nature portfolio

Peer Review File

U.S. rivers are transporting more suspended sediment, often in less time

Corresponding Author: Dr Admin Husic

This file contains all reviewer reports in order by version, followed by all author rebuttals in order by version.

Attachments originally included by the reviewers as part of their assessment can be found at the end of this file.

Version 0:

Reviewer comments:

Reviewer #1

(Remarks to the Author)
Comments for MS # COMMSENV-26-0263-T

General comments:
The manuscript presents a compelling and timely analysis of changing sediment transport regimes across U.S. rivers using a creative integration of high-frequency turbidity sensing, deep learning reconstruction, and interpretable attribution modeling. The central conclusion - that sediment export is becoming more temporally concentrated while often increasing in magnitude - is important for geomorphology, water quality, and infrastructure risk. The scale of analysis, methodological integration, and management relevance are well aligned with the expectations of Communications Earth & Environment. To meet the journal's standard, however, the study would benefit from deeper clarification of novelty, more explicit treatment of uncertainty and representativeness, tighter causal language, and several methodological and interpretive refinements.

Specific comments:
1. From a general perspective, the manuscript is strongest in its conceptual framing of sediment "burstiness" and in its use of daily reconstructions to quantify temporal inequality metrics such as the Gini coefficient and B90. This framing is appropriate for a broad Earth and environmental science readership. However, the novelty relative to recent continental and global sediment modeling studies should be emphasized. Several cited works have already applied deep learning to daily SSC and flux estimation at large scales, and some have examined trends in sediment magnitude. Your manuscript should more explicitly demonstrate what is uniquely enabled here, e.g., the reconstruction of multi-decadal daily sediment flux using turbidity-informed training targets, the explicit quantification of temporal inequality metrics as primary response variables, and the coupling of ML reconstruction with dynamic linear attribution models. Making this distinction clearer in the Introduction and again in the Discussion will help reviewers see the conceptual advance rather than only the scale advance.
2. A central issue that needs deeper treatment is representativeness and sampling bias. The study basins are predominantly in the eastern United States and are limited to sites with long, high-quality turbidity sensing and acceptable turbidity–SSC regressions. While this is understandable, it materially constrains the scope of inference. Communications Earth & Environment readers will expect a more quantitative accounting of how the 175 basins compare with the broader population of U.S. rivers in terms of climate, size, slope, land cover, and regulation status. A short additional analysis or figure comparing the distribution of your basins with the national gauge network would greatly strengthen the manuscript. The Discussion should also more directly acknowledge that arid, snowmelt-dominated, wildfire-affected, and heavily regulated western basins are underrepresented, and that conclusions should be interpreted as most applicable to sensor-monitored, predominantly eastern systems.
3. The credibility of the pre-sensor-era hindcasts is critical because the main trend claims extend back to 1985, while the sediment models are trained primarily for the 2007–2023 turbidity-informed period. You provide useful comparisons against earlier grab samples and discharge observations, and the cross-period skill appears reasonable in Extended Data Fig. 9. Still, for a high-impact journal, this validation should be expanded and made more central. In particular, you should test whether trend directions and magnitudes in flux, Gini, and B90 are robust when computed only over the sensor era versus the full hindcast era. You should also explicitly evaluate model performance in the upper tail of flux distributions, since burstiness metrics are dominated by extreme days. Demonstrating that the model reproduces the timing and rank of the top few percent of flux days would directly support your core conclusions.

4. Related to comment #4, uncertainty propagation into derived metrics is currently insufficiently addressed. Gini coefficients and B90 values are computed from reconstructed daily flux without explicit uncertainty bounds. Because ML errors are event-sensitive, uncertainty in daily flux can map nonlinearly into inequality metrics and trend tests. The study would be significantly strengthened by an ensemble or residual-resampling approach that produces uncertainty intervals for annual Gini and B90, and for their trends. Even if applied to a representative subset of basins, this would demonstrate robustness and elevate confidence in the reported fractions of sites with increasing burstiness or flux.
5. The turbidity-to-SSC conversion framework is reasonable and well-grounded in prior guidance, and the reported median R2 values are strong. However, the manuscript should more fully discuss non-stationarity and structural uncertainty in these relationships. Turbidity–SSC regressions can vary seasonally, with sediment source, and with particle-size distribution, and sensors can drift over multi-year deployments. Because these regressions are used to generate the effective training targets for the ML models, their uncertainty propagates throughout the workflow. A sensitivity analysis in which regression parameters are perturbed within their confidence bounds, and the downstream effect on SSC, flux, and Gini trends is evaluated, would materially strengthen the technical rigor. At minimum, the Methods and Discussion should more explicitly acknowledge these limitations and their likely direction of influence.
6. The modeling framework is generally well described, but several details required for full reproducibility are missing or inconsistent. The text reports different epoch counts for sediment flux training in different sentences, which should be reconciled. Key hyperparameters such as optimizer type, learning rate, early stopping criteria, and hyperparameter tuning strategy should be reported. It should also be clarified whether models were trained as fully pooled multi-basin models or with basin-specific fine-tuning. Given the growing emphasis on reproducibility in ML-enabled Earth science, adding these details will be important for peer review.
7. Your performance assessment using NSE and KGE is appropriate, and the spatial cross-validation experiments are a strength, as shown in the supplementary figures. However, because your conclusions depend heavily on extremes, the evaluation should include metrics focused on high-flow and high-flux events. For example, performance conditioned on the top decile or top percentile of observed flux, peak timing error, and bias in annual load totals would be highly informative. Without this, reviewers may question whether models that perform well on average conditions are sufficiently accurate for inequality metrics driven by extremes.
8. The attribution framework using hierarchical dynamic linear models and Shapley decomposition is innovative and appropriate, but the language describing results should be more cautious regarding causality. At several points, the manuscript states or implies that urbanization or forest loss “drives” increases in burstiness. While the modeling strongly supports association and relative contribution within the chosen predictor set, it does not establish causality in a strict sense. Revising the wording to emphasize attributed contribution or statistical association, and explicitly noting the possibility of omitted or correlated drivers such as channel engineering, bank stabilization, wildfire, or agricultural practice change, will improve scientific precision and credibility.
9. The statistical trend analysis should also be described in more detail. You report Mann–Kendall and Theil-Sen trends across many sites and variables, but it is not clear whether autocorrelation adjustments or prewhitening were applied, nor whether any correction for multiple testing across sites was considered. These details should be added to the Methods. Given the multi-site framework, a brief note on false discovery control or expected false positive rates would be appropriate.
10. The presentation of burstiness metrics is clear and effective, but readers would benefit from additional intuition about their sensitivity and complementarity. A short methodological note or supplementary analysis showing how Gini and B90 respond differently to changes in event frequency versus event magnitude would help interpret cases where the metrics diverge. Including uncertainty ranges on median B90 changes over time would also improve interpretability.
11. In the Discussion and Implications sections, the management relevance is strong and well-described, especially regarding narrowing intervention windows and BMP timing. To further meet the standard of Communications Earth & Environment, you could broaden the framing slightly to connect with global change themes, such as compound extremes, hydroclimatic intensification, and infrastructure resilience under non-stationary sediment regimes. A short paragraph on how this framework could be transferred to other regions given sufficient sensor data would increase generality and impact.
12. Finally, the data and code availability statements are good but could be strengthened by committing to release trained model weights, processed training targets (where permitted), and scripts for computing Gini and B90 metrics. Given the key role of ML reconstruction in the study, enhanced openness will be viewed favorably.

Reviewer #2

(Remarks to the Author)
See attached pdf.

Version 1:

Reviewer comments:

Reviewer #1

(Remarks to the Author)
Please see the attached PDF file.

Reviewer #2

(Remarks to the Author)
I congratulate the authors on a thorough revision; I particularly appreciated their addition of a more rigorous uncertainty

analysis. I have read through the rebuttal letter and associated changes and am mostly happy with the changes made. There remain however a few places where the authors have only responded to me, personally, without making changes to the paper. My comments are not just a result of my curiosity; instead, they were designed to help the authors make the work more robust and/or accessible to a general audience. I suggest the authors re-read the rebuttal and, for comments where they have not made changes to the paper, explicitly justify why they thought a clarification to the paper was not necessary. If I have asked a question or been confused by some aspect of the manuscript, it is likely other readers will have similar questions or confusions, and I hope the authors can use my comments to identify and ultimately avoid such issues for future readers.

Authors' Response to Review Comments:

In this document, the authors list the major comments made during the review of the original manuscript and, underneath each comment, **our responses are written in bold font**. References to in-text citations used in this response-to-reviewer file can be found at the end of this document.

Associate Editor Comments:

Your manuscript titled "American rivers are transporting more sediment in less time" has now been seen by 2 reviewers, and we include their comments at the end of this message. They find your work of interest, but some important points are raised. We are interested in the possibility of publishing your study in Communications Earth & Environment, but would like to consider your responses to these concerns and assess a revised manuscript before we make a final decision on publication.

In particular, please ensure that the revised manuscript meets the following editorial thresholds:

1. Explicitly articulate the novelty of your study relative to recent continental and global sediment modelling studies.
2. Provide a robust uncertainty analysis and a compelling validation of your model outputs.
3. Explicitly address issues of representativeness and sampling bias, in line with the reviewers' recommendations.

We therefore invite you to revise and resubmit your manuscript, along with a point-by-point response that takes into account the points raised. Please highlight all changes in the manuscript text file.

We thank the Associate Editor and the two reviewers for their careful and constructive engagement with our manuscript. The reviews helped us substantially improve the rigor and clarity of the work, and we have addressed every comment in detail in the point-by-point responses that follow. We have updated the manuscript title to better reflect the study's findings: *'U.S. rivers are transporting more suspended sediment, often in less time'*.

Before turning to the individual reviewer comments, we briefly summarize how the revised manuscript meets each of the three editorial thresholds raised:

1. **Most importantly, the central novelty of this work is the explicit treatment of sediment timing as a primary response variable. Existing continental- and global-scale sediment studies — including the satellite-based global mapping of Sun et al. (2025) and Dethier et al. (2022), the LSTM-based CONUS reconstruction of Song et al. (2024), the regional LSTM of Gupta & Feng (2025), and the pan-Arctic ML reconstruction of Tian et al. (2026) — quantify trends exclusively in sediment magnitude (concentration, flux, or annual yield). The temporal-inequality framework using the Gini coefficient and load-exceedance metrics has been applied to sediment and solute transport in regional snapshot studies (Jawitz & Mitchell, 2011; Preisendanz et al., 2021) but has never been used to quantify multi-**

decadal trends in sediment timing at continental scale. To our knowledge, ours is the first study to do so. This distinction is more than methodological. Our ΔFlux vs. ΔGini analysis (new Figure 5) shows that 31 of the 175 study sites exhibit significant increases in burstiness without significant increases in magnitude, and 22 sites show the inverse. A magnitude-only analysis would miss the timing signal at the 31 sites and conflate the regimes at the 22, mischaracterizing more than 30% of trending rivers. Furthermore, our HDLM-Shapley attribution shows that the dominant attributed predictor of timing (urbanization) differs from the dominant attributed predictor of magnitude (precipitation extremes), which means a magnitude-only framing would misidentify the leverage points for management. We have rewritten the closing of the Introduction and added a corresponding paragraph to the Discussion to make this contribution explicit, and we have revised the title to reflect that magnitude and timing trends are partially decoupled across the network.

2. **Robust uncertainty analysis and validation of model outputs. We have added a moving block bootstrap procedure that propagates LSTM prediction error into annual Gini and B_{90} estimates and through to their network-wide trend prevalences (Reviewer #1, Comment 4). The bootstrap-based medians of significant-trend prevalence (25.1% for flux and 26.3% for Gini across 1,000 realizations) confirm the point estimates reported in the main text. We have also added: (a) extreme-event performance metrics conditioned on the top decile and top percentile of observed flux (Reviewer #1, Comment 7), (b) a sub-period trend comparison demonstrating that hindcast-era and sensor-era trends agree in direction and magnitude (Reviewer #1, Comment 3), (c) a Mahalanobis-distance analysis confirming distributional similarity of LSTM inputs across hindcast and training periods (Reviewer #2 minor comments), and (d) a Livezey–Chen field-significance test confirming that observed network-wide trend prevalences exceed the 95th percentile of the spatial-correlation null distribution (Reviewer #1, Comment 9).**

3. **Representativeness and sampling bias. We have added a new representativeness analysis comparing our study basins to the GAGES-II national reference dataset using PCA on seven shared basin attributes (Reviewer #1, Comment 2). The analysis confirms that our sites occupy the bulk of the 95% data envelope of the national network while skewing toward lower-elevation, wetter, and more human-impacted catchments. We have correspondingly revised the Discussion to clarify that conclusions are most directly applicable to sensor-monitored, predominantly eastern systems undergoing human impact, and revised the title to 'U.S. rivers are transporting more suspended sediment, often in less time' to better reflect both the geographic scope and the partial decoupling of magnitude and timing trends across the network.**

Each of these additions is documented in the point-by-point responses below, and all corresponding manuscript changes are highlighted in the revised manuscript text file.

Reviewer #1 Comments:

General comments:

The manuscript presents a compelling and timely analysis of changing sediment transport regimes across U.S. rivers using a creative integration of high-frequency turbidity sensing, deep learning reconstruction, and interpretable attribution modeling. The central conclusion - that sediment export is becoming more temporally concentrated while often increasing in magnitude - is important for geomorphology, water quality, and infrastructure risk. The scale of analysis, methodological integration, and management relevance are well aligned with the expectations of Communications Earth & Environment. To meet the journal's standard, however, the study would benefit from deeper clarification of novelty, more explicit treatment of uncertainty and representativeness, tighter causal language, and several methodological and interpretive refinements.

We thank the reviewer for the thorough reading and critique of the manuscript. Their comments have helped improve its quality.

Specific comments:

1. From a general perspective, the manuscript is strongest in its conceptual framing of sediment "burstiness" and in its use of daily reconstructions to quantify temporal inequality metrics such as the Gini coefficient and B90. This framing is appropriate for a broad Earth and environmental science readership. However, the novelty relative to recent continental and global sediment modeling studies should be emphasized. Several cited works have already applied deep learning to daily SSC and flux estimation at large scales, and some have examined trends in sediment magnitude. Your manuscript should more explicitly demonstrate what is uniquely enabled here, e.g., the reconstruction of multi-decadal daily sediment flux using turbidity-informed training targets, the explicit quantification of temporal inequality metrics as primary response variables, and the coupling of ML reconstruction with dynamic linear attribution models. Making this distinction clearer in the Introduction and again in the Discussion will help reviewers see the conceptual advance rather than only the scale advance.

   Answer:

   We agree that the novelty of this work needs to be more sharply distinguished from the rapidly growing literature on continental and global sediment modeling. The central contribution of our study is the explicit quantification of multi-decadal trends in sediment timing — the temporal distribution of sediment delivery within each year — across U.S. rivers. To our knowledge, this has not been done before at this scale. Most relevant prior work focuses on sediment magnitude. Satellite-based global mapping (Sun et al., 2025; Dethier et al., 2022) provides multi-decadal records of suspended sediment concentration and flux at monthly resolution; LSTM-based reconstructions (Song et al., 2024; Gupta & Feng, 2025) provide daily concentration estimates from grab-sample training data; and the pan-Arctic study of Tian et al. (2026) documents multi-decadal flux change. Each addresses how much sediment is moving. None addresses how the delivery of that sediment is distributed through

time. The temporal-inequality framework itself is established in the regional literature (Jawitz & Mitchell, 2011; Preisendanz et al., 2021), but has been applied only as a snapshot characterization, not to multi-decadal trends, and not at continental scale.

Our work is the first to bring these two strands together: continental-scale, multi-decadal records of sediment delivery analyzed for changes in temporal distribution as well as magnitude. This distinction is consequential. Our ΔFlux vs. ΔGini analysis (new Figure 5) shows that magnitude and timing trends are partially decoupled: 31 sites exhibit rising burstiness without rising magnitude, and 22 show the inverse. A magnitude-only analysis would miss or mischaracterize the trend at more than 30% of these sites. Our HDLM-Shapley attribution further reveals that magnitude and timing trends have distinct dominant drivers (precipitation extremes versus urbanization), with different management implications. Thus, our contribution demonstrates empirically that magnitude and timing are co-evolving dimensions of the Anthropocene sediment cycle that prior continental-scale work has not.

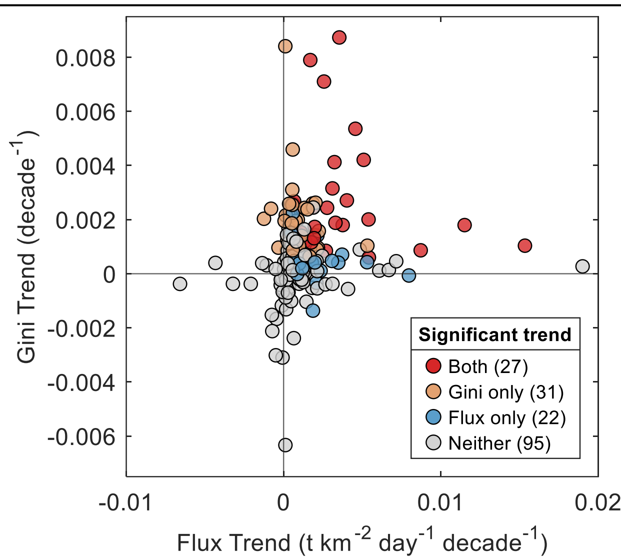


Figure. Partial decoupling of sediment magnitude and timing trends (1985–2023). Scatter plot comparing the decadal trend in mean annual sediment flux against the decadal trend in the Gini coefficient across all 175 study sites. Markers are color-coded based on the statistical significance of their respective trends (Mann-Kendall test, $p < 0.05$). Red markers ($n = 27$) represent basins experiencing concurrently significant increasing trends in both total flux and burstiness. Orange markers ($n = 31$) denote basins exhibiting a significant increase in burstiness only, while blue markers ($n = 22$) denote basins with a significant increase in total flux only. Gray markers ($n = 95$) represent sites lacking a significant increasing trend in either metric.

We have revised the closing of the Introduction and added a corresponding paragraph to the Discussion to articulate this framing. We retain the secondary methodological contributions — the use of high-frequency turbidity-derived training targets that enable timing analysis in the first place, and the coupling of LSTM

reconstruction with HDLM attribution — as supporting elements that make the timing analysis tractable, rather than as co-equal claims of novelty.

2. A central issue that needs deeper treatment is representativeness and sampling bias. The study basins are predominantly in the eastern United States and are limited to sites with long, high-quality turbidity sensing and acceptable turbidity–SSC regressions. While this is understandable, it materially constrains the scope of inference. Communications Earth & Environment readers will expect a more quantitative accounting of how the 175 basins compare with the broader population of U.S. rivers in terms of climate, size, slope, land cover, and regulation status. A short additional analysis or figure comparing the distribution of your basins with the national gauge network would greatly strengthen the manuscript. The Discussion should also more directly acknowledge that arid, snowmelt-dominated, wildfire-affected, and heavily regulated western basins are underrepresented, and that conclusions should be interpreted as most applicable to sensor-monitored, predominantly eastern systems.

   Answer:

   We thank the reviewer for this suggestion, and we now implement this comparison analysis and generate a new Supplemental Figure. To directly address representativeness, we identified 105 of our 175 study basins that have matching entries in the GAGES-II national reference dataset (~9,000 gauges) and conducted a PCA using exclusively GAGES-II attributes, allowing a like-for-like comparison entirely within the same data framework (Figure 1). Seven basin attributes were used: drainage area ($\log_{10}$-transformed), mean annual precipitation, mean elevation, terrain slope, forest cover, cropland fraction, and urban fraction. The two leading PCs explain 60.7% of total variance (PC1: 35.8%, PC2: 24.9%). PC1 primarily represents a gradient from steep, forested, high-elevation basins (positive end) to low-gradient, agricultural, low-elevation basins (negative end). PC2 separates wet, urban, large-area basins (positive end) from drier, smaller systems.

   As shown in the new figure, our study sites fall almost entirely within the 95% data envelope of the national network, demonstrating that they capture a realistic and representative cross-section of U.S. river conditions. However, the biplot also quantifies the spatial bias the reviewer noted: our sites heavily skew towards lower-elevation, wetter, and more human-impacted (urban and agricultural) watersheds (negative PC1 and positive PC2), while lacking representation in steep, pristine, high-elevation domains. We now explicitly report the spatial representation of our sites in the manuscript. We refined our concluding remarks to clarify that our findings regarding intensifying sediment burstiness should be interpreted as most directly applicable to sensor-monitored, predominantly eastern river systems undergoing human impact.

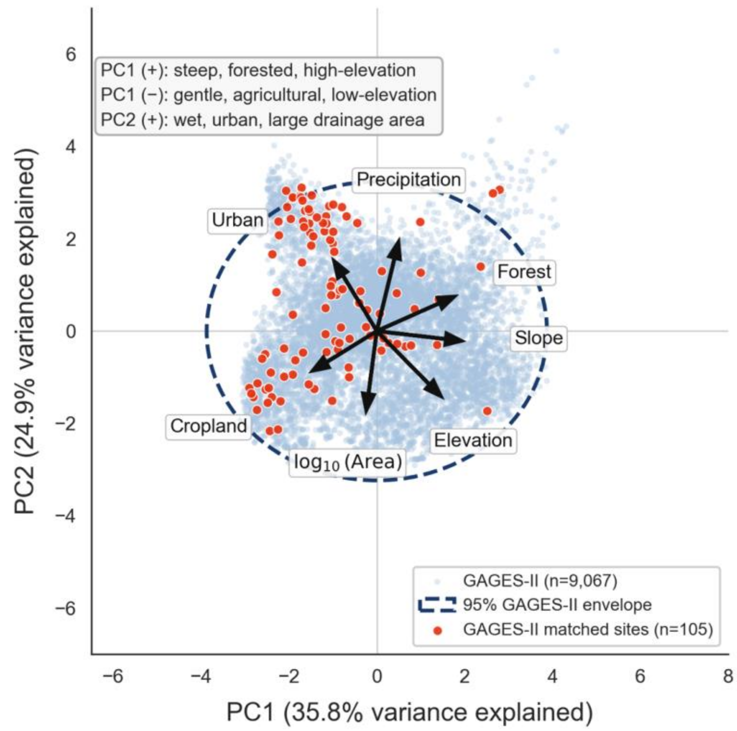


Figure. Representativeness of study basins (n = 105 of 175) relative to the national gauge network. Principal component analysis (PCA) biplot comparing the 105 GAGES-II-matched study sites (red circles) against the full GAGES-II reference network (n = 9,067; blue circles). The dashed ellipse denotes the 95% data envelope for the full GAGES-II network..

3. The credibility of the pre-sensor-era hindcasts is critical because the main trend claims extend back to 1985, while the sediment models are trained primarily for the 2007–2023 turbidity-informed period. You provide useful comparisons against earlier grab samples and discharge observations, and the cross-period skill appears reasonable in Extended Data Fig. 9. Still, for a high-impact journal, this validation should be expanded and made more central. In particular, you should test whether trend directions and magnitudes in flux, Gini, and B90 are robust when computed only over the sensor era versus the full hindcast era. You should also explicitly evaluate model performance in the upper tail of flux distributions, since burstiness metrics are dominated by extreme days. Demonstrating that the model reproduces the timing and rank of the top few percent of flux days would directly support your core conclusions.

   Answer:

   We have now computed Mann–Kendall trend tests across three analysis windows: the pre-sensor era (1985–2006, 22 years), the sensor era (2007–2023, 17 years), and the full record (1985–2023, 39 years) for both the sediment variables and their primary climate and land-use drivers. The results, summarized in a new Supplemental Table 1, show that the weaker significance in the sensor era is not a property of the hindcast but of record length.

   Sediment flux is significant at 15.4% of sites in the 22-year pre-sensor window and 4.6% in the 17-year sensor-era window, rising to 28.0% over the full 39-year record; Gini follows the same arc (15.4% → 1.1% → 33.1%). Hydrological signals have high interannual variability, and short records simply lack the statistical power for Mann–Kendall to resolve a monotonic trend — a limitation that applies equally to both sub-periods, not just the hindcasted one. Land use trends are strong due to the monotonic nature of land use change.

   The clearest evidence that this is a statistical power issue rather than a modelling artefact comes from comparing sediment with a primary driver: extreme precipitation (R95p). Extreme precipitation shows an essentially identical trajectory: significant at only 8.6% and 1.1% of sites in the two short windows (pre-sensor and sensor), then 58.3% over the full record. This parallel is logical, i.e., sediment is tracking flow which is tracking precipitation, and both require the longer baseline before the signal clears the noise threshold.

Table. Fraction of sites with statistically significant trends across sub-periods. Percentage of the 175 study sites exhibiting significant monotonic trends (Mann–Kendall test, α = 0.05) for climate and land-use drivers (top) and hydrologic and sediment response variables (bottom), computed separately for the pre-sensor era (1985–2006), the sensor era (2007–2023), and the full record (1985–2023). For urban fraction, values reflect the percentage of sites with significant increasing trends; for forest fraction, the percentage with significant decreasing trends (forest loss); for all other variables, values reflect significant increasing trends. Green values indicate higher fractions (≥20%) and red values indicate lower fractions (<20%), providing a visual summary of trend prevalence across periods.

Variable	Pre-sensor era 1985–2006 (22 yr)	Sensor era 2007–2023 (17 yr)	Full period 1985–2023 (39 yr)
Climate & land-use drivers			
R95p	8.6%	1.1%	58.3%
Temperature	4.6%	16.0%	60.0%
Urban frac.	98.9%	94.9%	98.3%
Forest frac.	71.4%	65.1%	78.3%
Hydrologic & sediment response			
Flow	17.1%	0.6%	22.9%
SSC	14.3%	6.3%	24.6%
Flux	15.4%	4.6%	28.0%
Gini	15.4%	1.1%	33.1%

Beyond significance prevalence, we also compared the magnitude and direction of trends between sub-periods. Theil–Sen slopes computed from the sensor era (2007–2023) align in sign with full-record slopes at 83 sites for flux, 98 sites for Gini, and 91 sites for B90 with a Spearman rank correlation 0.28 ($p < 0.001$) for flux, 0.32 ($p < 0.001$) for Gini, and 0.31 ($p < 0.001$) for B90 between the two sets of slopes. The hindcast does not introduce a directional bias relative to the sensor era; it primarily strengthens the statistical resolution of trends already present in the sensor record. We include these numbers in the text.

We respond to your final point about model evaluation for extremes in Comment 7.

4. Related to comment #4, uncertainty propagation into derived metrics is currently insufficiently addressed. Gini coefficients and B90 values are computed from reconstructed daily flux without explicit uncertainty bounds. Because ML errors are event-sensitive, uncertainty in daily flux can map nonlinearly into inequality metrics and trend tests. The study would be significantly strengthened by an ensemble or residual-resampling approach that produces uncertainty intervals for annual Gini and B90, and for their trends. Even if applied to a representative subset of basins, this would demonstrate robustness and elevate confidence in the reported fractions of sites with increasing burstiness or flux.

   Answer:

   To rigorously account for model uncertainty and its propagation into our burstiness metrics, we implemented a moving block bootstrap approach. We used a 10-day block size to preserve the short-term autocorrelation inherent in model residuals. For each site and year, we generated 1,000 synthetic flux time-series by sampling 10-day contiguous blocks from the residual pool and superimposing them on the LSTM predictions. From these ensembles, we derived the 95% confidence intervals for annual flux and Gini values. An example of how the bootstrapping process impacts annual flux and Gini values is shown in the figure on the next page.

   To confirm that the trend statistics are robust to model uncertainty, we re-ran the Mann–Kendall trend test on each of the 1,000 bootstrap realizations of the annual Gini and flux series at every site. The median percentage of sites with significant increasing trends was 25.1% [21.7−28.6%] for sediment flux and 26.3% [22.9−30.3%] for Gini. These ranges fall modestly below the point estimates of 28% and 33%, as expected given that residual noise attenuates trend signals in some realizations. Nevertheless, a consistent fraction of rivers show significant increasing trends across all 1,000 realizations, confirming that the reported network-wide trend prevalences are robust to within-site reconstruction uncertainty.

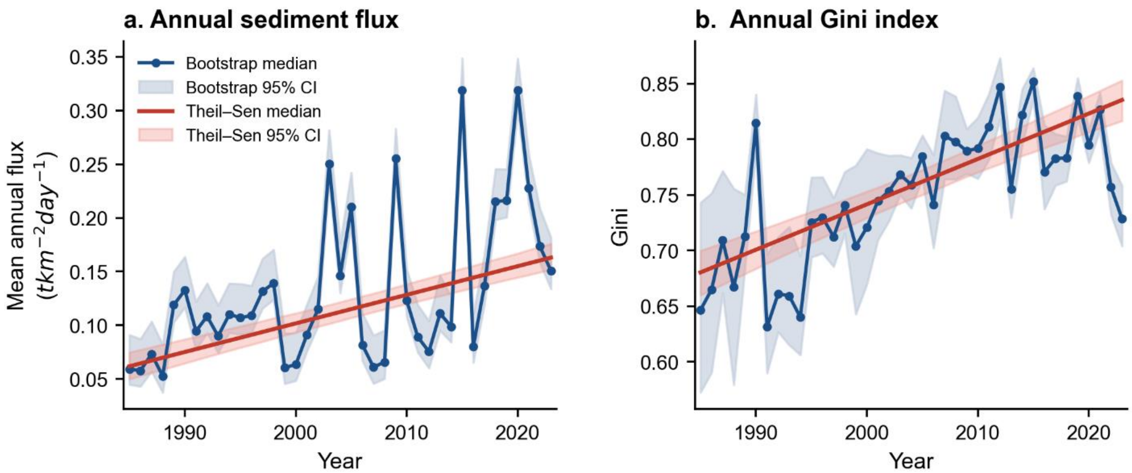


Figure. Demonstration of Moving Block Bootstrap (1,000 iterations, 10-day block size) for estimating uncertainty in annual sediment flux and Gini values. The dark blue line with points is the bootstrap median, and the solid red line is the Theil-Sen median trend line. Light blue and light red shaded regions represent 95% confidence intervals (CI) for the annual prediction and the trend line, respectively. The figure demonstrates that model results are robust to the introduction of uncertainty into the sediment flux time series by residual sampling.

5. The turbidity-to-SSC conversion framework is reasonable and well-grounded in prior guidance, and the reported median R2 values are strong. However, the manuscript should more fully discuss non-stationarity and structural uncertainty in these relationships. Turbidity–SSC regressions can vary seasonally, with sediment source, and with particle-size distribution, and sensors can drift over multi-year deployments. Because these regressions are used to generate the effective training targets for the ML models, their uncertainty propagates throughout the workflow. A sensitivity analysis in which regression parameters are perturbed within their confidence bounds, and the downstream effect on SSC, flux, and Gini trends is evaluated, would materially strengthen the technical rigor. At minimum, the Methods and Discussion should more explicitly acknowledge these limitations and their likely direction of influence.

   Answer:

   We agree that structural uncertainties in turbidity-SSC conversions (e.g., sensor drift, shifting particle sizes) propagate through the modeling chain. While a full parametric sensitivity analysis across 175 sites is beyond the scope of this revision, our newly added moving block bootstrap analysis (see Comment 4) indirectly captures this regression error and other sources of model misfit by resampling the final model residuals.

   Per your recommendation, we have expanded the Methods and Discussion to explicitly acknowledge these limitations and their likely direction of influence. In our log-log regressions, the slope acts as the scaling exponent. If static models systematically under-predict extreme peaks during floods (due to sensor saturation), the temporal distribution of transport is artificially "flattened." Because flattening extreme peaks mathematically lowers the Gini index (Voichick et al., 2018), we now explicitly note in the Discussion that our reported increasing trends in sediment burstiness likely represent a conservative lower bound of the true intensification occurring in these rivers.

   - **Voichick, Nicholas, David J. Topping, and Ronald E. Griffiths. "False low turbidity readings from optical probes during high suspended-sediment concentrations."** ***Hydrology and Earth System Sciences*** **22.3 (2018): 1767-1773.**

6. The modeling framework is generally well described, but several details required for full reproducibility are missing or inconsistent. The text reports different epoch counts for sediment flux training in different sentences, which should be reconciled. Key hyperparameters such as optimizer type, learning rate, early stopping criteria, and hyperparameter tuning strategy should be reported. It should also be clarified whether models were trained as fully pooled multi-basin models or with basin-specific fine-tuning. Given the growing emphasis on reproducibility in ML-enabled Earth science, adding these details will be important for peer review.

   Answer:

   We have updated the Methods section to provide a comprehensive technical specification of our modeling framework, addressing all the points raised . Specifically, we clarified the following.

   The LSTM models were implemented using the hydroDL framework with the Adadelta optimizer, using a default learning rate of 1.0 with a decay rate of 0.9. The models were trained using a fully pooled multi-basin approach; basin-specific fine-tuning was not applied. For the temporal validation multi-basin models, we used a fixed epoch setup rather than early stopping. This is because the pooled multi-basin training stabilizes loss at a regime-wide rather than site-specific minimum, and we verified plateau behavior visually. However, we did employ early stopping for the spatial cross-validation experiments to prevent overfitting, which we observed when using fixed epochs in that specific context. Hyperparameters were selected based on established literature for large-sample hydrology, rather than exhaustive grid search tuning. All models used RMSE as the loss function, a dropout rate of 0.3, and a minibatch size of 128.

   To ensure this information is easily accessible, we have added a new supplemental table (reproduced below as Table R2) summarizing these key hyperparameters for each model.

   We have also corrected the typographical error regarding the epoch counts in the Methods section. The text mistakenly listed flux twice with different epoch counts.

Table. Hyperparameters and training configuration for the streamflow, suspended sediment concentration (SSC), and sediment flux LSTM models.

Model	Hidden States	Sequence Length	Batch Size	Dropout Rate	Training Epochs
Flow	128	180	128	0.3	500
SSC	256	90	128	0.3	1000
Flux	256	90	128	0.3	500

7. Your performance assessment using NSE and KGE is appropriate, and the spatial cross-validation experiments are a strength, as shown in the supplementary figures. However, because your conclusions depend heavily on extremes, the evaluation should include metrics focused on high-flow and high-flux events. For example, performance conditioned on the top decile or top percentile of observed flux, peak timing error, and bias in annual load totals would be highly informative. Without this, reviewers may question whether models that perform well on average conditions are sufficiently accurate for inequality metrics driven by extremes.

   Answer:

   To address the concern that our Gini (burstiness) metrics might be influenced by model underperformance during extremes, we evaluated the model specifically on the top decile (Top 10%) and top percentile (Top 1%) of observed flows. Our results (Figure R5) show that model performance improves under high-flow conditions, with the median KGE increasing from 0.68 (full series) to 0.70 (Top 10%). Even at the absolute extreme (Top 1%), the model maintains a robust KGE of 0.51. While NSE drops at the 99th percentile (0.17) due to the high sensitivity of squared-error metrics to minor timing offsets in peak arrivals, the consistently high KGE components confirm that the model accurately captures the magnitudes and frequencies of the events that drive our inequality metrics.

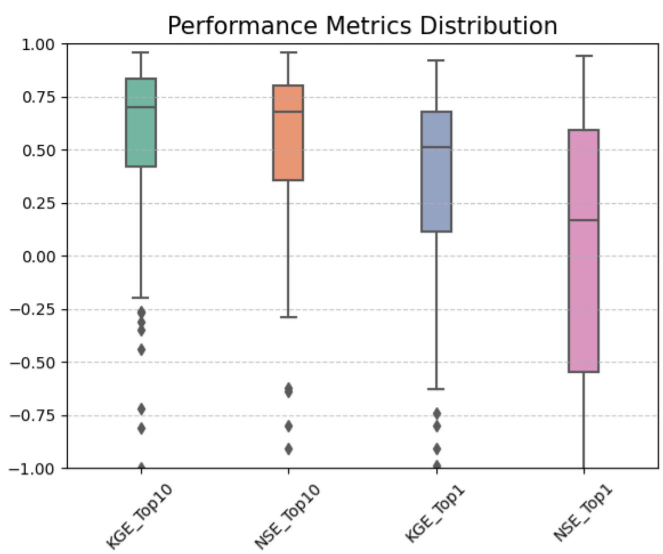


Figure. Model performance for extreme sediment flux events. Distribution of Kling-Gupta Efficiency (KGE) and Nash-Sutcliffe Efficiency (NSE) computed conditionally for the top 10% (KGE_Top10, NSE_Top10) and top 1% (KGE_Top1, NSE_Top1) of observed daily sediment flux days across all 175 sites.

8. The attribution framework using hierarchical dynamic linear models and Shapley decomposition is innovative and appropriate, but the language describing results should be more cautious regarding causality. At several points, the manuscript states or implies that urbanization or forest loss “drives” increases in burstiness. While the modeling strongly supports association and relative contribution within the chosen predictor set, it does not establish causality in a strict sense. Revising the wording to emphasize attributed contribution or statistical association and explicitly noting the possibility of omitted or correlated drivers such as channel engineering, bank stabilization, wildfire, or agricultural practice change, will improve scientific precision and credibility.

   Answer:

   Six specific language changes were made throughout the manuscript, we now use ‘attributed’ language rather than ‘driver’ language:

   i. **Section heading (L171) — "Urbanization and forest loss drive…" → " Increasing burstiness attributed to urbanization and forest loss"**
   ii. **Abstract (L18) — "drives increasing yields" → " they are governed by distinct attributed drivers: land-use change is the dominant predictor of temporal compression, while intensifying precipitation is the dominant predictor of rising yields "**
   iii. **Attribution results paragraph (L172–178) — replaced "most strongly governs" to “is most strongly attributed” and with attributed/associative language, and added a new sentence explicitly naming potential omitted drivers (channel engineering, bank stabilization, wildfire, agricultural practices).**
   iv. **Big Haynes Creek case (L184) — "confirms that…accounted for nearly all" → "jointly accounting for roughly two-thirds of the predictor-attributed change"**
   v. **Regime shift framing (L186–187) — "forced the regime shift" → "is most strongly associated with the observed regime shift"**
   vi. **Synthesis sentence (L194–196) — "it is the expansion of…that concentrates export" → "the expansion of impervious surfaces and removal of natural vegetation are the dominant attributed predictors" plus explicit acknowledgment of confounding factors.**

9. The statistical trend analysis should also be described in more detail. You report Mann–Kendall and Theil-Sen trends across many sites and variables, but it is not clear whether autocorrelation adjustments or prewhitening were applied, nor whether any correction for multiple testing across sites was considered. These details should be added to the Methods. Given the multi-site framework, a brief note on false discovery control or expected false positive rates would be appropriate.

   Answer:

   Autocorrelation: Prior to trend analysis, we evaluated the time-series for serial dependence (lag-1 autocorrelation). We found that autocorrelation was negligible across our annual metrics; therefore, prewhitening was not applied to the dataset.

   Multiple testing: To assess long-term hydro-environmental changes across the study sites, we employed a two-step trend evaluation framework. First, local trends for each metric were evaluated using a Mann-Kendall test ($\alpha = 0.05$) to identify significant increasing or decreasing trajectories at individual monitoring locations.

   Because spatial cross-correlation among interconnected or proximate watersheds can artificially inflate the regional false discovery rate, we subsequently evaluated the global metrics using the Livezey and Chen (1983) field significance test. This was implemented via a Monte Carlo approach (1,000 permutations) wherein the temporal domain (years) was randomly shuffled while strictly preserving the spatial cross-covariance structure across all sites. The observed total number of locally significant trends was then compared against the 95th percentile of the simulated null distribution to determine whether the regional occurrence of trends constituted a globally significant field response rather than spatially correlated noise.

   - **Livezey, R. E., & Chen, W. Y. (1983). Statistical field significance and its determination by Monte Carlo techniques. *Mon. Wea. Rev*, *111*(1), 46-59.**

   The results of this multi-site approach indicated globally significant trends for all four variables (flow, SSC, flux, and Gini). As an example, we have included a figure below demonstrating the test results for sediment flux. The 95% upper limit of the null distribution (the threshold for field significance) is 27 watersheds. Our observed number of watersheds with significant increasing trends is 51, which vastly exceeds this threshold ($p = 0.004$).

   We have added these methodological details regarding the autocorrelation checks and the Livezey and Chen field significance testing to the Methods section. We have also included the field significance histograms in the Supplement.

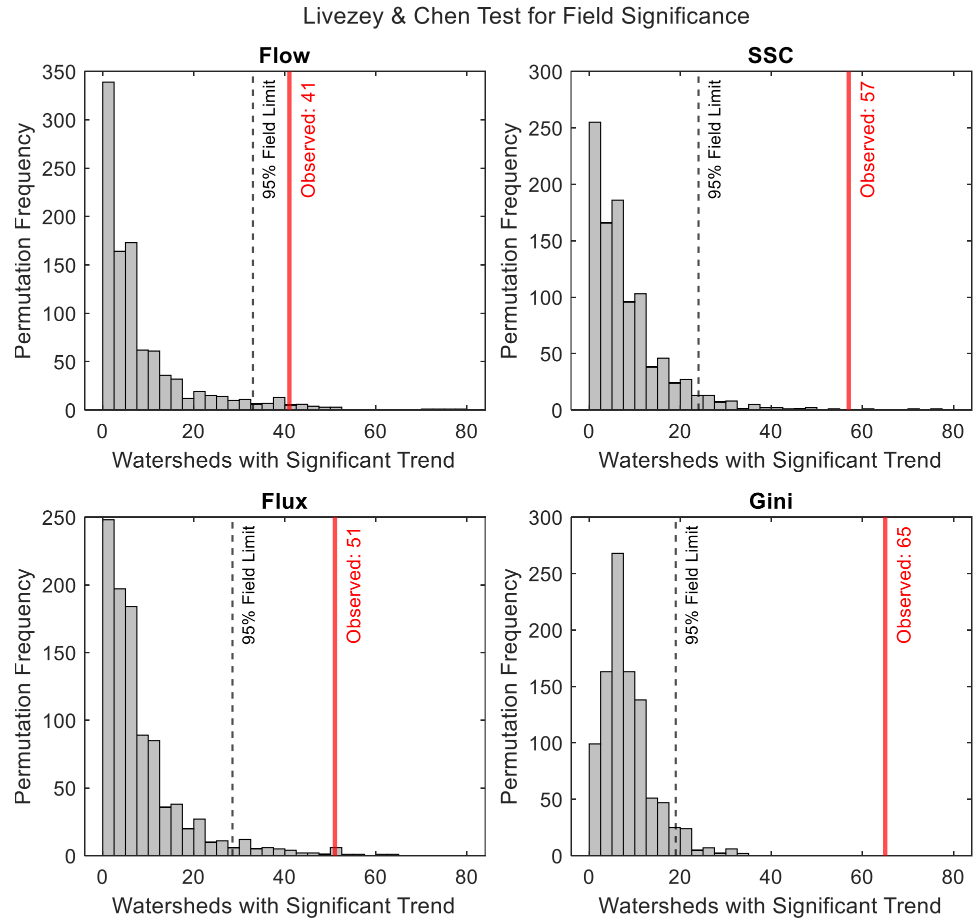


Figure. Field significance test results for trends in mean annual flow, suspended sediment concentration (SSC), sediment flux, and Gini coefficient across the study sites. The histograms represent the null distribution of the expected number of watersheds with significant trends, generated via 1,000 Monte Carlo permutations that shuffle the temporal domain while preserving the spatial cross-covariance structure. The dashed black line marks the 95th percentile threshold (95% Field Limit) of this null distribution. The solid red line indicates the actual observed number of watersheds exhibiting locally significant trends (Mann-Kendall test, $p < 0.05$).

10. The presentation of burstiness metrics is clear and effective, but readers would benefit from additional intuition about their sensitivity and complementarity. A short methodological note or supplementary analysis showing how Gini and B90 respond differently to changes in event frequency versus event magnitude would help interpret cases where the metrics diverge. Including uncertainty ranges on median B90 changes over time would also improve interpretability.

    Answer:

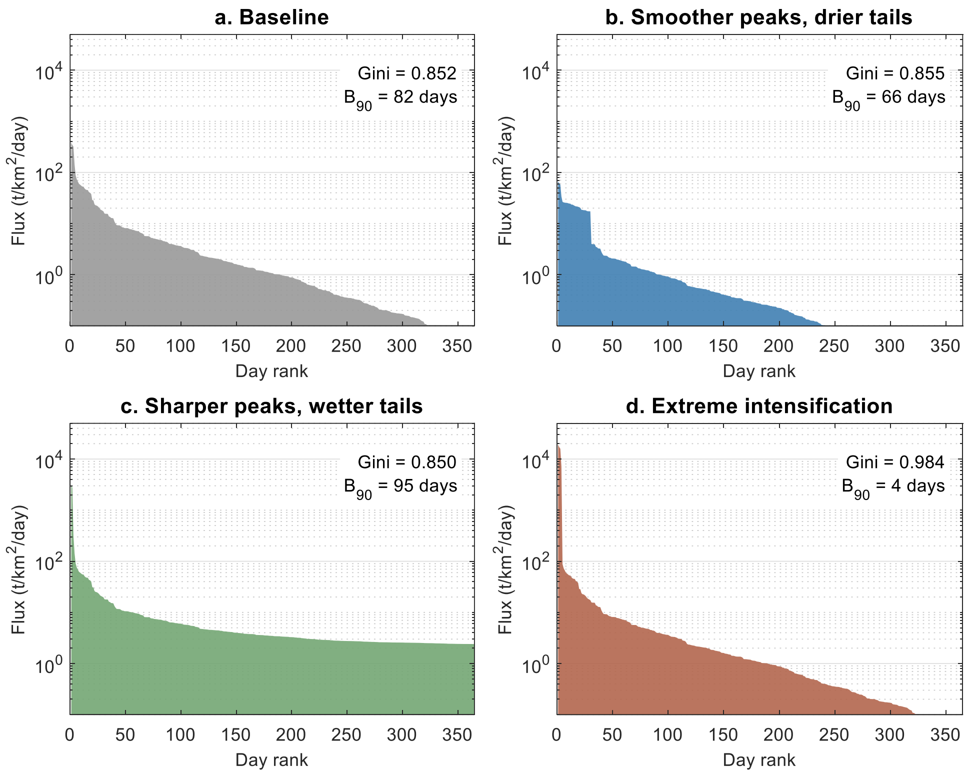
    We include a new methodological note in the supplement. To illustrate the sensitivity and complementarity of Gini and B90, we constructed four synthetic annual flux time series from a lognormal baseline and systematically varied event frequency and magnitude (Figure R6, on next page). The results demonstrate a clear divergence between the two metrics depending on the nature of the change, highlighting exactly why both metrics are required to fully characterize shifting sediment regimes. Here are the four scenarios we test:

    - **Scenario a: the baseline flux data set.**
    - **Scenario b: when extreme peaks are smoothly dampened and the low-flow tail is severely dried to compensate, the global inequality (Gini coefficient) remains identical to the baseline (ΔGini = 0.000), but B90 decreases. This illustrates that B90 isolates the local temporal compression of the remaining mass even when global inequality is mathematically conserved.**
    - **Scenario c: when top peak events are heavily sharpened but offset by the addition of a chronic, elevated baseflow, the Gini coefficient again remains identical to the baseline (ΔGini = 0.000), but B90 increases. This demonstrates how elevated background loads delay the 90% accumulation threshold despite the presence of extreme top-end spikes**
    - **Scenario d: metrics converge only when the load is simultaneously concentrated into fewer and more extreme days (e.g., the top 1% of events amplified 50-fold). Here, both metrics strongly agree to indicate a severe shift in burstiness, yielding the highest Gini and the shortest B90.**

    This distinction is directly relevant to our attribution results: the dominance of urbanization in driving Gini increases (Fig. 6) is consistent with a peak-magnitude amplification mechanism (Figure R6, scenario c), wherein impervious surfaces generate sharper, higher-intensity runoff responses. The comparatively weaker role of R95p in driving Gini — despite its strong influence on mean flux — is consistent with a frequency-amplification mechanism (scenario b) that increases total load without necessarily concentrating transport into fewer days.

    To quantify uncertainty around the reported contraction in the sediment transport window, we applied a moving block bootstrap procedure (1,000 iterations and 10-day block size) to derive 95% confidence intervals on the network-wide median B90 for each year. For each bootstrap, we took the difference between the median B90 days in 1985 and 2023. The effect was considered significant if the 95% CI on this difference did not include 0 days.

The paired difference in median B90 (and 95% CI) between 1985 and 2023 for all the bootstraps is 9.9 [4.6, 15.7] days, indicating that the results are significant. The 95% confidence interval on median B90s across sites were 69.7 [48.6, 89.8] days for 1985 and 59.8 [42.2, 79.3] days for 2023. These interval bounds contain the optimal LSTM model B90 estimates of 69 days in 1985 and 50 days in 2023.



Scenario	Physical Description	Gini	B_{90}	ΔGini	ΔB_{90}
a	Lognormal baseline distribution	0.852	82	--	--
b	Peak smoothly dampened, tail dried to compensate	0.855	66	**+0.003**	**-16**
c	Peak sharpened, high baseflow added to compensate	0.850	95	**+0.000**	**+13**
d	Top 1% amplified 50x, background unchanged	0.984	4	**+0.132**	**-78**

Figure. Sensitivity and complementarity of burstiness metrics to changes in event frequency versus magnitude. Ranked daily sediment flux distributions (sorted descending by day rank) for four synthetic scenarios derived from a lognormal baseline ($\mu = 0$, $\sigma = 2.0$). (a) Baseline distribution. (b) A scenario where extreme peaks are smoothly dampened and the low-flow tail is severely dried to compensate. The global inequality (Gini coefficient) is relatively unchanged from the baseline (ΔGini = +0.003), but B_{90} decreases, illustrating that B_{90} isolates the local temporal compression of the remaining mass even when global inequality is conserved. (c) A scenario where top peak events are heavily sharpened, but offset by the addition of a chronic, elevated baseflow. The Gini coefficient again remains identical to the baseline (ΔGini = 0.000), but B_{90} increases, demonstrating how elevated background loads delay the 90% accumulation threshold despite the presence of extreme top-end spikes. (d) An extreme intensification scenario where the top 1% of events are amplified 50-fold with no change to the background. Both metrics strongly converge to indicate a severe shift in burstiness, yielding the highest Gini and the shortest B_{90}.

11. In the Discussion and Implications sections, the management relevance is strong and well-described, especially regarding narrowing intervention windows and BMP timing. To further meet the standard of Communications Earth & Environment, you could broaden the framing slightly to connect with global change themes, such as compound extremes, hydroclimatic intensification, and infrastructure resilience under non-stationary sediment regimes. A short paragraph on how this framework could be transferred to other regions given sufficient sensor data would increase generality and impact.

    Answer:

    We agree that framing our results within the broader context of global change themes strengthens the manuscript's overall impact. To address this, we have expanded the Discussion section to explicitly connect our findings to hydroclimatic intensification and compound extremes. Specifically, we note that as extreme precipitation intensifies globally [1,2]. the co-occurrence of severe storms with degraded landscapes can generate massive sediment pulses that challenge current monitoring standards and management approaches [3]. We also added text discussing the implications for infrastructure resilience, highlighting that water treatment facilities and reservoirs designed around historical, stationary baselines may face increased stress as sediment delivery shifts toward higher magnitude episodes under non-stationary conditions [4,5].

    Finally, per your recommendation, we added a short concluding paragraph discussing the transferability of this framework. We highlight that non-stationary sediment regimes are emerging globally (such as those recently documented in pan-Arctic rivers [6]). We conclude by noting that wherever high-frequency in-situ sensing or remote sensing data are available to serve as SSC proxies (such as in a recent remote-sensing driven study; Lucchese et al., 2026), strategic expansion of these networks could enable this class of ML-reconstruction framework to diagnose shifting sediment regimes in other regions worldwide.

12. Finally, the data and code availability statements are good but could be strengthened by committing to release trained model weights, processed training targets (where permitted), and scripts for computing Gini and B90 metrics. Given the key role of ML reconstruction in the study, enhanced openness will be viewed favorably.

    Answer:

    We have updated our Data and Code Availability statements. We commit to releasing the trained LSTM model weights for all 175 study sites, the processed high-frequency training targets used for model development, and the Python scripts used to calculate the Gini coefficient and B90 metrics. These resources will be permanently archived on *Hydroshare*.

    The requested data are stored at this link. https://hydroshare.org/resource/bba202f5eda04ae0abfab695e8c4750e/. Upon acceptance of this manuscript, this link will be turned into a permanent DOI.

References

1. Fowler, H. J. et al. Anthropogenic intensification of short-duration rainfall extremes. Nat Rev Earth Environ **2**, 107–122 (2021).
2. Papalexiou, S. M. & Montanari, A. Global and Regional Increase of Precipitation Extremes Under Global Warming. Water Resources Research **55**, 4901–4914 (2019).
3. Zscheischler, J. et al. A typology of compound weather and climate events. Nat Rev Earth Environ **1**, 333–347 (2020).
4. Vogel, R. M. et al. Hydrology: The interdisciplinary science of water. Water Resources Research **51**, 4409–4430 (2015).
5. Milly, P. C. D. et al. Stationarity Is Dead: Whither Water Management? Science **319**, 573–574 (2008).
6. Tian, S. et al. Increasing river sediment concentration and flux across the pan-Arctic. Nat. Geosci. 1–8 (2026) doi:10.1038/s41561-026-01960-z.

Reviewer #2 Comments:

The authors use a chain of models to estimate and attribute trends in suspended sediment flux across ~40 years in the eastern US. The work is timely and exciting, exploring the relative influence of landcover versus hydroclimate change on suspended sediment dynamics in a period when both these drivers have shifted. While the methods are suitable and presentation is excellent, I feel there is work needed on quantifying uncertainty and its propagation across the model chain, on exploring the roles of basin size and streamflow extremes, and on extending the discussion.

We thank the reviewer for the thorough reading and critique of the manuscript. Their comments have helped improve its quality.

Major comments:

1. *Clarifying terminology around models and observations*

The authors use observations to fit linear-regression models, whose outputs become the target for a LSTM model and an HDLM model that use additional spatial-modelling outputs from other papers as input “data”. The words “simulated” and “modelled” are used throughout, but it was often unclear to me which of these models was referred to. I suggest specifying carefully throughout – I know this is frustrating but it will make things much easier for the reader to follow.

Similarly, the time-series outputs of the linear-regression models are often referred to with words such as “records” or “reconstructed” or “proxy”; even “observed” is sometimes used. Together, this language implies these model-outputs are observations instead. I urge the authors to carefully review the language used to describe the linear-regression outputs, ensuring that readers who skip the methods don’t get the false impression that the “reconstructed” SSC or flux time-series used to train the LSTM are in fact in-situ observations.

Finally, it’s worth noting that Daymet and other LSTM inputs (e.g. some from HydroAtlas) are outputs of spatial models; a quick note on this point will help readers to separate these model outputs from other data sources (e.g. SSC grab samples) that are direct observations.

Answer:

We have undertaken a careful, manuscript-wide revision of the language used to describe each component of the modeling chain. We have identified three classes of objects and specifically use the terminology associated with them:

- i. **Throughout the manuscript, we distinguish three classes of quantities. (1) Direct measurements — USGS gauge discharge, raw turbidity sensor readings, and discrete grab-sample SSC — are referred to as 'measured,' or 'discrete grab samples.' (2) The continuous 2007–2023 SSC series produced by the site-specific turbidity–SSC regression is referred to as 'turbidity-derived SSC.' (3) The continuous 1985–2023 flow, SSC, and flux series produced by the LSTM are referred to as 'LSTM-reconstructed' or, where context is unambiguous, 'reconstructed.' We also refer to the flux series Q_s (t) = SSC(t) × Q(t) — computed as the product of observed gauge discharge and turbidity-derived**

SSC, and used as the LSTM training target — as the 'observed sediment flux,' reflecting that its discharge component is a direct measurement and that no continuous sediment flux measurement exists against which the LSTM could otherwise be benchmarked. In the modeling literature, it is common to use this approach (turbidity-derived SSC × discharge = observed flux). We do this in the same way that discharge itself is a function/model of stage/height, but is still referred to as an observation.

- **ii.** **The word "data" is removed if referring to an model output, such as in the original Figure 5.**
- **iii.** **"Reconstruction" refers to LSTM outputs in all contexts.**
- **iv.** **Clarifying the nature of LSTM training targets. In the Methods section (Section: Continuous sediment time series generation and Model training), we have added explicit language clarifying that the training targets for the LSTM sediment flux and SSC models are the turbidity-derived regression estimates rather than direct in-situ observations.**
  - **a.** **Specifically, we now state: "We refer to these continuous series as turbidity-derived SSC and turbidity-derived flux throughout the manuscript, distinguishing them from the discrete in-situ grab-sample SSC measurements (referred to as "observed" or "discrete SSC") used to fit the regression. We similarly distinguish these turbidity-derived series, which serve as training targets for the LSTM, from the LSTM-reconstructed flow, SSC, and flux time series spanning 1985–2023, which are the primary subject of the trend and attribution analyses presented in this study. Where the active model is unambiguous from context we use "reconstructed" as shorthand for LSTM-reconstructed"**
- **v.** **We now note that Daymet and HydroATLAS are modeled estimates.**
- **vi.** **In Figure 2, for the "observed" flux time series, we add the following:**
  - **a.** ***"The observed time series refer to the daily sediment flux computed as the product of discharge and turbidity-derived suspended sediment concentration."***

2. *Deeper discussion of implications*

I feel the discussion could go deeper. I have a few suggestions in the paragraphs below, but I also wondered: what are the implications for sediment mass balance? Should we worry about sediment exhaustion or over-accumulation in sinks, in places with marked trends? What about implications for river lateral mobility? Obviously as a reviewer one immediately thinks about implications for one's own work. Evan Greenberg (e.g. Greenberg & Ganti, 2024) showed that river mobility increases with sediment flux, while analysis we did together (Leenman et al., 2025) shows that flow variability is another driver. If both sediment flux (Qs) and R95p are increasing (e.g. Fig 4a and Ext.Dat.Fig.5), could there be implications for river lateral mobility?

Answer:

We will first put our response to your particular interest in river lateral mobility. We think it is of interest to readers of our paper and have included it in our discussion. Because we don't directly measure this, we present it as a possibility

- *"Increasing flow variability and sediment supply may also accelerate lateral migration and floodplain reworking in many of the eastern basins where both forcings are intensifying [51,52]."*

Our results indicate that, across the 175 study basins, 28% show significant increases in mean sediment flux (Extended Data Table 1) and 58% show significant increases in extreme wet-day precipitation (R95p; Extended Data Table 1). The co-occurrence of increasing Q_s and increasing flow variability (R95p is a proxy for the frequency and intensity of flood-generating precipitation) in many eastern U.S. basins thus creates a compound forcing that, based on the frameworks of Greenberg & Ganti (2024) and Leenman et al. (2025), would be expected to accelerate lateral migration and floodplain reworking rates. We make this note in the updated discussion.

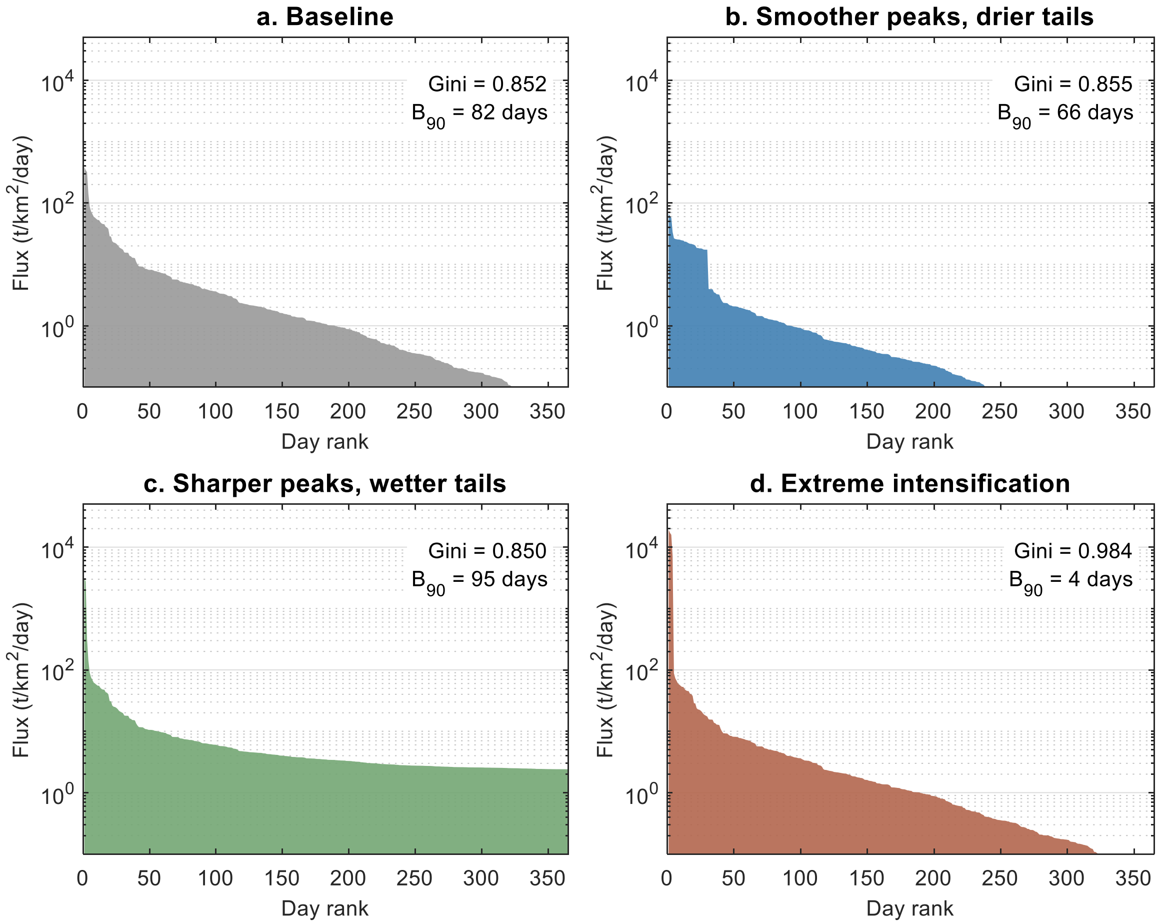
Regarding sediment mass balance and downstream sinks: our 175 study basins are predominantly small to mid-sized and located upstream of major reservoirs and estuaries. The compound forcing of increasing flux and increasing burstiness implies that downstream depositional environments — reservoirs, floodplains, and estuaries — will receive material in larger, more episodic pulses than historical regimes accommodate. In supply-limited basins, this could accelerate the depletion of upland sediment stores; in supply-rich systems, the same dynamic could overwhelm the sediment-management capacity of downstream reservoirs and trigger more frequent dredging or sluicing operations (Kondolf et al. 2014). Quantifying these basin-scale mass-balance implications is an important next step beyond the scope of the present study.

We have reworked the discussion, particularly the last section, considerably to integrate your requests and connect with the broader literature and implications.

3. *Comparison among different fluxes and basin sizes*

In Figure 1, I feel the sites selected should have approximately the same mean flux, to allow more meaningful comparison of high vs low B90. For instance, if one site has 10x fewer transport days but also has a 10x smaller annual flux than the other, we can hardly say it is a more hazardous site nor a more efficient sediment transporter. It's possible the authors have already selected sites with a similar Qs, in which case please annotate the figure to say so.

Answer: Figure 1 is intended as a conceptual illustration of the burstiness spectrum rather than a quantitative comparison of hazard or transport efficiency between two basins. We have updated the figure caption to make this purpose explicit. To address the reviewer's underlying concern about whether sites with similar mean flux can have different Gini/B90, we now include a new supplemental figure demonstrating exactly this scenario. It can be found below, see a detailed description of scenarios in our response to Reviewer 1 Comment 10.



Scenario	Physical Description	Gini	B_{90}	ΔGini	ΔB_{90}
a	Lognormal baseline distribution	0.852	82	--	--
b	Peak smoothly dampened, tail dried to compensate	0.855	66	**+0.003**	**-16**
c	Peak sharpened, high baseflow added to compensate	0.850	95	**+0.000**	**+13**
d	Top 1% amplified 50x, background unchanged	0.984	4	**+0.132**	**-78**

Figure. Sensitivity and complementarity of burstiness metrics to changes in event frequency versus magnitude. Ranked daily sediment flux distributions (sorted descending by day rank) for four synthetic scenarios derived from a lognormal baseline ($\mu = 0$, $\sigma = 2.0$). (a) Baseline distribution. (b) A scenario where extreme peaks are smoothly dampened and the low-flow tail is severely dried to compensate. The global inequality (Gini coefficient) is relatively unchanged from the baseline (ΔGini = +0.003), but B_{90} decreases, illustrating that B_{90} isolates the local temporal compression of the remaining mass even when global inequality is conserved. (c) A scenario where top peak events are heavily sharpened, but offset by the addition of a chronic, elevated baseflow. The Gini coefficient again remains identical to the baseline (ΔGini = 0.000), but B_{90} increases, demonstrating how elevated background loads delay the 90% accumulation threshold despite the presence of extreme top-end spikes. (d) An extreme intensification scenario where the top 1% of events are amplified 50-fold with no change to the background. Both metrics strongly converge to indicate a severe shift in burstiness, yielding the highest Gini and the shortest B_{90}.

In Figure 3, I was struck by the similarity in spatial patterns of Gini index and Qs. Are they correlated? If yes, is this (a) a real physical phenomenon vs (b) a result of how the Gini index is computed? If (a), this is exciting, and **worth further exploration in the discussion –** first, it's interesting to understand how Qs varies across hydroclimates (indexed by the Gini index, to some extent), and second, it implies that, even in the absence of any change to flux burstiness, the sites that move the most sediment do it in the shortest time, which has implications for managing sediment. If (b), it is possible that the apparent trends in Gini across ~1/3 of the sites arise simply from the observed trends in flux across ~1/3 of the sites. The authors could check for (b) by artificially inflating values in the Qs time series and exploring whether there is any impact on Gini.

Answer: Regarding the correlation between Gini and Qs in Figure 3: The Gini coefficient is a purely distributional metric derived from ranked daily fluxes within a year — it is mathematically independent of the total magnitude (scaling all daily values by a constant leaves the Gini unchanged; see the above Figure for this demonstration). Therefore, a spurious computational relationship between Gini and Qs magnitude (scenario b) can be ruled out. The observed spatial co-variation between mean Qs and Gini is a real physical signal (scenario a), reflecting a real co-occurrence in which the basins moving the most sediment also tend to move it in the most concentrated bursts, likely driven by combinations of intense rainfall regimes and land disturbance (e.g., deforestation and urbanization).

On a related note, in L137-138 I wondered if the ~1/3 of sites with increasing Gini are the same ~1/3 of sites with increasing Qs? If (a) above holds true, a scatter plot of ΔQs against ΔGini would address this question. I note that E.D. Fig 3 goes some way to answering this question with "no", which makes the title of the paper rather an oversimplification – it is *not* the same rivers transporting more sediment and in less time. A scatter of ΔQs against ΔGini, coupled with an earlier and more explicit discussion of this mismatch demonstrated in E.D.Figure 3, would allow the authors to get into these nuances, and would **make for some fruitful additional discussion**.

Answer: We scatter ΔQ_s against ΔGini in a new Figure 5. Among the 175 sites, 27 show statistically significant increases in both flux and Gini, 31 show significant Gini increases only, and 22 show significant flux increases only. While simultaneous changes in both magnitude and timing do occur, changes in one can occur independent of changes in the other. We have therefore revised the title to

'U.S. rivers are transporting more suspended sediment, often in less time'

which more accurately captures both the directional finding and the partial decoupling of magnitude and timing trends. We have also added a paragraph in the Results section that engages with this decoupling explicitly, distinguishing the three populations of sites and their dominant attributed drivers.

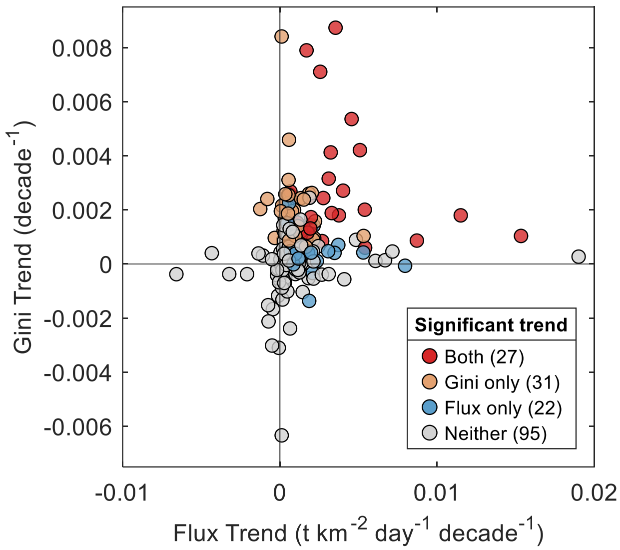


Figure. Partial decoupling of sediment magnitude and timing trends (1985–2023). Scatter plot comparing the decadal trend in mean annual sediment flux against the decadal trend in the Gini coefficient across all 175 study sites. Markers are color-coded based on the statistical significance of their respective trends (Mann-Kendall test, $p < 0.05$). Red markers ($n = 27$) represent basins experiencing concurrently significant increasing trends in both total flux and burstiness. Orange markers ($n = 31$) denote basins exhibiting a significant increase in burstiness only, while blue markers ($n = 22$) denote basins with a significant increase in total flux only. Gray markers ($n = 95$) represent sites lacking a significant increasing trend in either metric.

L142: “68% urban cover” – was this a small basin? It would be useful to know how ΔQs and ΔGini vary with basin size, as any scale-dependence would allow the reader to know how generalizable the results are. Moreover, a scale-dependence could be useful for management: e.g., if the ~1/3 of basins that gained in burstiness or Qs were also the smallest basins, this would highlight which catchments are most sensitive to changing land use or climate.

Answer: Regarding basin size at the Brushy Fork Creek example (L142): Yes, Brushy Fork Creek is a small basin (20 km²). To address the reviewer's broader question about scale-dependence, we computed the median drainage area for sites exhibiting significant trends. Median basin areas are 53 km² for sites with increasing flux, 54 km² for sites with increasing Gini, and 20 km² for sites with increases in both. The median size for basins with no significant trends is 268 km². This suggests that basins exhibiting simultaneous intensification of both burstiness and magnitude tend to be smaller, which is physically consistent with the greater sensitivity of small catchments to land-use change and flashy runoff generation. This concords with recent work that identified headwaters as susceptible to change (Golden et al., 2025). We have added this context to the manuscript, noting that the co-occurring signal is most pronounced in small, rapidly urbanizing basins (such as Brushy Fork Creek), while trends in either flux or burstiness alone span a broader range of basin sizes.

4. *Quantifying propagating uncertainty across the model chain*

As I understand, a linear-regression model (model 1) generates synthetic SSC time-series from turbidity time-series and SSC grab samples. The synthetic SSC is combined with observed(?) streamflow to model daily suspended sediment flux (model 2). Outputs from the Daymet spatial modelling effort (model 3) are combined with other large-sample estimates and input to the LSTM (model 4), whose target variables are the outputs of models 1-2. Finally, the longer term synthetic records generated by model 4 are used to fit model 5, the HDLM, which is used for attribution. I hope I have correctly understood the model chain based on its description in the methods & code; if not, please consider revising the method description. My main point here is that we need to see **how uncertainty propagates** across this model chain. Moreover, it is important to know how **sensitive** the LSTM and HDLM outputs are to uncertainty, and to decisions made in earlier models. For instance, how do the fits in E.D.Fig. 9 change if the R2 threshold of 0.4 for the SSC linear regressions is raised? Moreover, if the median site-level R2 for the HDLM's Gini index compared to model-2's synthetic record is only 0.41, how much confidence can we have in the HDLM's attribution results for the Gini index? These two examples highlight my more general opinion that questions of uncertainty and sensitivity need more explicit consideration.

Answer: The reviewer's careful description of the model chain is correct. Regarding uncertainty propagation, we now conduct an extensive uncertainty analysis on the daily LSTM sediment flux estimates through a residual sampling approach. We employ a moving block bootstrap and generate 1,000 synthetic flux ensembles per site-year. The flux estimates for these ensembles are perturbed by adding residuals (errors) during model training. We then propagate the LSTM reconstruction uncertainty through to annual flux, Gini, and B90 estimates and their trends (see the new Figure on the next page). This bootstrapping captures LSTM prediction uncertainty against the turbidity-derived training targets, including the contribution of any noise or residual structure in those targets that the LSTM was unable to fit. It does not separately resolve uncertainty in the upstream turbidity–SSC regression itself; we acknowledge this limitation in the Methods and discuss its likely direction of influence (Reviewer #1, Comment 5).

We assessed the relative fraction of change explained by the coefficient drift parameter and found the median change ranged from 0% (for SSC) to 6% (for Gini). Thus, our explicit predictors successfully explain 94%+ of the systemic changes over time.

Regarding the sensitivity of results to the R^2 = 0.40 turbidity–SSC threshold: while re-running the full modeling pipeline at a higher threshold is beyond the scope of the current revision, we note that this threshold is already conservative in practice. As shown in Extended Data Fig. 1d, the majority of our 175 retained sites have turbidity–SSC R^2 values well above 0.40, with a median R^2 of 0.79. The concern about low-quality regressions driving results therefore applies to a small minority of sites, and the network-wide conclusions are unlikely to be sensitive to this threshold.

On the HDLM R^2 of 0.41 for Gini: we acknowledge this is lower than for flux and flow, reflecting the greater year-to-year variability in the Gini index. However, the HDLM is used here as an interpretive surrogate to decompose relative driver contribution ranks, not for precise prediction. This is now made more clear in the methods. The Shapley attribution results reflect broad, consistent patterns across 64 sites with increasing Gini,

lending confidence that the identified driver hierarchy is robust despite moderate site-level HDLM fit.

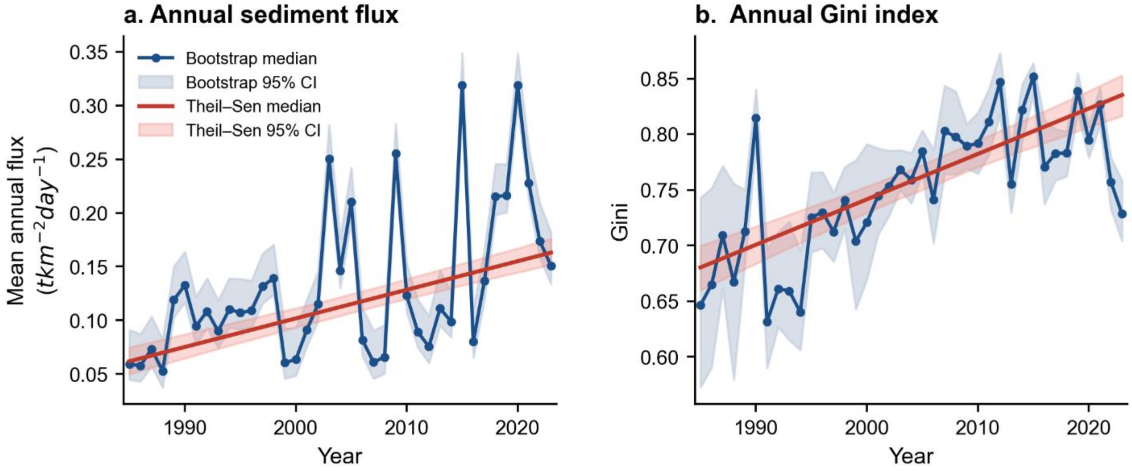


Figure. Demonstration of moving block bootstrap (1,000 iterations, 10-day block size) for estimating uncertainty in annual sediment flux and Gini values. The dark blue line with points is the bootstrap median, and the solid red line is the Theil-Sen median trend line. Light blue and light red shaded regions represent 95% confidence intervals (CI) for the annual prediction and the trend line, respectively. The figure demonstrates that model results are robust to the introduction of uncertainty into the sediment flux time series by residual sampling.

We also now include a figure showing more completely the model performance metrics and distributions for the HDLM model (next page).

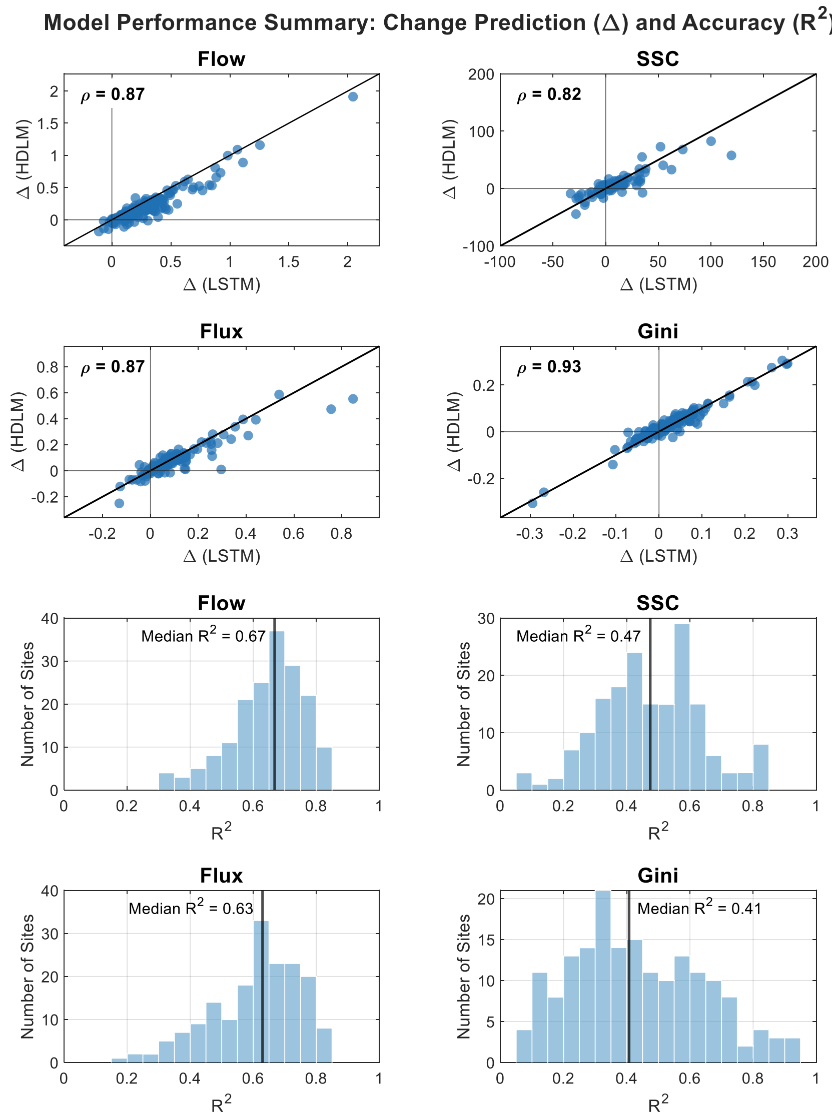


Figure. Summary of model performance for the HDLM across four hydrologic and sediment variables: flow, SSC, sediment flux, and the Gini coefficient. Scatter plots of LSTM-reconstructed versus HDLM-predicted changes (Δ) between baseline and recent periods for each variable, featuring a dashed 1:1 reference line and Spearman correlation coefficients (ρ). Histograms show the distribution of site-level model fit, as determined by R^2 of annual HDLM predictions versus LSTM reconstructions. A vertical line indicates median site-level model performance. HDLM = hierarchical dynamic linear model. LSTM = long short-term memory network.

Minor comments:

Fig 1 data selection: In Figure 1 and 2c-d, two contrasting degrees of Qs burstiness are compared, but I notice the left and right panels use data from 2020 and 2012 respectively. If the goal is to compare these two basins, I think we need to see them in the same year.

Answer: We selected the years that best represent each site's characteristic regime rather than matching them temporally. Forcing the same year would risk presenting an atypical example for one or both basins. We have updated the figure caption to clarify that these are illustrative end-members of the burstiness spectrum.

Assessing model performance across trends: the LSTM KGE and NSE for daily SSC and Qs are encouraging. Could the authors demonstrate (e.g. in the supplement) how KGE and NSE vary across different Qs *trends* ? Since the paper is about attributing trends in Qs, we need to be certain the LSTM performs equally well in sites with statistically significant Qs trends.

Answer: To ensure the robustness of our trend attribution, we analyzed the LSTM performance for basins with statistically significant flux and Gini trends. The results demonstrate that the model maintains high fidelity regardless of the trend magnitude.

- **For sites with significant flux trends: The median KGE (NSE) is 0.77 (0.78).**
- **For sites with <u>non</u>-significant flux trends: The median KGE (NSE) is 0.78 (0.78).**
- **For sites with significant Gini trends: The median KGE (NSE) is 0.73 (0.77).**
- **For sites with <u>non</u>-significant Gini trends: The median KGE (NSE) is 0.79 (0.78).**

On a similar vein, it worries me slightly that linear regressions were applied to model SSC back to 2007, with the LSTM then hindcasting further back to 1985. How confident can we be in LSTM estimates for the 1980s? Is LSTM uncertainty greater for years before 2007? The authors could colour points in E.D.Fig.9 by the "baseline" and "recent" time periods employed for attribution. It would also be useful to see versions EDFig9 made with just data from the baseline and recent periods, respectively, so we get an idea of how LSTM skill varies across the record from which trends are then measured.

Answer: We agree that it is crucial to establish trust in our model simulations during the hindcast period. While Extended Data Fig. 9 demonstrates that the LSTM accurately predicts the target outputs (SSC), we further validate hindcast credibility by verifying that the model inputs (climate and land use) during the 1985–2006 hindcast period do not systematically fall outside the range of environmental conditions the model was trained on (2007–2023).

To quantify this, we parameterized the multivariate feature space (precipitation, temperature, urban, agriculture, and forest fractions) of our training data and calculated the Mahalanobis distance for every site-year relative to the centroid of that training distribution. We then projected the hindcast (1985–2006) site-years into this same space to evaluate if the historical forcings were significantly further from the center of the training cloud (i.e., representing out-of-distribution inputs).

The new results are shown in the Figure below. We found no meaningful divergence between the distribution of model inputs in the hindcast and training periods. The median Mahalanobis distance for hindcast site-years was 1.85, compared to 1.91 for training site-years (p = 0.028, Wilcoxon rank-sum test). If the distributions were different, the hindcast period would have a significantly larger Mahalanobis distance. Between the direct SSC validation in Extended Data Fig. 9 and this new multivariate input analysis, we are confident that the LSTM is being deployed during a historical period whose environmental boundary conditions are fundamentally similar to those of the training period.

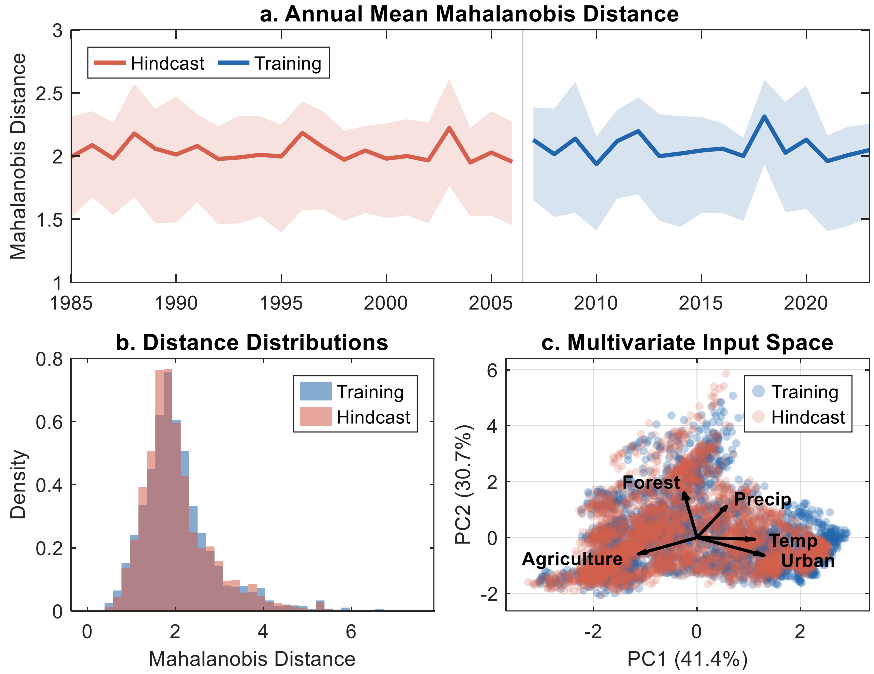


Figure. Evaluation of multivariate input space overlap between the LSTM training and hindcast eras. To validate the credibility of historical reconstructions, the 1985–2006 hindcast features were evaluated against the 2007–2023 training feature space using Mahalanobis distance. **(a)** Annual mean Mahalanobis distance across all sites, with shaded regions representing the interquartile range. **(b)** Probability density distributions of Mahalanobis distances for the training (blue) and hindcast (red) site-years, demonstrating that the historical inputs are well-contained within the modern training distribution. **(c)** Principal Component Analysis (PCA) biplot visualizing the multivariate input space. Points represent individual site-years projected onto the first two principal components (accounting for 72.1% of total variance). Black arrows represent the loading vectors for the five drivers.

Assessing trends or spatial patterns in extreme, not means: Figure 1 implies transport is event-driven. Why, then, look at mean annual Qs and water discharge (Qw) when looking for trends or spatial patterns? If urbanisation is altering runoff partitioning, we might expect the annual *mean* and annual *extremes* (e.g. Q99) of both Qs and Qw to exhibit differing behaviour – e.g. the mean may not change but extremes may become more "flashy" as runoff processes change. It may be useful to explore trends in extremes, not just in annual means, as (a) they tell us more about hazards, and (b) they may have a stronger signal.

Answer: We agree this would be an informative analysis. The Gini coefficient and B90 already capture much of the extreme-event behavior the reviewer is interested in (since both metrics are dominated by high-flux days), and our attribution analysis explicitly identifies R95p and urbanization as drivers most relevant to extremes. A separate trend analysis on Q99 of Q_s and Q_v is a natural extension that we view as a follow-up study, particularly given the additional uncertainty quantification it would require for high-percentile estimates.

In Ext.Dat. Fig 6, I was struck by the weak influence of R95p on Qw. It is possible that the 95th percentile of precipitation has a stronger influence on extreme, rather than mean, streamflow. Considering extreme streamflow percentiles instead of the mean may be informative.

Answer: To clarify, our analysis demonstrates a very strong influence of R95p on mean annual streamflow (Qw). As shown in the network-wide aggregates of Extended Data Figure 6 (Panels e and f), R95p is actually the single largest driver of increasing streamflow, accounting for 44% of the overall positive fractional share. Regarding the suggestion to model extreme streamflow percentiles: because our study's objective is to attribute changes in sediment temporal inequality directly to ultimate external drivers (precipitation and land use), explicitly modeling intermediate streamflow extremes falls outside the scope of our attribution framework.

On a similar note, I noticed in Ext.Dat.Fig 9 that the "model" (LSTM?) underpredicts extreme SSC and flux. This underprediction could result in underestimation of total flux some years, and could affect trend detection. I encourage the authors to (a) think about strategies to increase model skill for extreme values, (b) quantify for the reader, explicitly and in the main paper, how this underestimation of extremes affects the uncertainty of their final results, particularly trend detection, and (c) put the true R2 (currently in the caption) on panels a-b and e-f in this figure.

Answer: Yes, it is fairly typical for numerical models to have a bias whereby they underpredict high values and overpredict low values. Indeed, the underprediction does impact flux estimates. We now address the uncertainty introduced by under (and over) prediction with the block bootstrapping approach described earlier where we resample model residuals and perturb modeled time-series to get estimates on how much flux varies. To answer your points:

a. **Underestimation of extreme values is an inherent limitation of deep learning models, which we now mention in the discussion. Because these models learn best by matching patterns, they can identify connections between commonly**

occurring values (low to mid magnitude). However, for extreme values, which by definition don't occur frequently, the models have insufficient training data to learn from. One way that this can be reconciled is by fusing process-based and machine learning models using differentiable modeling, which provides a physical backbone to a hydrologic model but allows neural networks to learn some processes. This has yet to be done for sediment, but it is a promising area of future research.

- *"It is important to note that the intensification of sediment burstiness reported here likely represents a conservative lower bound of the true intensification because two methodological limitations (i.e., artificial flattening of extreme peaks by turbidity sensors [49] and underprediction of the rarest events by purely data-driven models [27]) combine to lower the computed temporal inequality."*

b. **We quantified our model performance for different extreme values in our response to Reviewer 1, Comment 7. Our model actually does very well for the top 10% of flow and flux values. However, it suffers for the top 1% because of the lack of training data and large penalties for missed timing of events (even if the magnitude is correctly predicted). Because the underprediction is consistent across the full record (1985–2023; see Extended Data Fig. 9 a vs b), it does not introduce a directional bias in trend detection, just a decrease in magnitude.**

c. **The figure has been updated to also include the true R2 (see on next page).**

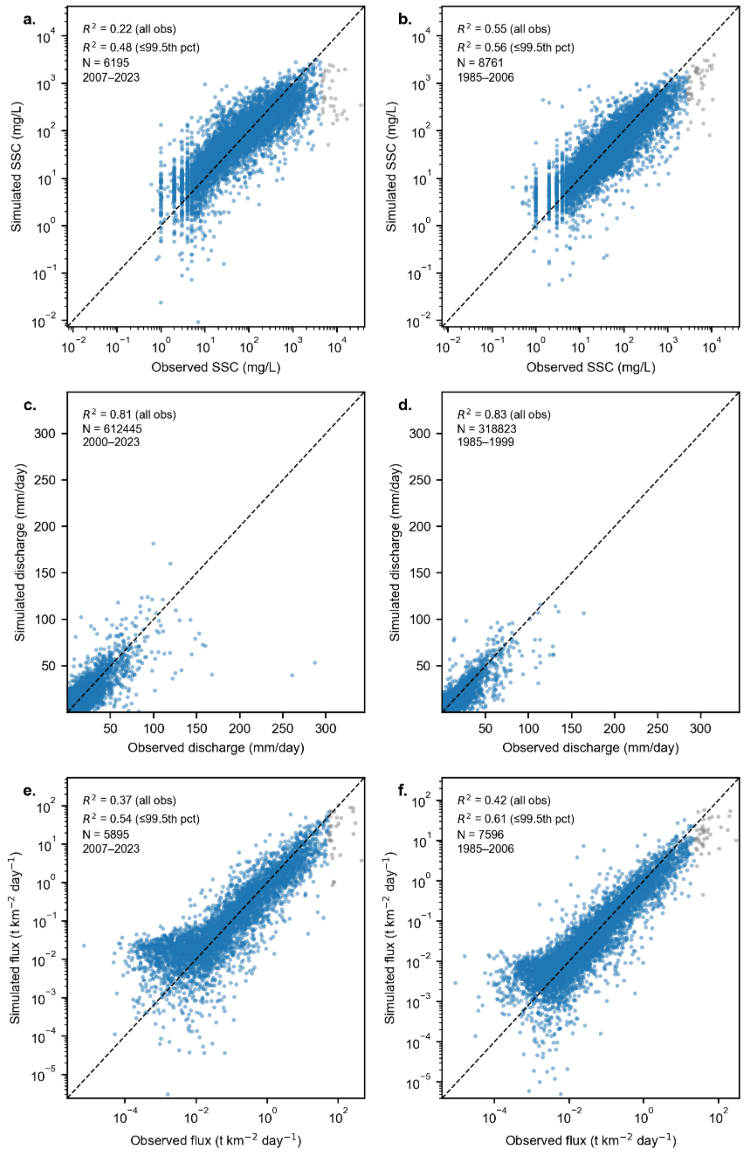


Figure. Model observation comparisons for suspended sediment concentration (SSC), discharge, and sediment flux across two time periods. Panels a–b show observed grab-sample SSC compared with model-simulated SSC for 2007–2023 and 1985–2006, respectively. Data are shown for the 66 sites that had multiple SSC grab samples for both the 1985–2006 and 2007–2023 periods. Panels c–d compare observed and modeled discharge for 2000–2023 and 1985–1999. Data are shown for the 76 sites that had multiple observed discharge datapoints for both the 1985–2006 and 2007–2023 periods. Panels e–f show observed versus modeled sediment flux for 2007–2023 and 1985–2006, using the same 66 sites as the SSC analysis. In all panels, each point represents a matched daily observation and model estimate, and the dashed line denotes the 1:1 reference line. For SSC (a–b) and flux (e–f), coefficients of determination (R^2) were computed using data up to the 99.5th percentile of observed SSC to reduce the influence of high-leverage extremes (gray markers; n = 32 for SSC, n = 30 for flux).

Overinterpretation of physical process from non-significant differences: In L96-98 and subsequent paragraphs (e.g. L 113-114 also stood out), the authors attempt to explain the physical processes responsible for differences between climates & land covers. However, most of these groups are not statistically significantly different in Fig. 3 b/d. I suggest the authors do not worry about interpreting differences that are not supported by statistical evidence.

Answer: We have softened the relevant passages by reframing observed median differences as patterns consistent with hypothesized mechanisms rather than as evidence for them. We retain the discussion of these patterns because they provide context for interpreting the trend results.

Streamflow regime as a driving mechanism for flux/Gini change: Lines 145-147 point out that the main changes to Qs documented here are in an area known to be undergoing streamflow change**. To what extent can the changes to Qs be explained purely by changes to streamflow?** I also wondered about this point when inspecting Ext.Dat.Fig. 4, which indicates that trends in mean flow and SSC are correlated. Because sediment transport is threshold-driven, I wondered whether changes in extreme streamflow percentiles might be directly responsible for the observed SSC/flux changes. I feel this physical driver is an important missing linkage in the attempts at attributing observed Qs trends to precipitation or land use, because streamflow elasticity to precipitation change may not be the same across landcover, basin size, etc, nor across the full flow duration curve (see e.g. Anderson et al., 2022; 2024; Hemshorn de Sanchez et al., 2025). Consequently, changes to R95p might not have the same influence on Qw (and therefore Qs) across time and space. It's evident from the explainability analysis that R95p is less influential than urbanisation, but what is not clear from this analysis is whether the weaker influence of R95p stems from (a) a weak influence of R95p on streamflow regimes, or (b) a weak influence of streamflow regime change on annual Qs. I feel the authors could get at this question by computing some simple metrics for extreme-streamflow periodicity. Plugging such metrics & their temporal trend into the explainability analysis and exploratory figures e.g. E.D.Fig 5, would help the authors to get at this missing piece in the cascade of physical drivers.

Answer: We are grateful for the suggestion of computing extreme metrics. We agree that changes to extreme streamflow percentiles act as the direct physical mechanism linking external stressors (climate and land use) to observed sediment flux changes.

While we considered including streamflow-based metrics in our explainability analysis, we ultimately excluded them to avoid issues of endogeneity and severe collinearity. Our hierarchical dynamic linear models (HDLMs) are specifically designed to attribute sediment trends to *ultimate* external boundary conditions (anthropogenic land-cover change and meteorological forcing). Because streamflow is an intermediate variable—directly driven by those same land-use and precipitation inputs—including it alongside them in the HDLM would obscure the attribution signal of the ultimate drivers.

However, we fully agree that this intermediate physical linkage is a crucial piece of the conceptual puzzle, particularly regarding how streamflow elasticity varies across land cover and the flow duration curve. To address this without altering the fundamental structure of the attribution model, we have added text to the Discussion that explicitly details this physical cascade. We now discuss how land use and R95p dictate

intermediate streamflow flashiness, which in turn drives the sediment threshold exceedance, directly incorporating the reviewer's suggested references (Anderson et al., 2022; 2024; Hemshorn de Sánchez et al., 2025).

On a related note, I wondered how well R95p would index hydro-climatic change for basins that are shifting from snowmelt to rainfall dominated hydrologic regimes – would shifting runoff timing/volumes be captured? This is yet one more reason why it may be helpful to explore shifts in streamflow regime/flashiness as a driver for the observed sediment flux changes.

The reviewer raises a valid point regarding R95p's limitations in capturing snow-to-rain transitions. Because our rigorously filtered dataset is predominantly located in the eastern and southern United States, this transition is less of a dominant driver here than it would be in western basins.

Attribution and space: Examining E.D.Fig. 5 made me wonder about the spatial pattern in variable importance. It would be instructive to see maps of e.g. dominant drivers or Shapley values, to interrogate whether there is any spatial pattern to the influence of urbanisation vs R95 etc.

Answer: We have made this and now present it on the next page. Apart from SSC, which is most clearly influenced by extreme precipitation trends, the other target variables are heterogeneously influence due likely to local controls rather than regional trends. E.g., a site could be urbanizing in a location where extreme precipitation is increasing, but the urbanization signal is strong – and vice versa. Because of the high degree of spatial heterogeneity and no clear pattern, this figure is included only here for the reviewer and not in the manuscript.

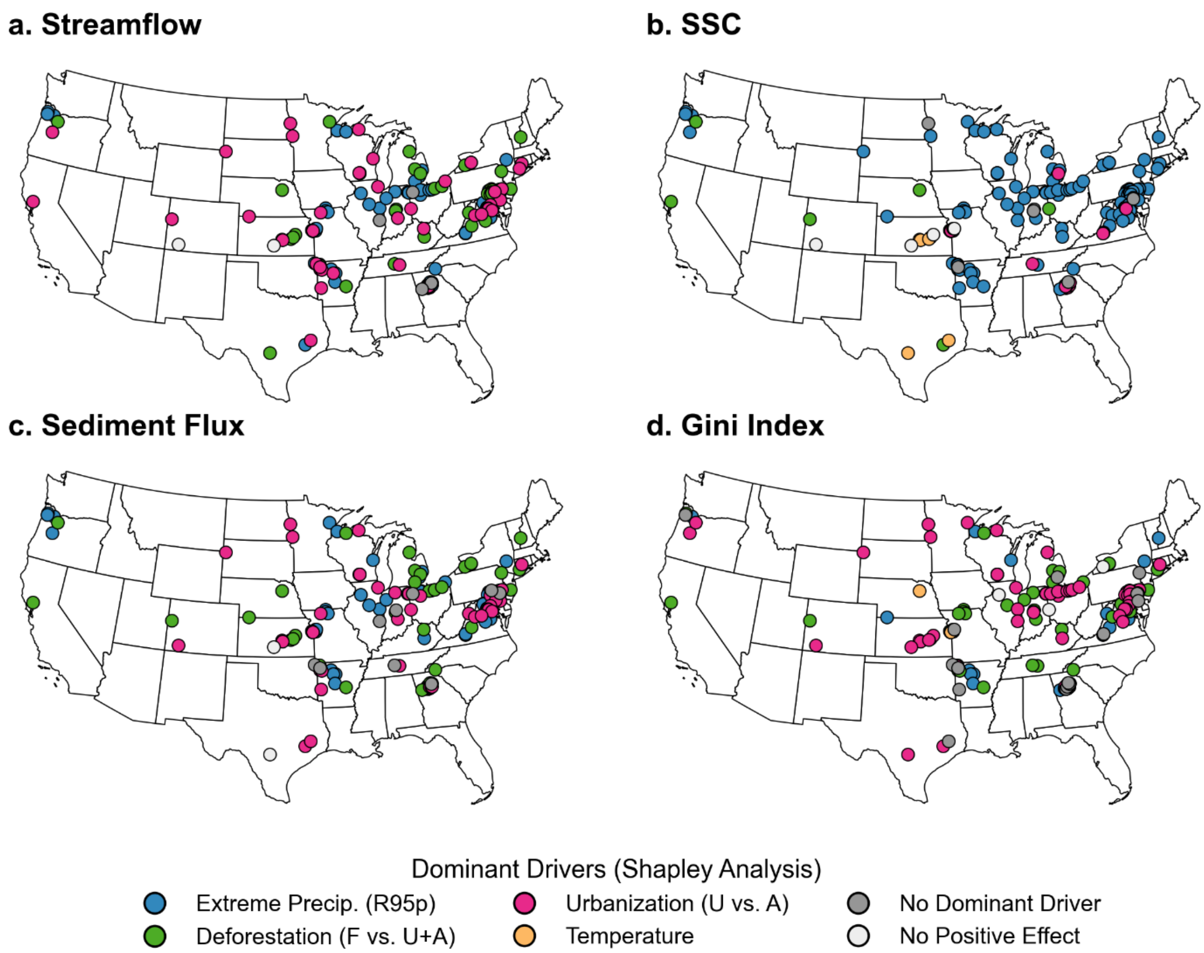


Figure. Dominant drivers identified via Shapley analysis for modeled increases across hydrological and sediment metrics. Dominant drivers are those that have at least a 50% share in the positive increase in a metric at a site. Per-site results are presented across the contiguous United States for (a) streamflow, (b) suspended sediment concentration (SSC), (c) sediment flux, and (d) the Gini index. Point colors indicate the dominant factor identified by the HDLM model contributing to the localized increase in each specific variable. Blue points denote extreme precipitation (R95p); green indicates relative deforestation (contrasting Forest versus Urban + Agricultural land cover); pink points show relative urbanization (contrasting Urban versus Agricultural land cover); and orange points indicate air temperature. Grey circles signify sites with increases where no single factor dominated (i.e., constituted more than 50% of the increase on its own), while open white circles denote sites where no drivers had a positive modeled effect. HDLM = hierarchical dynamic linear model.

On LSTMs/HDLMs: One of the reasons I signed this review is that I thought the authors should be aware of what I *don't* know. I have never used LSTMs or HDLMs so cannot provide any constructive comments for those aspects of the paper.

Answer: For a reviewer who is not familiar with these methods, you have made some very insightful observations and suggestions that have materially improved not just the article but also the implementation, evaluation, and interpretation of the models. Thank you!

Visualisation suggestions (mostly optional):

Figure 4: I suggest giving a hollow/white point for trends that are not statistically significant in a/c. This comment applies to other plots with trends, to focus the eye on statistically significant ones.

- **We keep the raw values here, as we color-code results by trend significance in the figure below.**

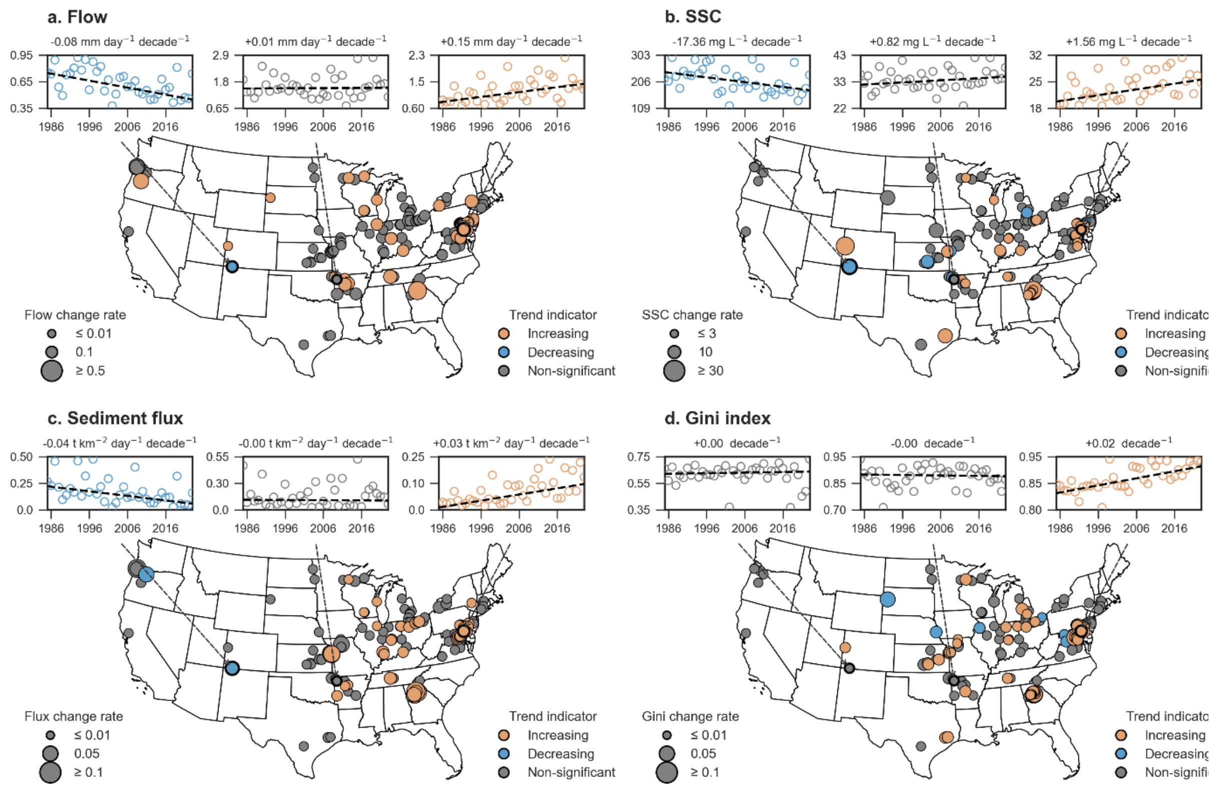


Figure. Long-term trends of hydrologic variables, sediment metrics, and Gini coefficient (1985–2023). Maps show the spatial distribution of increasing (orange), decreasing (blue), and non-significant (gray) trends for annual averages of key variables across CONUS. Panels include: (a) Flow (discharge), (b) SSC (suspended sediment concentration), (c) Sediment flux, and (d) Gini index. For each variable, trends from 1985 to 2023 are illustrated for three representative sites (USGS IDs: 09363500, 07048600, and 01585075). Theil–Sen slopes indicate decadal change rates, with color denoting trend direction.

Figure 5:

- Suggest annotating panel (c) with the line slope
- **We add the line slope in the caption.**
- Please add y-axis units for flux in (c)
- **We add the y-axis units in the caption as there is little space for lengthy flux units in the y-label.**

- Why choose 1989 and 2019 for (d)? I suggest using *either* the first + last years in (c), *or* the min + max flux in (c). I recommend changing the point colours for these years in (c) or annotating it, to highlight which years have their full time-series presented in (d).
- **These sites were chosen as the trend line passes approximately through them and they are representative of the general conditions during the beginning and end of the record.**

Figure 6 , and all other "attribution" figures

- Panel (a): it would be useful to show the value computed from the "reconstructed" Qs for years it is available, to enable comparison to LSTM/HDLM estimates.
- **In Fig. 6a, the open circles are the LSTM-reconstructed annual Gini values and the solid line is the HDLM smoothed prediction; the LSTM values *are* the reconstructed quantity for comparison, and they are already plotted.**
- It would be helpful to annotate panels (b) and (c) to show that they too pertain only to Big Haynes Creek, and not to the full dataset.
- **The titles are length as it stands. We make it clear in the figure caption that these panels refer only to Big Haynes Creek.**
- I can't see any difference between panel (b) for Figure 6 and Ext. Dat. Figures 6-8. Is this a mistake or is the signal truly identical across all variables?
- **This is not a mistake. Because we are demonstrating change at a single site (Big Haynes Creek), the predictors and their trajectories are the same. The same four predictors are used for the HDLM whether we are simulating Q, SSC, Qs, or Gini.**

Ext.Dat.Figure 1:

- I recommend showing additional versions of a-b for two more sites, taken from further down the CDF in panel (d). I recommend annotating (d) with points to indicate where on the CDF these timeseries + regressions have been selected from for (a)/(b). Please add *n* for (c). I note in the text the cutoff R2 for excluding sites was 0.4. I suggest annotating this threshold on (d) and also in the colour ramp for (c) – e.g. make points with R2 < 0.4 hollow.
- **We appreciate the suggestions for this figure. We declined to add additional time-series and regression panels (analogous to a–b) for sites further down the CDF in order to keep this figure focused on the high-quality reference example; the remaining sites can be inspected via the publicly archived data. The location of the time-series and regression is noted in the figure caption. We also now mention the R2 = 0.40 threshold explicitly in the figure caption.**

E.D. Figure 2: Remember to include units for the legends.

- **The units are shown in the y-label for the boxplots. There is insufficient space to put them with the colorbar for the map. Nonetheless, the two use the same units.**

E.D. Figure 3: I love the point size for trend magnitude approach here. Why not do so in Figure 5?

- **We wanted to keep the main-text figure relatively simple. The aim of Figure 5 is simply to show the existence of trends. Users are referred to this figure for magnitudes.**

E.D. Figure 4:

- Please include units for all map legends
- **Answered earlier.**
- Again, I suggest de-colouring points for trends that are statistically in-significant
- **Answered earlier.**
- Caption/L547: suggest specifying this is the trend in annual mean daily flow (yes?) & SSC
- **Units are shown in y-axis of boxplots.**

E.D. Fig. 6 and others: panels (d) and (f) have “flux” in the title but I think this needs updating? SM Figures S3-S4: In the caption, remind the reader which model these KGE/R2 metrics are for.

- **Good catch, we have fixed this.**

Line comments / semantics:

Title: “more *suspended* sediment” might be more accurate. In a similar vein, what about “Eastern US rivers…”. There are no rivers from south/central America included, and nor can one say “North American rivers” as Canada and Mexico are excluded. Moreover, the “more sediment in less time” holds mostly for eastern sites. I suggest specifying region, transport mode, and catchment scale.

- **We have added ‘suspended’ to the title for accuracy. We have also revised the title to acknowledge the partial decoupling of magnitude and timing trends (see also our response to Major Comment 3). The Discussion has been revised to flag that conclusions are most directly applicable to sensor-monitored, predominantly eastern systems undergoing human impact.**

20: Is high SSC really a hazard? The discussion for this point might shift towards related hazards for which high SSC “events” are a bellwether.

- **The reviewer is correct that SSC itself is not the hazard — rather, high-SSC episodes are bellwethers for downstream hazards including channel scour, water-treatment intake clogging, fish-gill damage, and reservoir sedimentation. We have revised the abstract framing accordingly to refer to ‘geomorphic and infrastructure hazards’ rather than implying SSC itself is hazardous.**

43: I think it’s a stretch to call these data continental-scale. Given that western US sites are largely excluded due to the rigorous filtering, and that no Canadian/Mexican/Alaskan sites are included, I suggest reframing “continental-scale” to something that better fits the dataset.

- **We have replaced ‘continental-scale’ with ‘large-scale’ or ‘conterminous U.S. scale’ throughout the manuscript to reflect the actual geographic extent of our network.**

48: “Interpretable” machine learning: in the abstract the authors use “explainability analysis” – I suggest being careful with consistent terminology around the explainable AI aspects of the paper.

- **Done.**

70-71: “spatial cross-validation”: in the relevant methods section, it might be useful to show how basins were selected for spatial cross-validation: e.g. a stratified sample across lithologies, land covers, basin slopes, climates? Completely random? This was more clearly explained in the code documentation but a little more detail in the paper would be helpful.

- **We note that selection was random.**

95: “Climate sets the baseline…”. A useful reference might be Chapman & Finnegan (2024).

- **We include this reference now.**

105-106: This sentence might be strengthened with graphical/quantitative evidence, e.g. perhaps the top 10 highest modelled-mean-flux sites are all in temperate, urban basins? This information may be readable from Fig 3 but I found the point sizes/colours hard to parse when printed.

- **This sentence has been removed to streamline the section.**

184: “Nearly all the increase” – I think readers may need a little help here interpreting what a Shapley value of 0.141 means. If I remember correctly, Shapley values are expressed in the units of the target variable. It would be good to indicate this on the y-axis for Fig 6c, even though Gini is dimensionless. In addition, perhaps the authors could clarify e.g. with a comment that, of the ~0.3 increase in Gini index at this site, 0.141 can be attributed to urbanisation. Some quantitative context like this will aid the reader with interpreting Fig 6 & Shapley values.

- **We now clarify this.**

202: This caption was the first time it became apparent that HDLMs were used. I recommend the authors mention them earlier, given that LSTMs are described in some detail in the main paper.

- **We now introduce HDLMs earlier.**

341: The comment on daily time resolution made me wonder whether the authors tested higher frequencies? In NZ where I work, streamflow can be very flashy so daily means underestimate peak flow of some events. A quick comment on how the authors decided daily frequency was sufficient would be useful.

- **Due to the resolution of the meteorological drivers (mostly at the daily timescale) available at the time the study was conducted, we were unable to refine analyses at sub-daily scale. However, this is something we are investigating for the future.**

362: “correlate strongly” – SSC and turbidity may correlate strongly in the reference cited here, but 50% of sites have an R2 < 0.8 in this dataset, which some may not call “strong” – particularly since the regression outputs are then used to train the LSTM. As part of the sensitivity analysis recommended in the major comments, I suggest tests with the threshold R2 cutoff in L 370. I’m curious to see how only using e.g. higher-quality synthetic records influences the results.

- **We thank the reviewer for raising this point, which gives us the opportunity to clarify our terminology and the trade-offs inherent in our dataset filtering.**
- **Regarding our characterization of the correlation as "strong," we base this on established literature standards for high-frequency sediment proxy sensing. For instance, Skarbøvik et al. (2023) classify an R2 > 0.60 as indicative of a relationship between SSC and turbidity in river systems. Given that our dataset achieves a median R2 of 0.79, we believe "strongly" remains an appropriate characterization for this specific context. We have added a sentence to the Methods section citing this literature standard to justify our terminology to the reader.**
- **Regarding the suggestion to test higher R2 cutoffs: while a formal re-run of the entire LSTM and HDLM pipeline with stricter thresholds falls outside the scope of this revision, we address the reviewer's underlying concern regarding regression-quality dependence through our earlier model residual uncertainty propagation. As noted in our response to Reviewer #1 (Comment 4), we have integrated a new moving block bootstrap analysis into the study. Because this procedure resamples the final model residuals, it inherently captures and propagates the accumulated variance from the entire workflow, including the initial noise and structural uncertainty from the proxy regressions.**

363 "Used" – how? Include the equation and assumptions for computing flux from concentration.

- **We add a note here of how the reconstructed SSC was turned into flux estimates.**

391-392: Here I was confused about why the LSTM was needed, given that raw streamflow was collated from NWIS and that flux/SSC were estimated from linear regressions. A clearer description of the LSTM *hindcasting* here would foreground the rationale for its use.

- **The LSTM is required to gap fill when neither SSC, turbidity, or flow data are available, which constitutes a considerable part of the hindcast and training periods.**

412-415: Please add the cross-validation results to the SM. 420 "The sediment model" – flux or concentration?

- **The cross-validation results appear in the Supplemental Material.**

422 "Strong performance": throughout the paper, please back-up statements such as this one, on model skill or fit, with quantitative evidence in text.

- **We relate our performance to that of other models in the literature and use Pandit et al., (2025) as a reference.**

- **Pandit, A. *et al.* Establishing performance criteria for evaluating watershed-scale sediment and nutrient models at fine temporal scales. *Water Research* 274, 123156 (2025).**

Also 422: “Comparable to existing models”. Is this a good enough benchmark in order to use the LSTM outputs for trend detection and analysis? As I mentioned earlier, I’d be keen to see KGE (for LSTM against direct observations) for *only* the sites with a significant trend in e.g. flux or Gini, to see whether these sites have a higher/lower KGE than the non-significant-trend sites.

- **We addressed this directly in our response to the previous comment: median KGE/NSE for sites with significant flux trends barely differs compared to sites without significant trends. The model performs comparably or better at trending sites, lending confidence that the trend-prevalence statistics are not artifacts of variable model skill**

494: For reproducibility, a table of the 175 gauge IDs is needed in the supplement. 503: For reproducibility, please publish the code online.

We have added a metadata Excel file to the HydroShare link that now contains our study results: https://hydroshare.org/resource/bba202f5eda04ae0abfab695e8c4750e/. Upon acceptance of this manuscript, this link will be turned into a permanent DOI.

References:

Anderson, B.J., Slater, L.J., Dadson, S.J., Blum, A.G. and Prosdocimi, I., 2022. Statistical attribution of the influence of urban and tree cover change on streamflow: A comparison of large sample statistical approaches. Water Resources Research, 58(5), p.e2021WR030742. https://doi.org/10.1029/2021WR030742

Anderson, B.J., Brunner, M.I., Slater, L.J. and Dadson, S.J., 2024. Elasticity curves describe streamflow sensitivity to precipitation across the entire flow distribution. Hydrology and Earth System Sciences, 28(7), pp.1567-1583. https://doi.org/10.5194/hess-28-1567-2024

Chapman, W. A. L., & Finnegan, N. J. (2024). The signature of climate in fluvial suspended sediment records. Journal of Geophysical Research: Earth Surface, 129, e2023JF007429. https://doi.org/10.1029/2023JF007429

Greenberg, E. and Ganti, V., 2024. The pace of global river meandering influenced by fluvial sediment supply. Earth and Planetary Science Letters, c34, p.118674. https://doi.org/10.1016/j.epsl.2024.118674

Hemshorn de Sánchez, A.L., Berghuijs, W.R., Van Loon, A.F., Hendriks, D. and van der Velde, Y., 2025. Sensitivities of mean and extreme streamflow to climate variability across Europe. EGUsphere, 2025, pp.1-26. https://doi.org/10.5194/egusphere-2025-5139

Leenman, A., Greenberg, E., Moulds, S., Wortmann, M., Slater, L. and Ganti, V., 2025. Accelerated river mobility linked to water discharge variability. Geophysical Research Letters, 52(2), p.e2024GL112899. https://doi.org/10.1029/2024GL112899

Authors' Response to Review Comments:

In this document, the authors list the major comments made during the review of the revised manuscript and, underneath each comment, **our responses are written in bold font**. References to in-text citations used in this response-to-reviewer file can be found at the end of this document.

Associate Editor Comments:

Dear Dr Husic,

Your manuscript titled "U.S. rivers are transporting more suspended sediment, often in less time" has now been seen again by our 2 reviewers. Based on their comments, we are happy to invite a final revision of your work. You will find the reviewers' reports below.

We thank the editor for facilitating our manuscript submission and inviting a final revision. We have addressed the reviewers' comments and completed.

Reviewer #1 Comments:

General comments:

The authors have done a commendable job addressing the structural limitations of data-driven modeling by introducing robust uncertainty propagation (the moving block bootstrap) and evaluating performance specifically during hydrologic extremes. The demonstration that sediment export timing (burstiness) is partially decoupled from transport magnitude is a critical contribution. From an earth-system and coastal-delivery perspective, however, the manuscript can be further elevated. High-resolution daily inputs from headwater catchments serve as the upstream boundary conditions for large-scale coastal margins and shelf sea bio-optical processes. To maximize the manuscript's impact for a broad Communications Earth & Environment readership, the framing should lean less into local watershed dynamics and more into the integrated physical-biological land-to-ocean continuum.

We thank the reviewer for the constructive framing and for recognizing the uncertainty and extreme-event analyses added in the previous round.

We have considered the recommendation to reorient the manuscript toward the integrated physical-biological land-to-ocean continuum, and we respectfully retain the current watershed-scale framing. The primary reason is that the study basins are by construction small-to-mid-sized, unregulated, and upstream of the reservoirs and estuaries that mediate delivery to the coastal ocean. Sediment leaving these lower-order watersheds is intercepted, lagged, and reworked by downstream storage before reaching any marine endpoint, so a direct claim that coastal zones receive these pulses would require the downstream routing and mass-balance analysis that is not a component of this work. Thus, foregrounding a land-to-ocean narrative would sit in tension with those representativeness limits. We would add that the Discussion already meets the journal's

breadth expectations by connecting our results to hydroclimatic intensification, compound extremes, non-stationarity, and infrastructure resilience.

Nonetheless, we recognize the reviewer's point that the present work has the potential to extend into a broader analysis that tracks sediment from headwaters to coastal shelves. Where the data allow, we have added two acknowledgments to the manuscript in response to Comments 1 and 2.

Specific comments:

1. While the analysis focuses on river basins primarily located in the eastern United States, these rivers empty into highly productive dynamic systems (such as the Western Atlantic shelf, the Gulf of Maine, and the Mid-Atlantic Bight). The discussion details the impact of burstiness on water treatment and localized infrastructure, but overlooks the downstream estuarine and marine endpoints. A 19-day network-wide median contraction in B90 means that coastal zones are receiving terrigenous material in massive, highly episodic pulses rather than steady seasonal inputs. In the Implications section, explicitly discuss how this temporal compression modulates coastal plume dynamics and physical-biological coupling. Extreme, short-duration sediment pulses alter light attenuation (photosynthetically active radiation) in coastal waters, directly influencing phytoplankton bloom timing, community structure, and localized marine ecosystem stress.

   We agree that the downstream endpoints matter and have added a sentence to the Implications. We have deliberately kept this to a boundary-condition statement rather than a claim about hypothesized plume dynamics or phytoplankton response, because our basins sit upstream of the impoundments and estuaries that would modulate any such signal and we have no coastal or bio-optical observations with which to specifically characterize light attenuation or ecosystem outcomes.

   The added sentence reads: "Consequently, in the downstream coastal systems that receive this sediment, the same shift toward pulsed delivery will impact sedimentation dynamics, nutrient delivery, and light regimes, potentially influencing algal bloom formation and management (Gillanders and Kingsford, 2002)."

   Reference: Gillanders, B., Kingsford, M. 2002. Impact of changes in flow of freshwater on estuarine and open coastal habitats and the associated organisms. In *Oceanography and Marine Biology*. pp. 233–309.

2. In Comment 11, you nicely connected your work to recent remote sensing studies (e.g., Lucchese et al., 2026). High-frequency in-situ turbidity sensing is a relatively recent development (primarily post-2007 in your study), requiring your LSTM to hindsand/extrapolate backwards into the 1985– 2006 era. Expand your discussion on transferability to note how historical satellite missions can act as a bridge. Long-term

satellite ocean color and surface reflectance archives (e.g., Landsat, early MODIS) provide multi-decadal, though intermittent, wide-swath proxies of suspended particulate matter (SPM) and turbidity. Emphasize that future frameworks could fuse in-situ sensor networks with satellite archives to constrain hindcast errors and extend this methodology to globally ungauged coastal rivers where in-situ data are absent.

We appreciate the suggestion and have extended the transferability discussion. However, we do note that the specific concern about constraining hindcast error is addressed within our study in two ways. First, a Mahalanobis-distance analysis showing that the pre-sensor inputs occupy the same multivariate feature space as the training period (Fig. S18). Second, a direct validation of hindcast flow, SSC, and flux against independent grab sample measurements not used in training (Fig. S19).

We do agree that long-term satellite archives and future products are a valuable route to extending this class of framework to ungauged systems, and we have added a clause to the transferability sentence in the Implications with a new reference (Feng et al., 2026).

The revised sentence reads: "Wherever high-frequency in-situ sensing or remote-sensing data exist as SSC proxies, the data-driven framework demonstrated here can be extended to diagnose shifting sediment regimes elsewhere. Long-term satellite archives can further constrain hindcast uncertainty and extend the approach to ungauged systems that lack in-situ records (Feng et al., 2026)."

> **Reference: Feng, D., Yang, X., Pavelsky, T.M., Allen, G.H., Bates, P., Gleason, C.J., Gardner, J., Lehner, B., Wang, J., Cooley, S. and Tarpanelli, A., 2026. Remote sensing and the new global river science. *Nature Water*, pp.1-20.**

3. The newly added analysis on Gini vs. B90 behavior using synthetic lognormal series (Figure R6) is exceptionally clear and illuminates the sensitivity of your metrics. You point out that urbanization increases Gini via a peak-magnitude amplification mechanism, whereas precipitation extremes (R95p) have a weaker effect on Gini because they act via a frequency-amplification mechanism.

Thanks for the comment regarding the clarity of the supplemental figure.

However, log-log regressions inherently struggle with sensor saturation during peak flood events. Add a brief caveat clarifying whether the "frequency-amplification" signature of precipitation could partially be a byproduct of sensor clipping. If the turbidity sensors flatten the true peak values of extreme rain events, the calculated Gini would artificially decrease for those events. Explicitly state if your attribution framework accounts for this potential data artifact or if it reinforces your conclusion that your reported trends represent a highly conservative lower bound.

We have added the requested caveat and agree that it reinforces our conservative framing. Our attribution framework does not separately correct for the clipping that you mention. Consequently, the analysis may understate influence of precipitation extremes.

We have dedicated the following paragraph to the implications section: “**The intensification of sediment burstiness reported here likely represents a conservative lower bound because of three structural realities of current data. First, turbidity sensors saturate at extreme concentrations, which tends to artificially flatten peak measurements [46]. Second, purely data-driven models tend to underpredict the magnitudes of the rarest events [52]. Third, aggregating sub-hourly streamflow and turbidity data into daily averages, in order to match the temporal resolution of the meteorological forcings, mathematically smooths temporal inequality [12]. Because these factors work in tandem to lower the magnitude of storm-driven peaks, they not only reduce the computed burstiness but also likely cause the model to under-attribute the specific role of precipitation extremes. Future advancements in sensor measurement capabilities, model architectures, and high-resolution meteorological data will enable us to capture the unattenuated scale of sediment transport.”**

4. The language adjustments listed in response to Comment 8 (moving from "drives" to "attributed predictors") are a massive step forward for scientific precision. On page 10, lines 248–250 of the merged manuscript, the text still states: "urbanization... acts as the primary driver for 43% of sites. Averaged across these rivers, urbanization accounts for 44% of the positive driver-attributed change...". The use of "primary driver" alongside "driver-attributed change" in back-to-back sentences dilutes your goal of using purely associative/attributed language. Ensure strict consistency by replacing "primary driver" with "dominant predictor" or "primary attributed factor" in line 242 to match the cautious tone established in your responses and in the abstract.

   This has been corrected.

5. Table R2 (Model Hyperparameters) is a great addition for reproducibility. Please ensure that the sequence length parameter unit (days) is explicitly labeled in the table header.

   This has been corrected.

6. In Figure 5, ensure the axis label formatting for the Flux Trend matches standard international oceanographic/hydrologic notation consistently—ideally explicitly utilizing negative exponents rather than mixed formatting inline.

   The x-axis label for the flux trend uses negative exponents.

Reviewer #2 Comments:

General comments:

I congratulate the authors on a thorough revision; I particularly appreciated their addition of a more rigorous uncertainty analysis. I have read through the rebuttal letter and associated changes and am mostly happy with the changes made. There remain however a few places where the authors have only responded to me, personally, without making changes to the paper. My comments are not just a result of my curiosity; instead, they were designed to help the authors make the work more robust and/or accessible to a general audience. I suggest the authors re-read the rebuttal and, for comments where they have not made changes to the paper, explicitly justify why they thought a clarification to the paper was not necessary. If I have asked a question or been confused by some aspect of the manuscript, it is likely other readers will have similar questions or confusions, and I hope the authors can use my comments to identify and ultimately avoid such issues for future readers.

We apologize for any oversights in ensuring that our responses to the reviewer were reflected with direct changes to the manuscript. We have gone through the previous round of revisions and identified six locations where intended changes were not made to the manuscript.

- **Major Comment 1 asked us to note that Daymet and the HydroATLAS attributes are themselves spatial-model outputs rather than direct observations. We agreed, but the note did not appear in the revised text. We have added a sentence to the Methods ("Model inputs") stating that these gridded products are spatially modeled and are treated as modeled inputs throughout, complementing the measured / turbidity-derived / reconstructed terminology added elsewhere in the Methods.**
    - **“We note that the Daymet meteorological forcings and the catchment descriptors are gridded products derived from spatial interpolation and modeling rather than direct in-situ measurements, and we treat them as modeled inputs throughout.”**

- **Major Comment 3 asked us either to match the example basins on mean flux and year or to state that the comparison is conceptual. We responded that Fig 1 shows illustrative end-members and that the years represent each site's characteristic regime, and said we would clarify this in the caption. That clarification was not carried into the revised captions. We have added it to the Fig 1 caption. The demonstration that similar-mean-flux systems can differ in Gini and B90 is provided in Fig. S16. Below is the Fig 1 addition:**
    - **"These two basins are shown as illustrative end-members of the observed burstiness spectrum. The year displayed for each basin was selected to represent that site's characteristic regime, so the two panels correspond to different years."**

- **Major Comment 3 noted the visual similarity between the spatial patterns of mean flux and Gini and asked whether it is physical or an artifact of how Gini is computed, adding that a real signal would be worth discussing. We confirmed to the reviewer**

that Gini is scale-independent, so the co-variation is physical, but did not add this to the text.

- **We have now added a short passage to the "Climatic and land use drivers" section stating that mean flux and burstiness co-vary in space, that this is not an artifact of magnitude because Gini is invariant to rescaling (Fig. S16), and that it reflects a physical co-occurrence in which the highest-yielding basins also concentrate transport most in time.**
  - **"Mean sediment flux and burstiness tend to co-vary in space such that the basins that move the greatest amount of sediment tend to do it in the most concentrated bursts (Fig. 3). This is most likely a physical co-occurrence rather than a mathematical artefact as the Gini coefficient is invariant to multiplicative rescaling of flux. Thus, basins subject to intense rainfall regimes and land disturbance experience both higher yields and greater concentration of transport in time."**

- **The reviewer also raised the streamflow linkage in several places: the major comment on comparison among fluxes, the minor comment on R95p's influence on Qw, and the line comment at L145-147. In our response we agreed that changes in extreme streamflow are the direct physical mechanism linking climate and land use to sediment, explained that we excluded streamflow from the HDLM to avoid endogeneity, and stated that we had added Discussion text detailing this cascade and incorporating two of the suggested references. On review, neither the text nor the references were present.**
  - **We have now added a paragraph at the end of the attribution section describing the cascade from ultimate drivers, through streamflow response, to sediment threshold exceedance, and noting that streamflow elasticity varies with land cover, basin size, and position on the flow-duration curve.**
    - **"The attribution above identifies the external drivers, but their influence on sediment is transmitted through an intermediate hydrologic response. Land use and extreme precipitation reshape the streamflow regime, and because sediment transport is threshold-driven, it is the resulting changes in high-flow behavior that most directly mobilize and concentrate sediment. The sensitivity of streamflow to precipitation varies with land cover, basin size, and position along the flow-duration curve [Anderson 2024; Hemshorn de Sánchez 2025]. Thus, a given change in extreme precipitation (R95p) need not translate into the same change in flow, and therefore in flux, across sites and time. This intermediate step helps explain why urbanization, which increases the flashiness of the runoff response, emerges as the dominant attributed predictor of temporal compression, whereas intensifying precipitation acts more diffusely on total magnitude."**
  - **We've also added a sentence in the HDLM methods stating why streamflow is excluded as an intermediate variable.**
    - **"We deliberately restricted the covariates to external boundary conditions (climate and land cover) and did not include streamflow, because streamflow is an intermediate variable driven by those same inputs."**

- **The two references:**
    - **Anderson, B. J., Brunner, M. I., Slater, L. J. & Dadson, S. J. 2024. Elasticity curves describe streamflow sensitivity to precipitation across the entire flow distribution. *Hydrology and Earth System Sciences, 28*, 1567–1583.**
    - **Hemshorn de Sánchez, A.L., Berghuijs, W.R., Van Loon, A.F., Hendriks, D. and van der Velde, Y., 2026. Sensitivities of mean and extreme streamflow to climate variability across Europe. *Hydrology and Earth System Sciences*, *30*, 2667-2683.**

- **The reviewer asked for a brief comment on why a daily timestep was sufficient (L341), noting that daily means can underestimate peak flow in flashy systems. We explained this only in the response letter.**
    - **We have added it to the Methods ("Model inputs"), stating that the daily analysis timestep is set by the resolution of the meteorological forcings.**
        - **"All reconstruction and trend analysis are performed at a daily timestep, dictated by the daily resolution of the Daymet meteorological forcings used to drive the models."**
    - **We also expand the Results and Discussion ("Implications") to discuss how daily aggregation may under-resolve peaks in a direction that reinforces our conservative interpretation.**
        - **"The intensification of sediment burstiness reported here likely represents a conservative lower bound because of three structural realities of current data. First, turbidity sensors saturate at extreme concentrations, which tends to artificially flatten peak measurements 53. Second, purely data-driven models tend to underpredict the magnitudes of the rarest events 54. Third, aggregating sub-hourly streamflow and turbidity data into daily averages, in order to match the temporal resolution of the meteorological forcings, mathematically smooths temporal inequality 12. Because these factors work in tandem to lower the magnitude of storm-driven peaks, they not only reduce the computed burstiness but also likely cause the model to under-attribute the role of precipitation extremes. Future advancements in sensor measurement capabilities, model architectures, and high-resolution meteorological data will enable us to capture the unattenuated scale of sediment transport."**

American rivers are transporting more sediment in less time by Sigdel & Husic in *Communications Earth & Environment*: review by Anya Leenman, Feb 2026

The authors use a chain of models to estimate and attribute trends in suspended sediment flux across ~40 years in the eastern US. The work is timely and exciting, exploring the relative influence of landcover versus hydroclimate change on suspended sediment dynamics in a period when both these drivers have shifted. While the methods are suitable and presentation is excellent, I feel there is work needed on quantifying uncertainty and its propagation across the model chain, on exploring the roles of basin size and streamflow extremes, and on extending the discussion.

Major comments:

Clarifying terminology around models and observations

The authors use observations to fit linear-regression models, whose outputs become the target for a LSTM model and an HDLM model that use additional spatial-modelling outputs from other papers as input "data". The words "simulated" and "modelled" are used throughout, but it was often unclear to me which of these models was referred to. I suggest specifying carefully throughout – I know this is frustrating but it will make things much easier for the reader to follow.

Similarly, the time-series outputs of the linear-regression models are often referred to with words such as "records" or "reconstructed" or "proxy"; even "observed" is sometimes used. Together, this language implies these model-outputs are observations instead. I urge the authors to carefully review the language used to describe the linear-regression outputs, ensuring that readers who skip the methods don't get the false impression that the "reconstructed" SSC or flux time-series used to train the LSTM are in fact in-situ observations.

Finally, it's worth noting that Daymet and other LSTM inputs (e.g. some from HydroAtlas) are outputs of spatial models; a quick note on this point will help readers to separate these model outputs from other data sources (e.g. SSC grab samples) that are direct observations.

Deeper discussion of implications

I feel the discussion could go deeper. I have a few suggestions in the paragraphs below, but I also wondered: what are the implications for sediment mass balance? Should we worry about sediment exhaustion or over-accumulation in sinks, in places with marked trends? What about implications for river lateral mobility? Obviously as a reviewer one immediately thinks about implications for one's own work. Evan Greenberg (e.g. Greenberg & Ganti, 2024) showed that river mobility increases with sediment flux, while analysis we did together (Leenman et al., 2025) shows that flow variability is another driver. If both sediment flux (Q_s) and R95p are increasing (e.g. Fig 4a and Ext.Dat.Fig.5), could there be implications for river lateral mobility?

Comparison among different fluxes and basin sizes

In Figure 1, I feel the sites selected should have approximately the same mean flux, to allow more meaningful comparison of high vs low B_{90}. For instance, if one site has 10x fewer transport days but also has a 10x smaller annual flux than the other, we can hardly say it is a more hazardous site nor a more efficient sediment transporter. It's possible the authors have already selected sites with a similar Q_s, in which case please annotate the figure to say so.

In Figure 3, I was struck by the similarity in spatial patterns of Gini index and Q_s. Are they correlated? If yes, is this (a) a real physical phenomenon vs (b) a result of how the Gini index is computed? If (a), this is exciting, and **worth further exploration in the discussion** – first, it's interesting to understand how Q_s varies across hydroclimates (indexed by the Gini index, to some extent), and second, it implies that, even in the absence of any change to flux burstiness, the sites that move the most sediment do it in the shortest time, which has implications for managing sediment. If (b), it is possible that the apparent trends in Gini across ~1/3 of the sites arise simply from the observed trends in flux across ~1/3 of the sites. The authors could check for (b) by artificially inflating values in the Q_s time series and exploring whether there is any impact on Gini.

On a related note, in L137-138 I wondered if the ~1/3 of sites with increasing Gini are the same ~1/3 of sites with increasing Q_s? If (a) above holds true, a scatter plot of ΔQ_s against ΔGini would address this question. I note that E.D. Fig 3 goes some way to answering this question with "no", which makes the title of the paper rather an oversimplification – it is *not* the same rivers transporting more sediment and in less time. A scatter of ΔQ_s against ΔGini, coupled with an earlier and more explicit discussion of this mismatch demonstrated in E.D.Figure 3, would allow the authors to get into these nuances, and would **make for some fruitful additional discussion**.

L142: "68% urban cover" – was this a small basin? It would be useful to know how ΔQ_s and ΔGini vary with basin size, as any scale-dependence would allow the reader to know how generalizable the results are. Moreover, a scale-dependence could be useful for management: e.g., if the ~1/3 of basins that gained in burstiness or Q_s were also the smallest basins, this would highlight which catchments are most sensitive to changing land use or climate.

Quantifying propagating uncertainty across the model chain

As I understand, a linear-regression model (model 1) generates synthetic SSC time-series from turbidity time-series and SSC grab samples. The synthetic SSC is combined with observed(?) streamflow to model daily suspended sediment flux (model 2). Outputs from the Daymet spatial modelling effort (model 3) are combined with other large-sample estimates and input to the LSTM (model 4), whose target variables are the outputs of models 1-2. Finally, the longer term synthetic records generated by model 4 are used to fit model 5, the HDLM, which is used for attribution. I hope I have correctly understood the model chain based on its description in the methods & code; if not, please consider revising the method description. My main point here is that we need to see **how uncertainty propagates** across this model chain. Moreover, it is important to know how **sensitive** the LSTM and HDLM outputs are to uncertainty, and to decisions made in earlier models. For instance, how do the fits in E.D.Fig. 9 change if the R^2 threshold of 0.4 for the SSC linear regressions is raised? Moreover, if the median site-level R^2 for the HDLM's Gini index compared to model-2's synthetic record is only 0.41, how much confidence can we have in the HDLM's attribution results for the Gini index? These two examples highlight my more general opinion that questions of uncertainty and sensitivity need more explicit consideration.

Minor comments:

Fig 1 data selection: In Figure 1 and 2c-d, two contrasting degrees of Q_s burstiness are compared, but I notice the left and right panels use data from 2020 and 2012 respectively. If the goal is to compare these two basins, I think we need to see them in the same year.

Assessing model performance across trends: the LSTM KGE and NSE for daily SSC and Q_s are encouraging. Could the authors demonstrate (e.g. in the supplement) how KGE and NSE vary across different Q_s *trends* ? Since the paper is about attributing trends in Q_s, we need to be certain the LSTM performs equally well in sites with statistically significant Q_s trends.

On a similar vein, it worries me slightly that linear regressions were applied to model SSC back to 2007, with the LSTM then hindcasting further back to 1985. How confident can we be in LSTM estimates for the 1980s? Is LSTM uncertainty greater for years before 2007? The authors could colour points in E.D.Fig.9 by the "baseline" and "recent" time periods employed for attribution. It would also be useful to see versions EDFig9 made with just data from the baseline and recent periods, respectively, so we get an idea of how LSTM skill varies across the record from which trends are then measured.

Assessing trends or spatial patterns in extreme, not means: Figure 1 implies transport is event-driven. Why, then, look at mean annual Q_s and water discharge (Q_w) when looking for trends or spatial patterns? If urbanisation is altering runoff partitioning, we might expect the annual *mean* and annual *extremes* (e.g. Q_{99}) of both Q_s and Q_w to exhibit differing behaviour – e.g. the mean may not change but extremes may become more "flashy" as runoff processes change. It may be useful to explore trends in extremes, not just in annual means, as (a) they tell us more about hazards, and (b) they may have a stronger signal.

In Ext.Dat. Fig 6, I was struck by the weak influence of R95p on Q_w. It is possible that the 95th percentile of precipitation has a stronger influence on extreme, rather than mean, streamflow. Considering extreme streamflow percentiles instead of the mean may be informative.

On a similar note, I noticed in Ext.Dat.Fig 9 that the "model" (LSTM?) underpredicts extreme SSC and flux. This underprediction could result in underestimation of total flux some years, and could affect trend detection. I encourage the authors to (a) think about strategies to increase model skill for extreme values, (b) quantify for the reader, explicitly and in the main paper, how this underestimation of extremes affects the uncertainty of their final results, particularly trend detection, and (c) put the true R^2 (currently in the caption) on panels a-b and e-f in this figure.

Overinterpretation of physical process from non-significant differences: In L96-98 and subsequent paragraphs (e.g. L 113-114 also stood out), the authors attempt to explain the physical processes responsible for differences between climates & land covers. However, most of these groups are not statistically significantly different in Fig. 3 b/d. I suggest the authors do not worry about interpreting differences that are not supported by statistical evidence.

Streamflow regime as a driving mechanism for flux/Gini change: Lines 145-147 point out that the main changes to Q_s documented here are in an area known to be undergoing streamflow change. **To what extent can the changes to Q_s be explained purely by changes to streamflow?** I also wondered about this point when inspecting Ext.Dat.Fig. 4, which indicates that trends in mean flow and SSC are correlated. Because sediment transport is threshold-driven, I wondered whether changes in extreme streamflow percentiles might be directly responsible for the observed SSC/flux changes. I feel this physical driver is an important missing linkage in the attempts at attributing observed Q_s trends to precipitation or land use, because streamflow elasticity to precipitation change may not be the same across landcover, basin size, etc, nor across the full flow duration curve (see e.g. Anderson et al., 2022; 2024; Hemshorn de Sanchez et al., 2025). Consequently, changes to R95p might not have the same influence on Q_w (and therefore Q_s) across time and space. It's evident from the explainability analysis that R95p is less influential than urbanisation, but what is not clear from this analysis is whether the weaker influence of R95p stems from (a) a weak influence of R95p on streamflow regimes, or (b) a weak influence of streamflow regime change on annual Q_s. I feel the authors could get at this question by computing some simple metrics for extreme-streamflow periodicity. Plugging such metrics &

their temporal trend into the explainability analysis and exploratory figures e.g. E.D.Fig 5, would help the authors to get at this missing piece in the cascade of physical drivers.

On a related note, I wondered how well R95p would index hydro-climatic change for basins that are shifting from snowmelt to rainfall dominated hydrologic regimes – would shifting runoff timing/volumes be captured? This is yet one more reason why it may be helpful to explore shifts in streamflow regime/flashiness as a driver for the observed sediment flux changes.

Attribution and space: Examining E.D.Fig. 5 made me wonder about the spatial pattern in variable importance. It would be instructive to see maps of e.g. dominant drivers or Shapley values, to interrogate whether there is any spatial pattern to the influence of urbanisation vs R95 etc.

On LSTMs/HDLMs: One of the reasons I signed this review is that I thought the authors should be aware of what I *don't* know. I have never used LSTMs or HDLMs so cannot provide any constructive comments for those aspects of the paper.

Visualisation suggestions (mostly optional):

Figure 4: I suggest giving a hollow/white point for trends that are not statistically significant in a/c. This comment applies to other plots with trends, to focus the eye on statistically significant ones.

Figure 5:

- Suggest annotating panel (c) with the line slope
- Please add y-axis units for flux in (c)
- Why choose 1989 and 2019 for (d)? I suggest using *either* the first + last years in (c), *or* the min + max flux in (c). I recommend changing the point colours for these years in (c) or annotating it, to highlight which years have their full time-series presented in (d).

Figure 6 , and all other "attribution" figures

- Panel (a): it would be useful to show the value computed from the "reconstructed" Q_s for years it is available, to enable comparison to LSTM/HDLM estimates.
- It would be helpful to annotate panels (b) and (c) to show that they too pertain only to Big Haynes Creek, and not to the full dataset.
- I can't see any difference between panel (b) for Figure 6 and Ext. Dat. Figures 6-8. Is this a mistake or is the signal truly identical across all variables?

Ext.Dat.Figure 1:

- I recommend showing additional versions of a-b for two more sites, taken from further down the CDF in panel (d).
- I recommend annotating (d) with points to indicate where on the CDF these timeseries + regressions have been selected from for (a)/(b)
- Please add *n* for (c)
- I note in the text the cutoff R^2 for excluding sites was 0.4. I suggest annotating this threshold on (d) and also in the colour ramp for (c) – e.g. make points with $R^2 < 0.4$ hollow.

E.D. Figure 2: Remember to include units for the legends.

E.D. Figure 3: I love the point size for trend magnitude approach here. Why not do so in Figure 5?

E.D. Figure 4:

- Please include units for all map legends
- Again, I suggest de-colouring points for trends that are statistically in-significant

- Caption/L547: suggest specifying this is the trend in annual mean daily flow (yes?) & SSC

E.D. Fig. 6 and others: panels (d) and (f) have “flux” in the title but I think this needs updating?

SM Figures S3-S4: In the caption, remind the reader which model these KGE/R^2 metrics are for.

Line comments / semantics:

Title: “more *suspended* sediment” might be more accurate. In a similar vein, what about “Eastern US rivers…”. There are no rivers from south/central America included, and nor can one say “North American rivers” as Canada and Mexico are excluded. Moreover, the “more sediment in less time” holds mostly for eastern sites. I suggest specifying region, transport mode, and catchment scale.

20: Is high SSC really a hazard? The discussion for this point might shift towards related hazards for which high SSC “events” are a bellwether.

43: I think it’s a stretch to call these data continental-scale. Given that western US sites are largely excluded due to the rigorous filtering, and that no Canadian/Mexican/Alaskan sites are included, I suggest reframing “continental-scale” to something that better fits the dataset.

48: “Interpretable” machine learning: in the abstract the authors use “explainability analysis” – I suggest being careful with consistent terminology around the explainable AI aspects of the paper.

70-71: “spatial cross-validation”: in the relevant methods section, it might be useful to show how basins were selected for spatial cross-validation: e.g. a stratified sample across lithologies, land covers, basin slopes, climates? Completely random? This was more clearly explained in the code documentation but a little more detail in the paper would be helpful.

95: “Climate sets the baseline…”. A useful reference might be Chapman & Finnegan (2024).

105-106: This sentence might be strengthened with graphical/quantitative evidence, e.g. perhaps the top 10 highest modelled-mean-flux sites are all in temperate, urban basins? This information may be readable from Fig 3 but I found the point sizes/colours hard to parse when printed.

184: “Nearly all the increase” – I think readers may need a little help here interpreting what a Shapley value of 0.141 means. If I remember correctly, Shapley values are expressed in the units of the target variable. It would be good to indicate this on the y-axis for Fig 6c, even though Gini is dimensionless. In addition, perhaps the authors could clarify e.g. with a comment that, of the ~0.3 increase in Gini index at this site, 0.141 can be attributed to urbanisation. Some quantitative context like this will aid the reader with interpreting Fig 6 & Shapley values.

202: This caption was the first time it became apparent that HDLMs were used. I recommend the authors mention them earlier, given that LSTMs are described in some detail in the main paper.

341: The comment on daily time resolution made me wonder whether the authors tested higher frequencies? In NZ where I work, streamflow can be very flashy so daily means underestimate peak flow of some events. A quick comment on how the authors decided daily frequency was sufficient would be useful.

362: “correlate strongly” – SSC and turbidity may correlate strongly in the reference cited here, but 50% of sites have an $R^2 < 0.8$ in this dataset, which some may not call “strong” – particularly since the regression outputs are then used to train the LSTM. As part of the sensitivity analysis recommended in the major comments, I suggest tests with the threshold R^2 cutoff in L 370. I’m curious to see how only using e.g. higher-quality synthetic records influences the results.

363 “Used” – how? Include the equation and assumptions for computing flux from concentration.

391-392: Here I was confused about why the LSTM was needed, given that raw streamflow was collated from NWIS and that flux/SSC were estimated from linear regressions. A clearer description of the LSTM *hindcasting* here would foreground the rationale for its use.

412-415: Please add the cross-validation results to the SM.

420 “The sediment model” – flux or concentration?

422 “Strong performance”: throughout the paper, please back-up statements such as this one, on model skill or fit, with quantitative evidence in text.

Also 422: “Comparable to existing models”. Is this a good enough benchmark in order to use the LSTM outputs for trend detection and analysis? As I mentioned earlier, I’d be keen to see KGE (for LSTM against direct observations) for *only* the sites with a significant trend in e.g. flux or Gini, to see whether these sites have a higher/lower KGE than the non-significant-trend sites.

494: For reproducibility, a table of the 175 gauge IDs is needed in the supplement.

503: For reproducibility, please publish the code online.

References:

Anderson, B.J., Slater, L.J., Dadson, S.J., Blum, A.G. and Prosdocimi, I., 2022. Statistical attribution of the influence of urban and tree cover change on streamflow: A comparison of large sample statistical approaches. *Water Resources Research*, *58*(5), p.e2021WR030742. https://doi.org/10.1029/2021WR030742

Anderson, B.J., Brunner, M.I., Slater, L.J. and Dadson, S.J., 2024. Elasticity curves describe streamflow sensitivity to precipitation across the entire flow distribution. *Hydrology and Earth System Sciences*, *28*(7), pp.1567-1583. https://doi.org/10.5194/hess-28-1567-2024

Chapman, W. A. L., & Finnegan, N. J. (2024). The signature of climate in fluvial suspended sediment records. *Journal of Geophysical Research: Earth Surface*, 129, e2023JF007429. https://doi.org/10.1029/2023JF007429

Greenberg, E. and Ganti, V., 2024. The pace of global river meandering influenced by fluvial sediment supply. *Earth and Planetary Science Letters*, *634*, p.118674. https://doi.org/10.1016/j.epsl.2024.118674

Hemshorn de Sánchez, A.L., Berghuijs, W.R., Van Loon, A.F., Hendriks, D. and van der Velde, Y., 2025. Sensitivities of mean and extreme streamflow to climate variability across Europe. *EGUsphere*, *2025*, pp.1-26. https://doi.org/10.5194/egusphere-2025-5139

Leenman, A., Greenberg, E., Moulds, S., Wortmann, M., Slater, L. and Ganti, V., 2025. Accelerated river mobility linked to water discharge variability. *Geophysical Research Letters*, *52*(2), p.e2024GL112899. https://doi.org/10.1029/2024GL112899

Comments for MS # COMMSENV-26-0263A

General comments:

The authors have done a commendable job addressing the structural limitations of data-driven modeling by introducing robust uncertainty propagation (the moving block bootstrap) and evaluating performance specifically during hydrologic extremes. The demonstration that sediment export timing (burstiness) is partially decoupled from transport magnitude is a critical contribution. From an earth-system and coastal-delivery perspective, however, the manuscript can be further elevated. High-resolution daily inputs from headwater catchments serve as the upstream boundary conditions for large-scale coastal margins and shelf sea bio-optical processes. To maximize the manuscript's impact for a broad Communications Earth & Environment readership, the framing should lean less into local watershed dynamics and more into the integrated physical-biological land-to-ocean continuum.

Specific comments:

1. While the analysis focuses on river basins primarily located in the eastern United States, these rivers empty into highly productive dynamic systems (such as the Western Atlantic shelf, the Gulf of Maine, and the Mid-Atlantic Bight). The discussion details the impact of burstiness on water treatment and localized infrastructure, but overlooks the downstream estuarine and marine endpoints. A 19-day network-wide median contraction in B_{90} means that coastal zones are receiving terrigenous material in massive, highly episodic pulses rather than steady seasonal inputs. In the Implications section, explicitly discuss how this temporal compression modulates coastal plume dynamics and physical-biological coupling. Extreme, short-duration sediment pulses alter light attenuation (photosynthetically active radiation) in coastal waters, directly influencing phytoplankton bloom timing, community structure, and localized marine ecosystem stress.

2. In Comment 11, you nicely connected your work to recent remote sensing studies (e.g., Lucchese et al., 2026). High-frequency in-situ turbidity sensing is a relatively recent development (primarily post-2007 in your study), requiring your LSTM to hindsand/extrapolate backwards into the 1985–2006 era. Expand your discussion on transferability to note how historical satellite missions can act as a bridge. Long-term satellite ocean color and surface reflectance archives (e.g., Landsat, early MODIS) provide multi-decadal, though intermittent, wide-swath proxies of suspended particulate matter (SPM) and turbidity. Emphasize that future frameworks could fuse in-situ sensor networks with satellite archives to constrain hindcast errors and extend this methodology to globally ungauged coastal rivers where in-situ data are absent.

3. The newly added analysis on Gini vs. B_{90} behavior using synthetic lognormal series (Figure R6) is exceptionally clear and illuminates the sensitivity of your metrics. You point out that urbanization increases Gini via a peak-magnitude amplification mechanism, whereas precipitation extremes (R95p) have a weaker effect on Gini because they act via a frequency-amplification mechanism. However, log-log regressions inherently struggle with sensor saturation during peak flood events. Add a brief caveat clarifying whether the "frequency-amplification" signature of precipitation could partially be a byproduct of sensor clipping. If the turbidity sensors flatten the true peak values of extreme rain events, the calculated Gini would artificially decrease for those events. Explicitly state if your attribution framework accounts for this potential data artifact or if it reinforces your conclusion that your reported trends represent a highly conservative lower bound.

4. The language adjustments listed in response to Comment 8 (moving from "drives" to "attributed predictors") are a massive step forward for scientific precision. On page 10, lines 248–250 of the

merged manuscript, the text still states: *"urbanization... acts as the primary driver for 43% of sites. Averaged across these rivers, urbanization accounts for 44% of the positive driver-attributed change..."*. The use of "primary driver" alongside "driver-attributed change" in back-to-back sentences dilutes your goal of using purely associative/attributed language. Ensure strict consistency by replacing "primary driver" with "dominant predictor" or "primary attributed factor" in line 242 to match the cautious tone established in your responses and in the abstract.

5. Table R2 (Model Hyperparameters) is a great addition for reproducibility. Please ensure that the sequence length parameter unit (days) is explicitly labeled in the table header.

6. In Figure 5, ensure the axis label formatting for the Flux Trend matches standard international oceanographic/hydrologic notation consistently—ideally explicitly utilizing negative exponents rather than mixed formatting inline.