product.

Answer: We thank the referee for raising uncertainty in the catchment balance approach and the need to consider this uncertainty whilst assessing different ET products. Our main focus in Figure 6 is at the Amazon basin scale (Fig. 6a), where it is clear that the ET products and models fall outside the uncertainty of the catchment balance approach. The robust seasonal cycle in catchment-balance ET is consistent with studies reporting increasing photosynthetic capacity or 'green-up' in the Amazon dry season, when light availability increases. We agree that more care has to be taken when estimating catchment-balance ET at smaller spatial scales, and as such we have decided to move panels 6b and 6c to the Supplementary Material. We have expanded the discussion to acknowledge uncertainties in the catchment-balance approach, particularly when applied over smaller areas, including adding the following statement:

"We note that relative uncertainties in ET estimated using the catchment balance approach increase at smaller spatial scales, precluding a more in-depth assessment of the different ET datasets at smaller scales."

My second major concern is that given the very coarse resolution of GRACE data (300 km, smoothed to 200 or 100 km), and the dependence of neighboring grid cells, how much GRACE signal adds value to the analysis? How large is the changes in storage as compared to the uncertainties of the other terms in the mass balance eq.?

Answer: We thank the reviewer for questions on our inclusion of GRACE data. Data from GRACE is crucial as it allows us to estimate monthly ET from the catchment-balance approach. This is because groundwater storage in the Amazon has a strong seasonal cycle, as shown in

Figures S3 & S4. Without the GRACE data, ET estimated as P–R is restricted to multi-annual means. In Figure 2, we compare ET estimated as P–R and ET estimated from catchment balance (including GRACE data) and demonstrate that including GRACE data does not bias the results. We already report uncertainties in ET estimated using catchment balance approach (see Table S2), allowing uncertainties in GRACE to be compared with other components of the water balance. The referee is correct that uncertainty in GRACE data is substantial. However, it is not much larger than uncertainty in the precipitation data (e.g. mean absolute uncertainty is 8.72 mm in GRACE vs. 6.76 mm for CHIRPS precipitation). Furthermore, at larger spatial scales, such as over the whole Amazon, our results were shown to be robust. Overall, we feel that inclusion of GRACE data, alongside a careful consideration of the uncertainties, is important and makes a useful contribution to our analysis.

Specific comments: Line 161-165: How large is the total bias at the spatial scales that is relevant to GRACE observations?

Answer: Error and bias in CHIRPS data is assessed and reported through comparison with in-situ rain gauges [we refer to Paredes-Trejo et al. (2017)]. It is challenging to assess error at larger spatial scales due to the lack of appropriate data at these scales. For this reason, we report error and bias in CHIRPS as reported previously in the literature.

Line 185-189: GRACE data contains three observations per month and reported as monthly data, same as runoff and precipitation in the current study. Not sure the need for interpolation here?

Answer: The GRACE data that we obtained contained data at irregular intervals (ten or eleven timesteps year) and thus interpolation was required to get monthly-resolution data. The reason for this is related to management of batteries on the GRACE satellites, causing gaps in the GRACE data every 5–6 months for a period of 4–5 weeks. We have added this reasoning to the text.

Line 220, not sure if this arbitrary exclusion of observations is justifiable (specially for the year 2018). Please report (explanation/graph) in what ways including/excluding these points affect your results.

Answer: Including or excluding these data had no effect on the main results reported in the paper, which focussed on the period 2003–2013 (common period across all datasets), but may have influenced trends in ET calculated over the full period. We recalculated the annual, JAS and JFM trends with and without these data points and found no statistically significant trend whether these years were included in the analysis or not. We added the following comment to the text:

"We tested the sensitivity of our interannual trend analysis to the removal of these data points and found it had no statistically significant impact on the reported results."

Line 226-228: Please explain in what ways inclusion of these sites (pasture) would affect your analysis.

Answer: We apologise for miswording this sentence. The pasture towers we excluded were in areas where the dominant land cover was forest, rather than pasture, and thus the towers were not representative of the surrounding land cover. We have corrected the text.

Figure 4 and 7: Are these data points extracted for the grid cells where flux towers are located (given the symbols on the plot). In general in most figures it is hard to understand over which spatial scales the ET estimates are aggregated.

Answer: Apologies for any confusion. Figure 4 shows the catchment-mean ET estimates for the Amazon and its ten sub-catchments that are shown in Figure 1, plotted against possible controlling variables. In Figure 7, ET is averaged over the area drained by Óbidos, which is also indicated in Figure 1. We indicate this information in the caption of each figure for clarity.

Figure 8: For all panels please provide the uncertainty band (at least for catchment mass balance ET).

Answer: We have now added shading to indicate the uncertainty in catchment-balance ET at the interannual timescale to all panels (see attached Figure).

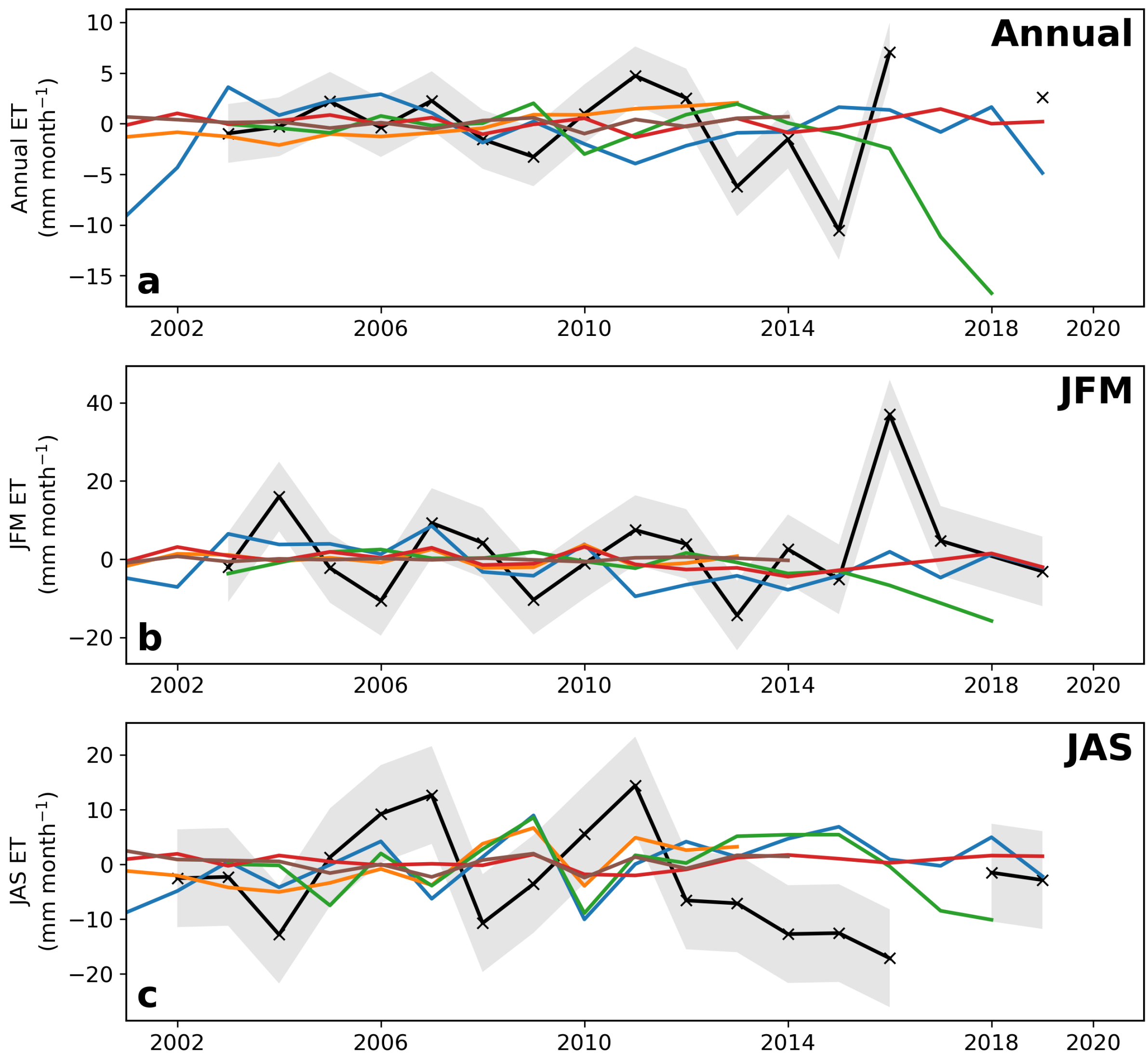

Figure 8 – Interannual variation in evapotranspiration from 2001 to 2019. Time series in ET over the Amazon from catchment balance (black, region drained by Óbidos, Fig. 1), satellites (MODIS, P-LSH, GLEAM), ERA5 reanalysis and CMIP6 models, for (a) the whole year, (b) January–March (JFM) and (c) July–September (JAS), normalised by the 2003–2013 climatological mean. Interannual trends are listed in Table 3. Grey shading indicates the interannual standard deviation in the catchment-balance approach.

Line 555-560: This is all known and well reported in the literature. What is new? A general comment: Why analysis of long-term monthly values and not presenting monthly data for all products?

Answer: We thank the referee for pointing out that we had not fully highlighted the novel aspects of our analysis. Importantly, our paper is the first to evaluate ET from CMIP6 models across the Amazon, comparing model output with satellite observations, reanalysis and catchment-balance estimates of ET. We show that the robust seasonal cycle in ET at the Amazon-basin scale is poorly represented by climate models, raising doubts over their ability to simulate future changes in Amazon hydrology. We add the following text to highlight the novel contributions of our work:

"Our analysis provided a first assessment of Amazon ET representation in the CMIP6 climate models, showing they struggled to capture major features of Amazon ET, including spatial and seasonal variability across the Amazon basin. Furthermore, CMIP6, which represents the latest generation of coupled climate models, showed little evidence of improvement in the representation of Amazon ET compared to CMIP5, highlighting the need for further process-based model development."

Our analysis focused on long-term monthly data. Future work could analyse monthly data in more detail, as suggested by the referee.

Other changes

We noticed that our Amazon LAI values were implausibly low (Amazon mean LAI value of 3.6 m^2/m^2), likely due to inadequate quality control during data processing. We have changed to use a quality-controlled MODIS MOD15A2H Collection 6 LAI dataset provided by Boston University (Amazon mean LAI value of 4.4 m^2/m^2). The main difference to the results arising from this change is that catchment-balance ET is no longer well related to spatial variation in LAI. The new figure and paragraph describing these results are copied below. There were no meaningful changes to any of the rest of the results.

"To understand the drivers of spatial variation in Amazon ET, we compared catchment-scale estimates against catchment-means of precipitation, surface radiation and LAI (Fig. 4). Since there were only eleven data points in the analysis (representing the Amazon and ten sub-catchments), statistical power was relatively low. However, we found spatial variation in catchment-balance ET showed some indication of an influence from radiation ($r=0.38$, $p=0.25$, Fig. 4h), but not precipitation ($r=0.14$, $p=0.68$, Fig. 4a) or LAI ($r=0.06$, $p=0.87$, Fig. 4o). This result tentatively suggests that spatial variation in radiation explains more of the spatial variability in ET across Amazon sub-catchments than other variables. None of the ET products and models analysed captured positive relationships between catchment-mean ET and radiation. ET from ERA5 and the CMIP ensembles instead showed negative associations with radiation (Fig. 4l–n), and, along with GLEAM ET, positive relationships with precipitation (Fig. 4d–g), indicative of water availability influencing spatial variation in ET (Fig. 4d–g). These results confirm that the reanalysis and climate models analysed here struggled to capture spatial patterns in Amazon ET due to misrepresentation of the controlling drivers, specifically the relative importance of precipitation and net radiation. ET from ERA5 and the models also showed positive correlations between LAI and ET (Fig. 4s–u), not seen in the satellite observations. However, it should be noted that satellite LAI was generally lower and showed less spatial variability than other LAI datasets over the Amazon (Fig. S8i–l), likely due to the satellite sensor being insensitive to variation in LAI over areas of dense tropical forest (Myneni et al., 2002, Yan et al., 2016a). This could hamper our ability to accurately assess the extent to which LAI influences spatial variation in ET."

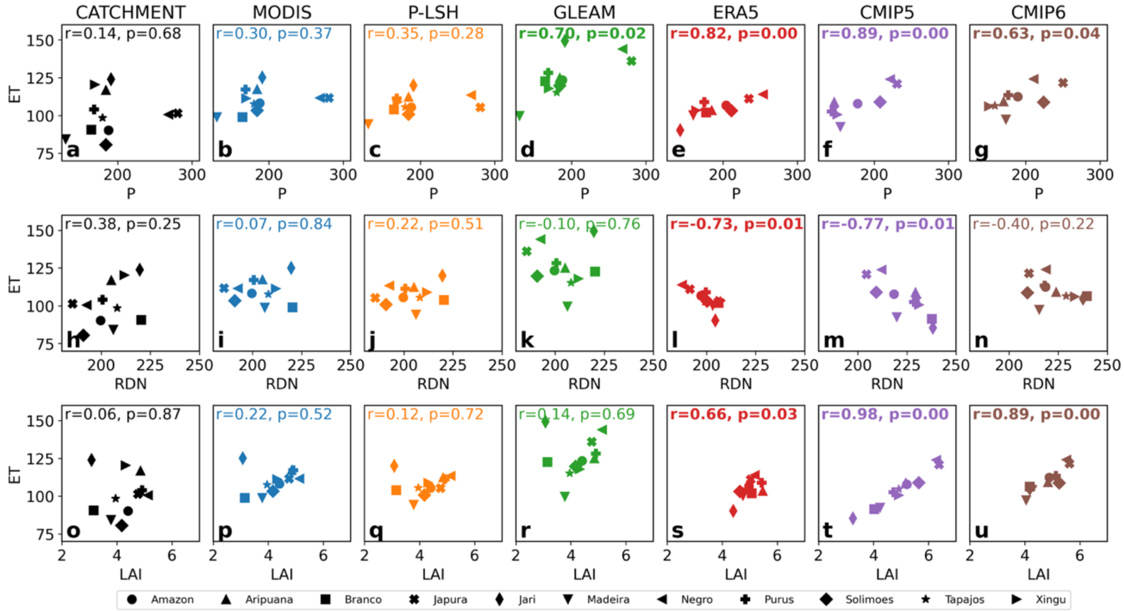

Figure 4 – Controls on spatial variation in Amazon evapotranspiration. Annual mean ET (in mm month^{-1}) for the Amazon and ten sub-catchments (Fig. 1) from catchment-balance, satellites (MODIS, P-LSH, GLEAM), ERA5 reanalysis, and climate models (CMIP5 and CMIP6), plotted against (a–g) precipitation (P, mm month^{-1}); (h–n) surface shortwave radiation (RDN, W m^{-2}); and (o–u) leaf area index (LAI, m^2 m^{-2}). Satellite ET data are plotted against P from CHIRPS, RDN from CLARA-A1 and LAI from MODIS; ERA5 and climate model ET are plotted against ERA5 and model P, RDN and LAI, respectively. Data are from 2003 to 2013, with the exception of CMIP5, where the data are from 1994–2004. Note that the axes do not start at zero.

References

Paredes-Trejo, F. J., Barbosa, H. A. & Lakshmi Kumar, T. V. 2017. Validating CHIRPS-based satellite precipitation estimates in Northeast Brazil. *Journal of Arid Environments,* 139**,** 26-40.