Response to the Editor

Dear Editor,

Thank you for your thoughtful feedback and for the opportunity to revise our manuscript. We appreciate the reviewers' comments and have carefully addressed all concerns raised. In particular, we have revised the model training and validation approach and included additional information to strengthen the robustness of the manuscript. We have also clarified and emphasized the innovative aspects and broader relevance of our findings throughout the revised manuscript. We have responded to each comment in detail and incorporated the suggested improvements to enhance the overall clarity and scientific contribution of the work. We hope the revised version now meets the standards for publication in Natural Hazards and Earth System Sciences.

Best regards,

Mariam Khanam

On behalf of all co-authors

Response to the Reviewers

We thank the reviewers for their detailed and constructive feedback, which has provided valuable guidance for improving our manuscript. Below, we address each comment point by point, incorporating clarifications and revisions where necessary to strengthen our study.

Note: Below is our response (italics) to each comment (regular font) from the reviewer

Reviewer 1:

1. A recent study has been performed such as https://doi.org/10.1016/j.jenvman.2024.121764

Response: We thank the reviewer for this suggestion. We have added this to the manuscript.

2. Line 137: Please modify the figure such that input (FGP, Rainfall, and Population) and output (LYI) of the XGBoost model can be distinguished.

Response: Thank you for this suggestion. I have modified the figure to clearly separate the inputs (FGP, Rainfall, Population) from the output (LYI) in the XGBoost model diagram.

3. Line 156: "variant" may be a better word choice.

Response: Thank you for pointing this out. We have replaced the current term with "variant" to ensure clarity and precision in our wording.

4. Line 172: I understand this concept and it is very interesting. It will be more interesting if the map of these categories (low, medium, high LYI) of the training dataset can be shown along with the map of the actual LYI.

Response: we thank the reviewer for this comment. As the manuscript is already dense with figures and plots, and the method aims to identify the categories, we prefer not to add the figure.

5. Line 238: Again, a map showing this index over the study area will be very interesting.

Response: Thank you for this comment. Please consider that in the revised manuscript, we have added Figure 7 showing an example for 1985 to 2020 changes.

6. Line 279: This is an interesting and crucial idea for your XGBoost model validation, but in my view, the method of dissecting your study area in half, using one half to develop the model and the other half for validation, would be the best way. Then, you can develop a confusion matrix between the simulated LYI category and the observed LYI category. Based on this result, you can even perform an uncertainty analysis of your final result (e.g., what is the probability that an area classified as high LYI actually has high LYI).

Response: Thank you for this comment. One should consider that we have measured LYI at the district and watershed scale only for Nepal, and not for the whole HMA. Dividing the data in half might fail to capture correctly the geographic variability of the area. As the training dataset is limited, we believe that the proposed approach is more robust. Please see below the part in the manuscript:

Line 286: We conducted thorough testing and validation of our model for Nepal, comparing the predicted value of LYI to the calculated Lifeyears Index (LYI) data from tabular values specific to the region. We trained the model and validated it only using the data for Nepal, at the district scale and then at the watershed scale. Overall, we opted for a 90-10 approach, for which 90% of the Nepal data were used for training and 10% for validation. Upon extending the model's applicability to the entire High Mountain Asia (HMA) region, we rigorously assessed the quality of our results by comparing the predicted social impact with that reported in established flood databases covering the region. To verify our findings, we compared the predictions at the HMA level with flood events reported in the Dartmouth Flood Observatory's (DFO) Global Active Archive of Large Flood Events, 1985–Present. This comprehensive database compiles information on major floods sourced from diverse channels such as news reports, governmental records, ground observations, and remote sensing data. Notably, the DFO dataset encompasses various flood types, including lowland floods and mountainous river floods characterized as fluvial and pluvial floods.

7. Line 305: An additional section that discusses the ranges of predictor variables for watersheds classified as high LYI would be interesting. This is because we cannot solely rely on the AI model, which operates as a black box.

*Response: We have added some discussion and figure 7 to "**3.1. Variability of the Predictors" section in** the manuscript:*

Line 341: Much of the population of Nepal tends to be concentrated in areas with higher FGP, as is typical for mountainous areas, where population and economic activities are mostly located in the river valleys. Globally, the floodplains of rivers are preferred living spaces for the population and provide

favorable locations for economic development. These areas are commonly exposed to floods, however, an increasing population, together with the changes in storminess, mean that the risks from flooding are expected to be higher. On average, the population increased significantly in watersheds that transitioned from low to medium (LtoM), medium to high (MtoH), or low to high (LtoH) flood risk categories (Figure 7: example variability from 1985 to 2020). This suggests that growing population density in certain watersheds may be contributing to increasing flood susceptibility. The CI (climate concentration index) slightly decreased over this period for some watersheds. However, watersheds experiencing population growth were more likely influencing the transition to a higher flood risk category. Although CI has not significantly increased, the interaction between land-use change, urban expansion, and demographic shifts may be playing a role in driving these transitions. Transitioning watersheds have a higher average FGP compared to the overall average FGP and tend to have a larger average watershed area compared to all watersheds. This indicates that larger watersheds are more prone to experiencing shifts FGP and in flood risk categories, possibly due to their ability to accumulate and distribute larger volumes of runoff and sediment. This supports the idea that intrinsic watershed characteristics (such as geomorphology and size) play a role in flood susceptibility alongside external factors like population growth and rainfall concentration index (CI). Area successfully predicted as at high risk (high LYI) in the most recent years, are areas showing high social vulnerability in terms of favorable Social Conditions (lack of communication, access to electricity and infrastructures, lower education, small children under 5); high percentage of migrating community and high risk of poverty and poor infrastructures (Aksha et al.,

2019).

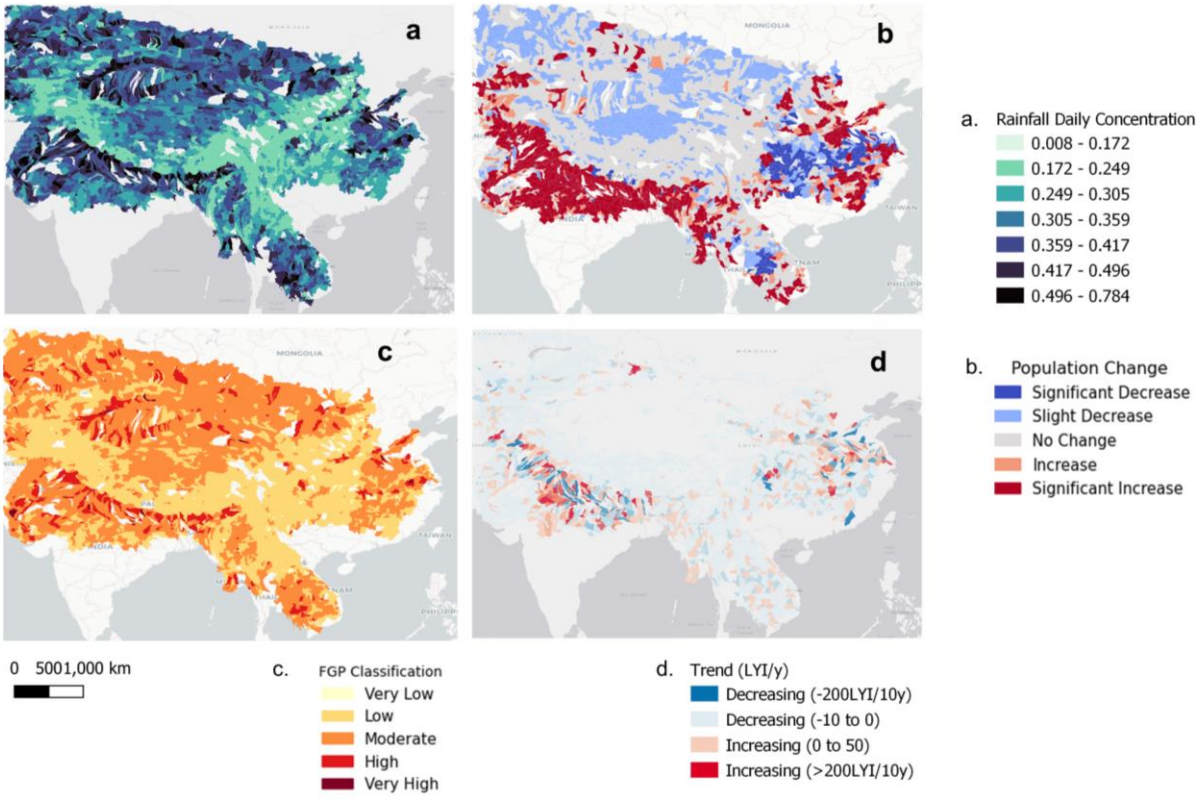


Figure 7: Average variability of the Rainfall CI (a), population change (b) compared to FGP (c) and LYI Trend (d) from 1980-2020

8. Line 306: Suggestion: This section digresses slightly from the main result of the paper. It may be better suited for methodology or a separate discussion section.

Response: Thank you for the suggestion. This section explains how the variability of all the predictors is connected and we consider this to be an important result to explain the correlation. Additionally, we have added a new figure to the section. We believe this section should be considered as a part of the results.

9. Line 330: Very interesting finding! You could write another paper that discusses this matter on a global scale.

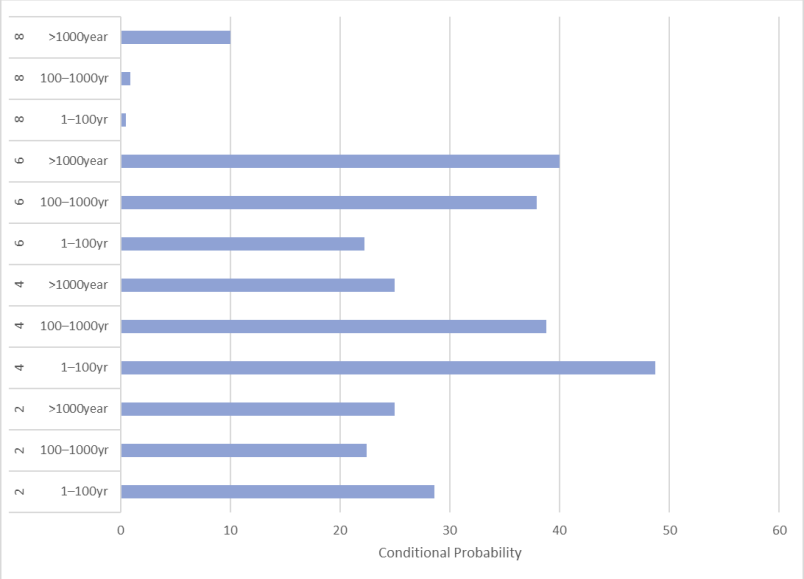
Response: We appreciate your enthusiasm for this finding. We agree that this topic has significant potential for a broader, global analysis, and we will consider pursuing this in a future publication.

10. Line 388: The letters "(b)" in the figure legend are overlapped on the other letters.

Response: Thank you for noticing this formatting issue. We have adjusted the figure 9, legend to remove the overlap and made some further changes as well.

11. Line 400: Very interesting result, but I would try to find a better way to visualize it using a figure.

Response: We thank the reviewer for this comment. The visual representation of these results could be as simple as a bar plot. We prefer however to keep this in a table form, to avoid adding too many figures to the manuscript.



12. Line 416: Same as above. Try to display this information using a figure.

Response: We thank the reviewer for this comment. The visual representation of these results could be as simple as a bar plot. We prefer however to keep this in a table form, to avoid adding too many figures to the manuscript.

13. Figure 10: While this figure is very interesting, its resolution is unacceptable. Please provide a high-resolution image.

Response: We appreciate this feedback. We have replaced the figure with a better resolution version which shows the HMA region.

14. Line 438: This is too much generalization. The primary reason for this result may be that long-term CI values are less variable than short-term CI values. Additionally, you are applying the model developed based on 35 years of rainfall data to data based on 5 years of rainfall. Please discuss or briefly state the limitation of this result.

Response: We thank the reviewer for this comment. Please consider that we calculated the CI for 5 years intervals in both cases. We clarified this in the manuscript. Indeed, there might be a difference if CI is computed over different time windows, but this is not the case of this specific study. We added a comment on this in the manuscript in the newly added paragraph about limitations.

Line 472: 3.7. Model constraints and limits

While this study demonstrates the promise of accurate flood impact prediction, the use of static Flood Geomorphic Potential (FGP) maps presents limitations. Flooding alters channel morphology and downstream topography, impacting future flood dynamics (Khanam et al., 2024). Therefore, dynamic flood topographies are essential for robust hazard assessment. Although high-resolution data post-extreme events can enhance prediction accuracy, the availability of such data is constrained by acquisition frequency. Hence, efforts to improve data availability post-disaster are crucial for enhancing

the reliability of predictive models. Researchers could also derive FGPs from enhanced high-resolution terrain data, such as those derived from LiDAR sources if available. In such cases, however, it is advisable to retrain the model and reassess the significance of this parameter in the updated model, as terrain resolution and survey techniques might determine a variability of the data, especially when dealing with hydrologic parameters (Sofia, 2020).

The climate index considered in this study might vary depending on the input dataset (Reanalysis VS measurements), as well as on the timescale of the analysis. When comparing results to this study, researchers should make careful consideration of the length of the time window used for this evaluation (5 years). If daily data are considered over shorter time windows (e.g., 1 year), the index itself might result in higher values, capturing only short-term variability due to specific isolated storms. Seasonal analyses, on the other hand, would capture more the concentration due to monsoon periods, or dry vs. wet months. The proposed multi-year analysis is in line with literature studies, pertaining climate change studies and studies on the effect of floodings (Sofia et al., 2019; Saki et al., 2023; Du et al., 2023).

Population data for this work relies on standard available datasets. When considering the method to predict future changes, outside the time range covered by the proposed model, headcounts alone cannot offer a full picture. It is crucial also to consider additional elements that could determine population shifts over time.

15. Figure 11: Please enhance the resolution of this figure as well.

Response: Thank you for bringing this to our attention. We have replaced the figure with a better resolution version.

Reviewer 2:

1. Comment: There are some major points that lead me to this decision. The most important is that you trained the XGBoost model on a single nation, achieving what seems to be overfitting, and then you deploy it to a much larger region without any validation. This leads to the possible conclusion that any results you report, while interesting, might be wrong or biased toward similar mechanics as you would have in Nepal.

> The choice of the indices expands the possibility of expanding the study. LYI in a different situation. We can not make decisions based on societal variables. It should not be used as the absolute labeling of the areas

Response: We thank the reviewer for this comment. We acknowledge that the dataset used for training the XGBoost model is geographically limited. The model parameters are determined over the training set. We have leveraged data augmentation techniques to augment training samples and balance difference classes in order to overcome the overfitting issue.

However, we emphasize that we have taken steps to validate the extended model using the DFO dataset, which records actual flood events. This validation helps ensure that the model's predictions are not solely dependent on the training region but can reasonably generalize to broader areas.

Regarding the concern of potential bias, we note that the variability in the dataset across different regions is comparable. Specifically, the ranges for the Climate concentration Index (CI) Index and Flood Geomorphic Potential (FGP) in Nepal and the HMA are as follows:

Nepal has

CI: 0.57 ± 0.15

FGP: 16.5 ± 18.3

And this variability encompasses overall the variability of the HMA dataset (CI: ranges from 0.3 to 0.8 while FGP ranges between 2.1 to 67.5). Please consider that our primary objective is not to establish an absolute classification of flood-prone areas based on societal variables but rather to provide a reasonable assessment of vulnerability in terms of classes of life year lost. While we recognize the limitations of our approach, our findings offer valuable insights into the regional variability of flood risk and its driving factors. We will clarify these points further in the manuscript to reflect the scope and intent of our study.

Line 506: Our goal is to provide a reasonable assessment of vulnerability through life years lost, rather than to definitively classify flood-prone areas by societal factors. Despite certain limitations, our findings offer valuable insights into regional flood risk and its key drivers.

2. Comment: Moreover, there is no clear description of how you trained and validated the model and how you assessed parameter importance.

Response: Thank you for pointing this out. In the revised manuscript we clarified the proposed approach.

Line 286: We conducted thorough testing and validation of our model for Nepal, comparing the predicted value of LYI to the calculated Lifeyears Index (LYI) data from tabular values specific to the region. We trained the model and validated it only using the data for Nepal, at the district scale and then at the watershed scale. Overall, we opted for a 90-10 approach, for which 90% of the Nepal data were used for training and 10% for validation. Upon extending the model's applicability to the entire High Mountain Asia (HMA) region, we rigorously assessed the quality of our results by comparing the predicted social impact with that reported in established flood databases covering the region. To verify our findings, we compared the predictions at the HMA level with flood events reported in the Dartmouth Flood Observatory's (DFO) Global Active Archive of Large Flood Events, 1985–Present. This comprehensive database compiles information on major floods sourced from diverse channels such as news reports, governmental records, ground observations, and remote sensing data. Notably, the DFO dataset encompasses various flood types, including lowland floods and mountainous river floods characterized as fluvial and pluvial floods.

3. Comment: Overall, the paper is also improvable in terms of writing, with several redundant phrasings and unclear sections.

Response: We appreciate this feedback. We will thoroughly review the manuscript to eliminate redundant language and clarify ambiguous sections wherever possible.

4. Comment: Figure 2: It is unclear why the LYI is both an input and an output of the model. Despite the caption saying to refer to the text, this should be made clearer in the image itself as well.

Response: We revised Figure 2 to clearly distinguish between input and output variables.

5. Comment: Line 145: What is the point of using district-level data if you then aggregate it at the watershed level, especially for the testing on the HMA region?

Response: We used district-level data to account for regional variations in population and socioeconomic impacts, which we then aggregated at the watershed level to match the spatial scale of our hydrological analysis. We will clarify this in the text.

6. Comment: Line 147: How do you weight different districts?

Response: The districts are not weighted per se. The aggregation from district to watershed is done by a weighted average, considering the extent of district area within the watershed as a weight. We clarified this in the manuscript.

7. Comment: Table 1: There's no need to repeat "Y =" in the description, as it already appears on the left.

Response: Thank you for catching this redundancy. We removed the extra "Y ="

8. Comment: Line 169: You often use capital letters for Low, Middle, and High risk. Consider using lowercase letters. This applies to the rest of the paper as well.

Response: We have considered the suggestion and revised the text.

9. Comment: Lines 194-195: In theory, you can fully automate from terrain data if you use contributing areas, as you can delineate them based on topography. You could argue that your method provides a better representation of bankfull discharge, but you would need to prove it by comparing the two approaches.

Response: We thank the reviewer for this comment. The method is fully automated from terrain. The FGP stems from the work of Samela et al, where they developed the index to identify flood prone areas based on the proposed geomorphic classifier. Their approach required as input a relationship connecting indeed drainage area to bankfull conditions. For the proposed approach, we automated this last part, meaning the definition of bankfull condition, and we did so by applying methods already existing in literature. The goodness of this method was already described in the referenced papers (Sofia et al. 2011, Sofia and Nikolopoulos 2020, Sofia et.al 2017). Validating bankfull geometry at the scale of HMA is not feasible, as measurements are not readily available. This, furthermore, goes beyond the scope of the work.

10. Comment: Line 209: You are comparing with HAND, which is not exactly a standard inundation model, as it models water depths with several assumptions. I think it is overall fine to compare with it, but mention at least its limitations.

Response: The HAND (Height Above Nearest Drainage) model is a widely used approach for estimating flood inundation extents and water depths. It operates on the principle of deriving relative elevations from a DEM, similar to our approach, which also relies on DEM-based analysis. While having assumptions may introduce some limitations in accurately capturing complex flood dynamics, HAND remains a useful and practical method for large-scale flood assessment due to its computational efficiency and compatibility with readily available topographic data. Given these similarities, we find it reasonable to include HAND as a comparative reference in our study while acknowledging its limitations. We will ensure that these aspects are clearly mentioned in the discussion.

11. Comment: Figure 3: Please reduce the size of the plot, as it does not need to be this large.

Response: Thank you for the suggestion. We have replaced the figure and reduced the size as much as possible.

12. Comment: Lines 211-212: Didn't you just say that you used a modified version of this index? If that is the case, then your claim is incorrect, as only the GFI has been validated, not the FGP.

Response: we thank the reviewer for this comment. Please consider that FGP stems from the work of Samela et al, where they developed the index to identify flood prone areas based on the proposed geomorphic classifier. Their approach required as input a relationship connecting indeed drainage area to bankfull conditions. For the proposed approach, we automated this last part, meaning the definition of bankfull condition, and we did so by applying methods already existing in literature. The goodness of this method was already described in the referenced papers (Sofia et al. 2011, Sofia and Nikolopoulos 2020, Sofia et.al 2017). Validating bankfull geometry at the scale of HMA is not feasible, as measurements are not readily available. This, furthermore, goes beyond the scope of the work. As an overall visual assessment, we proposed the HAND comparison, mentioning in the revised paper its limitations. Furthermore, we rephrased the sentences by changing the terminology from FGP to DEM-derived geomorphic index in this context.

It's worth noting that the DEM-derived geomorphic index has been previously published and applied in various contexts (Samela et al., 2017). While testing the quality of the DEM-derived geomorphic index lies beyond the scope of this work, its effectiveness for flood mapping has been well-established in previous studies (Manfreda et al., 2011, 2014; Manfreda & Samela, 2019; Samela et al., 2016, 2018), which have demonstrated the utility of the methodology, particularly in ungauged conditions, for preliminary identification of flooded areas in regions where conducting expensive and time-consuming hydrologic-hydraulic simulations may not be feasible. The goodness of the bankfull measurement system, furthermore, was already described in (Sofia et al. 2011, Sofia and Nikolopoulos 2020, Sofia et.al 2017).

13. Comment: Section 2.3.2: Please add the formula used for FGP calculation in the text, rather than only in the figure. You can leave the figure to explain the variables used.

14. Comment: Figure 4a: What is

$w=\alpha A\ \beta$? I understood from the text that the reference height was determined from the landscape rather than formulas.

Response 13/ 14: Thank you for this suggestion. We included the FGP calculation formula directly in the text of Section 2.3.2

Line 190: We opted for considering a variation of the Samela et al., (2017) which is a modified Geomorphic Flood Index (GFI) by Sofia, et al., 2017b & Sofia et al., 2015, thereby described as Flood Geomorphic Potential (FGP).

$$FGP = \ln (h_r / H) \qquad (2)$$

The index is calculated as the logarithm function of the bankfull elevations, H (estimated using a hydraulic scaling function, or HSF ($w=\alpha A^{\beta}$), based on bankfull width (w) and contributing area (A)) in the element of the river network closest to the point under examination and the elevation difference between these two points, h_r (Figure 4, Equation 2). The index was improved over a main aspect: the

automatic identification of the HSF directly from terrain data, applying the technique of (Sofia, et al., 2017b; Sofia et al., 2015) to retrieve the bankfull location automatically through the landscape. This has the advantage of allowing for full automation of the mapping starting purely from terrain data.

15. Comment: Figure 4b: Include in the caption that you are also showing the corresponding orthophotos of the sites considered for the flood maps.

Response: Thank you for this comment. We have added that to the caption.

16. Comment: Moreover, the legend is quite small; consider increasing it.

Response: Thank you for noticing this detail. We have increased the legend size.

17. Comment: Line 249: I don't understand the need to clarify that one variable is on the x-axis and the other on the y-axis. Consider removing this for clarity.

Response: Done.

18. Comment: Equation 2: I would move this equation to line 241, just before explaining it.

Response: We moved Equation 2 to align with the explanatory text.

19. Comment: Figure 5: It would be useful to color different areas under the curve with different colors or patterns to help the reader.

Response: Done.

20. Comment: Section 2.4.1: There is no indication of how the model was trained, validated, and tested; how many samples were used for each; or how results were assessed in terms of metrics.

Response: Thank you for your comment. We have modified the section with more details and also Section 3. We discuss further on the feature importance, model performance and validation in the section 3.

2.4.1 Validation of the System at the HMA Scale

We conducted thorough testing and validation of our model for Nepal, comparing the predicted value of LYI to the calculated Lifeyears Index (LYI) data from tabular values specific to the region. We trained the model and validated it only using the data for Nepal, at the district scale and then at the watershed scale. Overall, we opted for a 90-10 approach, for which 90% of the Nepal data were used for training and 10% for validation. In total, there are 1520 data points for 38 basins from 1981 to 2020. The model was trained on the data by removing one specific year's data (e.g., 2003, 2012, 2017, and 2020) and then test on this specific year's data. Average performance, e.g., precision, recall, F1 measure was reported. Upon extending the model's applicability to the entire High Mountain Asia (HMA) region, we rigorously assessed the quality of our results by comparing the predicted social impact with that reported in established flood databases covering the region. To verify our findings, we compared the predictions at the HMA level with flood events reported in the Dartmouth Flood Observatory's (DFO) Global Active Archive of Large Flood Events, 1985–Present. This comprehensive database compiles information on major floods sourced from diverse channels such as news reports, governmental records, ground observations, and remote sensing data. Notably, the DFO dataset encompasses various flood types, including lowland floods and mountainous river floods characterized as fluvial and pluvial floods.

21. Comment: This section is also very unclear when you describe the comparison with the DFO database.

Response: Thank you for this comment. We have tried to make the methodology clear as much as possible. We will appreciate if the reviewer can give us some specific indication on which part needs to be modified.

22. Comment: Figure 6 is not clear enough from the legend and caption.

Response: Figure 6 does not have any legend in it. We are not sure which figure are you referring to.

23. Comment: Section 3.2: How was this variable importance assessed?

Response: We have modified section 3.2 with the following explanation of the feature importance:

In this section, we present a variable importance comparison (Figure 8) based on the Feature Importance Score (F-score) in XGBoost. XGBoost provides F-score based on how frequently a feature is used in splitting the data across all decision trees. This is the number of times a feature appears in a split across all trees in the model. A higher value indicates that the feature was used more frequently in decision-making, suggesting it has a stronger influence on model predictions. The F-score indicated that population (Pop) was the most important variable, which was consistent with our expectation in the sense that the socioeconomic impact depends largely on the exposure. The climate variable (CI) happened to be the next important variable, showing the significance of the region's climate on the socioeconomic impact of flood occurrences.

24. Comment: Lines 344-346: This should go in the methodology section. Moreover, classification metrics with more than two classes should be better discussed, as they are not as straightforward.

Response: Thank you for this suggestion. We decided not to relocate the accuracy metrics results to the methodology section however, we have added the following explanation to the methodology:

Line 298: We performed a hyperparameter tuning using weighted accuracy (1-3-9 weighting scheme) for subsequently (low, med and high classes), prioritizing category "high". Initially, when XGBoost was trained, it achieved a 63% test accuracy, but its confusion matrix revealed that it struggled to correctly classify the most destructive category (category 3). Since this category was of primary interest, the model was refined using weighted accuracy, emphasizing its importance. A 5-fold cross-validation with 1000 iterations was conducted, and for each cross-validation, oversampling was applied to balance the dataset.

Line 389: The final tuned models achieved weighted accuracies between 52% and 58%, but significantly improved recall (71%), precision (73%), and F1-score (72%) for category "high". This means that out of 34 actual instances of the highest category, 24 were correctly predicted, and out of 33 predicted cases, 24 were accurate, confirming that the model effectively focused on the most critical category. This suggests that while the overall accuracy slightly decreased due to the re-weighting, the model's performance in identifying the most critical cases significantly improved.

25. Comment: Figure 7: The F score, which should go from 0 to 1, has values above 3000. Please correct the legend.

Response: The label "F SCORE" in this case does not represent the F1-score from classification metrics but rather the number of times a feature was used in splits. A more appropriate label for the x-axis would be: "Feature Importance (Number of Splits in XGBoost)". We modified the following in the manuscript:

Line 378: In this section, we present a variable importance comparison (Figure 8) based on the Feature Importance Score (F-score) in XGBoost. XGBoost provides F-score based on how frequently a feature is used in splitting the data across all decision trees. This is the number of times a feature appears in a split across all trees in the model. A higher value indicates that the feature was used more frequently in decision-making, suggesting it has a stronger influence on model predictions. The F-score indicated that population (Pop) was the most important variable, which was consistent with our expectation in the sense that the socioeconomic impact depends largely on the exposure. The climate variable (CI) happened to be the next important variable, showing the significance of the region's climate on the socioeconomic impact of flood occurrences.

26. Comment: Table 2: "a classification model" is too generic. Please specify that this is for your trained model, applied to the test dataset (or validation? This is not clear).

Response: We agree this should be clarified. We will specify in the table caption that these metrics refer to test dataset.

27. Comment: Line 366: This seems to indicate overfitting. Did you use a validation dataset to limit this?

Response: The reviewer is correct that this could be the case. For the overall validation, we compared the results with the DFO to highlight the actual quality of the proposed model.

28. Comment: Section 3.4: It may be valuable to add a correlation plot to understand if there is a match between DFO and LYI, rather than relying on a table that contains redundant information.

Response: We appreciate the suggestion, however it is not possible to create a real correlation measurement, because our system predicts classes of LYI (low med high), and there is no reference LYI at the watershed scale for the whole HMA. Hence why we opted for the proposed analysis, where we use the DFO reported losses as a proxy of the impact of measured floods.

Reviewer 3

We appreciate the reviewer's comments. We would like to highlight that all the raised comments follow exactly the public review, by Jakob F. Steiner, https://doi.org/10.5194/nhess-2023-120-RC1, which were extensively addressed in the revised submission and for which we provided a detailed response. We believe that this review was AI generated, and we respectfully disagree with this specific comment of rejection, as it was not pertaining to the revised manuscript submitted.

To ensure a thorough evaluation, we have conducted a detailed comparison between this review and our previous responses, demonstrating that these points were extensively addressed in our revised submission. Given this overlap, we believe the review may not fully reflect the updates and revisions incorporated into the manuscript. Therefore, we respectfully disagree with the recommendation for rejection, as it does not appear to consider the revised version of our work.

New Review Comments and Responses	Jakob F. Steiner Comments https://doi.org/10.5194/nhess-2023-120-RC1 and submitted Responses
1. Data Handling and Transparency:	
• *Unsubstantiated Data Scarcity Claims:* The manuscript repeatedly claims that the HMA region is "data-scarce," but it provides no specific justification for this assertion. Data from countries such as China, India, Pakistan, and Nepal is available, albeit not as openly accessible as in Europe or the United States. The blanket claim that the region lacks sufficient data without discussing what is missing or inaccessible is misleading.	**Previous Comment:** The lack of appreciation of existing data and simply depicting the target region as 'data scarce' to avoid scrutiny from what is known already. ***Response:*** *We thank the reviewer for this comment. Please note that our intent was not to avoid scrutiny of the data, but we understand that some of the statements were phrased in an ambiguous way, which we have revised to address accordingly. Line 53-58: Gathering data for the scale of the HMA region is a difficult task, as it requires collecting data from several countries and multiple sources, and this poses challenges due to the possible inhomogeneities of standards between different organizations. Especially in the context of the impact of floods using socioeconomic data, the analysis involves examining the number of fatalities, injured and people otherwise affected, as well as the financial damage that natural disasters cause, and this information is generally not always available, or it is collected at the local scale based on reported events.*
• *Insufficient Documentation of Data Sources:* The sources of key data, particularly socioeconomic and population data, are not well-documented. For example, census data from Nepal is mentioned, but there is no detailed explanation of how it was applied to watershed-scale flood modeling, which is critical for reproducibility. The use of knoema.com, a potentially unstable and non-reputable data source, further weakens the credibility of the research. It is crucial for scientific rigor that data sources be transparent, traceable, and stable, and that the uncertainties or limitations of these data be fully addressed.	**Previous Comment:** There is very poor documentation on where the exposure data is taken from and there is no way to make this traceable (no stable links, and also no attempt so far to make your own produced data available). ***Response:*** *We thank the reviewer for this comment. We will include a table with information on all the datasets used. Regarding the links being "not stable" – the data required for the index were accessed and we tested the links before submission. In the revised paper, we will clarify the date of the latest access so that the data is more clearly referenced.* ***Note:*** *We have in fact removed the knoema.com from the previous version of the manuscript.*
• *Socioeconomic Data Processing:* The spatial scaling of district-level socioeconomic data to the watershed level is a critical methodological step that is poorly explained. How was this data aggregated or distributed? What were the assumptions or limitations involved in using this data at the watershed scale?	**Previous Comment:** I also fail to see how you take census data to the watershed and how you align using Nepal government data with your approach to model at the watershed scale. ***Response:*** *We will clarify this in the manuscript. For Nepal, we considered a weighted spatial join between the watersheds and the districts. To each watershed, we attributed the statistics of the district intersecting it, weighted by the overlapping areas.*
2. Methodological Issues:	

• *Unclear Scope and Flood Types:* The paper does not clearly specify the types of floods being modeled. The introduction mentions multiple types of flooding, but the methodology seems focused on fluvial floods. This inconsistency in the types of flood hazards being assessed undermines the clarity and focus of the study. If the study is limited to fluvial flooding, this should be clearly stated in both the abstract and methods, and the introduction should avoid discussing other types unless they are directly relevant. • *Geographical Scope and Upscaling:* The model is trained on data from Nepal but then applied to the entire HMA region. However, the manuscript does not provide sufficient justification for this upscaling. HMA includes diverse climatic regions, ranging from arid areas in Central Asia to monsoon-dominated zones in the southern Himalayas. A model that works well in Nepal may not generalize to other parts of HMA without further validation. It would be more scientifically sound to either focus solely on Nepal or provide validation for other regions in HMA.	**Previous Comment:** At multiple points of the manuscript I was a bit confused on the scope. There is an introduction on all types of high flow events but the methods suggest you only look at fluvial floods with exceptionally high impacts. There is a relatively rapid investigation of the methods for watersheds that do lie to some part in Nepal compared against data only from areas within Nepal and then an upscaling to all of HMA, which in turn is not clearly defined in its scope or climatologies. I would strongly suggest to maybe limit the study to areas where data is available before scaling it up, allowing you more space for methodological and data based issues. ***Response:*** *We thank the reviewer for this suggestion. In general, this paper focuses on fluvial and pluvial flooding, and we will make this clearer in the introduction. Starting from this, for this work, we considered Nepal as our train site for two main reasons.* 1. *We had information at fine resolution regarding the flood events, in terms of the number of people, economic impact of the event, date of the event, and population data.* 2. *For Nepal, at the time of this paper we had access to the high-resolution 8m DEM from the previous NASA HIMAT effort. This DEM also covers other areas of the wider Himat region, but it presents some gaps. Nepal was completely covered, and we verified the homogeneity and quality of the data.* *The climatology in HMA is indeed variable. In Nepal, as well, we have regional climate variations largely being a function of elevation. For this work, for the main rainfall driver of the model, we focused on climate concentration. This index was proven to be highly linked to pluvial/fluvial flooding impacts in other regions of the world, including for example Italy (both in mountainous landscapes and floodplains (Sofia et al. 2019), the US (Saki et al. 2023), or China* (Du et al., 2023). *Climate concentration values are mostly related to the temporal variability of the rainfall, not to the total amount or the average yearly and seasonal statistics, and its variability captures well various climates (Monjo et al. 2016). For Nepal, as we showed in the paper, we have a gradient of CI values, and as ML models learn from the data they ingest, we believe the system can work across various regions from the climatic point of view. We will add comments on this in the paper, highlighting the strengths and weaknesses of the approach.*
3. Model Validation and Use of ERA5 Data:	

• *Validation Challenges:* The manuscript relies on ERA5 precipitation data for flood simulation in HMA, but ERA5 has well-documented limitations in representing precipitation in mountainous regions. Precipitation in these areas is highly variable, and ERA5's coarse spatial resolution may not adequately capture localized rainfall events. The authors should either provide a more thorough discussion of these limitations or supplement ERA5 data with region-specific datasets (e.g., from local meteorological agencies) to improve validation.	**Previous Comment:** While I understand that it would be well beyond the scope of this study to evaluate the suitability of ERA5 data for flood simulations (let alone in a mountain context where precipitation products are of poor quality) but it would be crucial to address this and dispel concerns from the get go by referring to discussions of this data in mountain regions as well as for flood mapping. ***Response:*** *This study is a part of the HiMAT project. There are a number of research groups working on different aspects of HMA. At the time we conducted this study, a subgroup of our team was working with ERA5. We wanted to utilize the available dataset and complement the existing study. We will add few comments on this, with highlights on HMA from other related works from the HiMAT team, such as* (Maggioni & Massari, 2018; Maina et al., 2023)
• *Model Validation with Flood Data:* The manuscript lacks sufficient validation of the ML model's predictive power. The use of the Dartmouth Flood Observatory (DFO) database, which focuses primarily on lowland floods, may not be the best choice for validating a model intended to predict flood impacts in mountainous regions. Additionally, the number of flood events recorded in the DFO database for Nepal is limited, making it questionable whether this provides robust validation for the entire HMA region.	**Previous Comment:** Apart from the Brakenridge citation not having a date nor being present anywhere in the references, and agreeing that in principle such a dataset would be an interesting set for validation, the fact that the whole dataset only has 46 events from Nepal since 2021 and <10 with the 1000 deaths plus displaced criterium you introduce below makes its use questionable considering this is the area you run your model in. Wouldn't data from Nepal (like https://bipadportal.gov.np/) be much more appropriate then? ***Response:*** *We have done our study for both Nepal and HMA from 1980-2020. To our best understanding, the dataset from emdat and DFO have the longest and most detailed series of point datasets for different events for the time period we are interested in. We appreciate the separate data source that you shared with us. Please note that we trained our model considering information for NEPAL from http://www.drrportal.gov.np/ which includes flood/heavy rain/flash flood events for all districts in Nepal from different sources. This database includes more than 46 events reported in the DFO. At the scale of HMA, there is no other available dataset reporting flood impacts, aside from DFO and EMDAT, to our knowledge, hence we considered these two, with their limits, to highlight how our model could help target priority areas where indeed events have happened, of a large impact, as highlighted by actual floods reported in these two independent datasets.* **Previous Comment:** Also this database captures lowland floods, rather than mountain floods, making me wonder whether the aim to characterize 'High Mountain Asia' floods is really the right scope here. Also the DFO reports single coordinates, are you then simply assuming the watershed that matches the coordinate is the only one

	affected? Likely the reported numbers refer to much larger areas, as the size of the watershed you chose is rather small (guessing from the Figure, it's not actually described anywhere!) ***Response:*** *The dataset is not only capturing lowland floods but also mountainous river floods that are characterized as fluvial floods. Also, the damage dataset can only be "point" data at a particular location. There may be one, many, or no point for the whole watershed. As we have described previously, we have used GIS techniques to distribute the damages for the watersheds. This is a common technique that is used widely.* *We will add information on the size of the watershed. (e.g., range of the watersheds' area)*
4. Presentation and Writing Quality:	
• *Grammar, Formatting, and Citations:* The manuscript contains numerous grammatical errors, incomplete references, and formatting issues. Several citations are listed as "n.d." or are missing from the reference list entirely. These issues, while minor in isolation, collectively reduce the professionalism of the manuscript and suggest a lack of attention to detail.	**Previous Comment:** *In numerous instances references are reported as 'n.d.' where they actually have a date and some are completely missing from the reference list.* **Response:** We will make sure to do a thorough check of the manuscript and correct these mistakes.
• *Figures and Captions:* The figures in the manuscript, particularly Figure 9, are inadequately explained. Captions are vague and do not provide sufficient information for readers to interpret the figures independently of the text.	**Previous Comment:** *Figure 9: Panel a is elevation not rainfall as your legend suggests!* **Response:** There was a mistake in the legend... We will correct this.
5. Lack of Novelty in Results:	
• *Expected Findings:* Much of what is presented in the results seems to reflect known trends rather than novel insights. For example, the model's prediction that flood impacts are higher in areas with higher population density is unsurprising and does not provide new knowledge. The authors should focus on demonstrating how their ML model offers unique predictive power or novel findings that go beyond confirming expected patterns. • *Limited Discussion of Results' Implications:* The discussion of the results lacks depth, particularly in terms	**Previous Comment:** A main concern I have here is that I am still not very clear on where the observed events come from yu compare this to. I am also wondering if your Figure 8 simply only confirms one thing – that there are many people (an input to your model) where there are many people (a validation of your model). How does your model compare on actually coming up with an observed flood from the input 'ERA5 rain'? This concern then propagates into the result for the whole region,whereyou 'predict'the biggestimpacts with the highest population densities. That isn't quite so surprising and it is unclear to me how I can see the power of ML in these results. To be provocative, would the results have been different if you would have just distributed rainfall across the watersheds without a model in between? ***Response:*** *We thank the reviewer for his comment. We revised the manuscript thoroughly and we believe it is now clearer. Please note that we trained the model and validated it only using the*

of practical applications. How can the findings be used for real-world flood mitigation or policymaking? What specific new insights into flood risk do these results offer that were not already known?	*data for Nepal, at the district scale and then at the watershed scale. Overall, we opted for a 90-10 approach, for which 90% of the Nepal data were used for training and 10% for validation. Line 360 and following: Comparing predicted Lifeyears Index (LYI) flood impacts with observed data showed good correspondence between high-risk areas identified by the ML method and historical flood locations in Nepal. This suggests that the proposed approach effectively delineates flood risk on a national scale. Figure 8 illustrates this comparison, showcasing observed (empirically evaluated) and ML-predicted LYI values at both watershed (upper row) and district (lower row) levels. The 'observed' LYI values were empirically calculated from observational data (Table 1) and categorized into three groups: 'low', 'medium', or 'high', with basins/districts labeled as 'high' for LYI values exceeding 1000 years, 'medium' between 100 and 1000 years, and 'low' below 10 years. The 'predicted' values represent the outputs from the machine learning model. In Nepal, we achieved an overall training accuracy of 97% and a test accuracy of 63%. Notably, training the model at the watershed level yielded higher accuracy compared to the district level. This is attributed to watersheds being hydrologic units that integrate geomorphological and climatic properties, thus providing a more accurate representation of flood dynamics compared to administrative district boundaries. At the watershed level, nearly all year ranges exhibited a 100% match with observed impacts. In instances where the model's accuracy fell below 100% (e.g., 1985–90 and 1990–95), the LYI values in the affected watersheds were low, indicating that the predictors considered were more indicative of major flooding events. The superior accuracy achieved at the watershed level underscores the value of implementing the model at this scale when scaling up the system.*