

Dear editor and reviewers,

Thank you for your time and for the constructive feedback on our manuscript titled "*Breathing Storms: Enhanced Ecosystem Respiration During Storms in a Heterotrophic Headwater Stream*". We appreciate your thoughtful and insightful comments, which will help us improve the quality and clarity of the manuscript. Below, we respond to each comment. The original text of the decision letter is in *black and italics*, while our responses are in dark blue.

From RC1

Jativa et al. present an elegant study on stream metabolic rates during storm events from continuous data collection in a non-perennial Mediterranean stream. The framing of the story is logical and methods clearly address the narrative throughout the manuscript. Indeed, this contributes to a small but intriguing literature on resistance and resilience of ecosystem function in rivers. I have no large comments but raise a handful of additions to improve the clarity of the methods in the specific comments below and a comment on addressing temporal variability in metabolic patterns in rivers that could be expanded on in the introduction.

Many thanks for your positive and encouraging comments. We are glad that you find our study elegant and clear, and we have worked carefully to incorporate your comments in the revised version of the manuscript.

L30: 'regulates'

We have changed to "regulates" as suggested.

L50: 'triggers' and 'stoppers' seem like unnecessary potential jargon. Is there another schema or metaphor that could be used?

We appreciate your concern about potential jargon. However, and considering that Reviewer 2 was prone to these concepts, and actually asked to better frame and emphasize these ideas, we have chosen to keep the terms "*metabolic triggers*" and "*metabolic stoppers*" in the revised version of the manuscript. We believe that these are easy-to-catch concepts that can be used by other authors in future studies to describe the contrasting effects of storm events on stream metabolism. To ensure that these terms are interpreted correctly, we have added some clarification in the introduction to make the rationale behind clearer. Updated text (L52-56): "... increasing discharge can act either as a "*metabolic trigger*" or as a "*metabolic stopper*" depending on the balance between biological processes and physical disruptive forces. A storm event can trigger stream metabolic rates by supplying limiting resources, such as nutrients and organic matter. Conversely, a storm event can inhibit metabolic activity either through biofilm scouring or by increasing turbidity and reducing light availability."

L71: I have no disagreement with any of the introduction to this point, but I think the strong temporal variability in GPP and ER need to be emphasized as potential variability to deal with in identifying resistance or resilience. A wide range of recent literature have shown within year and across year variability in GPP that are influenced by river size, hydrologic variability, and light availability (e.g., Savoy et al. 2019; Marzolf et al. 2024). I would also recommend citing Lowman et al. 2024 as an example of identifying recovery of GPP in response to storm events across large scales.

Thanks for the suggested readings; these three papers are very interesting and relevant for our study. Following your recommendation, we have added these citations to the revised manuscript. Further, we agree that temporal variability in stream metabolic rates could influence the detection and interpretation of resistance and resilience to storm events. However, note that our estimates of resistance and resilience are expressed as relative changes of GPP and ER to pre-event metabolic rates. Thus, any potential temporal variability in metabolic rates did not influence our estimates.

L116: reviewer preference for ‘concentration’ instead of ‘levels’

We have changed to “concentration” as suggested

L122: odd wording. Maybe change to ‘we installed a monitoring station in the stream with upstream area of 9.9 km²’.

We have changed the sentence to “... we installed a monitoring station in the stream with an upslope contributing area of 9.9 km²”

L125: what is average depth in this case? In a stilling well or staff gauge? Or is this hydraulic depth of the 200 m upstream reach? Are the pools located in areas that may alter or disrupt advective flow and create longitudinal heterogeneity in DO patterns (Rexroade et al. 2025)?

We agree with the reviewer that stream depth might vary along the 200-m reach. In this study, average stream depth was estimated from water level measurements recorded by a pressure sensor installed in a stilling well. To verify the representativeness of this value, we conducted manual depth measurements at random transects across the reach every two weeks. We have clarified this procedure in the methods section. Updated text (L137-141):

“Stream water depth (h , in cm) was estimated from 10-minute pressure measurements using a HOBO water level logger installed in a stilling well, with atmospheric pressure corrections applied using a paired barometric logger located in a nearby tree. To verify the representativeness of these data across the 200-m study reach, manual depth measurements were conducted biweekly at randomly selected transects. These manual measurements were also used to calibrate the water level data, confirming that the sensor reliably captured overall depth variability during both baseflow and storm conditions.”

We also agree that pools can generate longitudinal heterogeneity in advective flow and that this could influence stream metabolic measurements, especially during low flow periods. To assess this effect, we installed two DO sensors along the reach during the transition from wet to dry conditions in 2023. These data showed similar DO patterns at the top and the bottom of the reach, suggesting that the pools did not disrupt advective flow at the scale of our metabolism measurements during that particular period. Given that our analyses focus on storm events, when longitudinal connectivity is higher than during the transition from wet to dry conditions, we are certain that the influence of slow-flow zones on DO dynamics was negligible in the present study.

L129: how was lux converted to PPFD? This is an increasingly common practice in the literature and readers would benefit from specifics on how this was done for use in their own studies.

In the revised Methods section, we have clarified this point as follows (L134-135):

“Lux values from the two sensors were averaged and converted to photosynthetic photon flux density (PPFD, in $\mu\text{mol m}^{-2} \text{s}^{-1}$) using a conversion factor of 0.0185, which is commonly applied for forested environments (Thimijan and Heins, 1983). The resulting PPFD values were then aggregated into daily totals.”

L150: What value of Q during the storm event was used in calculating RC? Or is it the total water flux during the storm (i.e., the integral of stream flow/total precipitation)? A few more details would be welcome as this is a potentially useful metric for others to use.

In the revised Methods section, we have clarified this point as follows (L160-162):

“The RC was calculated using the total water flux in mm during each storm event, estimated as the integral of Q over P. This cumulative approach accounts for antecedent moisture conditions and captures the stream’s overall hydrological response to each storm event, avoiding dependence on potentially unrepresentative single-point Q values.”

L155: this is a great presentation of metabolism data collection and modeling. One addition I would like to see is how mean depth was determined. Mean depth is the average cross-sectional depth of the upstream contributing reach, as is defined in this study as the 200 m upstream of their sensor installation. Mean depth is often the most difficult measure to obtain from a stream reach and across flow conditions but can be estimated in similar ways with rating curves and presumably available with the data collected for the propane injections. I would like to see 1-2 sentences added to this section describing how mean depth was determined. And another sentence on QAQC approaches to the continuous data and how DO_{sat} was calculated too (basically address how each of the inputs to streamMetabolizer were prepared).

Thanks for the positive comment. In a previous comment we already addressed how we measured and validated mean water depth and presented a corrected version of the text in the manuscript (L137-141). Moreover, we have added this clarification (L172-176):

“All inputs for the streamMetabolizer model (i.e., light, water temperature, depth, DO, and DO saturation) were derived from the high-frequency sensor data described in Section 2.2. DO saturation was calculated from 10-minute water temperature and barometric pressure values using the García and Gordon (1992) solubility equation. Sensor data were quality-checked by removing implausible values (e.g., negative DO, sensor spikes) and smoothed using LOESS regression (span = 0.03) to reduce noise and filter outliers.”

Finally, in the Supplementary Materials (Appendix 1), there is a section called “Data preparation” detailing the step-by-step QA/QC process.

L166: this is a great way of constraining K in the model inputs and a great example for future researchers to approach single-site evaluations. How well does the coverage of propane injections cover the hydroperiod in the stream? These injections are often biased towards lower flows for logistical reasons, but I wonder how well empirical measures were obtained at higher flows? And as you say in L174, getting metabolism estimates during highest flows is difficult or impossible based on data and/or the models failing to converge on days with high flows.

We were able to conduct propane additions across a wide range of flows, from 0.6 to 32 L s⁻¹. Logistic constraints would have been an issue, but the truth is that during the period in which we were able to conduct gas additions (2022-2024), there was an intense drought that precluded us from performing propane additions at higher discharges. We don't discard the idea of conducting more propane additions in the future to better constrain this parameter. Nevertheless, and for the sake of this study, note that we verified the accuracy of the empirically measured K₆₀₀ values by comparing them with independent estimates obtained from both the night-time regression method (Odum, 1956) and hydraulic geometry-based predictions (Raymond et al., 2012). This exercise showed that K₆₀₀ obtained from propane additions were similar to those estimated with the hydraulic geometry method (Raymond et al., 2012). Thus, we are confident about the robustness of our K₆₀₀-Q relationship. To further clarify this procedure, we have included the comparison curves showing K₆₀₀ values from propane additions alongside those predicted by the other two methods in the Supplementary Materials (Figure S1b).

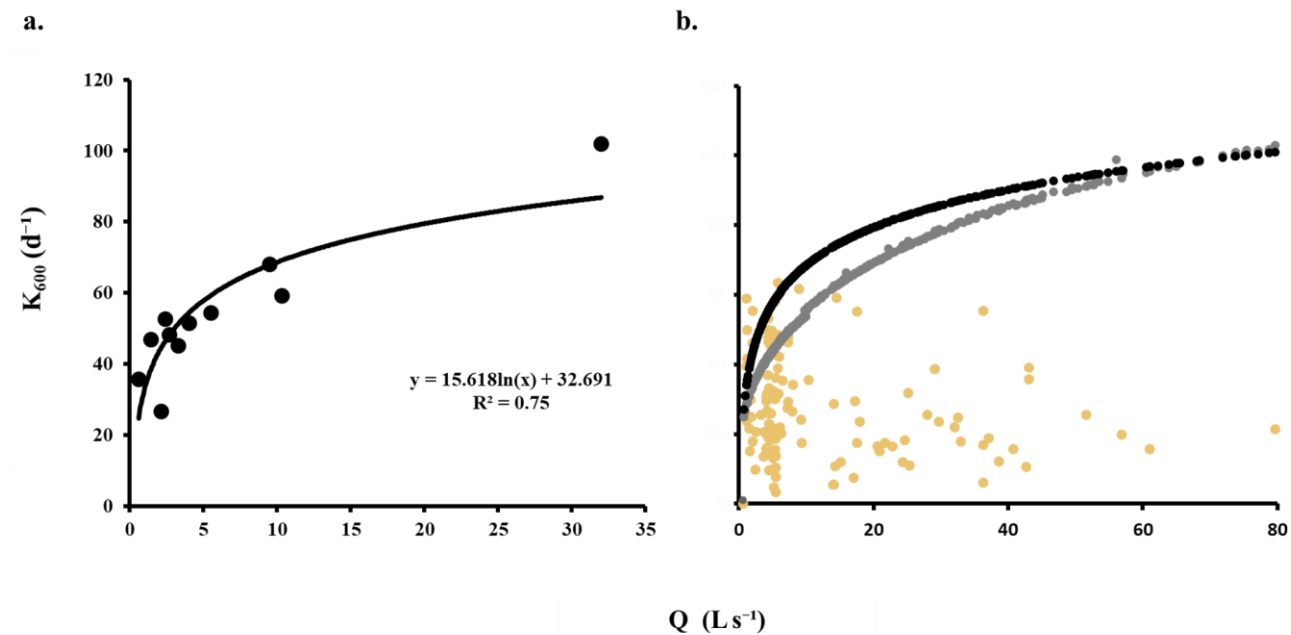


Figure S1. Mean daily discharge (Q) and gas exchange coefficient (K_{600}). (a) K_{600} estimates from 11 propane tracer additions using a mixed tracer injection method (Jin et al., 2012). A logarithmic relationship between Q and K_{600} was derived from these measurements. (b) K_{600} estimates for all days before, during, and after storm events using three methods: night-time regression (yellow dots, $n = 417$), hydraulic geometry following Raymond et al. (2012) (grey dots, $n = 774$), and modeled K_{600} values (black dots) calculated using the Q - K_{600} relationship derived from panel (a).

L206: subscript $PI_{\max}$ as is in L194

We have changed to “ $PI_{\max}$ ” as suggested

L280: might nit-pick on the ‘biota’ part of the response. Yes, organisms from bacteria to macro-fauna contribute to ecosystem metabolism, particularly ER, but with that statement, I would anticipate some measure of re-colonization of organisms post-storm events, whereas the response variable in this study is integrative ecosystem-scale metabolic function.

We acknowledge that our study evaluates stream metabolism as an integrative, ecosystem-scale function, rather than tracking biotic recovery directly. To better reflect the nature of our response variable, we have revised this sentence and refer to **stream processes** rather than biota, thus avoiding any potential misinterpretation (L302).

Figure 1) should the caption for the orange dot also include 'ER'?

We included only GPP for the orange dot because it represents a light-limited scenario, which affects GPP but not ER. Since ER is not directly influenced by light availability, it was only associated with the blue line.

Figure 4) It maybe my computer screen but it's difficult to see the non-filled circles against the filled circles. Might recommend a different, contrasting color. Also, purely aesthetic, but can the x-axis be extended to 1000? An additional component that may help the reader discern the relationship with flow: could a vertical line be added where the typical storm flow begins? Or where is the typical baseflow? This would create a part of the graph with baseflow or losing flow metabolism could be easily compared with gaining or stormflow metabolism. If there is not a single or narrow range of flows that separate base from storm flows, disregard this final comment.

Thanks for your suggestions. We have removed the points corresponding to estimates that failed quality checks, as they may be misleading to the reader.

Regarding the suggestion to include a vertical line to separate baseflow from stormflow conditions, we agree that such a reference could be helpful. However, given the wide range of storm magnitudes observed in our dataset, there is no consistent discharge threshold that marks the onset of stormflow. Instead, we have highlighted the days prior to each storm (i.e., baseflow conditions) using a different color in new Figure 4. This change allows for a clearer visual distinction between baseflow and stormflow conditions while accounting for the variability in stream discharge across events.

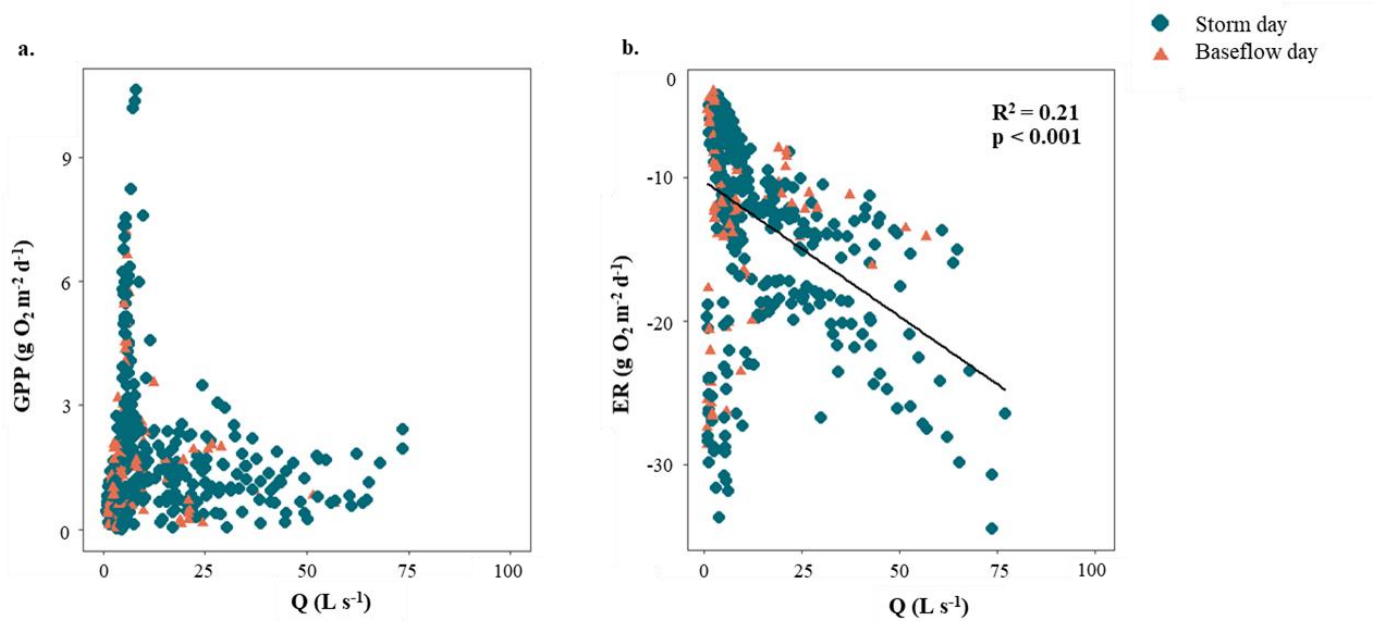


Figure 4. Relationship between mean daily discharge (Q) and daily a) gross primary production (GPP) and b) ecosystem respiration (ER) across all individual storm events (in blue) and for the days 1 to 6 prior to each event (orange). The black line represents the linear regression between variables (only shown when $p < 0.01$) and the shaded area indicates the 95% confidence interval. Note that regressions were based only on storm event days.

Figure 5) Just to be sure, the lines of best fit are coming from the methods text L193-199? What model comparison or evaluation was done to determine linear, logarithmic, or exponential was the 'best' fit to the data? Could all the evaluations be compiled into a supplementary table, perhaps with AIC and AICw values?

Thank you for this suggestion. We have included a supplementary table summarizing the model comparisons to enhance transparency (Table S4). After conducting AIC-based evaluations, we found that linear and logarithmic models were equally supported in two cases: the relationships between ΔQ and ΔER , and between ΔQ and RT_{ER} . These findings have been incorporated into the Results section, and Figure 5 has been updated accordingly to show both relationships. Additionally, we chose not to include exponential models, as they were not significant in any case and did not align with our expectations.

Table S4. Comparison of linear and logarithmic regression model performance comparing hydrological, environmental, and biological variables. Each row shows a pair of variables tested as predictors and responses. For each model type, the Akaike Information Criterion (AIC), p-value, and coefficient of determination (R^2) are reported. Predictor variables include maximum rainfall intensity (PI_{max}), storm duration (D), runoff coefficient (RC), change in discharge (ΔQ), daily discharge (Q), temperature (T), and light (PAR). Response variables include gross primary production (GPP), ecosystem respiration (ER), metabolic resistance as the change in metabolic rates between the storms and pre-storm conditions (ΔGPP , ΔER), and metabolic resilience as recovery time (RT_{GPP} , RT_{ER}). In bold are the models that were finally selected based on AIC and p-values (p value < 0.01). When the difference in AIC between the two models was less than 2, we considered them equally supported. Missing values (--) indicate cases where logarithmic transformation was not possible.

		Linear model			Logarithmic model		
Relationship explored		AIC	p value	R^2	AIC	p value	R^2
PI_{max}	GPP	126	0.035	0.13	129	0.183	0.05
PI_{max}	ER	232	0.015	0.17	234	0.034	0.13
D	GPP	131	0.632	0.01	130	0.424	0.02
D	ER	236	0.104	0.08	237	0.157	0.06
RC	GPP	131	0.775	0.00	131	0.659	0.01
RC	ER	238	0.386	0.02	239	0.650	0.01
ΔQ	GPP	130	0.312	0.03	131	0.660	0.01
ΔQ	ER	237	0.212	0.05	236	0.143	0.06
Q	GPP	1405	0.025	0.01	1409	0.803	0.00
Q	ER	2549	< 0.001	0.21	2586	<0.001	0.12
T	GPP	1903	0.06	0.01	1902	0.04	0.01
T	ER	3554	<0.001	0.06	3552	<0.001	0.06
PAR	GPP	128	0.097	0.08	125	0.017	0.16
PAR	ER	239	0.734	0.00	239	0.753	0.00
PI_{max}	ΔGPP	391	0.477	0.02	391	0.539	0.01
PI_{max}	ΔER	341	0.352	0.03	342	0.862	0.00
D	ΔGPP	391	0.672	0.01	391	0.526	0.01
D	ΔER	342	0.936	0.00	342	0.928	0.00
RC	ΔGPP	391	0.663	0.01	391	0.739	0.00
RC	ΔER	340	0.216	0.05	339	0.103	0.08
ΔQ	ΔGPP	387	0.034	0.13	388	0.071	0.10
ΔQ	ΔER	325	<0.001	0.37	325	<0.001	0.39
PI_{max}	RT_{GPP}	153	0.427	0.02	152	0.184	0.05
PI_{max}	RT_{ER}	150	0.864	0.00	150	0.563	0.01
D	RT_{GPP}	151	0.110	0.08	151	0.161	0.06
D	RT_{ER}	148	0.246	0.04	149	0.352	0.03
RC	RT_{GPP}	153	0.380	0.02	153	0.930	0.00
RC	RT_{ER}	147	0.132	0.07	148	0.221	0.05
ΔQ	RT_{GPP}	153	0.490	0.01	153	0.822	0.00
ΔQ	RT_{ER}	126	<0.001	0.49	128	<0.001	0.46
RT_{GPP}	ΔGPP	387	0.046	0.11	--	--	--
RT_{ER}	ΔER	321	<0.001	0.45	--	--	--

From RC2

General comments: The authors investigate how storm events influence stream metabolism, GPP and ER, in a headwater stream, by using high-frequency DO, hydrological, and environmental measurements to analyze 35 storm events, applying Bayesian modeling. A key strength of this study lies in its robust, high-resolution dataset, which allows for a detailed examination of metabolic dynamics. The clear finding is that most analyzed storms (those with $Q < 100$ L/s) act as "metabolic triggers" significantly stimulating ER and demonstrating a positive relationship between ER stimulation (ΔER) and storm magnitude (ΔQ). The second finding is also very nice information about the quantification of metabolic resilience, particularly the finding that ER recovery time increases with storm magnitude but appears to saturate around 6 days.

Despite these strengths, the manuscript requires major revisions to address the conceptual framework established in the Introduction fully and to enhance the robustness and transparency of its interpretations. Specifically, revisions should focus on (1) evaluating the concept of metabolic saturation introduced in the Introduction section, (2) addressing the implications of excluding high-flow data ($Q > 100$ L/s) for testing the "stopper" hypothesis and the overall representativeness of the findings, and (3-optional) acknowledging uncertainty related to gas exchange estimation during dynamic conditions.

Many thanks for your positive and insightful comments. We are glad that you find our dataset robust and our findings nice and clear. We have worked to improve the revised version of the manuscript following your suggestions and comments. Please, find below our responses to your specific queries regarding how to improve the discussion of the metabolic saturation concept and the implications of excluding the high-flow data.

Specific comments:

Lines 58: introduces an interesting question about River Network Saturation concept. However, the Results section, the authors only focus on the positive linear relationship found between ΔER and ΔQ , and the Discussion does not revisit whether the data showed signs of approaching or reaching this saturation/asymptote.

Was the ecosystem's processing capacity likely exceeded in the largest analyzed storms, or was the range insufficient to observe this? The authors may explore more Figure 5b, such as whether the observed range of storm magnitudes was likely sufficient or insufficient to induce metabolic saturation in this system. It seems that in Figure 5d, there is a visual saturation, but this is not the concept the authors introduced in the Introduction. Please clearly differentiate the observed saturation in recovery time from the lack of observed saturation in the magnitude of the ER response.

We appreciate this insightful comment. Following your suggestion as well as suggestions from R1, we re-analyzed the data considering AIC criteria. We found that linear and

logarithmic models were equally supported to understand the relationship between ΔQ and ΔER . These findings have been incorporated into the Results section, and Figure 5 has been updated accordingly to show both relationships. Updated text (L370-376):

“Moreover, the relationship between ΔER and ΔQ was equally well explained by both linear and logarithmic models, preventing us from identifying a clear saturation pattern in ER—unlike the saturation responses often reported for nutrient uptake with increasing discharge (sensu River Network Saturation concept; Wollheim et al., 2018). Nevertheless, we observed certain constraints in ER during storms. For instance, ER rates were never more negative than $-36.4 \text{ g O}_2 \text{ m}^{-2} \text{ d}^{-1}$, and the maximum ΔER was around 100%, indicating that storm-driven increases in ER never exceeded twice the prior baseflow rates. While these constraints do not confirm a saturation effect, they may reflect an upper limit of the heterotrophic response to storm disturbances in Fuirosos.”

Moreover, we agree with the reviewer that the saturation-like response observed in ER recovery time (RT_{ER}) is a distinct phenomenon, unrelated to the expectations derived from the River Network Saturation hypothesis. Rather than indicating saturation of metabolic response, this pattern more likely reflects a threshold in system resilience. We have made this point clear in the revised version of the manuscript. Updated text (L389-391):

“Further, this result suggests that, beyond a certain disturbance magnitude, the time needed for metabolic rates to return to baseline conditions stabilizes, potentially due to limits in biofilm recovery time or the temporal window of resource availability following storms.”

Line 170: All the estimates with $Q > 100 \text{ L/s}$ were excluded due to failed QC checks. I agree that the exclusion of high-flow data ($> 100 \text{ L/s}$) is based on the reported QC failures, but I am not sure if this action may prevent an empirical test of the "stopper" hypothesis. In the Introduction, lines 60-65, "Finally, during large storm events, [...] decreasing mean water residence time, scouring the benthic biomass, [...] reduce in-stream processing". These sentences refer to the "stopper" for the large storm events, but most valid estimates were skipped to check it. Therefore, the inability to assess larger events means the full spectrum proposed in Figure 1 cannot be validated. Here are some suggestions that only use the current dataset:

We thank the reviewer for raising this important issue and for the useful suggestions to address this point. We fully agree that the exclusion of high-flow data due to QC failures limits our ability to directly test the "stopper" end of the conceptual framework introduced in Figure 1. In order to improve this part of the discussion, we have taken steps to explore how prevalent these extreme discharges are and whether useful information can still be extracted from the high-discharge events. Specifically, we have revised these events using adjusted model constraints and providing full access to model outputs, regardless of the QC status. With this, we aim to shed light on the potential “stopper” behavior, even if not all data at high discharges meet the standard quality thresholds required by the Bayesian model. The

following responses describe the specific actions we have taken to address each of the reviewer's suggestions.

1) *Report the frequency/duration of flows > 100 L/s to know the unanalyzed portion.*

In our study, the frequency of high-flow days ($>100 \text{ L s}^{-1}$) represented only 26 out of the 567 storm-related days (4.6%). This limited representation constrained our ability to directly evaluate potential “metabolic stopper” effects. In the revised manuscript we have addressed this as follows Updated text (L368-370):

“During the study period, these high-flow days represented less than 5% of all the storm days analyzed and occurred during only 8 of the 53 identified storm events. Thus, large or extreme storms that might inhibit ER were relatively rare in this stream, which more commonly exhibited ER stimulation during storms.”

2) *Table S2 does not explicitly link these failures to discharge levels. Please report more details on the QC-Failed Outputs in Supplementary Information to know which QC criteria failed.*

We agree with the reviewer that this additional information will help to clarify why metabolism estimates could not be obtained for certain high-flow events and the model limitations under dynamic conditions. In the revised Supplementary Information, we have updated Table S3 with the range of discharge values with the specific QC criteria that were not met. Table S3 shows that all metabolism estimates for flows $>100 \text{ L/s}$ failed to meet one or more of the quality control criteria. Specifically, model failure at these high flows was driven by both poor model fit (e.g., 73% of high-flow days had $R^2 < 0.5$, and 73% had $\text{RMSE} > 0.4$) and biologically implausible results (e.g., 54% of days with $\text{GPP} < 0$ or $\text{ER} > 0$, and 69% with unrealistically high K600 values).

Table S3. Diagnostics assessing model performance, detailing the total days analyzed (including storm days and days prior to each storm event) and storm events affected by each criterion. The table includes the total available data, the number of imputed days using miceRanger, and occurrences of biologically implausible values (i.e., negative GPP or positive ER). It also reports instances of unsuccessful model convergence ($\hat{R}$ -hat > 1.2 and n_eff < 8000) and days with poor model fit (R^2 < 0.5, RMSE > 0.4). Finally, the number of days that passed all quality checks is indicated. The same diagnostics are shown for different ranges of discharge during storm events.

Quality test	Days analyzed	Number of storm events	Discharge Range During Storm events (L/s)			
			0.7–10	10.1–40	40.1–100	>100
Total data	698	53	347	135	57	26
Imputed data	48	6	32	0	2	14
GPP < 0, ER > 0	53	18	8	11	9	14
K₆₀₀ > 110	18	8	0	0	0	18
$\hat{R}$-hat > 1.2	48	9	18	10	0	0
n_eff > 8000	0	0	0	0	0	0
R^2 < 0.50	147	31	56	18	7	19
RMSE > 0.4	92	26	30	23	7	19
Passed quality check	542	35	274	106	37	0
% Success	72%	66%	79%	79%	65%	0%

3) Figure S1 about Q-K600 relationship is very informative, but the highest discharge measured during these injections appears to be only around 32 L/s. Applying the derived Q-K600 relationship via extrapolation beyond the measured range (~32 L/s) during dynamic storm flows (up to 100 L/s) introduces uncertainty. Is it a reason for the model failing at high discharge? I recommend the SI provide a discussion about why the model likely failed QC at high flows in this system while contrasting with successful high-flow modeling in larger systems (e.g., Diamond et al., 2025a, 2025b)

Thank you for this observation. As we addressed in our responses to R1, we acknowledge that we were unable to conduct propane additions at discharges greater than 32 L s⁻¹ due to prolonged drought during the study period. This limitation introduces some uncertainty when extrapolating the Q-K₆₀₀ relationship to high-flow conditions. In that response, we explained how we compared the propane-based K₆₀₀ estimates with independent values from the night-time regression method and the hydraulic geometry approach to assess the robustness of the derived relationship. Please refer to that comment for a detailed explanation.

To further clarify why model performance deteriorated at high flows, Table S3 shows that most days failing QC did so because the model produced unrealistic estimates (e.g., negative GPP or positive ER) or because modeled DO patterns diverged substantially from observed diel dynamics. These issues likely stem from sensor displacement or burial during turbulent flow, or from diel variability in metabolic (GPP, ER) or physical (K_{600}) parameters not captured by the model. This point is now included in the discussion (L365-368):

“Fuirosos is a small stream (median $Q = 12 \text{ L s}^{-1}$; median depth = 7.5 cm), where floods exceeding 100 L s^{-1} often lead to overbank flooding and displacement or burial of sensors. These complex hydrodynamics and field constraints can disrupt DO signals and complicate gas exchange inference, ultimately violating key assumptions of metabolic models. These complex hydrodynamics and field constraints can disrupt DO signals and complicate gas exchange inference, ultimately violating key assumptions of metabolic models.”

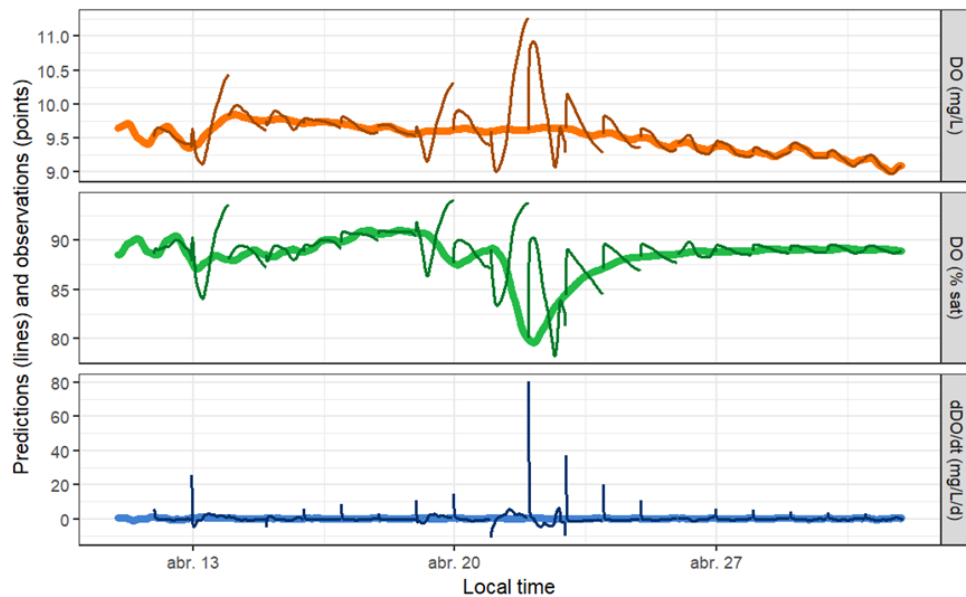
4) *I would like to see output distributions (credible intervals/ranges) for GPP/ER/ K_{600} for all these high-flow runs ($Q > 100 \text{ L/s}$). Even though the median values failed for QC, using the credible intervals may give us some helpful information, such as the system is more "stoppers" or more "triggers" behavior at these high flows.*

--> I suggest that authors may explicitly state the "stopper" hypothesis remains empirically untested by reliable data from this study for higher flows if using credible intervals output still does not give any further information.

We thank the reviewer for this constructive suggestion. We will include in a public repository (HydroShare) the full set of metabolism model outputs from streamMetabolizer, including all daily GPP, ER, and K_{600} estimates, regardless of whether they passed the quality check. This will allow readers to evaluate the full distribution of outputs, including those from high-flow events ($>100 \text{ L s}^{-1}$), and to interpret model behavior beyond the subset of accepted values.

Following the reviewer's recommendation, we revisited the 26 high-flow days ($Q > 100 \text{ L s}^{-1}$) by re-running the metabolism model using adjusted constraints, specifically narrowing the feasible range for depth and K_{600} based on our empirical data. We found that for most of these high-flow days, model outputs remained unreliable even after these adjustments, typically due to persistent convergence issues or implausible parameter estimates (see example figure below). However, in 8 of the 26 days, where the original model failures were specifically due to K_{600} values exceeding 110 d^{-1} , the constrained model produced estimates that met our convergence and plausibility criteria. For these days, high Q ranged from 100 to 1370 L s^{-1} and ER values were similarly elevated (range: 20 to $41 \text{ g O}_2 \text{ m}^{-2} \text{ d}^{-1}$) as on days with large ΔQ increases, providing no evidence for a metabolic stopper effect within our data.

Fig R1. Temporal patterns of dissolved oxygen dynamics during a period including three consecutive high-flow days ($Q > 100$ L/s). The panels show (top) dissolved oxygen concentration (DO, mg/L), (middle) DO saturation (%), and (bottom) the rate of change in DO (dDO/dt , mg/L/d). Bold lines represent observed values, while lighter lines indicate model predictions generated using StreamMetabolizer. The figure illustrates both the general agreement and discrepancies between observed and modeled values under high-discharge conditions.



Line 174: "did not passed" -> did not pass

We have changed to "did not pass" as suggested

Line 155 and 175: Add a brief example of ΔER calculation with negative ER. This is quite confusing when saying to increase or decrease ER while the ER is negative.

We have added a clarification in the Methods section (L296-204) to explain that for the ΔER calculations, we used absolute values. In this sense, an increase in ER (i.e., more negative) indicates a stimulation of respiration.

Line 270-275: Consistent adding $\Delta MET/\Delta GPP/\Delta ER$ definition. Better clarifying axis labels in Figs 5 & 6.

We have changed the captions of both figures to make it more clear.

"Figure 5. Relationship between changes in discharge (ΔQ , defined as the difference between peak flow and baseflow) and stream metabolic responses. Panels (a) and (b) show the change in gross primary production (ΔGPP) and ecosystem respiration (ΔER), respectively, used as proxies for metabolic resistance. Panels (c) and (d) show recovery time

for GPP (RT_{GPP}) and ER (RT_{ER}), used as proxies for metabolic resilience. Color coding indicates whether metabolic activity was stimulated (blue), suppressed (red), or showed no significant change (yellow) in response to the storm event. Solid and dashed lines represent the best fit models, only when statistically significant ($p < 0.01$) and the shaded area indicates the 95% confidence interval. For panels (b) and (d), there were no differences in the goodness of fit between the logarithmic (log) and linear (lm) models (Table S4).”

“Figure 6. Relationship between resistance and resilience of gross primary production (GPP) and (b) ecosystem respiration (ER) during storm events. Resistance is the relative change in metabolic rates compared to prior baseflow conditions (i.e., ΔGPP and ΔER). Resilience is the recovery time of metabolic rates to prior conditions (i.e., RT_{GPP} and RT_{ER}). Color coding indicates whether metabolic activity was stimulated (blue), suppressed (red), or showed no significant change (yellow) in response to the storm event. The black line represents the linear regression between variables (only shown when $p < 0.01$) and the shaded area indicates the 95% confidence interval.”

Line 55/88: Clarify "recovery time" vs. River Network Saturation.

Thanks, this distinction is important. As mentioned earlier, rather than indicating saturation of metabolic response, the recovery time pattern reflects a threshold in system resilience. We have added a clarification in the discussion (L389-391).

Line 352: "ER recovery times at ca. 6 days (Fig. 5a)" -> it should be Fig. 5d

We have fixed it as suggested

Diamond, J. S., Bernal, S., Boukra, A., Cohen, M. J., Lewis, D., Masson, M., Moatar, F., & Pinay, G.: Stream network variation in dissolved oxygen: Metabolism proxies and biogeochemical controls, *Ecological Indicators*, 131, 108233. <https://doi.org/10.1016/j.ecolind.2021.108233>, 2025.

Garcia, H. E. and Gordon, L. I.: Oxygen solubility in seawater: Better fitting equations, *Limnology and Oceanography*, 37, 1307–1312, <https://doi.org/10.4319/lo.1992.37.6.1307>, 1992.

Raymond, P. A., Zappa, C. J., Butman, D., Bott, T. L., Potter, J., Mulholland, P., Laursen, A. E., McDowell, W. H., and Newbold, D.: Scaling the gas transfer velocity and hydraulic geometry in streams and small rivers, *Limnology and Oceanography: Fluids and Environments*, 2, 41–53, <https://doi.org/10.1215/21573689-1597669>, 2012.

Odum, H. T.: Primary Production in Flowing Waters, *Limnology and Oceanography*, 1, 102–117, <https://doi.org/10.4319/lo.1956.1.2.0102>, 1956.

Wollheim, W. M., Bernal, S., Burns, D. A., Czuba, J. A., Driscoll, C. T., Hansen, A. T., Hensley, R. T., Hosen, J. D., Inamdar, S., Kaushal, S. S., Koenig, L. E., Lu, Y. H., Marzadri, A., Raymond, P. A., Scott, D., Stewart, R. J., Vidon, P. G., and Wohl, E.: River network saturation concept: factors influencing the balance of biogeochemical supply and demand of river networks, *Biogeochemistry*, 141, 503–521, <https://doi.org/10.1007/s10533-018-0488-0>, 2018.