First, we want to thank the reviewer for the insightful detailed comments and recommendations. We appreciate your outlook on the potential of our study.

Following the suggestions, we will revise the manuscript accordingly. Below we provide responses detailing the revision actions we plan to take to address the reviewers' comments.

Note: Below is our response (italics) to each reviewer's comment (regular font)

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.

Response 1: *We thank the reviewer for all the insightful comments: in our revised manuscript we will incorporate the recommended changes to represent the novelty and originality of our work more clearly. We will particularly focus on explaining the methods and rationale behind every step in the procedure. We will do our best to bridge the stated aim, interpretation, and limitations of the study following the reviewer's suggestions.*

Regarding the coverage of stream gages, we agree on the intrinsic limits of the dataset, based on the fact that there is variability across CONUS regarding the spatial and temporal coverage of stream gages. These limits in general have been addressed in literature and are very well summarized in the publication by Kiang et al., (2013).

Overall, gage coverage is higher in the Eastern United States compared to the Western United States. The arid Southwestern United States, Alaska, and Hawaii show the lowest spatial coverage, and these regions,

except for Hawaii, often have short streamflow records. Gage statistics quality, according to Kiang et al. 2013 also varies across the country, mostly due to variations in hydrology. Notably, the arid and semiarid areas in the Central and Southwestern United States exhibit higher interannual variability in flow, leading to greater uncertainty in flow statistics. These findings will be discussed further in our results section. Despite these observations, it's important to note that any research relying on gaging sites faces challenges of potential over or underrepresentation, which we will also emphasize in the discussion of the limitations and advantages of our proposed model.

The definition of extreme in this article is unclear and it is unknown to what extent the included storms are extreme or quite frequent.

Response 2: *We will revise the manuscript and we will revise the wordings*

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.

Response 3: *We thank the reviewer for this comment. Indeed, channel changes can be due to other geographically significant events (e.g. landslides, debris flow etc), however, such occurrences could also be triggered by storm events. At this stage, we have a complete database of storm properties, but we did not include a complete analysis of additional events such as mass movements. This goes beyond the scope of this work. In the revision, however, we will also highlight this point.*

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.

Response 4: *We thank the reviewer for this comment. We believe that adding the other variables gives a better context to the impact of the storms. Overall, we have done a feature importance analysis and selected only those drivers that are the most important and influential. The feature importance itself, highlighted which variables mattered the most in predicting the residuals. We decided to use hydrologic and geomorphological variables because landscape properties are also linked to the potential effects of storms.*

The following subsections include general comments on various sections of the paper. Subsequently, specific comments are provided on a line-by-line basis.

<u>Introduction</u>

the short paragraphs appear and read choppy. Consider combining paragraphs where subject matter allows.

Response 5: *We will read through the manuscript carefully and introduce edits as needed.*

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.

Response 6: *Thank you for this recommendation. We have thought about this comment and decided to find an alternative to the word "extreme events" in the revised manuscript.*

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.

Response 7: *We thank the reviewer for this comment. We believe that as USGS stream gage information could potentially be transferred from nearby stream gages if there is sufficient similarity between the gaged watersheds and the ungaged watersheds of interest, our model could also be applied to ungaged sites.*
However, one must always keep in mind that the successful 'translation' to ungaged environments depends on the correlation of the stream gages in the surrounding areas. For example, there are areas of CONUS (mostly mountainous) that show highly correlated stream gages (Kiang et al., 2013), whereas the Central United States and coastal areas of the Southeastern United States show much uncorrelated gages. Therefore, the goodness of the information transfer might not work as well. Also, transferability would be most likely to be successful when basin attributes show high similarity and storm properties are within the range of variability of the training set used for this work. We will add some consideration about this in the manuscript

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.

Response 8: *Regarding the coverage of stream gages, we agree on the intrinsic limits of the dataset, based on the fact that there is variability across CONUS regarding the spatial and temporal coverage of stream gages. These limits in general have been addressed in literature and are very well summarized in the publication by Kiang et al., (2013).*

Broadly speaking, the Eastern United States has better coverage compared to its Western counterpart. Particularly, the arid Southwestern United States, Alaska, and Hawaii show notably lacking spatial coverage. Except for Hawaii, these regions also tend to be covered by shorter streamflow records. Discrepancies in hydrology contribute to variations in the statistical uncertainty calculated across different parts of the country (Kiang et al., 2013). *The Central and Southwestern United States, characterized by arid and semiarid conditions, generally display higher interannual variability in flow, resulting in increased uncertainty in flow statistics. In the revised manuscript, we will incorporate these comments. Despite these distinctions, it's essential to recognize that any research relying on gaging sites faces similar limits and is overall affected by potential over or underrepresentation of flows. This aspect will be emphasized further in the revised manuscript, in the section about limitations and advantages of the proposed model.*

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.

Response 9: *We thank the reviewer for this comment. Indeed, we focus mainly on sudden shifts, rather than on permanent shifts. The main reasons for this were - 1. Short-term conveyance capacity changes are not considered in typical flood hazard assessments and could substantially overstate or understate flood threats at any particular time for subsequent floods; 2. there is a plethora of complex and sometimes not linear-processes and coupled feedback that we would need to 'model' in the training set, to provide a comprehensive benchmark to identify permanent shifts vs sudden ones, and this could be a potentially interesting research that could be tackled by further studies building on our model, but at this stage it goes beyond the scope of this work. We will highlight this point better in the manuscript.*

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.

Response 10: *We will add a discussion section as per reviewer's suggestion*

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.

Response 11: *yes, this is correct. Giving this comment, we will rephrase this sentence*

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.
Response 12: *We will revise this part of the manuscript*

Line 34: This is, in effect, what fluvial geomorphology is, and this sentence is somewhat redundant with the rest of the paragraph.
Response 13: *We will revise this part of the manuscript*

Line 39: also critically modify the landscape and climate(???)
Response 14: *We will rephrase the sentence for better clarity.*

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.
Response 15: *Yes, this is correct. We will rephrase it to 'flood risk estimation'. We will also improve the wording of this sentence.*

Flood risk measurement has traditionally been based on flood frequency, derived from variability in streamflow, assuming constant channel capacity (Merz et al., 2012; Slater et al., 2015). The relationship between magnitude and frequency is also generally built upon the peak flow distribution, whereas peaks

are discretized as either annual maxima or peaks over threshold, but mostly assuming that river capacity remains constant over the investigation records.

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.

Response 16: *Thank you for the references. We will add these to the manuscript and rephrase the text.*

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.
Response 17: *We will revise this part of the manuscript*

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.
Response 18: *We will consider the suggestion and revise this part of the manuscript*

Line 49: please give an example of some flood properties.
Response 19: *We will clarify that we refer to inundation extent and depth*

Line 54: magnitude, frequency, and risk.
Response 20: *We will rephrase*

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.
Response 21: *For this work, we investigated sudden changes of positive to negative residuals. We acknowledge that these might not be permanent changes. Given the complexity of processes involved in the 'restoration' of river forms, or the permanence of a channel shift, we decided to focus this work on the sudden changes, as proposed by Slater et al. 2015 in her work. With this idea, with the proposed method we highlight rivers more prone to changes in the aftermath of a storm, highlighting potential increased flood hazard in the case of subsequent storms.*
In literature, using Slater's concept, the work by Ahrendt et al 2022 offers an overview of historic long-term and short-term conveyance changes for WA, whereas the work of Li et al 2020 highlighted how relatively modest long-term changes in river channel capacity are composed of numerous short-term transients which are of much larger magnitude. We referred to this work in our manuscript and will add some considerations on the fact that this work only considers sudden changes but not their persistence in time.

Line 58: are the trends in stage or erosion/deposition or both comparable to trends in peak streamflow?

Response 22: *Other works in literature highlighted that some channel changes could provoke a higher change in flood hazard than shifts in discharge alone (Slater et al. 2015, 2016, Li et al. 2020, Ahrendt*

et al 2022). For this work, we did not assess changes in streamflow, as we are training the model based on storm properties, and not on long term discharge properties.

Figure 1 would benefit from a scale bar.
Response 23: *We will add this in the revised figure*

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?
Response 24: *Indeed, the changes could be downward if erosion occurred. We imply that an upward revision is a proxy for an increase in flood hazard, whereas a downward revision potentially could mean a reduced hazard. Our analysis is consistent with other works in the literature relating shifts in the stage-discharge relationship as a proxy for flood hazard.*

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".
Response 25: *We will consider the suggestion and revise this part of the manuscript*

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.

Response 26: *We have excluded the gages that do not have continuous data for the tie frame from 2002-2013. We will clarify this better in the manuscript. For the work, aside from considering consistent gages present in the Shen et al. Database, and covered by stream measurements, we applied the same criteria as Slater et al. (2015), who only considered field measurements in which the discharge is within one percent of the product of channel velocity and cross-sectional channel area, as reported by the USGS, and those made in close proximity to the gage station.*

Line 155: stage, water level, or water surface elevation is more clear than "levels"
Response 27: *We will rephrase this based on the reviewer's suggestion*

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?

Response 28: *The figure was shown as an example of shifts present in the measurement data. For this gage, the flood stage is at 10ft, and the peak discharge of the 2007 event was 11.51 ft, and the Quinnipiac River itself (at the gage right upstream of the one in the picture) measured the maximum discharges for the period of record of the station during the 2007 flood. Aside from the information provided by USGS on that event, we do not have direct knowledge of the event itself so we cannot make a precise statement on the reason behind the similarities highlighted by the reviewer.*

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.
Response 29: *We will move this part to the text*

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.

Response 30: *The figure reports the channel width as reported in the gage measurements. We will clarify this in the manuscript.*

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.

Response 31: *There was a mistake in the text, the line should have read 'and channel conveyance (Figure. 3d).' We will rephrase this in the text.*

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.

Response 32: *According to the information of the gage, the measurements did not shift in location. For the work itself, consistently with Slater and the open codes provided in her work, we removed all field measurements made in a different (or potentially different) location, and all field measurements made in icy conditions, as these might affect measurements of channel geometry. We will highlight this more clearly in the manuscript.*

Line 178: does a frequency of 520 events at a gage disqualify them as "extreme"? This seems like a high frequency.

Response 33: *Indeed, the reviewer is correct. The magnitude of these events varied in time, and in the revised manuscript we will refrain from defining 'extreme' the events.*

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.

Response 34*: Yes, this is correct. We will clarify this in the manuscript as suggested.*

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.

Response 35*: Thank you for this comment. We have not considered this at this stage. For the revised work, we will add a note if we see any meaningful results.*

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).

Response 36*: Thank you for the suggestion. We will try to revise this as suggested*

Table 1: Should the RFACT (Rainfall runoff factor) be classified as hydrologic instead of geomorphic?

Response 37*: There is a mistake. We will correct this.*

Line 196: As per previous comments, how do we know they are "severe"? Do the median characteristics reflect this?

Response 38*: We will rephrase this in the manuscript*

Line 326: comparing the predicted residual with the average residual - Why was this done? Was it for validation?
Response 39*: The reviewer is correct. We did this to validate the model.*

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.
Response 40*: We will rephrase this*

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.

Response 41*: We will rephrase this and clarify.*

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.
Response 42*: We will rephrase this.*

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.

Response 43*: We will rephrase this adding the comments regarding the distribution of gages across CONUS.*

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.

Response 4*: We will rephrase this*

Technical Corrections

Response: *We will carefully fix all the technical corrections mentioned by the reviewer and revise the manuscript.*

References:

Kiang, J. E., Stewart, D. W., Archfield, S. A., Osborne, E. B., Eng, K., & Survey, U. S. G. (2013). A national streamflow network gap analysis. In *Scientific Investigations Report*. https://doi.org/10.3133/sir20135013