bg-2020-304 Resistance and resilience of stream metabolism to high flow disturbances

General comments:

The authors analyse the effects of 15 isolated storm events on stream metabolism, focusing on subsidy-stress hypotheses and drivers of response and recovery. The authors make use of a five-year high-temporal resolution dataset to address their four hypotheses. While the general ideas and approach are interesting and worthy of study, the current manuscript requires major revisions before publication. Some of the suggested revisions are substantial changes to analysis and fall more into the "reject and resubmit" category, but the authors of course can provide adequate justification for not conducting these changes.

Specific comments:

Major:

1. After reviewing the approach and the data, and because it underlies all results in the paper, I think the authors need to present a stronger rationale (which currently does not exist)—or change the approach— for how they arrived at their use of arbitrary discharge thresholds (e.g., 50%, 10%) in discriminating the "isolated events". An option could be a simple sensitivity analysis. Additional rationale should support the use of cumulative daily discharge, as opposed to other commonly used metrics in hydrology for event detection. I further think using cumulative discharge may be obscuring some results and indeed missing many events. I first started down this path of inquiry because only 15 isolated events over five years seemed to be a small sample size. Would the same thresholds result in different events if applied to depth or even comparing maximum daily discharge as opposed to cumulative daily discharge? This, I realize, may be a bit of a task because it requires an entirely new analysis, but I think that the authors need to consider this route and defend their assertions more fully. If more events could be used based on a simple adjustment like threshold choice, there could be a much more robust sample size to draw inference from, and would make this a much stronger paper.
2. Similarly, I think this paper could be much stronger by including as many events as possible, regardless of whether they are "isolated". As discussed in the Introduction, the pulsing paradigm/pulsing equilibrium would seem to include the pulses of hydrologic

fluxes, whereas the presented approach discounts them. I understand that perhaps the authors were particularly interested in capturing “resilience” metrics, which may require a period of calm after the storm, so to speak. But, one can imagine a much richer analysis if, for example, the authors calculated some kind of “resistance” metric for as many events as possible, but parsing which ones were preceded by large events. And, for “resilience”, the authors could still calculate the time to return to pre-(initial)event conditions, but just parse which of these “initial” events had subsequent events. Without much effort, the authors could even estimate the subsequent rate of events and its influence on “resilience”. One can imagine a figure of, for example, ΔGPP vs. ΔQ where points are colored by their recovery time and sized by their subsequent rate of events. I mention these suggestions because the current methods seemed disingenuous in taking an arbitrarily “neat” approach to this potentially very fruitful test of the pulsing paradigm. Another important point in this regard is how to take into account *when* a rain event occurs during the day. For example, if a rain event occurs at 23h, it seems like your approach considers its effect for the previous day, when it is probably more appropriate to consider its effect starting for the following day. (This is understood by the authors in their approach in section 2.5.1, but their simple correlation approach does not effectively get at this idea). The current approach also likely discounts many possible events for this reason. I am aware that this may be a big ask of the authors, and, if different routes are taken, I still suggest that they provide stronger rationale for the (apparently) arbitrary decision for identifying events.

3. In the same vein, the authors should provide some kind of justification or sensitivity analysis for both of their critical choices in calculation of their resistance/resilience metrics. The first is the choice of “…three days prior to define a range of antecedent metabolism for each isolated flow event.” (Lines 140–141). The second is the choice of defining “X prior [as] the maximum or minimum value of GPP or ER from the antecedent range…” (Lines 147–148). Why not the median or mean, which would represent more of the “equilibrium” of the previous period? And why not use the most similar previous day in terms of driving forces–in particular, light availability (this seems especially relevant for the recovery interval!). I’m sure the authors considered such options in their initial work, but they need to do more work to convince the audience of

their presented approach—or take a different one if the evidence from sensitivity analyses suggests that they should. Both of these choices are major factors in the subsequent analyses because they define the metrics used, and because these choices do not appear to have literature support/precedent, they need to have clear rationale.

4. Lines 154–156: "To quantify the resilience of GPP and ER, we estimated recovery intervals (RI) by counting the number of days until metabolic rates returned to within the range of pre-event values, signifying a return to antecedent dynamic equilibrium (Figure 3)." This is a good illustration of a potential issue/untapped possibility with the current approach. If you look at the data for the event shown in Figure 3, depth increases in that event by approximately 0.12 m, which decreases light availability by approximately 13% (according to exponential attenuation). This is nearly exactly the difference between GPP on 7 February and 9 February, both of which had nearly identical incoming light signals (making them comparable).
5. Lines 157–160: "To ensure additional flow events did not obscure the recovery interval of GPP or ER, we stopped counting RI the day before the next event (i.e., if another flow event happened four days later, we stopped counting RI at 3 days), and have noted this in our results as days+ and used different symbols in data figures." Why? As far as I can tell, the authors throw these data points out in their analysis, and only reference them in Table 3 (which already uses asterisks to note the issue). Is this to note that the system was on its way to "recovery"? Maybe it would be better to just show a recovery rate, instead of a time, which could result in more data points being included. So, instead of the time it takes to get back to some baseline (which I argue above is a bit arbitrary), you can calculate the rate of increase in GPP over a period (which could be equal to the baseline period that you settle on). Let's say an event occurs and on that day GPP was 5 g O_2 m^{-2} d^{-1}; the subsequent days maybe it's 8, then 10 g O_2 m^{-2} d^{-1}. The rate of increase could then be 2.5 g O_2 m^{-2} d^{-2} ((10 – 5) / 2). Then, even if a subsequent event occurs, you can still compare the rate of increase before that event. A rate also seems like it could be more comparable/scalable across systems in contrast with a number of days. I don't presume to have the best idea here, but I think an approach like the one outlined above could increase inclusion of useful data points, and thereby lead to more useful inferences.

6. Lines 165–166: "We assessed three categories of potential predictors of metabolic resistance and resilience: antecedent conditions, characteristics of the isolated flow event, and characteristics of the most recent prior flow event." Antecedent conditions and characteristics of the recent prior flow event (especially the latter) are unrelated to any stated hypothesis and appear to come out of nowhere. There needs to be clear rationale in the Introduction that leads the reader to understand why you are doing this.
7. Generally speaking, I had difficulty with the entire Results section, which I think needs a complete rewrite. Some specific details are presented below, but I glossed over several in the interest of time. This section needs to link to stated hypotheses (in the order that they are stated in the Introduction) and test them directly without including spurious tests and weak assertions.
8. Figure 5 as presented is not informative. What do the authors want the reader to understand from this figure? Is the R^2 based on a linear regression for all of the points or just the black circles? What is the slope of the regression and the p-value? How does the slope compare to the 1:1 lines? The second panel (right, ER vs GPP recovery interval) is not related to any stated hypothesis. The text discussing this Figure does not support the points on the figure, particularly for the high stated value of ER stimulation = 0.22 (Lines 189–190: "…The magnitude of departure for ER (M ER) ranged from -0.59 to 0.22, with a median of 0."). Looking at this figure also raises red flags about how the authors defined stimulation/repression. How do the extremely small changes in magnitude shown here compare to the uncertainty in GPP/ER, which are never discussed or propagated through any of these analyses? For example, is a 1% increase (i.e., M = 0.01) detectable if uncertainty is considered? The authors should improve this figure substantially, or remove it/leave it as a table. One possible idea is to color or size points based on the event size. Moreover, based on this figure, I am not sure I believe the results on Line 193–195 (italics mine): "Although GPP exhibited stronger responses across isolated flow events than ER, M GPP and M ER were positively correlated (R 2 = 0.39, p = 0.007, Figure 5) *and not significantly different (p = 0.06, α = 0.05)*." Just an eyeball test makes this seem unreasonable. ER magnitudes on average are about 0
9. Figure 6 is not easily understood and appears to simply repeat the information on Table 4 in a cluttered way. What key piece of information is the reader supposed to understand

from this? The results of the controls on process response in this section 3.3 is quite difficult to connect with any prior hypotheses and leaves the reader uninformed. There are two figures and a table with only six sentences to describe them in this section. One of the stated hypotheses (H2) is never even formally tested here, and only the resistance metric is tested for H3 (somewhat, in Figure 7), which included both resistance and resilience metrics.

10. As far as the Discussion and Conclusion, I have many comments, but the issues all stem from previous issues relating to hypotheses, methods, and results. If the authors apply any of my suggested revisions to their approach, they will inevitably have to rewrite these sections. So, I have not provided many specific comments out of the interest of time, but a few key ones are here. I again suggest to organize the Discussion (and entire document) in order of the hypotheses as they are presented and as makes logical sense. As written, the Discussion jumps around in its assertions and ideas. Finally, much of the content in these sections is hypothetical and rhetorical, with little critical analysis of the results actually presented in the manuscript and how they relate to the broader literature.

Minor:

1. Ideas of pulsing steady state could be clarified a bit with regard to the study design and terminology throughout. In the Introduction, the authors note "Frequent disturbances generate oscillations that form a pulsing steady state (sensu Odum et al. 1995) that includes ambient variability in processes (Resh et al., 1988; Stanley et al., 2010)." (Lines 21–23). So, flow disturbance regime defines the pulsing steady state of lotic systems. But, the authors then use–incorrectly I think–the periods outside of flow disturbance to define a "pulsing steady state" (or at times, "pulsing equilibrium", like in Figure 1, and "dynamic equilibrium", like in Figure 3, and "antecedent equilibrium" on Line 187), to which they then compare to periods *with* flow disturbance. The approach is clear, but there is some circular reasoning with respect to the definition of pulsing steady state. I recommend perhaps using different terminology for these two concepts. One idea could be to use something like "ambient equilibrium" for metabolism under baseflow conditions, and "pulsing equilibrium" to refer to the larger scale, (inter)annual behaviour of as originally conceived by Odum. I think these small changes would improve the clarity of the study design and arguments within.

2. Similarly, I do not think that "resilience" is appropriately used throughout the manuscript, first defined by the authors on Lines 59–60: "We can also quantify post-disturbance ecosystem responses by estimating resilience: the time it takes for a process returns to equilibrium following a disturbance (Carpenter et al., 1992)." We have first of all the issue of "return[ing] to equilibrium", which is not so clear based on the previous definition of a pulsed equilibrium that includes disturbance. In a system organized by regular disturbance regimes, the idea of resilience to that same disturbance regime is a bit convoluted. In contrast, the idea of a "recovery interval" to previous ambient conditions is clear and appropriate. Resilience in this context might make more sense if there were alternative metabolic equilibria that the stream could occupy, where each of these equilibria were tolerant to different levels disturbance. Ultimately, this is a choice of language and does not affect the analyses presented and if the authors opt to keep their current choice, I suggest spending some more time to expand these ideas/defend their use out in the Introduction and Discussion.
3. Line 73: "(H0) some flow events will not push GPP and ER outside of their pulsing equilibrium." Should this by "(H4)"? Or is this some kind of null hypothesis? Consider renumbering, or placing this at the beginning of the sentence–seems strange to go from 1–3 then back to 0.
4. Lines 70–71: "…(H2) there will be a stimulation of GPP and ER at intermediate flow disturbances due to an influx of limiting carbon and nutrients…". Is this stream known to be limited by carbon and nutrients? What is the timeframe for stimulation? It seems like the influx of carbon and nutrients would pass through the system quite quickly in this small stream, and would not be easily acquired/processed by organisms. In larger systems with long recession curves, I think this perspective can make sense, but this hypothesis does not seem well supported in the Introduction as currently written.
5. Lines 71–72: "…(H3) metabolic resistance and resilience will change with the size of the event, with larger flow disturbances inducing more stress due to enhanced scour…" The point about scour here seems important. Scour is a function of shear stress, which itself is a linear function of depth. The authors focus on discharge as their subsidy/stress driver, but I wonder if water depth would be more appropriate? Because depth only increases to the square-root of discharge (for a large range of depth-discharge in their Supplemental

data), a quadrupling of discharge only results in a doubling of benthic shear stress. I don't expect for the authors to redo any analyses with this perspective, but I do think this kind of information would be useful to include especially in the Discussion so that future works would consider this as well. It also could be used as a future framework to further test the idea of subsidy/stress balance. Depth is a first-order control on both light availability and shear stress at the benthos, making it a more appropriate indicator of stress than discharge.

6. The light data (first referred to on lines 92–93) appear to be in units of µA according to the supplementary material ("ODonnellHotchkiss_SuppData_ReadMe.pdf", under point "1"). I am not familiar with this unit (is it micro-amperes?) for sunlight, and I think this needs some clarification. The light data in the data file itself appear to range between 0 and 1, but the *streammetabolizer* model take data in PAR (units µmol m^{-2} s^{-1}), which can be upwards of 1000 by noontime. I'm sure this is not a major issue, but I do not think the results will be replicable as currently presented—those units, if fed into *streammetabolizer*, will lead to very strange outputs I think. The sensor used (according to O'Donnell and Hotchkiss 2019) is a Campbell CS300, which should output data in typical units like W m^{-2}.
7. Line 140: "To acknowledge the pulsing, day-to-day variability…" I don't think "pulsing" is appropriate or needed here.
8. Line 152: "…suppression…" please check for the consistent use of suppression and repression (and others) throughout.
9. In section 3.2 "Metabolic resistance and resilience", it would be very helpful to explicitly organize/label these paragraphs according to your numbered hypotheses from the Introduction. For example, Lines 187–192: There is no directly stated connection between any of the statements presented here and the actual hypotheses.
10. Lines 194–196 bring up another issue with the idea of "magnitude" (italics mine): "*M GPP was less than M ER for nearly all flow events*, except for one in which M GPP and M ER were both zero and two where M GPP and M ER were both small (Figure 5, Figure 195 A19)." The general idea of magnitude is that is not directional. I would argue that the magnitude of GPP response was *greater* than that of ER, and that they both had

similar directional change (decrease in process magnitude). Consider different language throughout.

11. Lines 198–199: "Similarly, the only other event that stimulated GPP (M GPP = 0.03) had no ER response, suggesting many flow disturbances may decouple GPP and ER." This seems like an unsupported assertion (which should be in the Discussion, if anywhere) based on one event with an extremely small signal.
12. Table 3: n/a is not clearly defined.
13. Lines 208–209: "Although GPP and ER are linked processes, the variables that were moderate or strong predictors of resistance or resilience ($r > 0.5$)." Why is 0.5 the threshold for being a strong predictor? That's only 25% of the variance explained.
14. Lines 210–211: "Because the median RI ER was zero, bivariate correlations could not be used to determine potential predictors of ER resilience." Another reason to consider rate instead of day count.
15. Lines 214–215: "Overall, there were multiple environmental controls on metabolic resistance or resilience that were strongly correlated with either GPP or ER, but no significant drivers of both GPP and ER resistance and resilience." This is not supported by the figure or the table.
16. Line 219: "Notably, ER was more resistant than GPP (Figure 1)." Figure 1 is a conceptual figure and does not support this statement.
17. Line 239–240: "In assessing metabolic responses and recovery from smaller flow events relative to the dynamic equilibrium of metabolism at baseflow, we found some of the shortest metabolic recovery intervals recorded in the literature (Figure 8; Table A1)." Do these other studies use the exact same methodology as you? How are they comparable? Are they similar sized streams? You should compare and contrast more here.
18. Line 259–260: "Contrary to our predictions, the size of the most recent antecedent flow disturbance had a positive relationship with M GPP and M ER (Figure A19)." Where is this prediction?

19. Technical corrections:

1. Equation 1 (Line 110) seems boiler-plate and unnecessary.
2. There are extra parentheses in Figure 2c description for "((m^3 d^{-1}))"
3. Figure 3 should describe what the error bars are on the GPP estimates.

4. Lines 163–164: "Quantifying how different antecedent conditions induce variable responses from GPP and ER is critical to furthering our understanding of stream ecosystem responses to flow disturbances." This belongs in the Introduction, not the Methods.
5. Lines 167–168: "Antecedent medians for turbidity were estimated from seven days prior due to missing sensor data." This is not clear, please explain what this means. There was always missing data for turbidity within the three days prior to an event? I can't imagine turbidity changes very much at baseflow.
6. Lines 190–191: "Three of 15 flow events stimulated ER, 5 repressed ER, and ER did not deviate from the antecedent equilibrium for 7 events (i.e., M ER was 0)." It's more common to use numerals for numbers greater than 10, and to spell the numbers out for numbers less than 10.