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 manuscript presents a machine learning-based approach to predict socioeconomic flood impacts in the High Mountain Asia (HMA) region, utilizing channel properties, storm characteristics, and socioeconomic indicators. The study addresses a pressing issue of flood vulnerability in data-scarce regions by combining geomorphic and climatic data to estimate flood risk. The introduction of a Lifeyears Index (LYI) to assess long-term impacts adds an innovative dimension to the study. However, while the topic is relevant and timely, the manuscript suffers from significant methodological, data handling, validation, and presentation flaws, which diminish its overall scientific contribution.

Strengths:

The focus on flood risk in High Mountain Asia (HMA), a region highly vulnerable to hydro-meteorological extremes, is highly relevant in light of climate change and its associated impacts on vulnerable populations. Floods in HMA are a significant concern, particularly given the region's complex topography, glacial melt dynamics, and varying monsoonal patterns.

The paper attempts to incorporate machine learning (ML), specifically XGBoost, into flood risk modeling. While ML is becoming increasingly common in environmental sciences, its application in flood impact assessment for data-scarce regions remains a novel area of research. The methodology has the potential to offer more flexibility in prediction compared to traditional models.

The use of the LYI metric is commendable, as it provides a more comprehensive way to assess both the immediate and long-term socioeconomic impacts of floods. This metric is a valuable addition, especially as it incorporates the broader social cost of disasters, similar to the Disability Adjusted Life Years (DALYs) used in health studies.

Major Weaknesses:

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.
- **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.
- **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?

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 (e.g., pluvial, fluvial, glacial lake outburst floods), 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.

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.

- **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.

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.

- **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. For example, Figure 9(a) is mislabeled as “rainfall” when it appears to depict elevation. Each figure should be carefully checked for accuracy, and captions should be expanded to clarify what the figure represents and how it relates to the findings.

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 of practical applications. How can the findings be used for real-world flood mitigation or policy-making? What specific new insights into flood risk do these results offer that were not already known?

Recommendations for Improvement:

1. **Clarify and Focus the Scope**:
    - Clearly specify the types of floods being modeled (e.g., fluvial only) and ensure that the methodology aligns with this scope. Avoid mentioning other flood types unless they are directly relevant.
    - If the study is focused on Nepal, limit the geographical scope accordingly. Alternatively, provide validation for the model's application across the broader HMA region.
2. **Improve Data Transparency and Documentation**:
    - Provide clear documentation of all data sources, including stable links, proper citations, and explanations of how the data were processed. Address the limitations of using certain datasets and provide a more balanced discussion of data availability in HMA.
    - Explain in detail how socioeconomic and population data were integrated into the model, particularly how district-level data was applied at the watershed scale.
3. **Strengthen Model Validation**:
    - Provide a more robust validation of the model using appropriate regional datasets. Supplement ERA5 data with local meteorological data where possible, and address its limitations for flood modeling in mountainous regions.
    - Reconsider the use of the DFO database for model validation. Instead, use more region-specific flood data, particularly for high-altitude areas, to better assess the model's predictive accuracy.
4. **Enhance Writing and Figure Presentation**:
    - Revise the manuscript to eliminate grammatical errors, correct formatting, and complete missing references. The manuscript should be proofread thoroughly before resubmission.
    - Improve figure captions to ensure that all visual data are clearly explained and relevant to the text. Ensure that all figures are correctly labeled and accurately represent the data being discussed.
5. **Increase the Novelty and Practical Implications of the Findings**:
    - Emphasize any novel findings from the model, particularly any unexpected results or new insights into flood risk in HMA. Highlight how the machine learning approach provides advantages over traditional models.
    - Provide more discussion on how the results can be applied in real-world flood mitigation efforts, disaster management, or policy-making in the HMA region.

Final Decision: Rejection

While the topic is highly relevant and the use of machine learning has potential, the manuscript in its current form is not ready for publication. The significant issues with data handling, methodological clarity, and model validation undermine the credibility of the findings. Additionally, the paper lacks sufficient novelty and depth in its results, and the writing and presentation need substantial improvement.

Given these foundational problems, I recommend **rejection**. However, I encourage the authors to address the major issues highlighted in this review. With significant revisions, including better data transparency, clearer scope, and stronger validation, the paper could be reconsidered for submission to a future issue of the journal.

Note to the Special Issue Editors: The paper aligns with the theme of hydro-meteorological hazards and socioeconomic vulnerability. However, due to the significant methodological and presentation flaws, it does not meet the standards for publication in this special issue at this time. The authors should be encouraged to revise their work, especially focusing on addressing the issues raised regarding data transparency, model validation, and scope clarity, before resubmitting to the journal or special issue.