NHESS-2023-120

Predictive Understanding of Socioeconomic Flood Impact in Data-Scarce Regions Based on Channel Properties and Storm Characteristics: Application in High Mountain Asia (HMA)

The authors present a machine learning approach to describe flood impact at the regional scale for individual watershed, based on the example of High Mountain Asia. They do so with a simple XGBoosting approach and reproducing life years lost per watershed.

General comments

The study has merit in its outline, as a better understanding of flood impacts at this scale is definitely needed and I agree that doing this with a generally simple approach for a general kick start to the options of AI in this domain makes sense. However, I have a number of concerns on how the study is built and executed, which I believe are at this stage too big to recommend the manuscript for publication. I detail the concerns below and would encourage the authors to rethink their strategy before moving to an eventual submission. I fully understand that this is a submission from an ECR and I want to complement you on the aim and pulling this together – definitely work that should be pursued and there is a lot of demand for outcomes of such approaches! I would have hoped to see more scrutiny here before a submission from the more experienced co-author team.

My concerns range from (a) general sloppiness of manuscript writing (many simple editing mistakes that can always happen for drafts but should not occur for a submitted manuscript over (b) the lack of appreciation of existing data and simply depicting the target region as 'data scarce' to avoid scrutiny from what is known already to (c) a lack of proper documentation of data sources on the exposure side as well at times confusing jumping between topical (what types of floods) as well as spatial (national, watershed, HMA wise) domains. I briefly summarize these concerns below and then present a list of line indexed responses for the complete manuscript.

a) Sloppy mistakes
   In numerous instances references are reported as 'n.d.' where they actually have a date and some are completely missing from the reference list. There are many instances with missing spaces as well and figure captions are often incomplete. Please be careful on such matters before submitting
b) General statement on 'no data'
   You make general unsupported statements on the region being data scarce on hydromet data. That is decidedly not the case. While data may often not be readily accessible, it is available and many studies have been published on this, especially for China and India and data is generally reachable from China, India, Pakistan, Afghanistan as well as Central Asian states. The data that is available you do away with as 'not trustworthy' in a single sentence. This coming from an all-US based author team is problematic and I guess you could imagine how stunned a reviewer from the US (or Europe) would be if a Chinese author would make that claim before proceeding to apply ML on all of the US or Europe. You will need to make a clearer description (with references) on what is lacking and how your approach fills that gap.
c) Poor documentation of socio-economic data
   As I detail below 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, see comment on Availability statement). 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 (which do not follow national borders).

d) General scope and methodology
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.

Specific comments:

There are multiple citations as 'n.d.' while actually they are published and have a year – please check your references carefully.

L31f: Be careful in your framing – population growth does not increasing likelihood of flooding, it increases flood risk! Also, in the abstract and your general analysis, you focus on precipitation as a flood driver but here then passingly mention glacial melt as well – those are very different flood drivers and would be crucial to be clear what kind of flooding you wish to tackle here.

L54: '*HMA does not have enough hydrological stations for region-wide flood monitoring*' is a huge statement to make without a citation – what is an appropriate number? Also most countries in HMA, especially China, India, Pakistan and Nepal have large and dense network of hydro(-met) monitoring, which they also use for forecasting. That is not as open as in the US, but the statement that there is 'not enough' needs to be qualified. You then claim '*Moreover, the available meteorological datasets may not be sufficiently trustworthy.*', which again lacks any qualification. Imagine me making that statement for a European or North American country, that would be thrown out. The region has a large amount of met data (see e.g. the overview figure in (Nepal et al. 2023)) and if you do not trust the data you need to justify why.

L61: '*The use of remote sensing technology for disaster studies in HMA is comparatively new*' – I also do not quite agree. Remote sensing itself isn't very old and it has been used in HMA for many studies already (which maybe anyway would need some acknowledgement here).

L87: You focus here a lot on monsoon changes with intense precipitation – but if you actually focus on HMA (rather than just the Hindukush Himalaya) there are a lot of other processes – Westerlies in Central Asia, Eastern Monsoons in the Upper Yangtze etc. Maybe it is required to reconsider the total spatial scope of the study here?

L92: You now finally get to actual numbers of potential affected, but leave it to the reader to get the data from EMDAT. It would be prudent to explain here (or rather in the introduction) what the actual numbers are and for what types of hazards, to then narrow down and which ones you actually focus.

Figure 1: Up to this point there was no clear description how the watersheds are selected, i.e. what boundary you used for HMA. This needs to be provided to give context to why so many watersheds outside HMA are also included.

L116: At this point you mention that you will predict impacts of 'floods', i.e. all of them? The way you describe your research you are narrowing this down on pluvial floods, as glacial lake outburst floods or debris flows etc need very different driver analysis. Can you be precise here? In L185 you then suddenly just focus on 'fluvial flooding', so is it just that you focus on?

L120ff: This part is crucial as you present the socioeconomic data and how you treat it. However there are a few issues that would need to be addressed with respect to traceability and presentation of data used.

- You refer to data sources that are questionable, the knoema.com page is not stable and it is unclear from where their data is sourced or where it is known needs to be documented here.
- You refer to general government and Worldbank websites (like http://drrportal.gov.np/) that exist but what data you took from there at what point in time remains unclear. Copernicus Journals subscribe to FAIR practices, that includes the documentation of third party data used in a publication.
- You introduce a lot of data as well as parameters from literature (like T and e) without any questioning of their accuracy, uncertainty etc. This would propagate and need to be addressed, especially as you seem to upscale from this approach with a few numbers on Nepal government websites to all of HMA.
- You calculate these values for Nepal as a whole but then work on the watershed scale – how is this compatible?
-

Figure 3: I am not sure whether these are now LYl only due to floods or all disasters. Considering that there are no jumps for earthquake events like 2015, I assume this has been calculated for floods only? Then this needs to be made very clear in the caption rather than just calling it 'disasters'.

L176 + Figure 4: What is HAND in Figure 4. Is this from (Delalay et al. 2018)? The publication is not open access and only limited to Sindupalchowk, how does it go to all of Nepal? What does it actually map?

L194f: 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.

L210: As for the other socioeconomic data above, the description of population data here remains lacking. For Nepal you only refer to the Census Bureau, which does not report distributed data or data by watershed (so how was that brought in line with inundation maps) and you also do not specify where on the general page you retrieved the data from. You then refer to the GHSL but do not provide a citation or link where this data was retrieved. Distributed data in Asia is generally of problematic and definitely not homeogenous quality, hence a discussion of how this was dealt with need a much more thorough description than the short paragraph here without any references. A detail but you also call it LYI (capital I) here while it should be LYl (lower case L)!

L227ff: You discuss your first results here on the F score and model performance discussion – this should come under Results and Discussion respectively, not Methods! Figure 6 as well as Table 1 also lacks a description of variables and results presented. Unclear how this should be interpreted.

L248f: 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? 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!)

L261: You include a crucial boundary condition of your model here, i.e. '1000 deaths plus displaced'. Does this mean your model will only be useful in this domain? It would be crucial to report how many such events have actually happened in your domain then. Also how is the adding up of 'dead and displaced' justified? These are quite 'different' responses to a flood.

Figure 9: Panel a is elevation not rainfall as your legend suggests!

L265ff: To be honest I am not entirely sure how I should interpret Figure 7 – doesn't it just confirm that people live close to wide river channels? Then there is really no link to atmospheric characteristics as you claim in L270. There is a lot of discussion already as well on convection patterns all stemming from other literature and not really relevant to what I read in the Figure.

L295f: A main concern I have here is that I am still not very clear on where the observed events come from you 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, where you 'predict' the biggest impacts 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?

L349F: The figures you note here do not show what is described in the text.

L356f: I lack some context here - <10% of watersheds see an increase, are all other stable or see a decrease? How can you differentiate here between hazard (rain) and exposure (population) as a driver of change? How do you explain that increase has slowed after 2010 significantly? And how is it possible that in the 1995-2010 jump the number of increasing watersheds is similar to the just 5 year jump between 1990 – 1995? Isn't that completely counterintuitive?

L406: While in general 'an intention to make data available' shouldn't be followed, for a journal like NHESS this is definitely not acceptable. Data availability needs to be clearly described (or arhued why this is not the case).

Technical corrections (Minor issues):

L14: 'from flooding and debris flows'

L33: missing space

L58: 'is foremost'?

L82: This is not correct, HMA includes the Tibetan Plateau, Tien Shan, Pamir, Qilian Shan etc etc, while the 'Hindukush-Himalaya', as the name suggests only comprises the southern fringe of HMA. HMA furthermore includes Myanmar.

L85: Missing citation of the population number – also does that include all of the above countries or just areas above a certain elevation?

L88: 'located in this region'

L95: 'HMAT'?

Figure 2: Issue in caption after 'section 2.2.1'

L177: Spacing issues like here are found throughout the manuscript and need to be carefully checked before submission!

References:

Delalay, Marie, Alan D. Ziegler, Mandira Singh Shrestha, Robert James Wasson, Karen Sudmeier-Rieux, Brian G. McAdoo, and Ishaan Kochhar. 2018. "Towards Improved Flood Disaster Governance in Nepal: A Case Study in Sindhupalchok District." *International Journal of Disaster Risk Reduction* 31 (October): 354–66. https://doi.org/10.1016/j.ijdrr.2018.05.025.

Nepal, S, J F Steiner, S Allen, F M Azam, S Bhuchar, H Biemans, M P Dhakal, et al. 2023. "Consequences of Cryospheric Change for Water Resources and Hazards in the Hindu Kush Himalaya." In *Water, Ice, Society, and Ecosystems in the Hindu Kush Himalaya: An Outlook*. Kathmandu, Nepal: ICIMOD.