We respond to the three major points raised by reviewer 1 below after the original text from reviewer 1 (in bold). We have not included a response to the editor because they did not highlight specific topics to focus on or identify additional needs for revisions beyond those listed by reviewer 1.

**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.**

**RESPONSE:** We appreciate that sensitivity analyses are useful tools in method development and validation, but it is unclear what the reviewer is asking for in a sensitivity analysis or different type of storm threshold cutoff and what would resolve their concern in a revised manuscript. Our limited understanding of reviewer vision based on what they wrote requires us to (a) start over with all analyses but without guidelines for comparison when deciding which threshold choice or (b) quantitatively justify to the reviewer why we kept previous thresholds if we don't re-do all analyses to include smaller "storms" or a different definition of baseflow for prior day calculations (different than <10% change in Q on prior days). Because we do not have a clear direction from the reviewer or a methods precedent in the literature, we have respectfully chosen not to re-do our analyses in one or more ways that may or may not satisfy reviewer questions.

Our early analyses of different flow thresholds were not at the level of doing a full analysis to delineate storms and analyze metabolic resistance and recovery for different delineation thresholds, but we did test how different % discharge (Q) change thresholds altered the number of isolated storm events we would be able to analyze. That preliminary analysis increased our storm threshold to 50% Q change to include more isolated storm events and added the requirement of less than 10% Q change during what we categorized as baseflow pre-storm days to remove days with smaller storms prior to larger storms from our analysis. For the 15 storms analyzed, we captured a wide range flow changes: 53 - 1105% change in cumulative daily Q.

We also note that there is not a standardized method for characterizing metabolic responses to flow disturbances, and we chose distinct cutoffs based on our ***site-specific*** understanding of changes in hydrology during baseflow and higher flows. As one example of this, we present the relevant methods from Reisinger et al. (2017, Ecosphere), one of the few other manuscripts that

analyzed storm-induced changes in stream metabolism after Hurricane Sandy and a subset of smaller storms: "*For the storm events using the four additional sites, we sorted through a daily discharge and metabolism record spanning April–November 2015 to identify high-flow events with enough time at baseflow between events to allow for metabolic recovery. We only included events with at least four days of baseflow prior to the flood with relatively stable GPP and ER, coupled with enough time following the flood for GPP and ER to recover to pre-flood rates. We selected nine stream–storm events to include in addition to the Sandy data.*"

Given what we present above, past responses to reviewers on this topic, and a brief communication with the editor, we did not go through a total re-analysis of all of our data without a specific justification, resource, or expected outcome for this suggestion. We sincerely hope that our revised manuscript will not be penalized by this choice. We were simply not in a position to start over without clear guidance or justification and could not see how this exercise would be aligned with our objectives for this work.

**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).**

**RESPONSE:** In attempt to clarify this approach and our motivation for this approach in a different way: we used the highest (maximum) and lowest (minimum) mean estimate of daily GPP and ER during the 3+ pre-storm baseflow days. We did this because day-to-day variation in metabolism is high, and we believed that taking the mean or median of these 3+ days of median GPP and ER was doing a disservice to the characteristics of our study site.

However, based on the above request from reviewer 1 and to keep with older methods used by other researchers, we reanalyzed metabolism changes during (M, magnitude of departure) and after (RI, recovery interval) storms compared to the mean GPP or ER during the previous baseflow period (i.e., mean GPP over 3 days or mean ER over 3 days). Below we show updated plots (Fig 3,5,6,7 and appendices), tables (table 3,4), and text.

We revised the methods text for "MaxMin" or "Mean" in section 2.5 (lines 270-319 in the revised manuscript):

**"2.5 Characterizing metabolic resistance and resilience**

"To acknowledge the ambient day-to-day variability of GPP and ER, we used metabolism estimates from three days prior each isolated flow event to calculate a mean value of antecedent metabolism. We quantified metabolic responses to flow disturbances by comparing the pre-event metabolic means with event and post-event metabolism rates. To assess resistance, we estimated the metabolic magnitude of departure (M) during events to quantify the resistance of GPP and ER to higher flow disturbances. We calculated M per isolated flow event by comparing the difference between GPP and ER to the nearest value of the antecedent range (Equation 3; Figure 3),

[Eqn 3 – code copied from LaTeX version for track change]

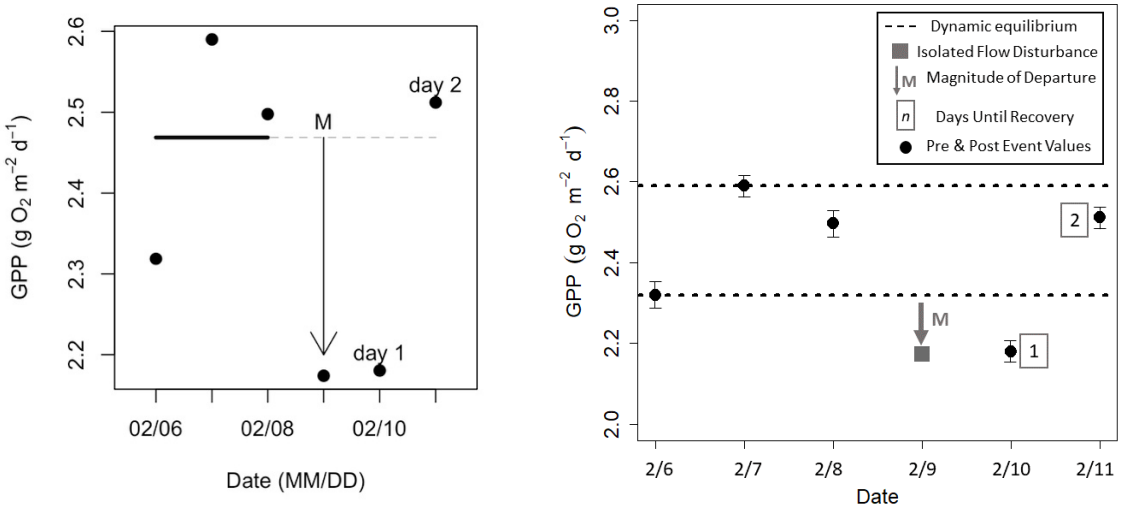
M = 1 - \frac{X_{event}}{X_{prior}}

where $X_{event}$ is either GPP or ER (g $O_2$ $m^{-2}$ $d^{-1}$) on the day of the isolated flow event. $X_{prior}$ is the mean value of GPP or ER from the antecedent range, and whether M is positive or negative depends on if the isolated flow event resulted in a stimulated (increased) or suppressed (reduced) metabolic response. For instance, if GPP declined during a flow event, M was calculated as the difference between GPP for the isolated flow event and the mean GPP from the antecedent 3-day range (Figure 3). If GPP or ER on the event day did not fall above or below the antecedent mean, M was zero, thus indicating high resistance. A negative M represents a suppression, and a positive M a stimulation, of GPP or ER relative to the antecedent mean.

To quantify the resilience of GPP and ER, we estimated recovery intervals (RI) by counting the number of days until metabolic rates returned to or exceeded pre-event mean GPP or ER, signifying a return to antecedent conditions (Figure 3). If metabolism (mean and 2.5-97.5% credible intervals) during the isolated flow event did not fall outside of the antecedent mean, the RI was zero days (metabolism cannot recover if it never shifts outside ambient values). 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 three days). To test for statistically significant differences between ER and GPP recovery intervals ($RI_{ER}$ and $RI_{GPP}$) and ER and GPP magnitude of departure ($M_{ER}$ and $M_{GPP}$), we ran Welch's t-tests in R (R Core Team, 2018)."

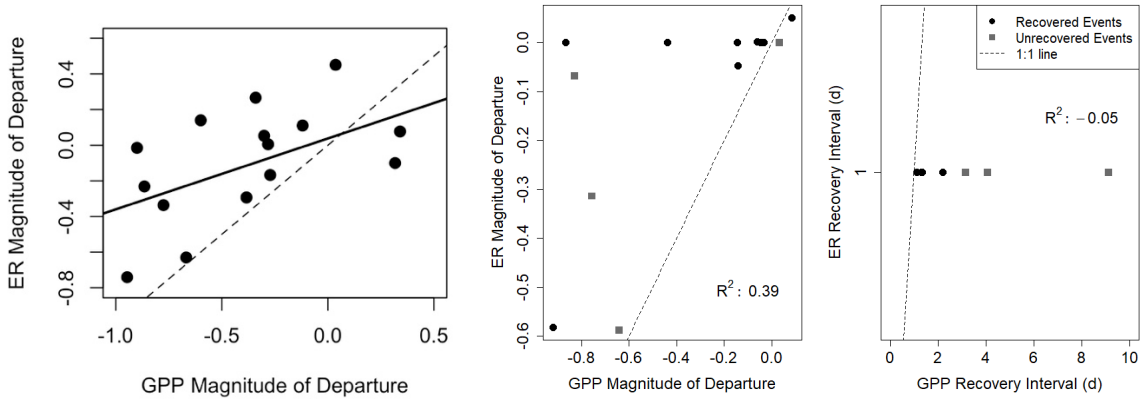
We updated the following figures, tables, and text to include results based on analyzing M and RI compared to a 3-day mean instead of highest or lowest mean GPP or ER prior to each storm. We also used this as an opportunity to improve some of our data visualizations.

(1) Figure 3 now removes the window of baseflow variability from our calculations (original dashed lines) and shows the calculation of M and RI relative to median prior GPP. Updated/original figure are below for comparison.



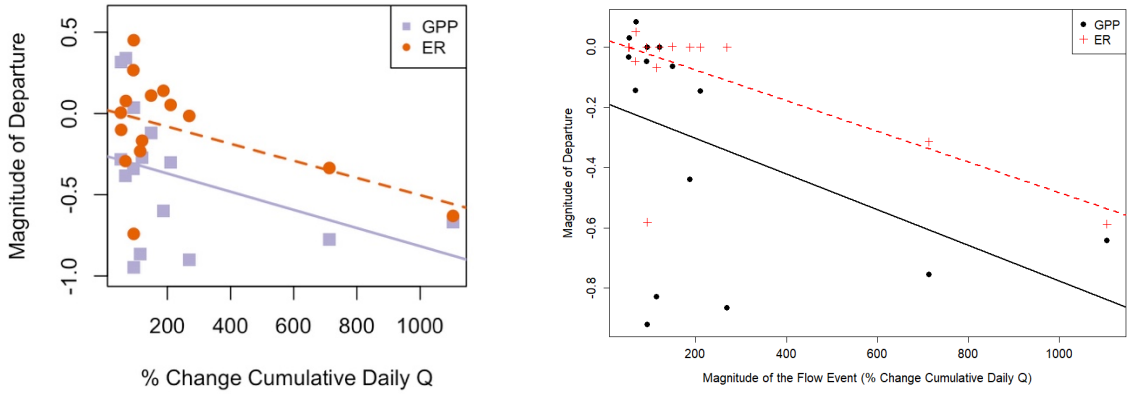

Above: Revised Figure 3 (left panel) illustrating the method requested by reviewer using the median instead of the min/max daily median (original figure showing this method is the right panel).

(2) Figure 5 now includes M-GPP and M-ER calculated using the "mean" approach instead of the "MaxMin" approach. Updated/original figure are below for comparison.



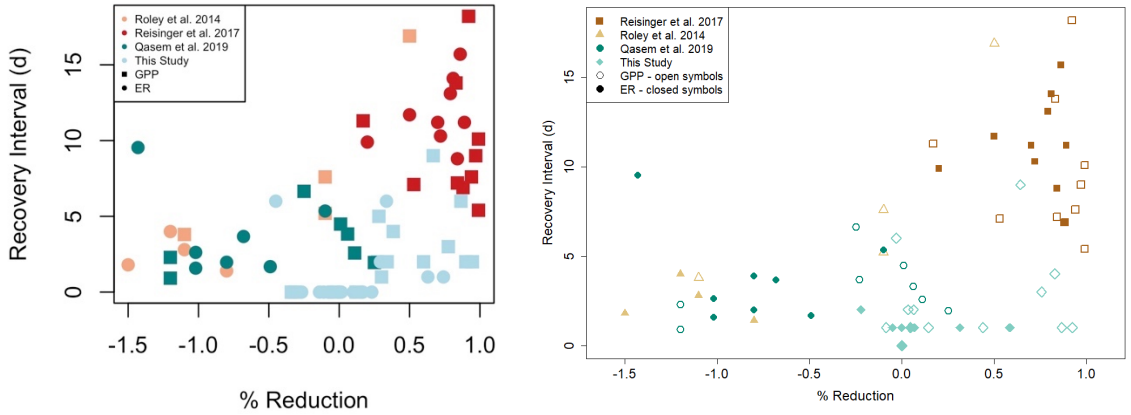

Above: Revised Figure 5 (left panel) with updated approach to calculate M. Original figure with these results, including M and RI (removed as requested by reviewer) is shown in the righthand two panels.

(3) Figure 6 now includes % Q change versus updated estimates of M-GPP and M-ER calculated using the "mean" approach. Updated/original figure are below for comparison.




Above: Revised Figure 6 with updated M from revised method (left panel) in comparison to original figure showing these results (right panel).

(4) Figure 7 now plots our data as RI from the "mean" approach, not "MaxMin". Updated/original figure are below for comparison.




Above: Revised Figure 7 (left panel) with RI calculated from mean prior GPP and ER, not min/max (as in original figure in right panel).

(5) Table 3 now includes updated estimates of M, RI calculated using the "mean" approach, not "MaxMin". Updated Table 3 is below:

**Table 3.** Magnitude of departure (M, unitless) and recovery intervals (RI, days) of gross primary production (GPP) and ecosystem respiration (ER) during and after fifteen isolated flow events between 2013-01-08 and 2018-04-14. A negative M represents a suppression, and a positive M a stimulation, where GPP or ER increase relative to the prior mean GPP or ER calculated over three days. Estimates of M differed between GPP and ER ($t(26.3)=2.15$, $p=0.04$), while the RI for GPP and ER were not significantly different ($t(25.8)=-1.22$, $p=0.23$). The two instances where GPP did not recover during the isolated flow event analyzed are noted with an "NA" and the number of days without recovery (X+) that could be counted before the next high flow event occurred.

| **Date** | **$M_{GPP}$** | **$RI_{GPP}$ (d)** | **$M_{ER}$** | **$RI_{ER}$ (d)** |
|---|---|---|---|---|
| 2013-03-12 | -0.78 | NA (6+) | -0.34 | 6 |
| 2013-03-31 | -0.60 | 2 | 0.14 | 0 |
| 2013-05-23 | 0.34 | 0 | 0.08 | 0 |
| 2013-06-02 | -0.34 | 2 | 0.27 | 0 |
| 2015-02-02 | -0.30 | 1 | 0.05 | 0 |
| 2015-05-17 | 0.04 | 0 | 0.45 | 6 |
| 2015-09-03 | -0.27 | 2 | -0.17 | 0 |
| 2016-04-01 | -0.38 | 4 | -0.29 | 2 |
| 2016-04-07 | -0.28 | 5 | 0.01 | 0 |
| 2016-04-22 | -0.87 | 6 | -0.23 | 0 |
| 2016-08-21 | -0.95 | 2 | -0.74 | 1 |
| 2017-02-09 | -0.12 | 0 | 0.11 | 0 |
| 2017-08-21 | -0.67 | NA (9+) | -0.63 | 1 |
| 2017-09-06 | -0.90 | 2 | -0.01 | 0 |
| 2017-10-16 | 0.32 | 0 | -0.10 | 0 |
| Average | -0.38 | 2.5 | -0.09 | 1.1 |

(6) Table 4 now includes results from updated correlation analyses based on M and RI calculated using the "mean" approach, not "MaxMin". Updated Table 4 is below:

**Table 4.** Pearson correlations (r) between predicted drivers of gross primary production (GPP) and ecosystem respiration (ER) magnitudes of departure (M) and recovery intervals (RI) of isolated flow events. Predictor variables with moderate or stronger relationships (r > 0.5; Hinkle et al. 2003) are bolded. p-values are included in parentheses.

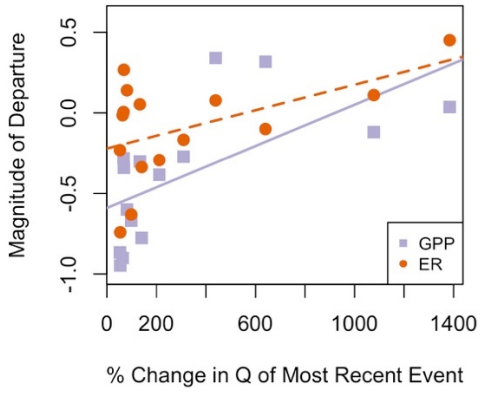
| Predictor variable | r, $RI_{GPP}$ | r, $M_{GPP}$ | r, $RI_{ER}$ | r, $M_{ER}$ |
|---|---|---|---|---|
| **Isolated flow event of interest** | | | | |
| Daily median light | 0.19 (0.51) | 0.17 (0.55) | -0.10 (0.74) | -0.06 (0.84) |
| Daily peak discharge | **-0.65 (0.01)** | -0.23 (0.42) | 0.36 (0.18) | -0.39 (0.15) |
| Daily median temperature | 0.10 (0.72) | -0.02 (0.94) | 0.00 (1.00) | -0.29 (0.30) |
| Event median discharge (Q) | 0.14 (0.63) | -0.13 (0.65) | 0.50 (0.06) | 0.09 (0.75) |
| % change in Q during event | **0.71 (0.00)** | -0.40 (0.14) | 0.30 (0.28) | -0.49 (0.07) |
| Season | 0.02 (0.93) | -0.10 (0.73) | -0.19 (0.50) | -0.27 (0.34) |
| Time of peak Q | 0.14 (0.61) | -0.06 (0.82) | 0.07 (0.81) | -0.04 (0.89) |
| Turbidity | 0.46 (0.13) | -0.41 (0.19) | 0.26 (0.41) | -0.07 (0.83) |
| **Most recent flow event** | | | | |
| Days since last event | 0.05 (0.86) | -0.07 (0.82) | -0.12 (0.67) | -0.08 (0.78) |
| Last event cumulative daily Q | -0.40 (0.14) | 0.49 (0.06) | -0.21 (0.45) | 0.14 (0.62) |
| % change in Q during last event | **-0.56 (0.03)** | **0.63 (0.01)** | 0.38 (0.16) | **0.51 (0.05)** |
| **Antecedent conditions** | | | | |
| Antecedent GPP | **0.62 (0.01)** | **-0.54 (0.04)** | 0.13 (0.64) | -0.29 (0.29) |
| Antecedent ER | -0.21 (0.46) | 0.00 (1.00) | 0.21 (0.46) | 0.33 (0.23) |
| Antecedent median gas exchange | 0.26 (0.36) | -0.07 (0.81) | -0.11 (0.70) | -0.41 (0.13) |
| Antecedent median light | 0.06 (0.82) | 0.03 (0.92) | 0.06 (0.83 | 0.26 (0.36) |
| Antecedent median Q | -0.22 (0.41) | 0.44 (0.10) | 0.21 (0.44) | 0.47 (0.08) |
| Antecedent median water temperature | 0.07 (0.79) | -0.02 (0.95) | -0.09 (0.75) | -0.29 (0.29) |
| Antecedent median turbidity | 0.12 (0.69) | -0.02 (0.95) | 0.11 (0.71) | -0.29 (0.34) |

(7) We altered text providing ranges in results that reflect the updated analysis. We also removed text stressing the importance of accounting for day-to-day variability when testing the responses of metabolism to flow changes using our proposed "MaxMin" approach, and instead of highlighting the novelty and utility of our "MaxMin" approach for hydrologically variable streams, now discuss that including metabolism estimates from multiple days prior to the storm will be important for future work on this topic given day-to-day variability in metabolism.

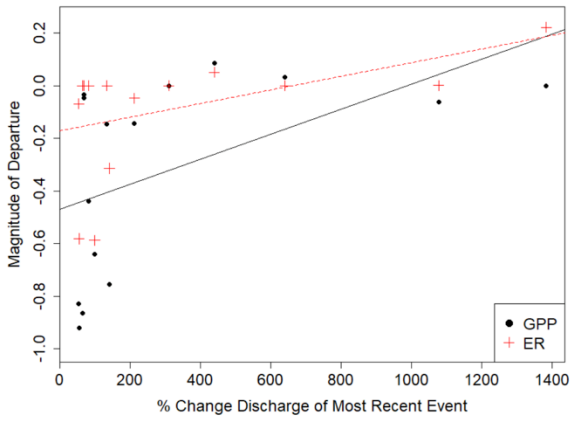
(8) Appendix plots were updated to include (a) revised M and RI as described above and (b) storm-specific plots with GPP and ER credible intervals, as described below in the response to point 3 below.

Updated Figures A04-A18 are included below with response #3.

Updated Figure A19:




Original Figure A19:




Finally, we removed Figures A20 and A21 to focus on the presentation of keys based on prior reviewer comments and the trajectory of this manuscript which no longer referenced either figure after revisions.

**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.**

**RESPONSE:** Thanks for catching this. The appendix plots of flow and metabolism for each isolated flow event was using the SE of the posterior mean, not the credible intervals. While error bars on our appendix plots to not change any of our calculations and data interpretations, they absolutely should have been credible intervals. Because we updated how we calculated M and RI based on point 2 above, we: updated these plots to replace the min/max lines that reflected our initial approach with dashed lines for prior mean GPP and ER; added a vertical

dashed red line to note the high flow day in all three panels; left blank days with an incomplete dataset to track discharge, ER, and GPP; and added 95% credible intervals to GPP and ER means to honor the Bayesian parameter estimation approach used in StreamMetabolizer. We also used this as an opportunity to improve the overall presentation of these graphs and show cumulative daily discharge with daily metabolism estimates (instead of higher-frequency instantaneous values shown in earlier plots in the manuscript). Original and revised plots for each of the 15 isolated flow events are below.

Figure A04 - Storm 1:

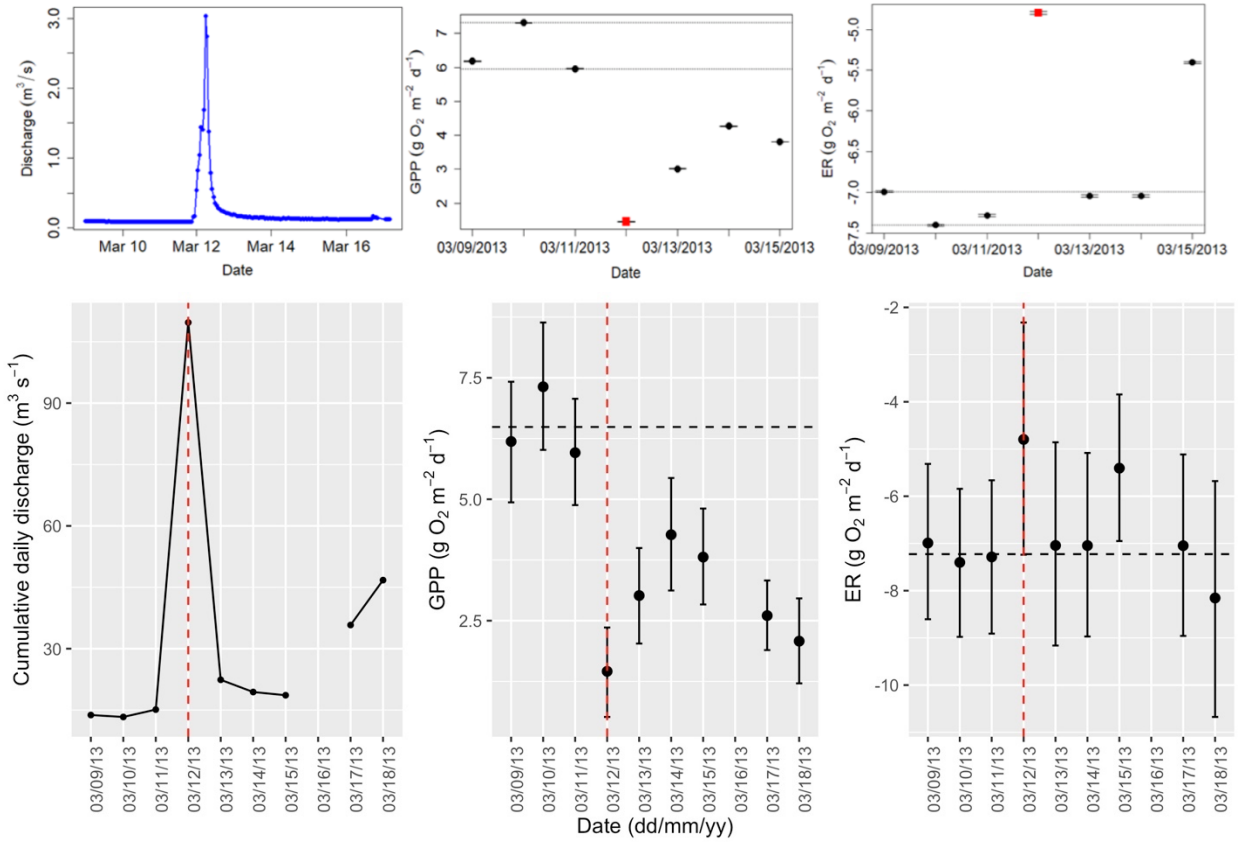
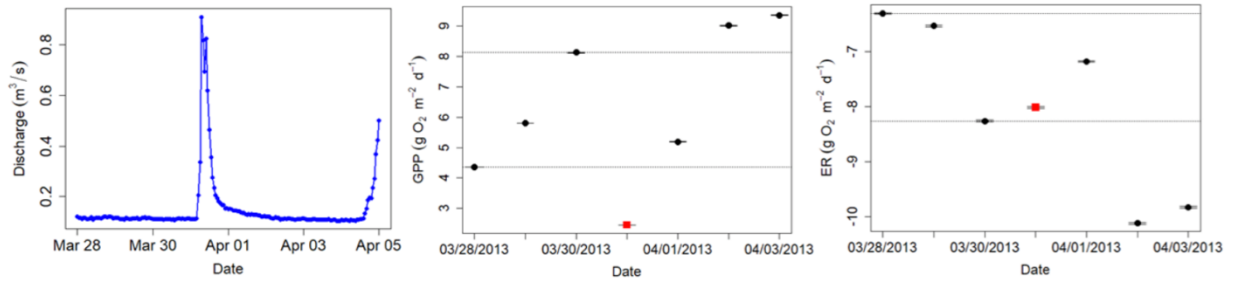



Figure A05 - Storm 2:



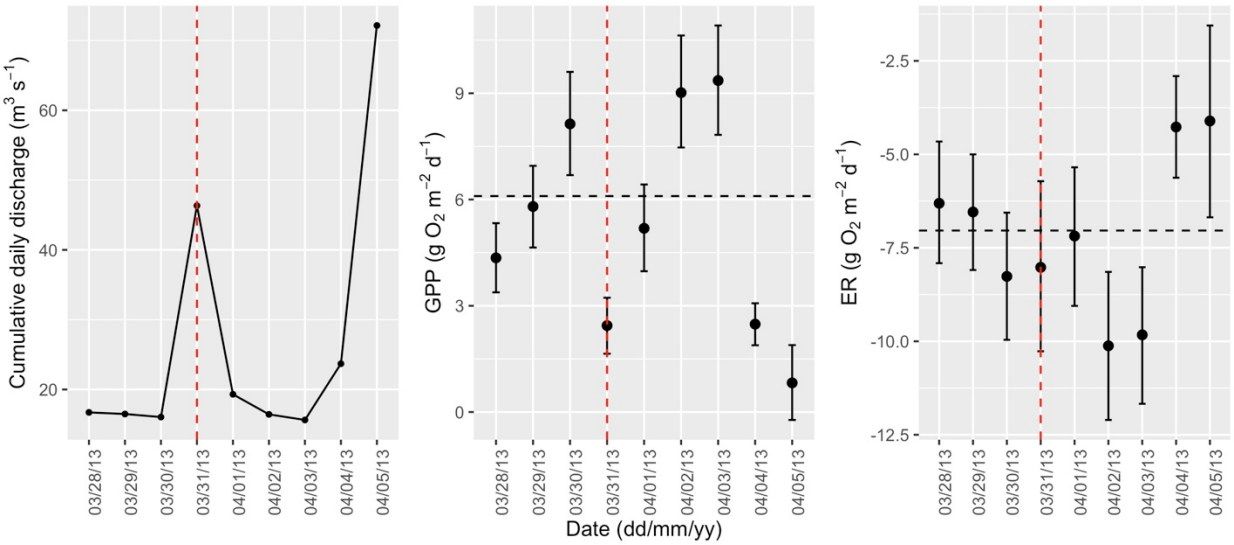



Figure A06 - Storm 3:

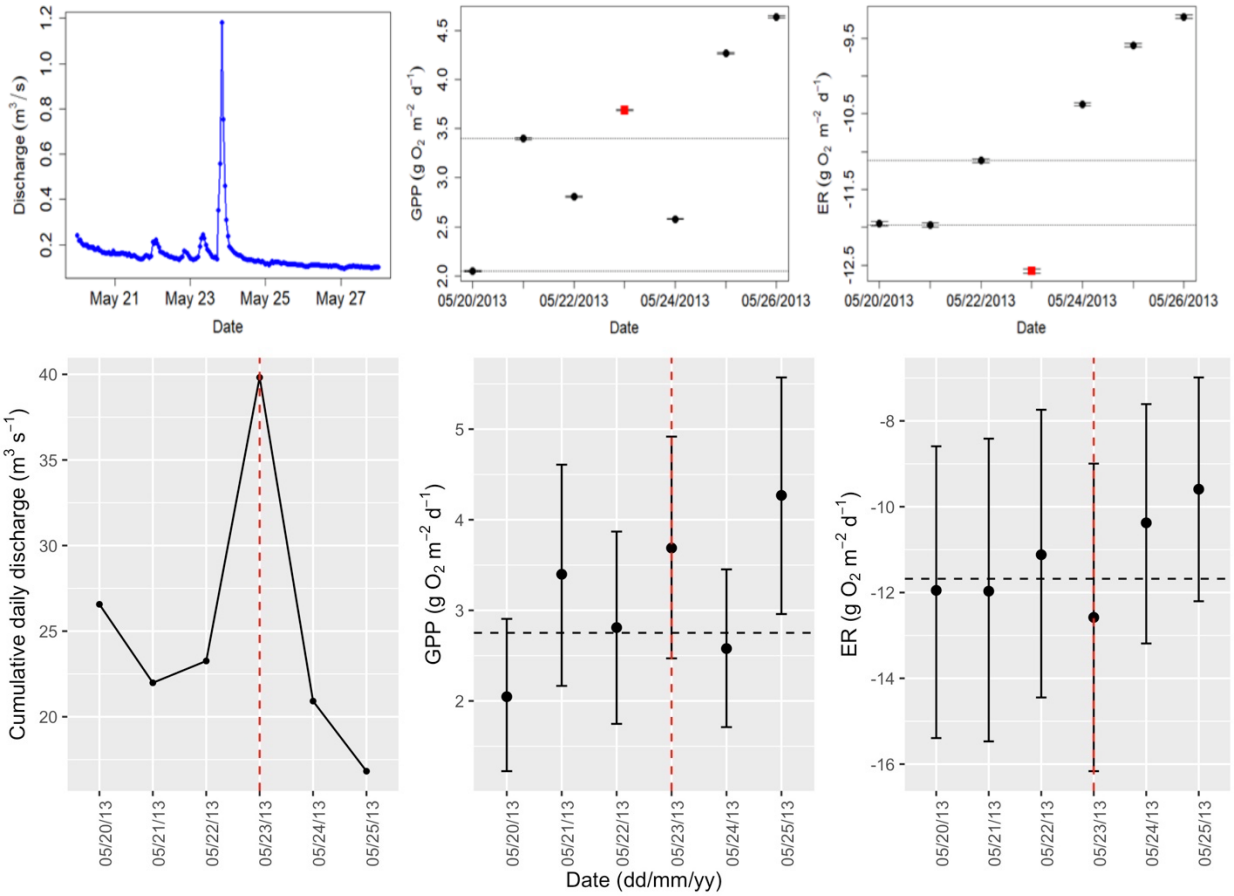



Figure A07 - Storm 4:

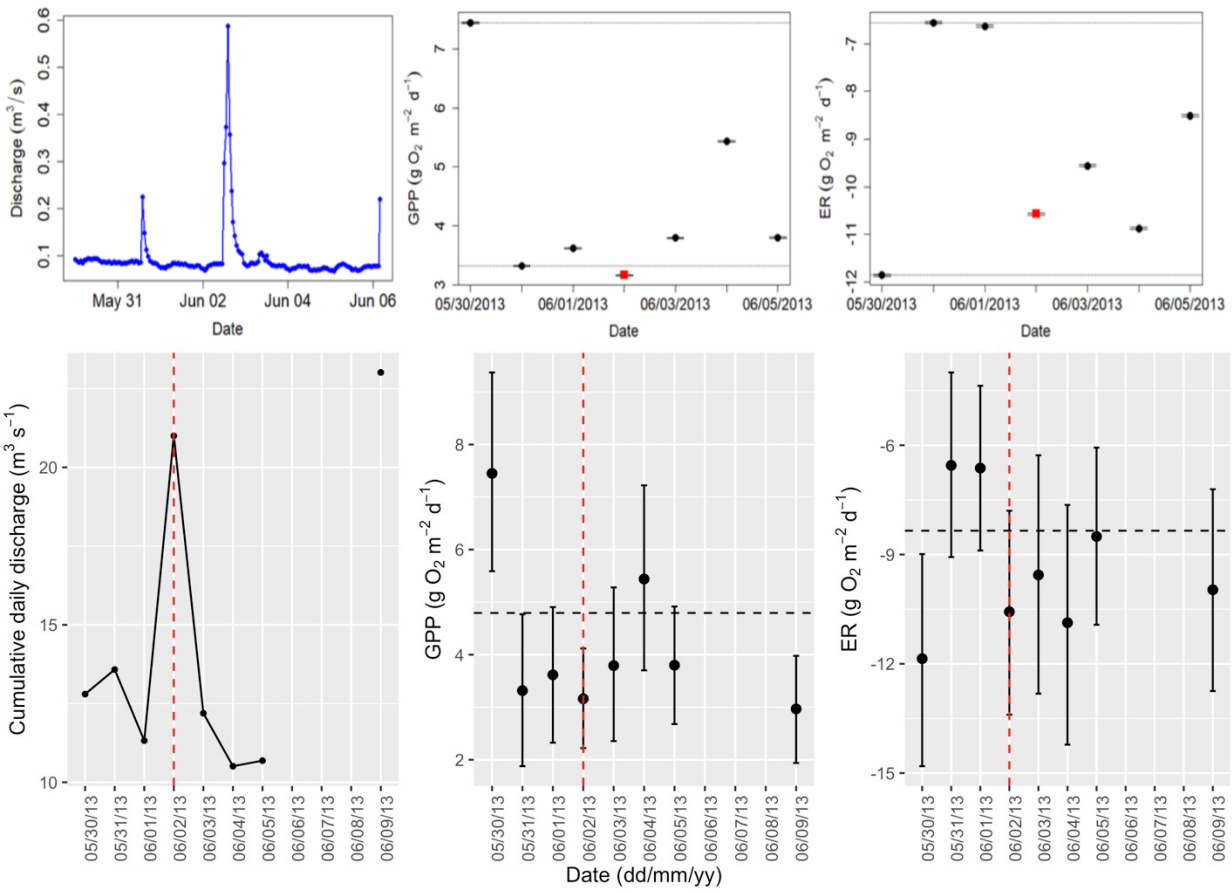
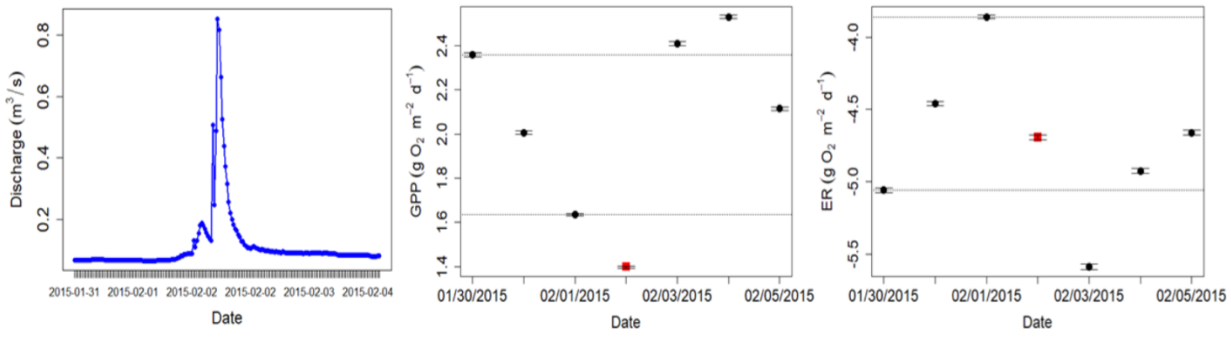



Figure A08 - Storm 5:



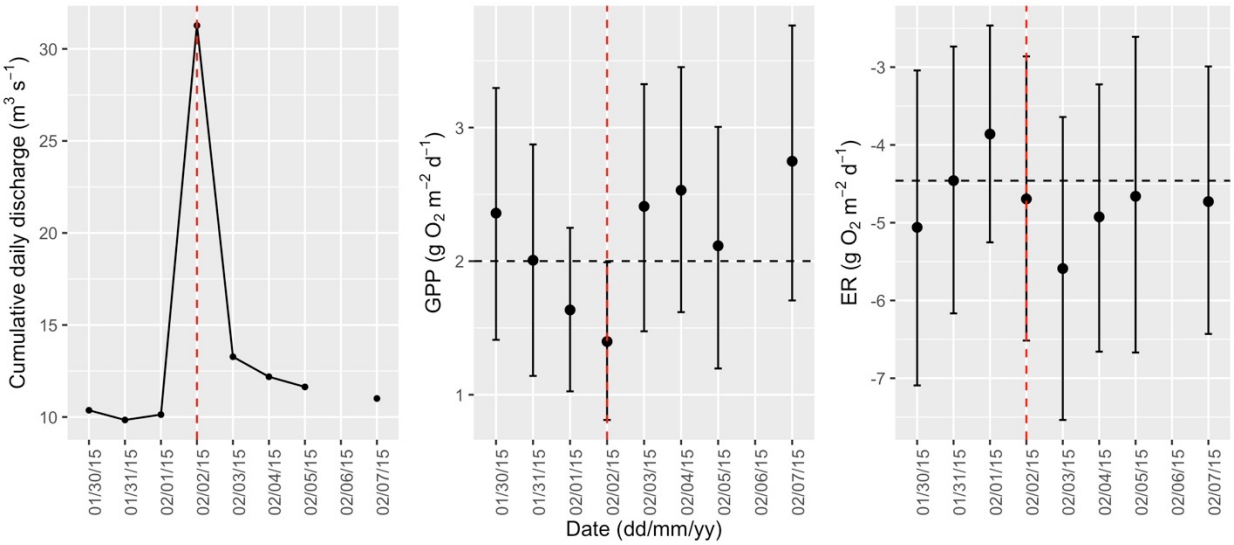



Figure A09 - Storm 6:

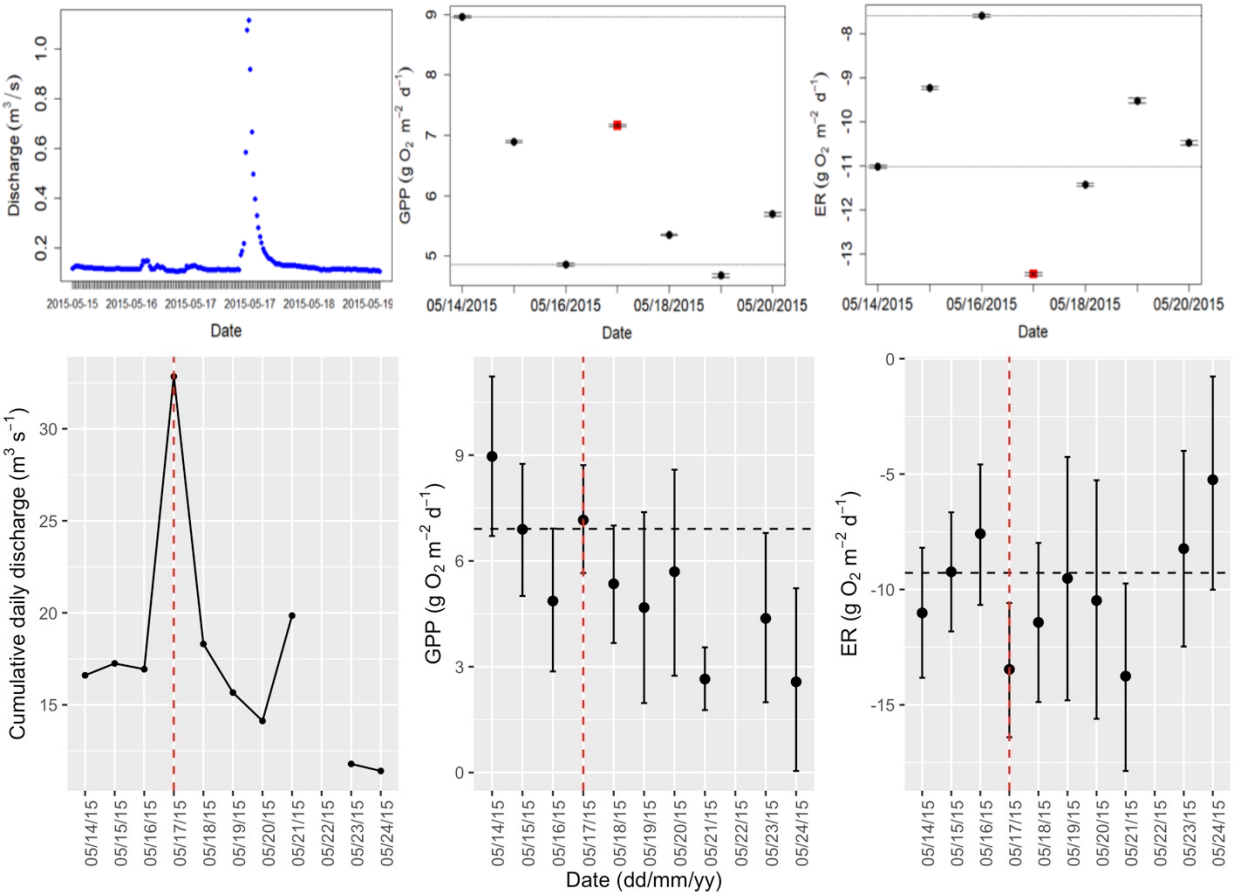



Figure A10 - Storm 7:

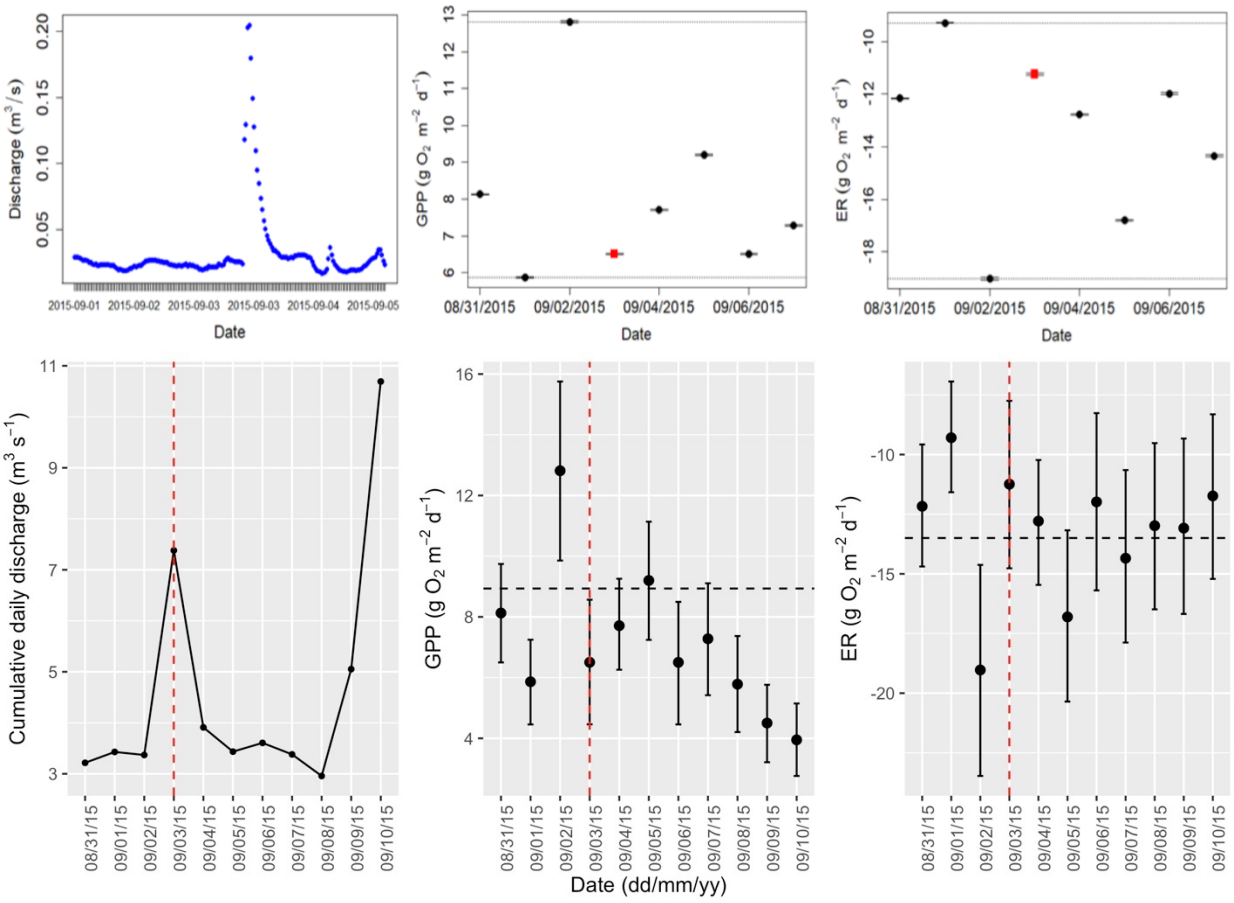
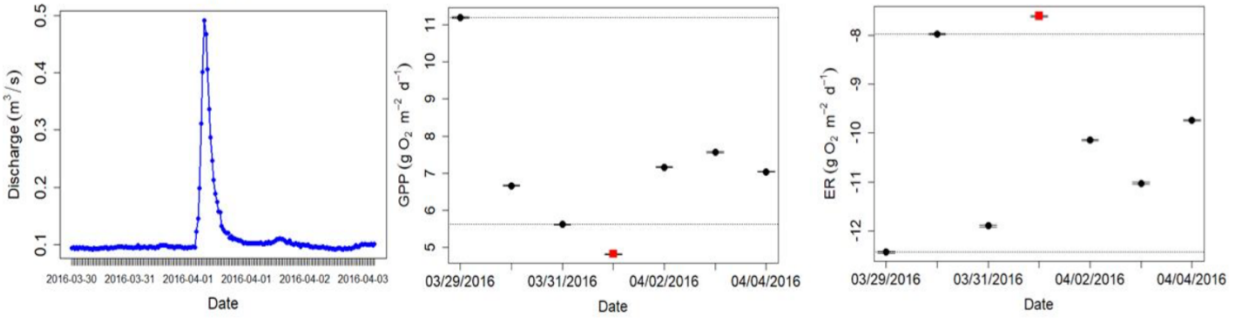



Figure A11 - Storm 8:



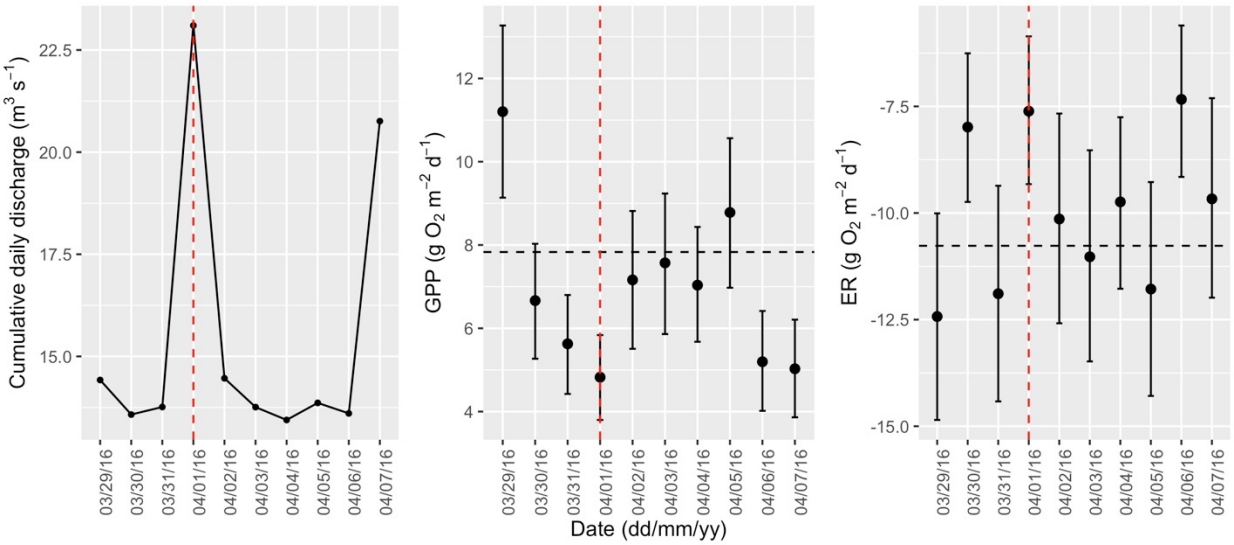



Figure A12 - Storm 9:

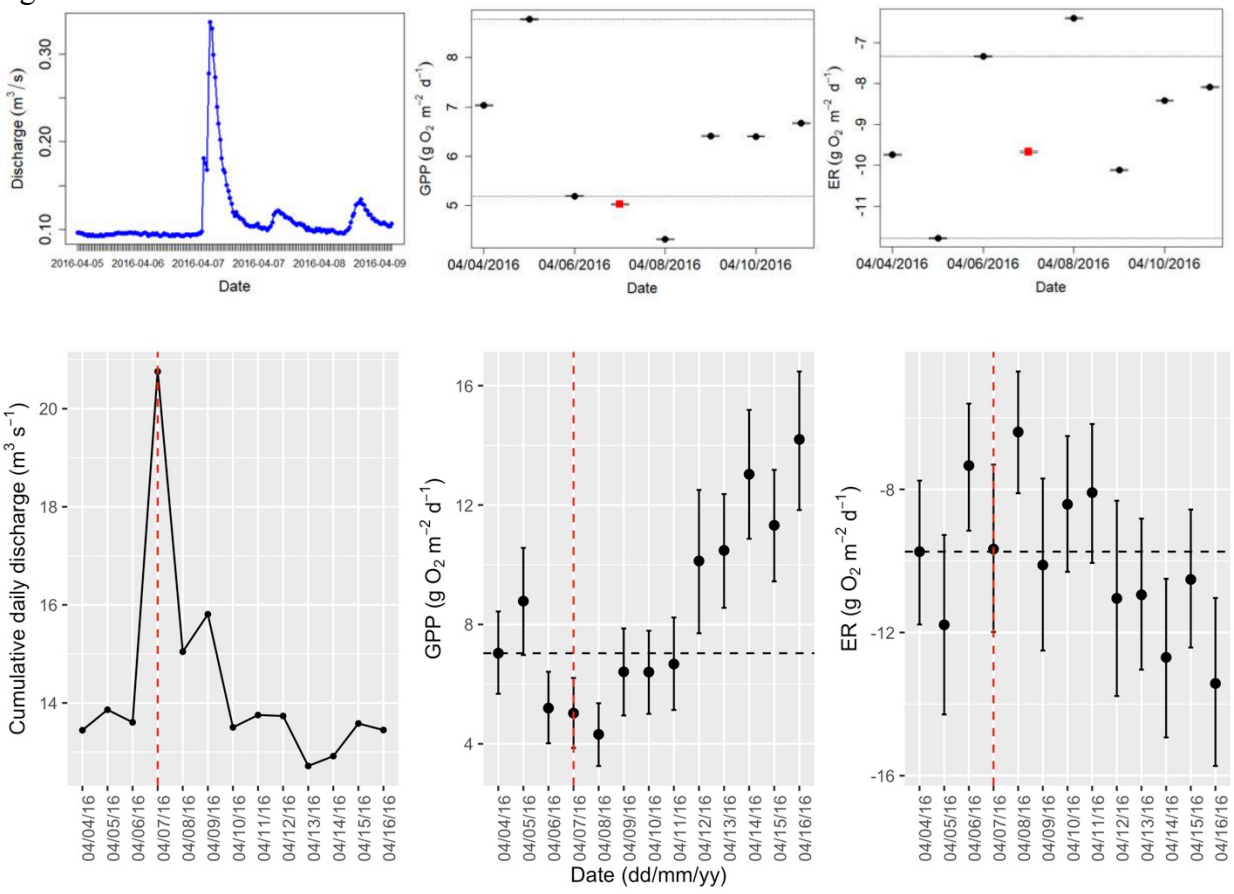



Figure A13 - Storm 10:

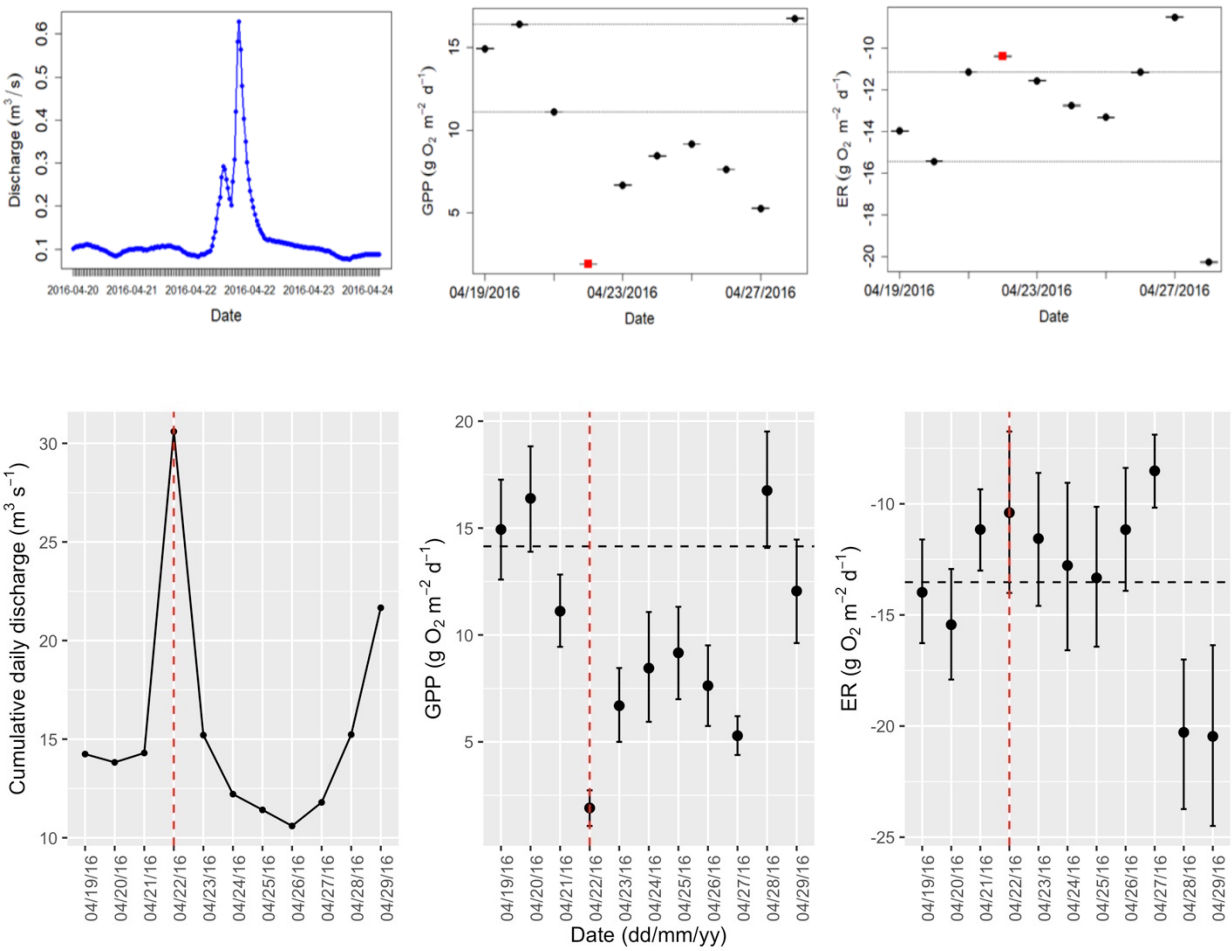
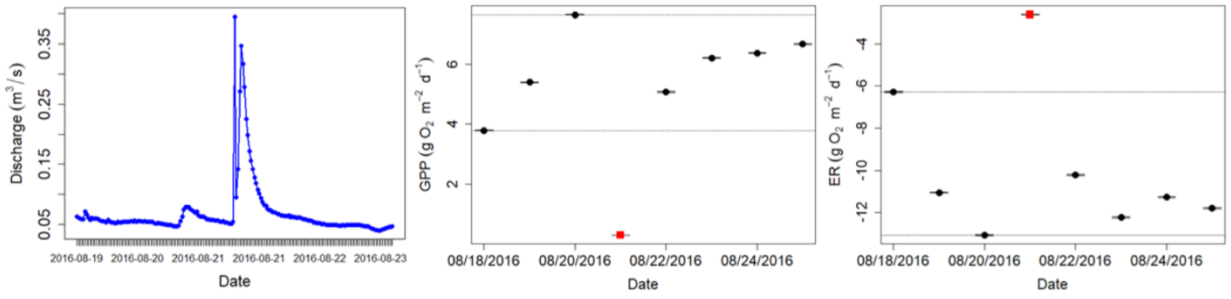



Figure A14 - Storm 11:



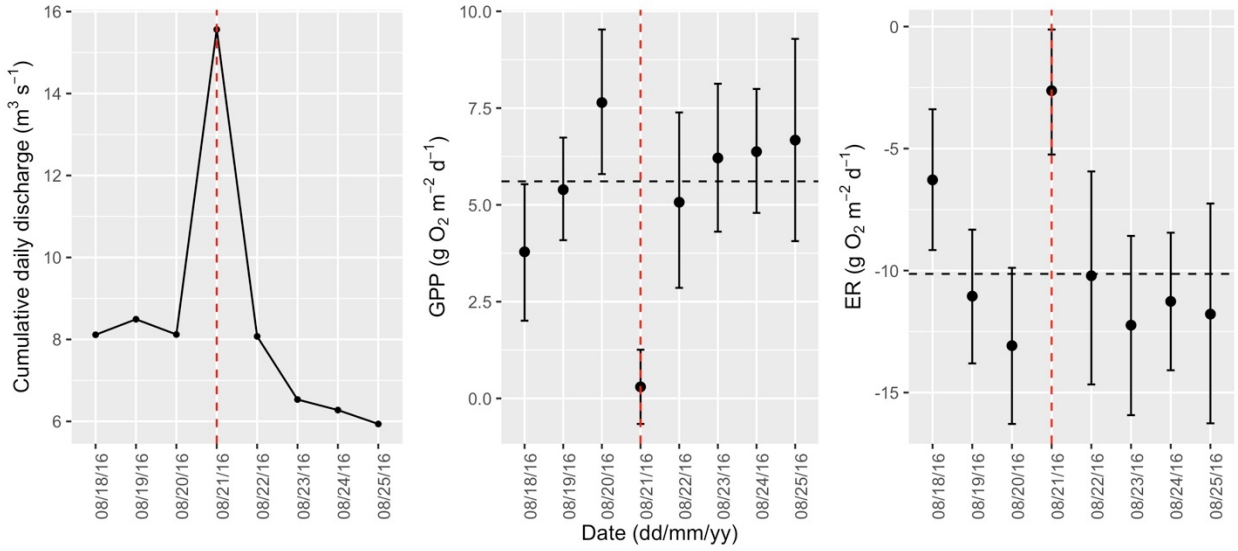



Figure A15 - Storm 12:

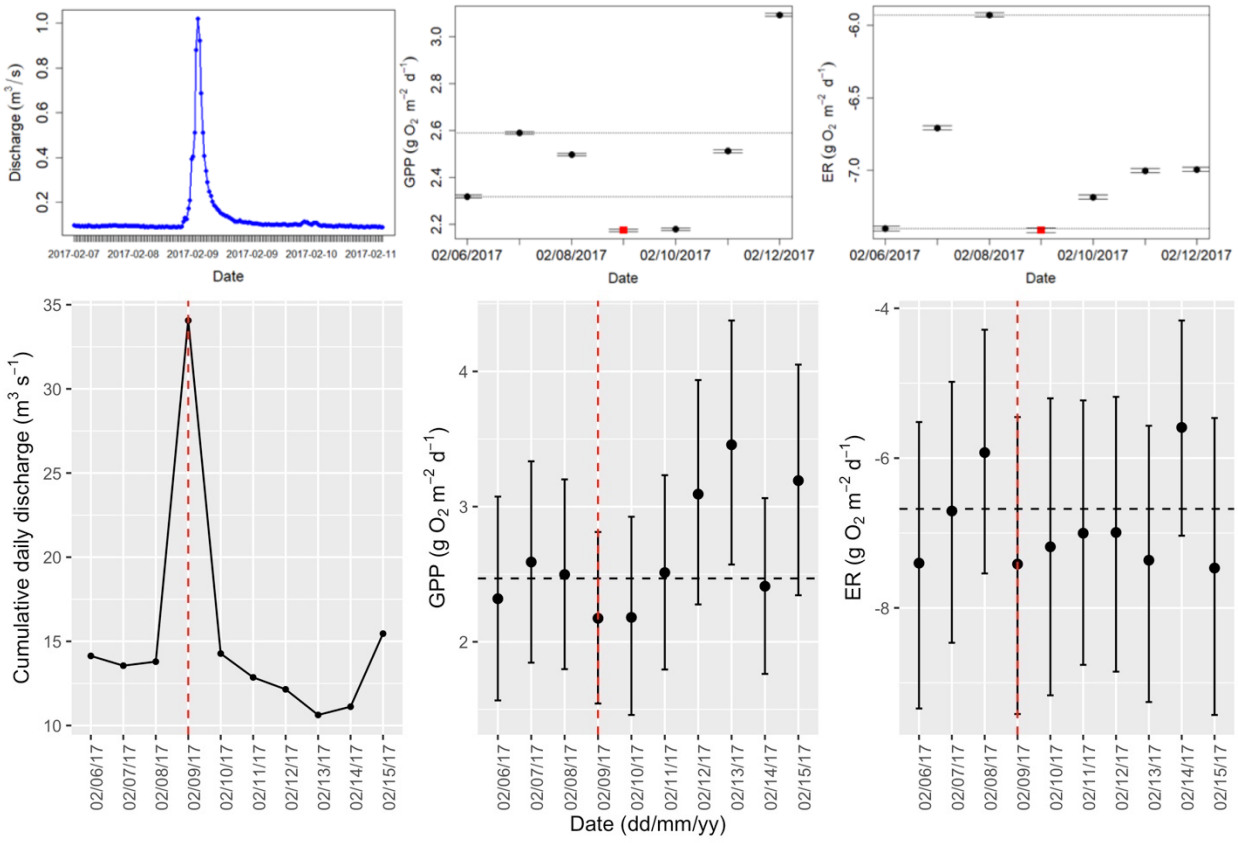



Figure A16 - Storm 13:

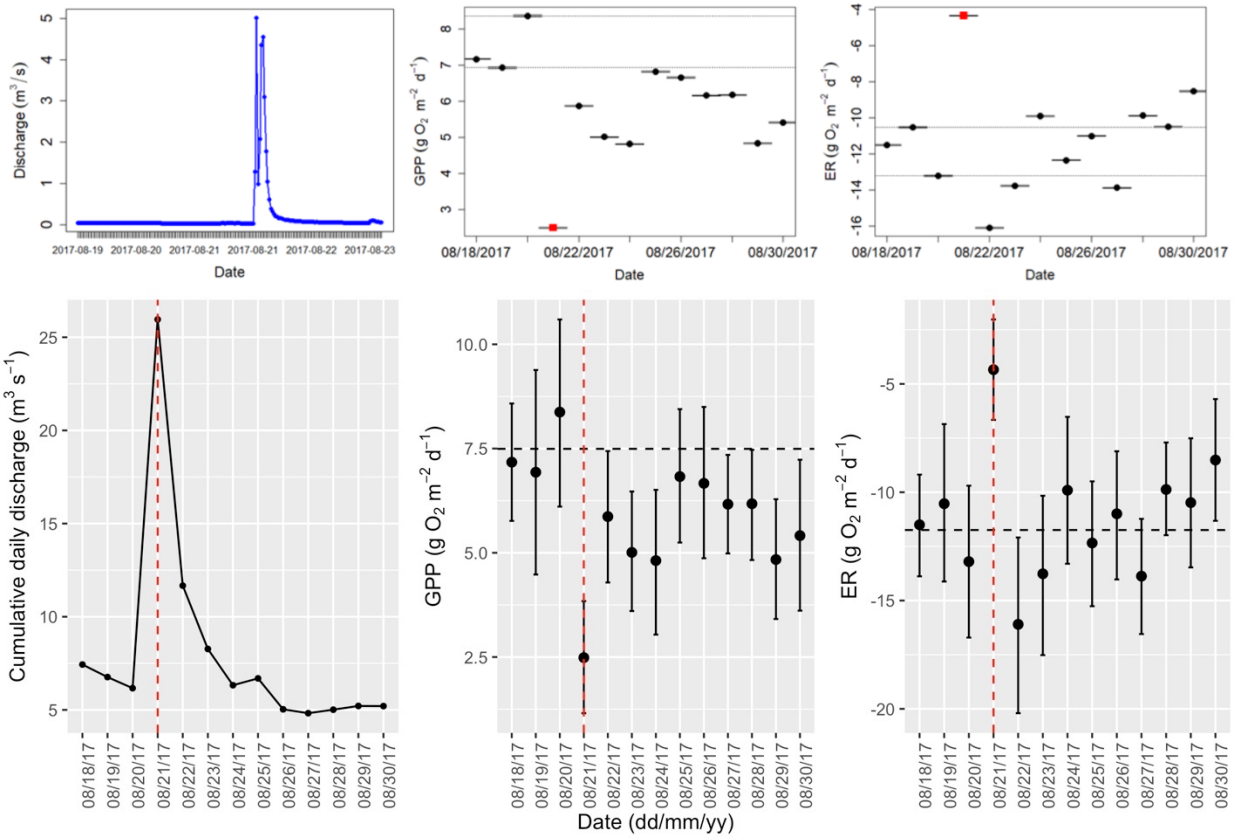



Figure A17 - Storm 14:

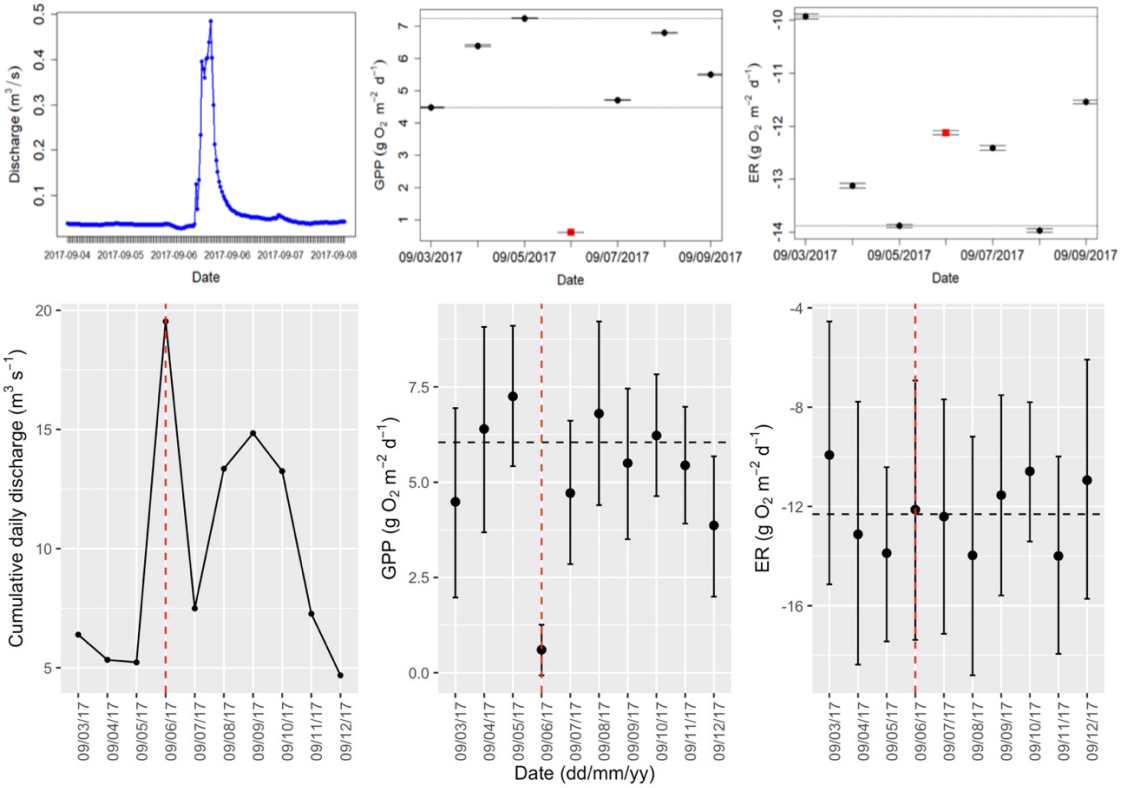
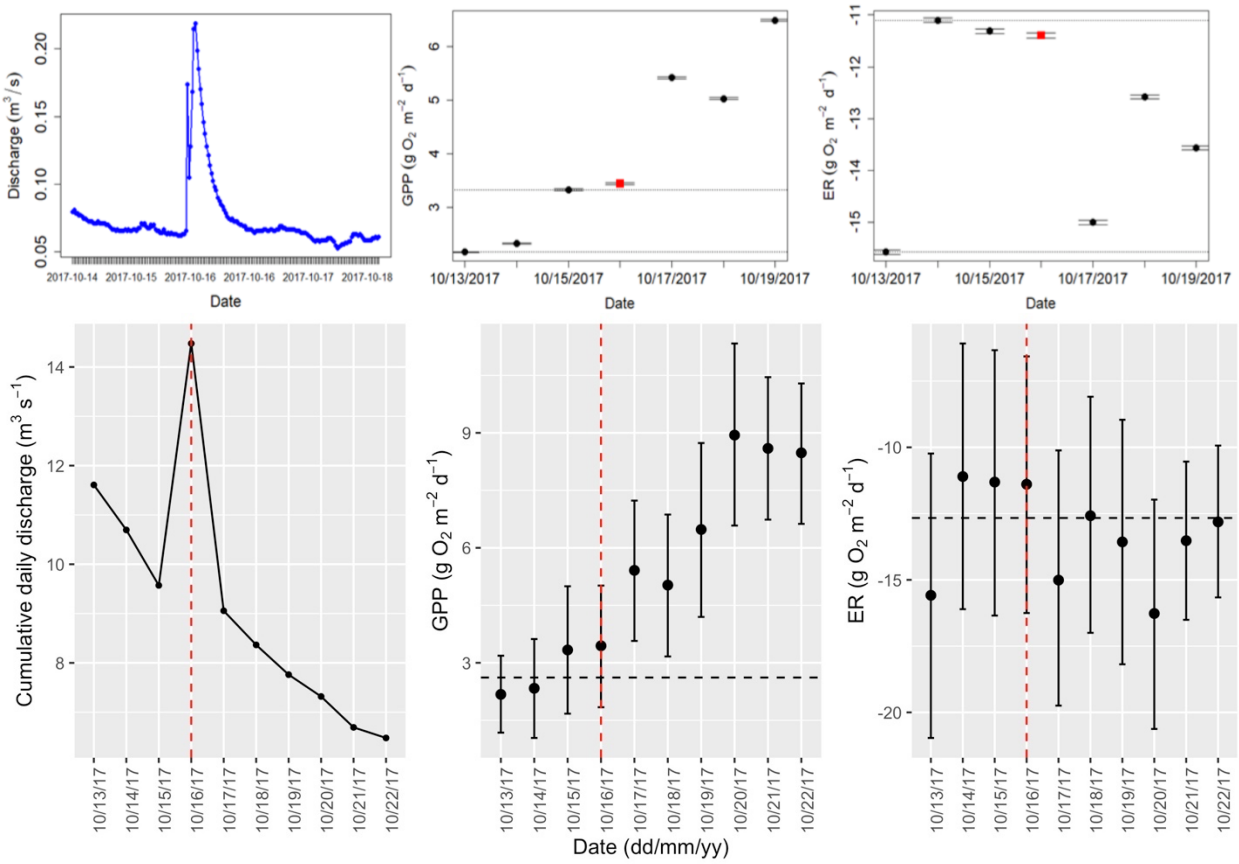


Figure A18 - Storm 15:




**Minor**

**Line 96 : Please specify if NO3 is as N or not.**

**RESPONSE:** Thanks for catching this too! Updated to confirm that this is mg N / L. New text reads: "Stroubles Creek has been designated an impaired waterway due to high sediment loading and has NO3 concentrations that typically exceed 1 mg/L N-$NO_3$."