Major

Overall, I appreciate the effort the authors made to clarify nearly every technical point requested by reviewers. However, it does not appear to me that the authors made any of the more major suggested changes to their analyses. Instead, they opted to defend the previous presentation of results. If this is OK with the Editor, it's OK with me. Below I present the critical examples:

1. Instead of showing or demonstrating a sensitivity analysis of their choice for the 50% cumulative discharge threshold, they simply added lines 160–164, which more or less say, "take our word for it". Even a simple discussion of results from a sensitivity analysis would suffice, but as written, the reader must accept the fact that the 10% and 50% choices were best. What about others who want to replicate the study? Should they use these values because they are shown here? I do not ask to be annoying to the authors and I trust they did their due diligence, but I think that by being more critical/quantitative in their choice here, they are also helping other researchers in their work, as well.
2. Similarly, I still do not understand the choice and rationale (Lines 185–187) for using the maximum/minimum as the departure baseline: "We use the maximum or minimum values instead of the median or mean because this approach allowed us to better capture the full range of average metabolism estimates in ways that summarizing pre-storm rates to means or medians would exclude". I apologize if I am being dense, but this is hard to follow. How do maxima or minima better "capture the full range of ***average*** [conditions]" (bold and italics mine) better than means or medians? This illustrates to me that the authors did not take into consideration both reviewers' valid point here, but instead discussed the issue away with one sentence. Would it be so hard to compare the results using mean previous conditions? It looks like it would drastically affect the results, implying strong sensitivity to method (see Figures A5–18).
3. The authors responded to the original suggestion about considering the effects of extremely small changes in magnitude as they relate to measurement uncertainty as follows: "We note that our metrics were indeed detectable relative to metabolism estimates (with low uncertainty, as shown in Fig 3, A4-A18, and supplementary data files). Consequently, we disagree that this approach "raises red flags." I have to disagree here, especially being familiar with uncertainty measurements in GPP and ER. I have looked at the supplementary files and the uncertainty measurements that the authors apparently used: the standard error of the mean. This value will always be extremely tiny because the authors are dividing the standard deviation of estimates by the square root of 2000 (the number of MCMC runs). A better uncertainty measurement would be the standard deviation, the IQR, or the 95% credible intervals, as are most commonly presented for the output of these Bayesian models. It is not reasonable to use the standard error of the mean in this regard and I find this to be a major issue with the approach. For example, the event on 2013–05–23 has overlapping credible intervals for the maximum GPP value on 2013–05–21 (2.2–4.6) and the event (2.5–4.9) And the event on 2013–06–02 has overlapping credible intervals for the maximum GPP value on 2013–05–31 (1.9–4.8) and the event (2.2–4.1). Most events are like this. Can you really distinguish between these with confidence? This is yet another reason why I think the mean of the previous three days would be better. Not to mention, the mean would better capture *ambient* ***equilibrium*** *conditions.*

Minor

Line 96 : Please specify if NO_3 is as N or not.