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 and provide an explanation of our plans to address them. 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 will work carefully to incorporate your comments in the revised version of the manuscript.

L30: 'regulates'

Thanks for noticing.

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 R2 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 as metaphorical and not as technical jargon, we will add a clarification sentence that makes the rationale behind the terminology clearer.

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. 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. Following your recommendation, we will carefully consider the best place to cite these previous studies and include additional text if needed, either in the introduction or the discussion, to better highlight the natural intrannual and interannual variability of metabolic rates and contextualize our findings.

L116: reviewer preference for 'concentration' instead of 'levels'

Thanks for noticing.

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

Thanks for noticing.

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 will clarify this procedure in the methods section.

We also agree that pools can generate longitudinal heterogeneity in stream flow, and that this could influence stream metabolic measurements, especially during low flow periods. To assess this effect, we installed multiple 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 at 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 likely 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 will clarify that lux values were converted to photosynthetic photon flux density (PPFD, µmol $m^{-2}s^{-1}$) using a conversion factor of 0.0185, which represents the approximate conversion in forested areas such as ours.

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.

We will include in the text that, for calculating RC, we used the total water flux during each storm event, estimated as the integral of discharge (Q) over the storm duration (total precipitation). This approach captures the cumulative effect of the storm on streamflow, rather than relying on a single value such as peak or mean discharge.

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 the revised manuscript, we will add a brief explanation of how we calculated all inputs for *streamMetabolizer* (light, water temperature, depth, DO, and DO saturation). In particular, depth was estimated by measuring pressure in the water column every 10 minutes using a HOBO water level logger and correcting these data for atmospheric pressure using a paired barometric logger. We then calibrated the water level data with manual depth measurements taken biweekly at the study reach throughout the whole sampling period.

We will also clarify how dissolved oxygen saturation (DO.sat, in mg L^{-1}) was calculated. To do so, we used the standard solubility function from García and Gordon (1992), which estimates DO.sat from 10-minute data on water temperature (°C) and barometric pressure (mm Hg).

Finally, we will include a sentence describing our QA/QC procedure for all high-frequency sensor data used as input to the metabolism model. Briefly, we conducted preliminary data cleaning by removing clearly erroneous values (e.g., negative DO concentrations or spikes

from sensor fouling) and then removing noise and outliers using the *loess* R package. Specifically, we applied a locally estimated scatterplot smoothing (LOESS) model with a span parameter of 0.03 to DO, water temperature, light intensity, water depth, and DO.sat variables, effectively smoothing fluctuations and replacing outliers. The detailed step-by-step QA/QC process will be described in the Supplementary Materials.

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 K_{600} 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 K_{600} obtained from propane additions were similar to those estimated with the hydraulic geometry method (Raymond et al., 2012), and thus, we are confident about the robustness of our K_{600}-Q relationship. To further clarify this procedure, we will include the comparison curves showing K_{600} values from propane additions alongside those predicted by the other two methods in the Supplementary Materials.

L206: subscript PImax as is in L194

Thanks for noticing.

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 will revise this sentence and refer to **stream functional processes** rather than biota, thus avoiding any potential misinterpretation.

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 will extend the x-axis to 1000 to enhance visual clarity, and we will remove the points corresponding to estimates that failed quality checks, as they may be misleading.

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 will highlight the days prior to each storm (i.e., baseflow conditions) using a different color in the figure. This change will allow for clearer visual distinction between baseflow and stormflow conditions while accounting for the variability in stream discharge across events.

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?

The lines of best fit shown in Figure 5 correspond to the model types described in lines 193–199, selected based on the best fit to the data. Following your recommendation, we will include a supplementary table summarizing the model comparisons to increase transparency.

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 will work 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. In the revised discussion, we will clearly state that we did not observe saturation in ER stimulation (ΔER) across the range of storm discharges

analyzed. The River Network Saturation concept has been tested and empirically proved in other fluvial systems, mostly using in-stream nutrient processing rates (Wollheim et al, 2018). Following your suggestions, we will re-analyze the data to determine if we did not observe saturation because (a) discharge never acts as a "stopper" in our system or (b) we could not estimate metabolic rates at high flows. For instance, we plan to examine the relationship between Q and ΔER using asymptotic models or breakpoint analyses to assess if there really is a lack of evidence for saturation across the observed range. Finally, we will add a discussion regarding the River Network Saturation in the new manuscript.

Moreover, we agree with the reviewer that the saturation-like response observed in ER recovery time is a separate phenomenon, not related to the expectations derived from the River Network Saturation hypothesis. We will make this point clear in the revised version of the manuscript. Specifically, we will emphasize that this pattern likely reflects a threshold in metabolic resilience, whereby the system returns to pre-storm conditions within approximately one week, regardless of further increases in storm magnitude. This concept is introduced in line 83 of the introduction, as we anticipated that recovery time may reach a threshold corresponding to the time required for biofilms to rebuild after large storm disturbances or the extended influence of nutrient and organic matter inputs, We will clarify the difference between these two phenomena throughout the revised version of the manuscript.

Line 170: All the estimates with Q>100L/s were excluded due to failed QC checks. I agree that the exclusion of high-flow data (>100 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 will take steps to explore how prevalent these extreme discharges are and whether useful information can still be extracted from the high-discharge events. Specifically, we are revisiting 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 will take to address each of the reviewer's suggestions.

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

In the revised Results section, we will include a summary of the occurrence of high-flow conditions. Specifically, we will report that, only 26 out of the 567 storm days analyzed exceeded 100 L s^{-1} (4.6%). These days corresponded to 8 individual storm events out of the 53 analyzed (15.1%). We agree that these numbers will help contextualize the proportion of storms that could not be analyzed due to model limitations in our study stream.

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.

In the revised Supplementary Information, we will include a new table that links the range of discharge values with the specific QC criteria that were not met. This table will summarize which quality checks of the seven considered failed across discharge intervals. 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.

Table R1. *Summary of model performance diagnostics, showing the number of days affected by each evaluation criterion across different discharge ranges. The table reports the total number of available days, instances of unsuccessful model convergence (n_{eff} < 8000 or $\hat{R}$ > 1.2), poor model fit (R^2 < 0.5 or RMSE > 0.4), and biologically implausible estimates (e.g., negative GPP or positive ER). The final column indicates the number of days that failed the quality criteria.*

Q (L/s)	# of days	n_eff > 8000	$\hat{R}$ -hat > 1.2	R^2 < 0.50	RMSE > 0.4	GPP < 0, ER > 0	K600 > 110	Total of failed days
0.7 - 10	347	0	18	56	30	8	0	73
10.1-40	135	0	10	18	23	11	0	29
40.1-100	57	0	0	7	7	9	0	20
>100	26	0	0	19	19	14	18	26
							Total Failed days	**148 (26%)**

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 RC1 (Comment L166), 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-K_{600} relationship to high-flow conditions. In that response, we explained how we compared the propane-based K_{600} 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 R1 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.

It is also important to note that Fuirosos is a small, shallow headwater stream with a median discharge of 12 L s^{-1} and a median depth of 7.5 cm. Flow events exceeding 100 L s^{-1} are rare and represent an order-of-magnitude increase over baseflow, often resulting in overbank flooding and complex hydrodynamics that can disrupt DO signals and violate model assumptions. In contrast, studies such as Diamond et al. (2025) involve larger, deeper systems where similar flow increases produce less drastic hydromorphological changes.

As recommended, we will include a brief discussion of these points, along with the comparison plots, in the Supplementary Information.

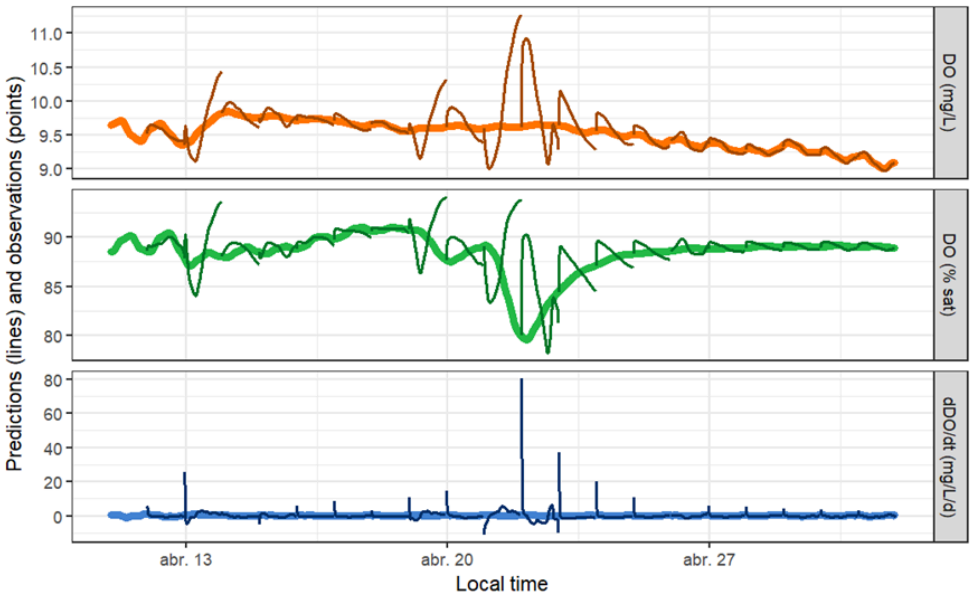
4) I would like to see output distributions (credible intervals/ranges) for GPP/ER/K600 for all these high-flow runs (Q>100 L/s). Even though the median values failed for QC, using the credible intervals may give us some helpful information, such as the system is more "stoppers" or more "triggers" behavior 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 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 L s^{-1}), and to interpret model behavior beyond the subset of accepted values.

Following the reviewer's recommendation, we will also re-run the high-flow events (n = 26 days in total) using adjusted model constraints, specifically by setting feasible upper limits for depth and K_{600} based on our empirical data. This approach may allow us to "rescue" some estimates that were previously rejected due to the uncertainty associated with these two parameters. From preliminary trials, we have found that for most high-flow days where the quality checks failed, model outputs remained unreliable even after doing some adjustments for depth and K_{600} (see example figure below). However, in a subset of cases where the failure was due to K_{600} values exceeding 110 d^{-1}, these adjustments has allowed the model to converge with credible estimates. In the revised discussion, we plan to include this exercise to shed some additional light on our saturation hypothesis.

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
Thanks for noticing.

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 will add a clarifying sentence (and a simple example) in the Methods section to explain that ER values are negative by convention, and that an increase in ER (i.e., more negative) indicates a stimulation of respiration.

Line 270-275: Consistent adding ΔMET/ΔGPP/ΔER definition. Better clarifying axis labels in Figs 5 & 6.

We will ensure consistent use and clear definitions of ΔGPP, and ΔER as the changes in gross primary production and ecosystem respiration during storm events relative to baseflow conditions in the main text and the caption of both figures 5 and 6.

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

Thanks, this distinction is important. In the revised manuscript, we will make clear that the River Network Saturation concept relates to the potential limit of the system to process materials during high flows—essentially a matter of how much metabolism changes can occur in response to disturbance. In contrast, the recovery time refers to how long it takes for the system to return to baseline metabolic conditions after a storm, which is related to the resilience of the system. In both cases, we can observe asymptotic behaviour but the mechanisms are different in each case.

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

Thanks for noticing.

References

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.

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.