nature portfolio

Peer Review File

Deep learning completes US flood hazard maps revealing millions exposed to previously unrecognized risk

Corresponding Author: Dr Rudi Stouffs

This file contains all reviewer reports in order by version, followed by all author rebuttals in order by version.

Version 0:

Reviewer comments:

Reviewer #1

(Remarks to the Author)
This paper presents a generative AI method to infer unmapped areas in FEMA's NFHL. It highlights the importance of applying AI to fill gaps in NFHL coverage and to analyze the potential risks and socioeconomic factors associated with these unmapped regions. The study is scientifically sound and of generally good quality. I recommend a "Major revision" at this stage.
1. "In contrast to FEMA's localized mapping efforts, computational flood models are able to simulate flood risks through stochastic processes and produce continuous flood maps, even globally." Do you actually mean "computational flood models" or "non-computational flood models"? Based on my understanding, it should be "non-computational flood models." If it truly refers to "computational flood models," this statement conflicts with the content in the Results section.
2. A limitation of the FEMA model is that it only represents the 100-year flood inundation area. Consequently, your AI model can only infer the 100-year flood inundation extent as well. This should be clearly stated in your discussion section.
3. Since the FEMA program began long ago, many of the latest inundation maps in certain areas have not been updated for over a decade. With extreme weather events occurring more frequently, many agencies and engineering firms now use different hydrological conditions (e.g., rainfall or upstream flow rates) than those applied 10–20 years ago. This can lead to boundary condition disparities in FEMA maps, which will also be reflected in your training data. How was this issue addressed? I did not see any discussion of this point in the current version.
4. Since your method divides the entire block into sub-context tile sets, the water stage may be inconsistent at the connection points between tiles. How did you address this issue?
5. It would be great if you could provide detailed neural network architecture for local enhancer and global generator.
6. The "model architecture" section is too general, and I would strongly suspect that readers can replicate the process.
7. It would be great if you could provide at least the full name of TP, FP, TN, FP.
8. All the demonstration figures show flood inundation maps in mountainous areas, where slopes are generally quite steep. AI-modeled flood maps typically achieve higher accuracy in such regions compared to very flat terrain. I suggest the authors include demonstrations for low-relief, plain areas as well.
9. From Figure 10, the AI model appears to perform very poorly in certain areas (IOU < 0.2), and the median accuracy metric is not high. The illustrated cases from Michigan, Texas, and New Jersey would be considered only "acceptable" by most industry standards. I suggest that the authors discuss both the extreme underperforming cases and the overall performance, and clearly state that the current proposed AI-modeled results do not achieve high accuracy.
10. Several statements in the manuscript appear to overstate the performance of the proposed AI model. To be honest, the current error levels may not meet FEMA's standard requirements.

(Remarks on code availability)
The code is not available before publication. I believe the current method section is not detailed enough to replicate the results.

Reviewer #2

(Remarks to the Author)
General Comments:
Overall, this study addresses an important and timely topic; expanding flood hazard mapping into unmapped areas using AI methods. However, there are several critical gaps that should be addressed before this work is suitable for publication. First,

the introduction does not engage with the large and growing literature on the social dimensions of flood hazard vulnerability in areas unmapped by FEMA. There is an opportunity to situate your study within this body of work and discuss how your findings compare to prior research. Second, some statements in the introduction (e.g., regarding paywalls, model accessibility, and model accuracy) require clarification or more precise framing to avoid misinterpretation. Third, methodological details are lacking in key areas, such as how FEMA data gaps were determined, whether pluvial flooding is accounted for, and how coastal versus inland flooding differences are treated. Finally, the discussion section would be strengthened by explicitly comparing your model validation, exposure estimates, and socio-demographic findings to other similar studies. These changes would help clarify the novelty and implications of your work and ensure it is accurately positioned in the broader research landscape.

More specific comments:
In the introduction, the authors missed an opportunity to discuss the growing literature on the social dimensions of flood hazard vulnerability in areas unmapped by FEMA. For example, see Pricope et al., 2022; Smiley et al., 2020; Flores et al., 2023, 2025; Huang et al., 2020, Clark et al., 2025, Yu et al., 2025. The authors even state that in recent years, new flood models and research has revealed additional insights on the spatial and social equity within the U.S. But they only cite one Wing et al., paper and don't even discuss the findings of that work. I would strongly suggest that the authors do a more comprehensive literature review and discuss what other research has found, especially in terms of the social equity dimensions.
1. Pricope, N. G., Hidalgo, C., Pippin, J. S., & Evans, J. M. (2022). Shifting landscapes of risk: Quantifying pluvial flood vulnerability beyond the regulated floodplain. Journal of Environmental Management, 304, 114221.
2. Smiley, K. T. (2020). Social inequalities in flooding inside and outside of floodplains during Hurricane Harvey. Environmental Research Letters, 15(9), 0940b3.
3. Flores, A., Collins, T., Grineski, S., Amodeo, M., Porter, J., Sampson, C., & Wing, O. (2025). Federally-overlooked flood risk inequities in the conterminous United States. Scientific reports, 15(1), 10678.
4. Flores, A. B., Collins, T. W., Grineski, S. E., Amodeo, M., Porter, J. R., Sampson, C. C., & Wing, O. (2023). Federally overlooked flood risk inequities in Houston, Texas: novel insights based on dasymetric mapping and state-of-the-art flood modeling. Annals of the American Association of Geographers, 113(1), 240-260.
5. Huang, X., & Wang, C. (2020). Estimates of exposure to the 100-year floods in the conterminous United States using national building footprints. International Journal of Disaster Risk Reduction, 50, 101731.
6. Clark, A. S., Collins, T., Grineski, S., Brewer, S., & Flores, A. (2025). Comparative assessment of residential property values at risk to flooding: The case of Utah, USA. International Journal of Disaster Risk Reduction, 118, 105247.
7. Yu, Y., Flores, A., Connor, D., Meerow, S., Braswell, A. E., & Leyk, S. (2025). Flood risk and the built environment: big property data for environmental justice and social vulnerability analysis. Population and Environment, 47(1), 14.

Introduction: There is no evidence that continuous flood models exacerbate information inequity; although there is a large literature related to this for air pollution (e.g., sensor deserts).

To my knowledge, the Bates et al flood models (that the authors seem to be speaking of, although it's not totally clear), are accessible to academic institutions via data sharing agreements, and are only locked behind paywalls for commercial users. Still, that doesn't address the issue of accessibility to the public, however, the claim about being locked behind paywalls is not totally accurate and could be misconstrued.

Results
Authors say that pixel precision is up to 81%; Wing et al and Bates et al find similar validation values when validating against FEMA; yet the authors state in the introduction that these studies still have a lot of room for improvement. I would rephrase accordingly or at least explain how this accuracy compares to other flood models validated against FEMA maps.

They grayed out census tracts: do they have no FEMA data at all or do some of these areas indeed have FEMA flood maps that just aren't digitized and available via the National Flood Hazard Layer. How was this accounted for in the analysis? If this was not accounted for, then it would be inaccurate to say that these areas have no FEMA data at all.

Figure 3: the smaller maps are impossible to read.

Group 4 results: does this account for differences in coastal flooding vs. inland flooding? Typically, research finds that FEMA coastal zones are more likely to be population by higher income groups because of the associated environmental amenities.

While the results are interesting and so is the methodology, especially the use of AI, how does this analysis account for pluvial flooding, if at all? If not, then it's not really improving the NFHL either, which is one of the major criticisms of FEMA data.

There is zero mention of how results in terms of validation, exposure numbers, and socio-demographics align with other studies who have done similar research. This is a major oversight in my opinion.

(Remarks on code availability)

Version 1:

Reviewer comments:

Reviewer #1

(Remarks to the Author)
The authors addressed all my comments and concerns, therefore, I would suggest "accept" at this stage.

(Remarks on code availability)
I had some simple tests on the code provided. It functions well in my test.

Reviewer #2

(Remarks to the Author)
Section 3.3 Limitations and Future Works: the authors partially addressed my comment about the limitations of FEMA data, but don't explicitly state that FEMA data do not account for pluvial flooding which is an emerging hazard due to the increases in extreme precipitation events. I think this is important to note rather than just saying it only delineates fluvial and coastal flooding.

Flores et al. also use high resolution (10-meter) with gridded population data (30-meter); very similar to what is being done here. So, the text about providing a crucial advantage in terms of granularity and verifiability is somewhat inaccurate.

Authors did a good job comparing their numbers to other studies; but still in terms of the socio-demographic findings, they only compare their results to Wing et al 2022. However, I can think of at least 5 over national level studies examining the socio-demographic characteristics of people in FEMA flood zones and they aren't cited.

(Remarks on code availability)
All code appears to be publicly assessable and reproducible. The code does provide a README file. I did not install and run the code.

Open Access This Peer Review File is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made.
In cases where reviewers are anonymous, credit should be given to 'Anonymous Referee' and the source.
The images or other third party material in this Peer Review File are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.
To view a copy of this license, visit https://creativecommons.org/licenses/by/4.0/

Authors' Response to Reviewers

Addressing the accuracy-scale-cost trilemma: Bridging inequitable access to flood information in the United States with AI generated flood maps

Nature Communications, NCOMMS-25-29998A

RC: ***Reviewers' Comment***, AR: Authors' Response, ☐ Manuscript Text

Reviewer #1

RC: ***This paper presents a generative AI method to infer unmapped areas in FEMA's NFHL. It highlights the importance of applying AI to fill gaps in NFHL coverage and to analyze the potential risks and socioeconomic factors associated with these unmapped regions. The study is scientifically sound and of generally good quality. I recommend a "Major revision" at this stage.***

AR: Thank you very much for your thorough review and valuable comments on our manuscript. We sincerely appreciate your recognition of the significance and quality of our work, as well as your support for this research. We have carefully addressed each of your points and incorporated additional details and revisions throughout the manuscript accordingly.

RC: ***"In contrast to FEMA's localized mapping efforts, computational flood models are able to simulate flood risks through stochastic processes and produce continuous flood maps, even globally." Do you actually mean "computational flood models" or "non-computational flood models"? Based on my understanding, it should be "non-computational flood models." If it truly refers to "computational flood models," this statement conflicts with the content in the Results section.***

AR: We appreciate this request for clarification. In the original manuscript, we intended "computational flood models" to mean large-scale, physics-based flood simulation models that produce continuous flood hazard maps (as opposed to FEMA's localized static maps, which might also involve certain computerized processes). These include hydrologic/hydraulic models that run simulations (for example, the continental-scale models by Wing et al. 2017 and Bates et al. 2021 in our references).

Our methodology differs from these physics-based simulations models because it is based generative AI that learns the best way to translate terrain models into flood models without the need for human intervention.

To clarify this, we made the following changes to the manuscript:

> However, similar to the high cost of acquiring FEMA flood maps, high-resolution computational models are only made possible with significant financial investments. While several of these global-scale models are available to academic researchers via data-sharing agreements, they often exist as commercial products that are not as readily accessible to the general public or integrated into local-level planning documents in the same way as the public-facing, regulatory NFHL. As such, while these data have enabled new insights on flood-related socio-economic impacts, limited access to the underlying data restricts the reproducibility of the findings and communities at risk still have no access to such

data for resilience planning (Wing et al., 2020; Shu et al., 2023).

RC: ***A limitation of the FEMA model is that it only represents the 100-year flood inundation area. Consequently, your AI model can only infer the 100-year flood inundation extent as well. This should be clearly stated in your discussion section.***

AR: We agree that this is an important limitation to state explicitly. Our model is trained on FEMA's 100-year flood zones (1% annual probability floodplains) and therefore its inferences are inherently tied to this specific return period. It does not directly provide additional information about other return periods (e.g. 50-year or 500-year floods), nor does it simulate flood depth or timing.

We have added a clear statement about this constraint in the Discussion section. Our model is designed to learn the complex patterns from the data provided. Based on the performance and scalability demonstrated in our experiments, we are confident that the model can adapt to additional return periods (e.g., 500-year) or different flood types (e.g., pluvial) if such data were provided as training targets.

This flexibility underscores the potential of our methodology as a low-cost tool to 'scale-up' or interpolate regional flood maps (e.g., from new, localized pluvial flood studies) to larger, unmapped areas. We will revise the Discussion to make this point clear.

Limitations and Future Works

A key characteristic of our generated maps to note is that they reflect the specific scenarios present in the training data. Since the FEMA NFHL primarily delineates the 100-year fluvial and coastal floodplain (1% annual exceedance probability), our model learns to infer this specific inundation extent. This distinction is important, as our approach is highly adaptable. ...Based on the performance and scalability demonstrated, we hypothesize that the model could be trained to generate maps for other return periods (e.g., 500-year) or even different flood types (e.g., pluvial) if such datasets (Pricope et al., 2022) were available for training. This opens up significant opportunities for future research, positioning our methodology as a powerful, low-cost means to 'scale-up' and interpolate specialized, regional flood studies to larger areas.

RC: ***Since the FEMA program began long ago, many of the latest inundation maps in certain areas have not been updated for over a decade. With extreme weather events occurring more frequently, many agencies and engineering firms now use different hydrological conditions (e.g., rainfall or upstream flow rates) than those applied 10–20 years ago. This can lead to boundary condition disparities in FEMA maps, which will also be reflected in your training data. How was this issue addressed? I did not see any discussion of this point in the current version.***

AR: We thank the reviewer for raising this critical issue. The reviewer is correct that the NFHL is temporally heterogeneous, containing a mix of both recently updated maps and older maps that do not align with the latest terrain data.

This heterogeneity in the training data is, in fact, a key element that demonstrates our model's robustness and generalization. The national-scale model learned from all available data pairs, including (a) high-quality, updated maps that align with current terrain, and (b) low-quality, outdated maps that do not.

By training on this massive and "noisy" dataset, the model learned to generalize the fundamental relationship between geomorphology and flood inundation, rather than simply memorizing existing (and sometimes incorrect) boundaries.

We provide direct visual evidence of this generalization capability in Figure 5b. As shown in those examples, the official FEMA flood map is abruptly cut off in some examples, likely due to an old administrative or mapping boundary. Our model, however, recognizes that the terrain features (e.g., the river valley) continue, and it correctly extends the flood zone based on the more recent 3DEP terrain information.

Therefore, while we did not "address" this with a data pre-processing step (which would be subjective and difficult), the model's ability to generalize is the solution. It learns from the "good" maps how to "correct" the "bad" ones.

We agree with the reviewer that this point was not clear, and we will add a new paragraph to the Results section (Section 2.3) to explain this characteristic of the model.

> **Learning for generalization**
>
> This observation highlights a key strength of our model: even though it was trained on a massive and "noisy" dataset containing a mix of both high-quality, updated maps and older, outdated ones, the model is able to generalize the underlying relationship between the terrain context from modern 3DEP terrain data and the FEMA flood inundation patterns. As shown in the figure, the model, having learned from correct examples elsewhere, disregards erroneous boundaries of the FEMA floodzones and can generate hydrologically consistent flood zones that follow the terrain. Thus, the model is not merely 'replicating' outdated maps elsewhere; it is a robust system that can identify and, in many cases, correct these historical or administrative inconsistencies based on current topographical data.

RC: ***Since your method divides the entire block into sub-context tile sets, the water stage may be inconsistent at the connection points between tiles. How did you address this issue?***

AR: We thank the reviewer for this important question regarding tile boundaries. We apologize for not making our approach clearer. In fact, we did not require an overlapping-tile or blending strategy. Instead, as briefly mentioned in Section 4.1, our model is trained on a dataset of tiles from multiple zoom levels (zoom 10 to 13) based on the WMTS tiling convention. This multi-scale training allows the model to learn the context of terrain features and water channels at different scales. By seeing both the fine-grained detail at high zoom and the broader context at low zoom, the model learns to generate hydrologically continuous features that seamlessly connect at the tile edges. The visual evidence in our results (e.g., Figures 5, 8, 9, 11) confirms the lack of significant edge artifacts even though all images are stitched together from individual tiles.

We will add a sentence to Section 4.1 to make this point more explicit for the reader.

Original Text:

> ...At lower zoom levels, each tile captures more information which allows the model to be more contextually aware such that the generated images can be stitched together seamlessly."

Revised Text:

> At lower zoom levels, each tile captures more information which allows the model to be more contextually aware such that the generated images can be stitched together seamlessly. This multi-scale training approach reduces the prevalence of edge artifacts; by learning the continuity of features at multiple scales, the model can generate more seamless maps without requiring a post-processing overlap or blending strategy. We found that the best training set should contain tiles from multiple zoom levels...

RC: ***It would be great if you could provide detailed neural network architecture for local enhancer and global generator.***

AR: We thank the reviewer for this suggestion. We agree that the description could be more tightly linked to the figure. The architecture for the global generator (G1) and local enhancer (G2) is described in Section 4.2, paragraphs 2 and 3, and each layer is visually detailed in Figure 7. To remove any ambiguity, we have corrected a typo in the text that likely caused the confusion (the text stated 512x512 resolution, while the figure correctly shows 1024x1024). We have updated the text to align with the figure, which should make the existing description clearer.

Original Text: "Our model is capable of generating high-resolution images at 512x512 pixel resolution."

Revised Text:

> Our model is capable of generating high-resolution images at 1024x1024 pixel resolution, as detailed in Figure 7.

RC: ***The "model architecture" section is too general, and I would strongly suspect that readers can replicate the process.***

AR: We agree with the reviewer that replicability is paramount. The most effective way to ensure this is by making our code fully available. We have updated Section 5 (Data and Code Access) to provide clear, separate links to the two repositories: one for reproducing the national analysis, and one containing the model code for training and inference.

> All datasets used in this study are publicly available from their original sources (as referenced in the text). To ensure full replicability of our method and findings, the complete source code is available in two public repositories. The code for the national-scale flood analysis, which reproduces the figures and statistics in this paper, is available at [https://github.com/Iceofsky/national-flood-analysis]. The code for training and inferring new flood maps with our generative model is available at [https://github.com/Iceofsky/FloodMapper].

RC: ***It would be great if you could provide at least the full name of TP, FP, TN, FP.***

AR: Our apologies for the typo in the draft (which should be FN) and for not defining these standard terms in the manuscript. We will correct this.

Proposed Revision (Methods Section 4.3):

> Given the unsupervised nature of the model, the selected metrics are not optimized during training. Rather, these numbers offer us insight into how the generated flood regions compare to the ground truth in different geographic contexts. The recall and precision are calculated from the components of the confusion matrix:
>
> - True Positives (TP) (pixels correctly classified as flood),
> - False Positives (FP) (pixels incorrectly classified as flood),
> - True Negatives (TN) (pixels correctly classified as non-flood), and

- False Negatives (FN) (pixels incorrectly classified as non-flood)

RC: ***All the demonstration figures show flood inundation maps in mountainous areas, where slopes are generally quite steep. AI-modeled flood maps typically achieve higher accuracy in such regions compared to very flat terrain. I suggest the authors include demonstrations for low-relief, plain areas as well.***

AR: We thank the reviewer for this important observation, which highlights a key point of clarification. The reviewer is correct that the visuals in Figures 8, 9, and 11 appear to show high-contrast, steep terrain. However, this is an feature of our data augmentation, and does not represent the degree of slopes.

As we state in Section 4.1, "Additional visual enhancements, such as color gradients and hill shades, are applied to the context tiles to help with model training...". This hillshade enhancement, while crucial for helping the model learn, creates a high-contrast 'embossed' look that exaggerates small elevation shifts.

In fact, our study areas (such as the sample in Shiawassee County Michigan in Figure 10) and case studies (like Grand Rapids in Figure 11) are predominantly low-relief or rolling-hill regions, not mountainous. To prevent this misinterpretation for future readers, we will add a clarifying sentence to Section 4.1 where we first introduce these visual enhancements.

Revision (Methodology Section 4.1, Paragraph 1, at the end):

Original Text:

> ...Additional visual enhancements, such as color gradients and hill shades, are applied to the context tiles to help with model training and metric computation.

Revised Text: "

> Section 4.1 Additional visual enhancements, such as color gradients and hill shades, are applied to the context tiles to help with model training and metric computation. It is important to note that these enhancements, particularly the hill shades, create a high-contrast 'embossed' effect (e.g., in Figures 8, 9, 11) that visually exaggerates small shifts in elevation and does not mean that the absolute elevation of the terrain is significant.

RC: ***From Figure 10, the AI model appears to perform very poorly in certain areas (IOU < 0.2), and the median accuracy metric is not high. The illustrated cases from Michigan, Texas, and New Jersey would be considered only "acceptable" by most industry standards. I suggest that the authors discuss both the extreme underperforming cases and the overall performance, and clearly state that the current proposed AI-modeled results do not achieve high accuracy.***

AR: Thank you for the candid feedback. A balanced discussion of the model's performance variability is essential.

The quantitative metrics reported in the experiments must be interpreted with nuance, as the "ground truth" NFHL data is itself heterogeneous and contains outdated or erroneous boundaries (as shown in Fig 5b). Like what we have discussed above, the model learns the generalizable patterns between flood zone delineation and the given terrain context which means that a low IOU score can result from the model correcting an error in the Ground Truth instead. This is further tested with Grand Rapids, MI in Figure 11 where even though the metric came lowest among the three test regions (mIOU at 0.38), Figure 11 demonstrates that the model can still extrapolate flood patterns seamlessly into large, unmapped areas. This highlights the limitation of IOU as the sole measure of success.

We have revised the end of Section 4.4 (Model performance) to provide this crucial context.

Original Text:

> "However, it is to be noted that this might also be caused by... [and the final paragraph] ...improving the model's performance further."

Revised Text: "

> The mIoU scores indicate overlap between the predicted and ground truth flood regions. We see that the model's performance in Michigan has the lowest performance of 0.38, with the lower end reaching 0.0 in extreme cases. This indicates that there might be a high proportion of false negatives and false positives compared to the FEMA ground truth.
>
> However, this variability must be interpreted with nuance. First, as discussed in Section 4.3, our generative model is not explicitly optimizing for pixel-level IOU, but rather learning the underlying relationship between terrain and inundation. Second, as shown in Section 2.3, the 'ground truth' NFHL data is often outdated, so some discrepancies occur where the model 'corrects' erroneous administrative boundaries or outdated flood channels based on the new terrain information (Figure 5), leading to a low IOU score despite a more hydrologically plausible result.
>
> This discrepancy between quantitative metrics and practical utility is best demonstrated by the Michigan case study. As shown in Figure 13, the model is still highly effective at the primary goal of the study: generalizing and seamlessly extrapolating flood patterns into large, previously unmapped areas like Grand Rapids. In the figure, we observe that the results generated for the study area seamlessly extrapolate from the mapped flood regions to the unmapped surroundings of Grand Rapids, MI. This study proves that, in cases where a large portion of the data is missing in a remote region, we can still train models in a data-rich region and extrapolate in a data-poor region with a certain degree of confidence. The closer the similarity between the regions, the better the performance of the synthetic results.

RC: ***Several statements in the manuscript appear to overstate the performance of the proposed AI model. To be honest, the current error levels may not meet FEMA's standard requirements.***

AR: We agree that some of our language may be overstated. Our model is not intended to meet FEMA's regulatory standards, but rather to serve as low-cost, yet effective informational tool to fill gaps and help stakeholders identify potential risk in unmapped areas to plan for resilience starting *today*. To ensure this nuance is carried through the entire paper, we performed a thorough review of the manuscript—particularly the Title, Abstract, Introduction, and Conclusion—to moderate our claims.

Proposed Revision 1 (Title):

Original Title:

> Resolving the accuracy-scale-cost trilemma: Bridging inequitable access to flood information in the United States with AI generated flood maps

Revised Title:

Addressing the accuracy-scale-cost trilemma: Bridging inequitable access to flood information in the United States with AI generated flood maps

Proposed Revision 2 (Abstract):

Original Text:

...generates flood maps at 30 meters resolution in the unmapped areas with up to 81% and 65% in national scale studies.

Revised Text:

...with an average precision of 81% and recall of 65% when evaluated against existing FEMA maps in selected study areas.

Proposed Revision 3 (Discussion, Section 3, new paragraph at end):

This opens up significant opportunities for future research, ... That being said, it is important to contextualize the role of these AI-generated maps. They can serve as an public guide to reveal flood-prone areas in unmapped regions and are should not be intended as a direct replacement for official, regulatory FEMA flood zones. Instead, they serve as a crucial, large-scale risk information tool. Their primary utility is to quantify exposure and direct limited mapping resources to the most vulnerable unmapped communities, allowing them to begin resilience planning *today*.

More importantly, we also want to use this opportunity to include an extended Model Validation section that explains the trade-offs between higher accuracy versus better generalization of the model. Therefore, we have expanded the Methods section for a in-depth exploration of the performance of our model in different test areas. We have included a comparison of model results between the national model used in the analysis of our paper versus 3 locally trained models. The locally trained models obtains visibility higher performance across all metrics but are unable to perform as well when applied to far away context. The finding suggest that for localize interpolation of incomplete flood maps, training of a specialized model on local geomorphology is superior than a national, more generalist model.

Please see below for an excerpt from the newly edited Methods section. The full edit with the newly added tables is available in the submitted manuscript.

4.5 Model Validation

4.5.1 Model excels at mapping unmapped regions

.... Table 2 shows the performance of the models in the 'In-sample Performance' columns. In this experiment, the test regions are selected within the training boundary, which is a good proxy to how the model can infer unmapped areas that are nested within a larger, mapped area. ...

4.5.2 Model adapts to new contexts

... In addition to extending the NFHL flood maps to areas with *similar* terrains, the greater challenge of the model is to extrapolate the learned parameters to *dissimilar* (Out-of-sample) regions. We validate the performance by training the models with partial data within each state and deploying the models to regions *out* of the training boundary with different geographical context. ....

4.5.3 Performance tradeoffs between national and local models

.... In the previous section, we observed that models trained only on data from a single region (local models) suffered performance degradation when extrapolating to other regions. The degradation increases as the test regions become more different from the trained regions which means that these 'locally trained' models would struggle when extrapolating in completely new environments (e.g. mapping flood regions in California using a model trained in Michigan). We also observed that the higher training samples in Texas provided additional robustness in generalizing their prediction in unseen regions. Thus, to create the most robust model for nation-wide analysis, we train a single national model on all available FEMA data, totaling 300,000 images, covering all regions that are mapped in the NFHL dataset.

... Table 3 compares the performance of the national model with the 3 'local models' on the same test regions. From the table, we see that the national model, trained on all possible combinations in the NFHL, outperforms the local models with performance gains over the local models in all regions and for all metrics. New Jersey experienced the highest performance gain with an increase of +0.16 in precision, and other regions mainly experienced various degrees of performance gain.

.... To conclude, the flood maps produced by the AI model show similarities to the ground truth FEMA flood maps, and this characteristic allows anyone to reproduce a continuous flood map that is consistent with the FEMA methodology to certain degree of accuracy for the first time. Depending on the scale of analysis, the trade off between performance versus generalizability should be considered. From the examples and the comparison with the local models above, we conclude that the 'general' model trained on all available FEMA data is best suited for nation-wide analyses, having better performance when compared to out-of-sample regions in the local models. Thus, the Results section of this paper utilizes the output of the 'general' model for all national demographic and impact studies. That being said, a locally trained model would be more suitable for regional studies in the future.

RC: ***(Remarks on code availability) The code is not available before publication. I believe the current method section is not detailed enough to replicate the results.***

AR: We understand the importance of code availability for transparency and reproducibility. In the revision, we have updated the manuscript to link to our open access FloodMapper repository with guides on data preparation, training, and inference, along with sample data, and a series of jupyter notebooks that reproduce the figures and results in the paper.

All datasets used in this study are publicly available from their original sources (as referenced in the text). To ensure full replicability of our method and findings, the complete source code is available in two public repositories. The code for the national-scale flood analysis, which reproduces the figures and statistics in this paper, is available at [https://github.com/Iceofsky/national-flood-analysis]. The code for training and inferring new flood maps with our generative model is available at [https://github.com/Iceofsky/FloodMapper].

Reviewer #2

RC: ***General Comments: Overall, this study addresses an important and timely topic; expanding flood hazard mapping into unmapped areas using AI methods. However, there are several critical gaps that should be***

addressed before this work is suitable for publication. First, the introduction does not engage with the large and growing literature on the social dimensions of flood hazard vulnerability in areas unmapped by FEMA. There is an opportunity to situate your study within this body of work and discuss how your findings compare to prior research. Second, some statements in the introduction (e.g., regarding paywalls, model accessibility, and model accuracy) require clarification or more precise framing to avoid misinterpretation. Third, methodological details are lacking in key areas, such as how FEMA data gaps were determined, whether pluvial flooding is accounted for, and how coastal versus inland flooding differences are treated. Finally, the discussion section would be strengthened by explicitly comparing your model validation, exposure estimates, and socio-demographic findings to other similar studies. These changes would help clarify the novelty and implications of your work and ensure it is accurately positioned in the broader research landscape.

AR: Thank you very much for your thorough review and valuable comments on our manuscript. We sincerely appreciate your recognition of the significance and quality of our work, as well as your support for this research. We have carefully addressed each of your points and incorporated additional details and revisions throughout the manuscript accordingly.

RC: ***In the introduction, the authors missed an opportunity to discuss the growing literature on the social dimensions of flood hazard vulnerability in areas unmapped by FEMA. For example, see Pricope et al., 2022; Smiley et al., 2020; Flores et al., 2023, 2025; Huang et al., 2020, Clark et al., 2025, Yu et al., 2025. The authors even state that in recent years, new flood models and research has revealed additional insights on the spatial and social equity within the U.S. But they only cite one Wing et al., paper and don't even discuss the findings of that work. I would strongly suggest that the authors do a more comprehensive literature review and discuss what other research has found, especially in terms of the social equity dimensions.***

AR: Thank you for providing this highly relevant list of citations, which has fundamentally improved the framing of our paper.

We acknowledge that our original introduction failed to engage with the large and critical body of literature on the social equity dimensions of "federally-overlooked" flood risk. We have performed a major revision of the Introduction based on a comprehensive review of the suggested literature. Our new introduction now reframes the NFHL gap as "systematic, inequitable blind spot" that stems from a combination of high costs, administrative boundaries, and the omission of certain hazard types. In addition, we use the cited literature to define the "federally-overlooked" risk and its massive scale (affecting 26 million people) , and to detail the "compounding vulnerability" (social, structural, and financial) that falls on vulnerable communities.

> The incompleteness of the FEMA flood map .... This is not just a random gap, but a systematic blind spot. Recent research has defined this as "federally-overlooked" flood risk (Flores et al., 2023), which is estimated to affect 26 million people nationally (Flores et al., 2025). Without the mandatory public insurance in the unmapped areas, ... This observation is confirmed at a national scale, where studies consistently link lower income to higher risk in these unmapped zones (Flores et al., 2025). Event-based analyses of disasters like Hurricane Harvey found that racial inequalities in damage were "primarily driven by" impacts *outside* the official floodplains, often in Black and Hispanic neighborhoods (Smiley, 2020). .... This creates a "double jeopardy" (Flores et al., 2023), as these overlooked, at-risk structures are often "of substandard quality, poorly maintained, and uninsured" (Yu et al., 2025), representing billions in unprotected financial liability with minimal insurance coverage (Clark et al., 2025).

RC: ***Introduction: There is no evidence that continuous flood models exacerbate information inequity; although***

there is a large literature related to this for air pollution (e.g., sensor deserts).

AR: Yes, there are currently no evidence that inaccessible flood models exacerbate information inequity. It was an inferred consequence. The paragraph that speculate the consequence has been deleted.

RC: ***To my knowledge, the Bates et al flood models (that the authors seem to be speaking of, although it's not totally clear), are accessible to academic institutions via data sharing agreements, and are only locked behind paywalls for commercial users. Still, that doesn't address the issue of accessibility to the public, however, the claim about being locked behind paywalls is not totally accurate and could be misconstrued.***

AR: Yes, we have highlighted in the new edit that large computational models are in fact available to researchers. We acknowledge that advanced hydrodynamic models (like Wing et al., 2022) are what enable this research, and our work is complementary and could serve as a means to democratize access outside of academia.

> While several of these global-scale models are available to academic researchers via data-sharing agreements, they often exist as commercial products that are not as readily accessible to the general public or integrated into local-level planning documents in the same way as the public-facing, regulatory NFHL.

RC: ***Results Authors say that pixel precision is up to 81%; Wing et al and Bates et al find similar validation values when validating against FEMA; yet the authors state in the introduction that these studies still have a lot of room for improvement. I would rephrase accordingly or at least explain how this accuracy compares to other flood models validated against FEMA maps.***

AR: We apologise for the confusion. You are right to point out this apparent contradiction. Our original phrasing in the introduction was imprecise and created confusion, for which we apologize. The "room for improvement" we mentioned was not intended to be about validation metrics, but about the "accuracy-scale-cost trilemma" as a whole.

While invaluable, many of those models are computationally expensive (cost) and are not designed to seamlessly fill gaps in the existing public regulatory framework (scale/applicability). Our Generative AI approach is presented as a novel method to address this specific gap: providing a low-cost, scalable tool to supplement the existing NFHL.

To fix this, we have taken two actions:

We have deleted the problematic sentence from the introduction that stated these other models "still have a lot of room for improvement" regarding their validation metrics. This steers our narrative and clarifies that our contribution is not a more accurate model, but a novel, low-cost, and scalable methodology to interpolate missing areas in the FEMA flood zones.

RC: ***They grayed out census tracts: do they have no FEMA data at all or do some of these areas indeed have FEMA flood maps that just aren't digitized and available via the National Flood Hazard Layer. How was this accounted for in the analysis? If this was not accounted for, then it would be inaccurate to say that these areas have no FEMA data at all.***

AR: Based on our version of the NFHL database dated October 2023, the census tracts that are grayed out have no flood maps. We have changed the text and also the diagram to refer to grayed out areas as tracts with no flood zone information in the NFHL database.

RC: ***Figure 3: the smaller maps are impossible to read.***

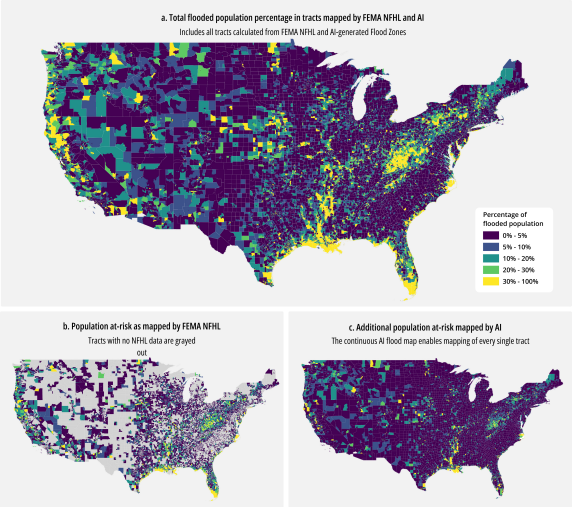

Figure 1: Additional populations at-risk can be highlighted by the AI-generated flood map, resulting in a comprehensive assessment for every single tract in CONUS.

AR: Thank you for the feedback. Based on edits to the text in the previous comment, we have also updated the diagram for more visibility and added descriptive texts to guide readers.

> Given the high resolution of the generated map, we can aggregate the H3 hexagons to the Census Tract level and calculate the percentage of exposed population for every tract to understand how the AI flood map adds to completing the existing FEMA flood database. Figure 3 presents a series of maps that visualize the population in flood zones as a percentage of the total population of each census tract. This allows us to understand how much the AI generated maps can contribute to completing the missing flood information.
>
> Figure 3b presents the current status quo, showing the percentage of at-risk population as mapped by the FEMA ground truth flood maps. Census Tracts with *no* FEMA flood maps accessible from the NFHL database are grayed out. In this map, we see the commonly flooded regions, such as areas around West Virginia, Louisiana, and Florida, having a larger proportion of populations at risk.
>
> Figure 3c shows the additional flooded population revealed by the AI in all census tracts. Due to the continuous nature of the AI flood zones, we see that the AI flood map has revealed additional population at risk, both in tracts already mapped by FEMA and also in tracts that are overlooked by FEMA.
>
> Adding the data from Figures 3b and 3c together, we obtain Figure 3a, which shows the total percentage of the population at risk in each census tract. This map depicts a much more comprehensive assessment of the exposed population in the country as it *combines* the flood zones mapped by FEMA and the additional flood zones mapped by the AI model.

RC: ***Group 4 results: does this account for differences in coastal flooding vs. inland flooding? Typically, research finds that FEMA coastal zones are more likely to be population by higher income groups because of the associated environmental amenities.***

AR: This is an insightful question. The reviewer is correct that our model does not differentiate between coastal and

inland/fluvial flood zones. The FEMA NFHL contains both, so our model learns patterns from a composite of all available zones.

However, our findings for "Group 4" (Completed by NFHL, the red CBSAs) do not support the hypothesis that this group is skewed by high-income, coastal-amenity populations. As shown in our manuscript (Figure 4d), the CBSAs in Group 4 are not geographically concentrated in the high-income coastal regions the reviewer describes. Instead, our analysis shows that these cities are predominantly micropolitan areas .

Furthermore, our socioeconomic analysis found this group has significantly lower median household income, the lowest income growth, and a higher Gini index compared to other groups. Looking at the map, these Group 4 areas are more likely cities with smaller populations (making them easier for FEMA to map completely) or are in regions that are genuinely not as flood-prone (meaning our AI did not uncover significant additional flood zones because there aren't any).

RC: ***While the results are interesting and so is the methodology, especially the use of AI, how does this analysis account for pluvial flooding, if at all? If not, then it's not really improving the NFHL either, which is one of the major criticisms of FEMA data.***

AR: This is a crucial point. The reviewer is correct: our model does not account for pluvial (rainfall-driven) flooding.

This is because our model is trained on the FEMA NFHL, which itself largely omits this specific risk. This is a constraint we inherit directly from the available national-scale training data, not a fundamental limitation of the AI methodology itself.

Therefore, we must clarify our contribution. Our model "improves" the NFHL not by adding new types of flood physics (like pluvial), but by extending its current focus (fluvial/coastal) to fill the vast, unmapped spatial gaps. It fixes the spatial incompleteness, not the underlying methodological ones.

We argue this highlights a key advantage of our approach: its adaptability. Future work could leverage this same generative method to "scale-up" and interpolate more localized, high-fidelity pluvial flood models, should that data be available for training. We have added this clarification to our Discussion section to make this limitation and its context explicit.

> **Limitations and Future Works**
>
> A key characteristic of our generated maps to note is that they reflect the specific scenarios present in the training data. Since the FEMA NFHL primarily delineates the 100-year fluvial and coastal floodplain (1% annual exceedance probability), our model learns to infer this specific inundation extent. We emphasize that this is a constraint of the input dataset, not a limitation of the generative methodology itself.
>
> This distinction is important, as our approach is highly adaptable. ...Based on the performance and scalability demonstrated, we hypothesize that the model could be trained to generate maps for other return periods (e.g., 500-year) if such datasets were available for training. This opens up significant opportunities for future research, positioning our methodology as a powerful, low-cost means to 'scale-up' and interpolate specialized, regional flood studies to larger areas.

RC: ***There is zero mention of how results in terms of validation, exposure numbers, and socio-demographics align with other studies who have done similar research. This is a major oversight in my opinion.***

We agree completely, and we thank the reviewer for this critique. We have added a new, dedicated section

to our Discussion (titled "4.X Comparison with National-Scale Flood Exposure Studies") to perform this comparison and contextualize our findings.

In this new section, we: 1. Compare Exposure Numbers: We directly compare our finding of 11 million unmapped people to the 26 million "federally-overlooked" residents from Flores et al. (2025). We explain that our number is a conservative baseline (as it's trained on the NFHL's fluvial/coastal data) , but that its real value is in its granularity and verifiability. 2. Highlight Granularity: We make the case that our 11 million figure is derived from a 30-meter pixel-level analysis , not census-tract aggregation . This is what allows us to flag specific buildings , which is a more actionable insight. 3. Emphasize Accessibility: We frame our open-source methodology as a direct solution to the inaccessibility of the "closed-off" computational models , making our 11 million estimate a verifiable, public-facing contribution. 4. Align Socio-Demographic Findings: We show how our own socio-demographic results align with the nuanced findings of Wing et al. (2022)

Comparison with National-Scale Flood Exposure Studies

Our results, when placed in the context of the recent literature, provide a complementary view on the scale and nature of unmapped flood risk. Our analysis identified 11.04 million people and 4.1 million buildings exposed to flood risks in areas unmapped by the NFHL. This exposure number can be compared to the 26 million residents estimated to live in "federally-overlooked" 100-year floodplains by Flores et al. (2025).

The differences highlight the distinct contributions of each methodology. First, the Flores et al. (2025) estimate is derived from advanced hydrodynamic models that explicitly include pluvial risk (Flores et al., 2025), whereas our model was trained on the NFHL, which largely omits this (Pricope et al., 2022). Our 11 million figure should therefore be interpreted as a conservative, national-scale baseline of the unmapped *fluvial and coastal* risk.

Second, our estimate provides a crucial advantage in **granularity and verifiability**. Our 11 million figure is derived from a direct, 30-meter pixel-level analysis overlaid with gridded population data. This high-resolution approach, in contrast to analyses aggregated at the census-tract level, is what enables our study to match every building to the highly detailed flood map and quantify risk for specific assets and areas.

This distinction gets to the core of our paper's "accuracy-scale-cost trilemma". While the advanced hydrodynamic models are invaluable for scientific research, their results are often not publicly accessible or verifiable . Our methodology, by contrast, is fully open-source . Our 11 million estimate is therefore a publicly-verifiable, high-resolution assessment of the gaps in the *current regulatory map*. We welcome future work comparing our open dataset directly against these high-fidelity, closed-access models.

Finally, our socio-demographic findings also align with this body of research. Our analysis of over 900 CBSAs revealed that cities with more complete FEMA mapping (Groups 3 and 4) have significantly lower median household incomes . This result is consistent with the national-scale findings of Wing et al. (2022), who found that *present-day* flood risk is disproportionately concentrated in poorer, whiter communities (reflecting legacy development patterns) (Wing et al., 2022). Our finding suggests that FEMA's prioritization efforts, despite their well-documented spatial gaps, have been partially effective in first mapping these known, economically vulnerable areas.

References

O. E. J. Wing, N. Pinter, P. D. Bates, C. Kousky, New insights into US flood vulnerability revealed from flood insurance big data, Nature Communications 11 (2020) 1444. Publisher: Nature Publishing Group.

E. G. Shu, J. R. Porter, M. E. Hauer, S. Sandoval Olascoaga, J. Gourevitch, B. Wilson, M. Pope, D. Melecio-Vazquez, E. Kearns, Integrating climate change induced flood risk into future population projections, Nature Communications 14 (2023) 7870. Publisher: Nature Publishing Group.

N. G. Pricope, C. Hidalgo, J. S. Pippin, J. M. Evans, Shifting landscapes of risk: Quantifying pluvial flood vulnerability beyond the regulated floodplain, Journal of Environmental Management 304 (2022) 114221.

A. B. Flores, T. W. Collins, S. E. Grineski, M. Amodeo, J. R. Porter, C. C. Sampson, O. Wing, Federally overlooked flood risk inequities in Houston, Texas: Novel insights based on dasymetric mapping and state-of-the-art flood modeling, Annals of the American Association of Geographers 113 (2023) 240–260.

A. Flores, T. Collins, S. Grineski, M. Amodeo, J. Porter, C. C. Sampson, O. Wing, Federally-overlooked flood risk inequities in the conterminous United States, Scientific Reports 15 (2025) 10678.

K. T. Smiley, Social inequalities in flooding inside and outside of floodplains during hurricane harvey, Environmental Research Letters 15 (2020) 0940b3.

Y. Yu, A. Flores, D. Connor, S. Meerow, A. E. Braswell, S. Leyk, Flood risk and the built environment: big property data for environmental justice and social vulnerability analysis, Population and Environment 47 (2025) 14.

A. S. Clark, T. Collins, S. Grineski, S. Brewer, A. Flores, Comparative assessment of residential property values at risk to flooding: The case of Utah, USA, International Journal of Disaster Risk Reduction 118 (2025) 105247.

O. E. J. Wing, W. Lehman, P. D. Bates, C. C. Sampson, N. Quinn, A. M. Smith, J. C. Neal, J. R. Porter, C. Kousky, Inequitable patterns of US flood risk in the Anthropocene, Nature Climate Change 12 (2022) 156–162. Publisher: Nature Publishing Group.

Point-by-Point Response to Reviewers

We sincerely thank the reviewers for their positive feedback and constructive comments. Their insights have helped us refine our manuscript, particularly in clarifying the model's limitations, the specific advantages of our open-source approach, and the broader socio-demographic context of our findings. Below is our point-by-point response detailing how we have addressed each comment in the revised manuscript.

Reviewer #1

Reviewer #1 (Remarks to the Author): *The authors addressed all my comments and concerns, therefore, I would suggest "accept" at this stage.*

Response: We thank the reviewer for their careful reading of our manuscript and are grateful for the suggestion to accept.

Reviewer #1 (Remarks on code availability): *I had some simple tests on the code provided. It functions well in my test.*

Response: We appreciate the reviewer taking the time to test our code and confirm its functionality and reproducibility.

Reviewer #2

Reviewer #2 (Remarks to the Author): *Section 3.3 Limitations and Future Works: the authors partially addressed my comment about the limitations of FEMA data, but don't explicitly state that FEMA data do not account for pluvial flooding which is an emerging hazard due to the increases in extreme precipitation events. I think this is important to note rather than just saying it only delineates fluvial and coastal flooding.*

Response: We agree with the reviewer that explicitly mentioning the omission of pluvial flooding is critical, given its increasing importance due to climate change. We have revised Section 3.3 (Limitations and Future Works) to explicitly state that the FEMA NFHL does not account for rainfall-driven pluvial flooding, and thus our model (which is trained on FEMA data) inherently shares this limitation.

Updates in the manuscript (Section 3.3):

"Because the FEMA NFHL primarily delineates the 100-year fluvial and coastal floodplain, our maps do not account for pluvial (rainfall-driven) flooding, an increasingly critical hazard due to climate change."

Reviewer #2: *Flores et al. also use high resolution (10-meter) with gridded population data (30-meter); very similar to what is being done here. So, the text about providing a crucial advantage in terms of granularity and verifiability is somewhat inaccurate.*

Response: We thank the reviewer for pointing this out. We completely agree that the models used by Flores et al. have comparable or even superior spatial granularity. We have adjusted the manuscript in Section 3.2 to remove the claim of having a granularity advantage. Instead, we now emphasize that our distinct contribution lies in **transparency and verifiability** via our fully open-source approach, contrasting it with advanced hydrodynamic models whose high-resolution outputs are often closed-access or difficult to verify publicly.

Updates in the manuscript (Section 3.2):

"Our approach offers distinct advantages in transparency and verifiability, directly addressing the accuracy-scale-cost trilemma in flood mapping... While advanced hydrodynamic models are invaluable for scientific research, their high-resolution results are often closed-access and not easily verifiable. In contrast, our 30-meter, building-level methodology is fully open-source and reproducible, providing a publicly-accessible assessment of gaps in the current regulatory map."

Reviewer #2: *Authors did a good job comparing their numbers to other studies; but still in terms of the socio-demographic findings, they only compare their results to Wing et al 2022. However, I can think of at least 5 over national level studies examining the socio-demographic characteristics of people in FEMA flood zones and they aren't cited.*

Response: We thank the reviewer for highlighting the need for a broader literature comparison. We agree that our findings should be contextualized within the larger body of national-level socio-demographic flood research. We have substantially expanded Section 3.2 to include citations and comparisons to several key national-level studies, including Qiang (2019), Tate et al. (2021), Flores et al. (2025), and Wing et al. (2020).

Updates in the manuscript (Section 3.2):

"Finally, our finding that median household incomes are lower in cities with more complete FEMA mapping aligns with the broader socio-demographic literature. Studies by Wing et al. (2022), Qiang (2019), and Tate et al. (2021) consistently demonstrate that flood risk is disproportionately concentrated in economically disadvantaged and socially vulnerable communities. Flores et al. (2025) and Wing et al. (2020) further corroborate these vulnerability patterns across different flood types and historical claims. Collectively, these studies support our conclusion that FEMA's mapping prioritization, despite its spatial gaps, has been partially effective in initially targeting economically vulnerable areas."

Reviewer #2 (Remarks on code availability): *All code appears to be publicly assessable and reproducible. The code does provide a README file. I did not install and run the code.*

Response: We appreciate the reviewer verifying the public accessibility and documentation of our code repository.