General Comments

This study quantified residuals in average stage-discharge rating curve from manual field measurements at U.S. Geologic Survey stream gaging stations. The residuals about the average stage-discharge curve quantify changes in channel capacity by evaluating the change in discharge required to achieve a certain water stage. For each measurement, at each gage, the authors quantified a set of geomorphic, hydrologic, and atmospheric variables. Included in these variables were individual storm properties. Assign of storm properties for each measurement were quantified by considering a lag time and computing the median storm property for all storms within that lag time. Lag times of 15, 30, 90, 180, and 365 days prior to the stage-discharge measurement were considered. The authors then trained and validated a machine learning model to predict residuals based on the suite of geomorphic, hydrologic, and atmospheric variables. They evaluated abrupt loss of channel capacity by identifying shifts from a positive residual to a negative residual about the average stage-discharge curve. Only residuals outside the 95% confidence bounds of the stage discharge curve were considered. They quantified the likelihood of change after a storm as the percentage of residuals that underwent a shift from positive to negative and where the residual was outside the 95% confidence bounds of the average stage discharge curve. The authors also identify correlation and important variables for accurately predicting residuals from the machine learning model.

The overall method, data, and evaluation technique has the potential to provide a valuable contribution to predicting variability in channel capacity through residuals of the average stage-discharge curve. Inquiry into relevant scientific questions are presented. However, the current interpretation and analysis makes assumptions that may not be valid, lacks clarity, and requires more direct links between cause and effect than are stated within the article. Therefore, the article requires major revisions, including specificity of research aims, results interpretation, consideration of applied terminology, and acknowledgement of additional limitations. For instance, the first aim of the paper is to map the spatial variability of geomorphic response to extreme storm events, but the authors fail to acknowledge or address spatial correlation and bias in the stream gaging network. The definition of extreme in this article is unclear and it is unknown to what extent the included storms are extreme or quite frequent. The second aim is to understand the impacts of these storms on the stage-discharge relationships at gaged sites as a proxy for changes in flood hazard. However, this makes the assumptions that the storms alone are responsible for any observed changes in the residuals. While possible, other geomorphically significant events could have occurred that are unaccounted for. Further, the authors include other metrics in addition to the storms for predicting residuals, which makes it difficult to separate the impact of other drivers from the storms. The following subsections include general comments on various sections of the paper. Subsequently, specific comments are provided on a line-by-line basis.

Introduction

- the short paragraphs appear and read choppy. Consider combining paragraphs where subject matter allows.
- In the introduction, the authors imply that “extreme” storms or events are predominantly responsible for abrupt shifts in channel capacity and thus flood hazards. It is important to recognize that extreme storms/events likely contribute significantly to the population of abrupt shifts in channel capacity. However, there might be more frequent events that contribute to

these changes as well, particularly depending on channel response potential (e.g., a sand bed river with high sediment supply and non-cohesive banks vs. a gravel bed river with heavily vegetated banks.) Thus, it is recommended to re-consider the use of "extreme" and apply more focus on "abrupt" channel changes to more accurately state the study objectives. For instance, it is not clear to what degree the population of storms included in the analysis is composed of "extreme" storms and what classifies those storms as extreme.

- Line 102: How might this tool be used at ungagged sites without the detailed and rich dataset available? If applicable, it would be beneficial to highlight the use and importance of the tool in the conclusions.

Materials and Methods

- The authors should acknowledge the bias of stream size representation and spatial density in the gaging network and how this might impact spatial interpretation of results. Some sizes and areas are vastly under- and over-represented.
- The method for computing likelihood of change ignores monotonic trends in decreasing capacity – increasingly negative residual. If the residuals become more and more negative, it indicates channel capacity is decreasing, but this is not accounted for by only counting shifts from positive to negative. This limitation should be acknowledged. To some degree, the reported method only accounts for oscillating shifts – positive residual to negative residual then positive residual to negative residual.

Results Analysis

- Why did the authors choose to provide a results analysis section instead of organizing as results and discussion. The overall coherence and understanding of the results would be improved by breaking the results analysis section up into a results and discussion section.

Specific Comments

- Line 30: It is not entirely clear what is meant by traditional "cause-effect" studies. I presume the authors are referring to changes in peak flows due to changes in causal mechanisms such as climate, land use, etc.
- Line 32: How are might they over- or under-estimate actual damage, and what damage? Perhaps a follow-up example or additional explanation would clarify this sentence.
- Line 34: This is, in effect, what fluvial geomorphology is, and this sentence is somewhat redundant with the rest of the paragraph.
- Line 39: also critically modify the landscape and climate(???)
- Line 40: I am not sure flood risk is something that we measure more so than we estimate. Flood risk in fact can be highly uncertain Further, it is not only based on flood frequency, but the relationship between magnitude and frequency as is typically described by a distribution of peak flow, which are discretized as either annual maxima or peaks over threshold. Not just based on flood frequency.
- Line 41 - 43: Recent works have employed methods that incorporate changing channel capacity:
    - Stephens, T. A., & Bledsoe, B. P. (2023). Flood Protection Reliability: The Impact of Uncertainty and Nonstationarity. *Water Resources Research*, *59*(2), e2021WR031921.

    - Stephens, T. A., & Bledsoe, B. P. (2020). Probabilistic mapping of flood hazards: Depicting uncertainty in streamflow, land use, and geomorphic adjustment. *Anthropocene*, *29*, 100231.
- Line 44: This is poor wording, the amount of water that flows through the river systems during floods could in fact change in some situations. Revise to a more correct sentence or consider removing the first portion.
- Line 47: I presume by the use of frequency, the authors are describing the discharge magnitude of the flood. Instead of frequency, consider revising to magnitude, flow, or discharge since they are referring to the size and not how often it floods during a single event.
- Line 49: please give an example of some flood properties.
- Line 54: magnitude, frequency, and risk.
- Line 55: Do the changes have to be rapid? What about long term trends that are not accounted for. Consider shifts in the mean vs. monotonic trends. Sometime flood hazard maps are not updated for a decade or more, beckoning a definition of rapid in this context.
- Line 58: are the trends in stage or erosion/deposition or both comparable to trends in peak streamflow?
- Figure 1 would benefit from a scale bar.
- Line 70: How do we know these are “sharp”, and how do we know the revisions are “upward”? Couldn’t they be downward if erosion occurred?
- Line 95: “Despite some limitations” is used to start the previous sentence. Consider removing from one of the sentences. This sentence would read more formally by re-writing to remove the words “we” and “us”.
- Line 148: please define gaps in the measurements. The manual field measurements may follow irregular frequency. Therefore, what constituted a gap? Minor gaps or missing data in the regular stage-flow measurements by the gage may not have a substantial impact on the analysis.
- Line 155: stage, water level, or water surface elevation is more clear than “levels”
- Figure 3a would be improved by indicating the flood stage. Near a stage of 2m, there is not much difference in the pre and post 2007 measurements. Is this due to overbank flow?
- Figure 3: “In (b), some outlier residuals are evident, likely due to shifts in measurement locations. These points were filtered out before performing the ML training.” Belongs in the text rather than the figure caption.
- Figure 3 c and d caption: Is it in fact channel area and width or wetted area and width? The use of channel over wetted mean two different things. The wetted area and width can change for a single channel geometry. Please clarify at line 160 as well.
- Figure 3d: Should the y-axis label and caption be area or volumetric rate? Contradicts what is reported at lines 160 – 162. For area use area. For capacity use flow rate. Please clarify.
- Figure 3: Please note that Figure 3c and possibly 3d (depending on capacity or area) could fluctuate due to differences in measurement location, which can vary substantially from measurement to measurement. Even if measurement locations are close in distance, they may be upstream or downstream of a bridge. These factors must be considered when comparing widths to evaluate changes in the channel.
- Line 178: does a frequency of 520 events at a gage disqualify them as “extreme”? This seems like a high frequency.

- Line 181: The authors might improve clarity by explicitly stating each gage measurement contained 5 different median storm characteristics – 1 median storm characteristics for the five different lag times considered. If I am interpreting this correctly.
- I understand it would be difficult to graphically convey this in the paper, but I am wondering if the authors investigated the sensitivity of median storm characteristics to lag time. I wonder how much difference there is here. It is not essential, but if available, a note on this would be interesting.
- Table 1 would be more easily viewed in landscape layout and perhaps broken into 3 different tables. One table for each variable classification (geomorphic, hydrologic, and atmospheric).
- Table 1: Should the RFACT (Rainfall runoff factor) be classified as hydrologic instead of geomorphic?
- Line 196: As per previous comments, how do we know they are “severe”? Do the median characteristics reflect this?
- Line 326: comparing the predicted residual with the average residual - Why was this done? Was it for validation?
- Line 332: Some change is neglected in this computation: negative to positive, positive to more positive, and negative to more negative. Therefore, this sentence is somewhat inaccurate.
- Line 332 to 335: How was the confidence interval for the stage-discharge relationship computed?
- Line 344: Does this show the importance of geomorphology of the watersheds or bias in the number of variables selected to represent each variable class? In this interpretation, the authors have neglected the fact that there are different numbers of variable classes. Simply the inclusion of more in once class than the other does not directly translate to its importance in this case. The following sentence does fit the authors interpretation.
- Line 351: drainage density is correlated with other variables as well, such as precipitation.
- Line 360: There is no evidence that flow regulation structures are the cause for these findings. It might suggest it if hydro_disturb_index only reflects flow regulation structures, but it could also include urbanization.
- Table 2: The caption should state what the Corr and RMSE compare.
- Line 374: It would be helpful to know something about the distribution of residuals to provide context to the RMSE magnitudes.
- Line 391: Is it the spatial “spread” or spatial “patterns”?
- Line 418: Low flows are more of a hydrologic property rather than a morphodynamic property.
- Line 450: Vulnerability was not defined, quantified, or reported anywhere prior to this.
- Line 473: In this sentence, it is not clear how the FHF increased logarithmically. How do we know this from the data presented?
- Line 481: Directly comparing regions does not account for spatial correlation or representation bias in the gaging network. Some areas/regions and streams are more represented than others making a comparison between regions misleading.
- Line 494: Only a portion of the streams in the Atlantic Plain are tidally influenced by the ocean. Further, an even smaller portion of the gages are. This sentence is not supported and speculative at best.

- Line 508 – 509: How does it confirm this? Please explicitly link the supporting evidence with this statement. As it is written, the previous sentences do not provide any clear evidence of this.
- Line 511: A citation would strengthen this statement.
- Line 515: many different sediment types may exist within a physiographic region. Sand bed channels, gravel bed, cobble, etc. The Appalachian highlands for example. Broadly inferring sediment types by physiographic regions is tenuous. I would suggest trying to relate sediment types to other variables such as drainage area and slope. See:
    - Montgomery, D. R., & Buffington, J. M. (1998). Channel processes, classification, and response. *River ecology and management: lessons from the Pacific Coastal Ecoregion*, 13-42.
    - Flores, A. N., Bledsoe, B. P., Cuhaciyan, C. O., & Wohl, E. E. (2006). Channel-reach morphology dependence on energy, scale, and hydroclimatic processes with implications for prediction using geospatial data. *Water Resources Research*, *42*(6).
- Line 545 – 555: the connections between the centroid of perception, flash floods, and residuals is not made clear here. The centroid of precipitation is an important variable in the analysis for predicting residuals. How does this tie to flash floods? Please explain more clearly.
- Line 575 – 580: this not necessarily true. Just because the channel conveyance capacity is exceeded does not mean the channel is expected to change. The flood must result in a geomorphically significant conditions of hydraulic and sediment supply conditions. The authors mentioned previously the importance of hysteresis in sediment deposition. This sentence over-simplifies and incorrectly categorizes a complex, situationally unique, and nuanced process.
- Line 599: The use of “future” here is misleading since the authors explicitly evaluate short-term or abrupt shifts. The temporal persistence of that shift is not addressed.
- Line 600: As it stands, specific impacts from individual drivers is insufficiently addressed. A more accurate representation of the analysis would be to say that the method identified important drivers for predicting residuals from the average stage discharge curve. From my understanding, the analysis does not necessarily reveal the actual impact of specific variables on the predicted residual.
- Line 608: Did the authors mean to say channel capacity here instead of “river discharge”?
- Line 615: More specifically, the risk of immediate reduction in channel capacity. The authors did not evaluate increases in channel capacity.
- Line 616: Knowing the temporal persistence of these changes would provide insight to the feasibility of these updates or alternative methods for quantifying flood risk if the process is highly variable in time.

Technical Corrections

- Line 284 and 285: The numbers indicating the numbered list should contain parentheses after the number e.g.: 1) text…., 2) text…,
- Line 341: typo “. analysis.”
- Line 387-388: Revise: “We have got”. Improved writing would be something like “There are 12 clusters of gages”
- Line 517: typo. Remove period.
- Line 542: add a “the” between “in” and “literature”.

- Line 543: Missing a space after the parentheses.
- Line 551: New paragraph? Indent if so.