Reviewer 1

Reviewer #1: This study presents a semi-supervised segmentation framework based on satellite images and partially labeled data to improve the detection of small and informal settlements in Metropolitan Lima, which is often missed by global datasets. Results show that the city expanded by about 76 km² between 2016 and 2025, with a significant share of new development occurring in areas exposed to tsunami, landslide, and seismic hazards, highlighting growing risk in hazard-prone zones. Overall, this study is well-designed and comprehensive, and the findings are meaningful for risk-informed and resilient urban management. However, I still have several comments and suggestions for improving the current work.

Main Reply:

We truly thank the reviewer for the very deep comments provided to us. In the following, we address each comment individually. For each reviewer's remark, we present two sections: (i) our response to the comment and (ii) a description of the corresponding modifications made to the manuscript. All changes introduced in response to the reviewer's comments are highlighted in blue in the revised manuscript.

Comment 01:

Lines 17-18: It is suggested to make it clear that the correlation refers to Spearman's correlation coefficient, and present the corresponding p-value that indicates its statistical significance.

Answer:

The information the reviewer is referring to comes from a United Nations Human Settlements Programme's (UN-Habitat) report, which is a policy synthesis report. It summarizes relationships, but unfortunately does not present a full statistical methodology. Thus, it does not mention whether the correlation is Pearson or Spearman. The correlations are presented as descriptive indicators supporting the narrative.

Additionally, we corrected a typo in the originally reported coefficients. The positive correlation previously stated as 0.82 has been corrected to 0.42, and the negative correlation previously reported as −0.64 has been corrected to −0.53, in accordance with the values presented in the source report.

Changes in manuscript:

Lines 17-21

ural and climate-related hazards (UN-Habitat, 2024). Today, more than one billion people live in informal settlements—nearly one-quarter of the global urban population—where access to basic services, secure land tenure, resilient infrastructure, and political representation remains limited. Using the Notre Dame Global Adaptation Initiative's (ND-GAIN) country index (Chen et al., 2023), UN-Habitat (2024) reported that, at the country level, higher shares of informal settlements are associated with greater climate vulnerability ($r = 0.42$) and lower climate readiness ($r = -0.53$). These findings highlight that the most excluded populations are simultaneously the most exposed and the least equipped to adapt or recover from environmental disasters.

Comment 02:

Lines 41-42: What is the difference between the proposed semi-supervised segmentation framework and those in the literature?

Answer

To clarify the contribution of this study with respect to previous semi-supervised segmentation approaches, we revised the Introduction section. In the updated manuscript, we now report several existing semi-supervised learning strategies.

We clarify that the proposed study does not introduce a new segmentation architecture. Instead, the contribution lies in the adaptation and integration of PU-based learning strategies for the specific problem of mapping urban growth in Metropolitan Lima.

Unlike several existing semi-supervised segmentation approaches that rely on subsets of images with dense pixel-level annotations, the proposed framework exploits partially labeled samples and a closed boundary to generate large quantities of non-urban data surrounding the urban fringe. Besides, the framework constrains the range of plausible positive prior values using spatial information from the study area.

Changes in manuscript:

Lines 42-56

Despite these advances, most existing approaches rely on fully supervised learning, which demands extensive and highly accurate training data. However, in many regions—such as Peru—reliable datasets are limited or incomplete, constraining the effectiveness of machine learning-based urban footprint mapping (Kuffer et al., 2016). To overcome these limitations, recent studies have explored semi-supervised learning approaches (Patel et al., 2021; Saha et al., 2021; Shi et al., 2022; Yu et al., 2023). Semantic segmentation requires dense pixel-level annotations, which are time-consuming and expensive to produce. Semi-supervised learning addresses this limitation by exploiting both labeled and unlabeled data during model calibration. Several strategies have been proposed in the literature. Consistency regularization approaches employ teacher–student frameworks to enforce stable predictions under perturbed inputs (Patel et al., 2021; Guo et al., 2024). Other methods use pretrained models to generate reliable pseudo-labels from unlabeled samples and progressively expand the training dataset (Wang et al., 2022). These approaches typically require a subset of images with dense pixel-level annotations together with additional unlabeled samples. Positive-unlabeled (PU) learning has also been explored for binary classification problems where only positive samples are available (Li et al., 2021). Unlike previous approaches, PU learning does not require densely annotated negative samples (De Comité et al., 1999; Letouzey et al., 2000; Kiryo et al., 2017). However, its performance strongly depends on the estimation of the positive prior probability. To promote the development of semi-supervised learning techniques, Castillo-Navarro et al. (2022) introduced the MiniFrance suite, a large dataset designed for semi-supervised remote sensing applications. However, the dataset is composed of urban and rural environments from France, which differ substantially from the fragmented and rapidly evolving urban patterns commonly observed in cities of the Global South, such as Lima.

Lines 78-83:

In this study, we present a practical semi-supervised framework based on positive-unlabeled (PU) learning to improve urban footprint mapping along the boundaries of Metropolitan Lima, Peru. The approach integrates open-source datasets, including Sentinel-2 imagery, OpenStreetMap (OSM), and publicly available governmental data, to generate partially labeled training samples. A closed-boundary strategy is introduced to obtain large quantities of non-urban samples surrounding the urban fringe, while the range of plausible positive prior values is constrained using spatial information from the study area. The proposed methodology is scalable and transferable for monitoring urban expansion in data-scarce regions. The remainder of this paper is organized as follows. Section 2 describes the proposed approach, Section 3 presents the experimental results for Metropolitan

Comment 03:

Line 45: “SAR” stands for Synthetic Aperture Radar? It is better to use the full name for its first appearance in the manuscript.

Answer

The reviewer is right, we have written the full term “Synthetic Aperture Radar (SAR)” at its first occurrence.

Changes in manuscript:

Line 58:

Other efforts to map settlements on a global scale have been made. The World Settlement Footprint (WSF) products provide 10-m resolution built-up extents using optical and Synthetic Aperture Radar (SAR) data (German Aerospace Center, 2023), while the Global Human Settlement Layer offers built-up surfaces derived from Sentinel-2 imagery (Pesaresi et al., 2024). Furthermore, Microsoft’s Global Building Footprints deliver building-level maps at continental scales. Such datasets are essen-

Comment 04:

For figures with maps, it is suggested to add “N” to the north arrow, and add labels and units like “longitude (°)” and “latitude (°)” to the axes.

Answer

We have updated all map figures to include the letter “N”, included coordinates, and specified the units in the captions.

Changes in manuscript:

Figura 01:

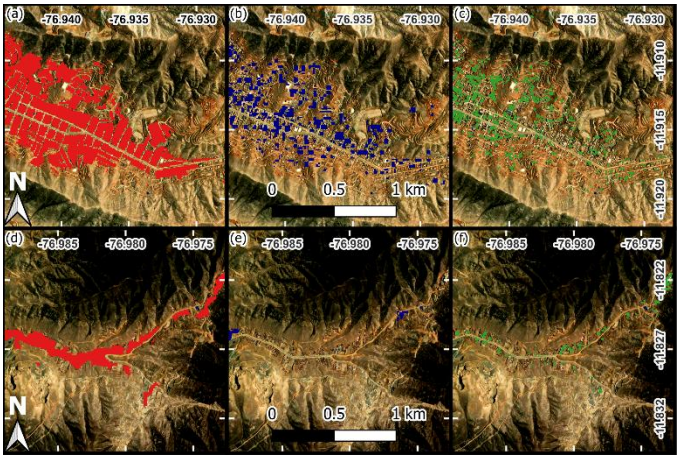


Figura 07:

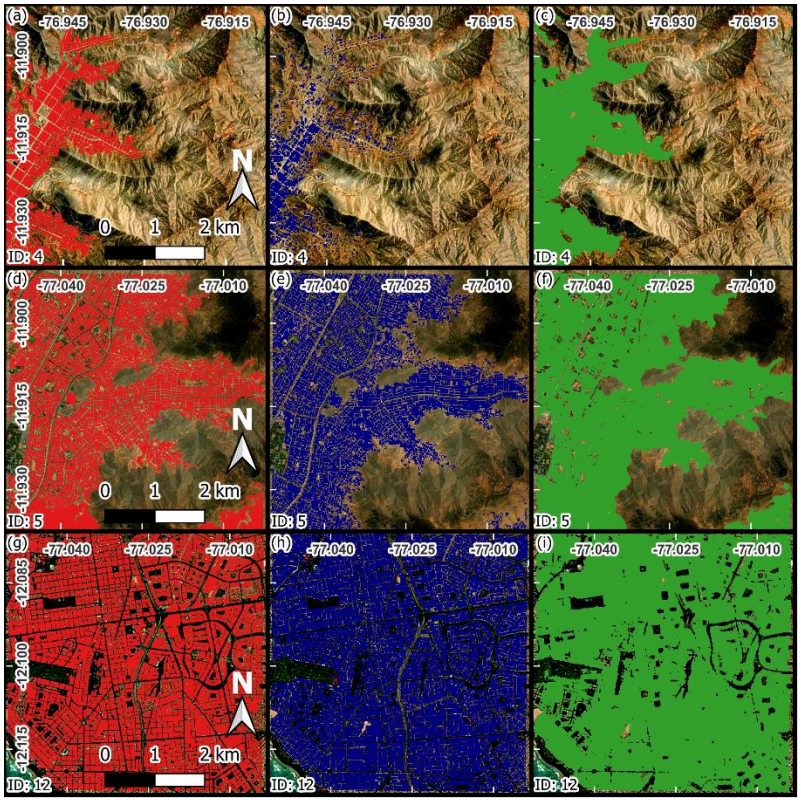


Figura 08:

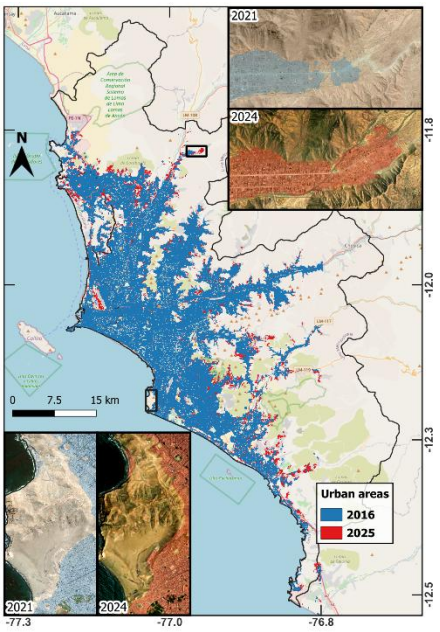
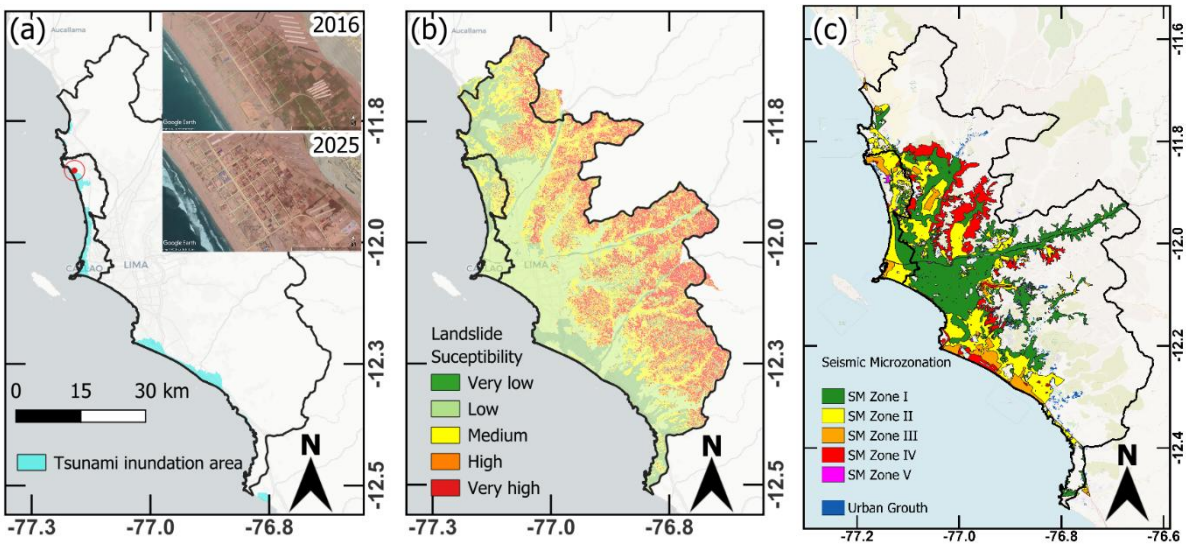


Figura 11:



Comment 05:

Equation (1): Why does the power number of e include a coefficient of “5”?

Answer

We included a coefficient of 5 to scale the output of the loss function. Consider the case where B_{ij} = -1, and Yij ranges from -1 to 1. Then, the loss:

$$l(-1, Y_{ij}) = \frac{1}{1 + e^{-Y_{ij}}}$$

ranges from about 0.27 to 0.73 (See solid blue line in Figure R1). On the other hand, the loss

$$l(-1, Y_{ij}) = \frac{1}{1 + e^{-5Y_{ij}}}$$

ranges from about 0 to 1 (See dashed orange line in Figure R1). As can be observed, this scaling increases the contrast between correctly and incorrectly classified samples.

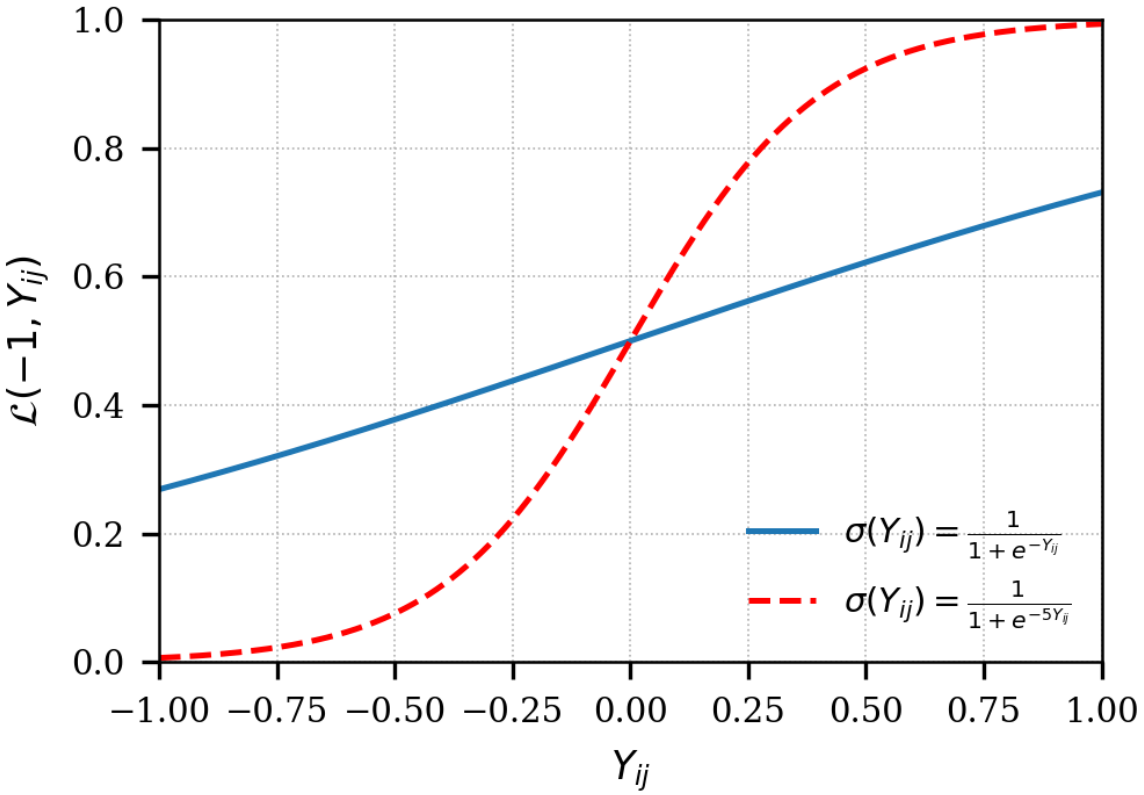


Figure R1.

Changes in manuscript:

Lines 117-118:

$$l(B_{ij}, Y_{ij}) = \frac{1}{1 + e^{5Y_{ij}B_{ij}}} \quad (1)$$

Note that we used an exponential factor of 5 to increase the dynamic range of the loss function when $Y_{ij} \in [-1, 1]$. This scaling improves the separation between correctly and incorrectly classified samples by pushing the loss values closer to 0 or 1. If only training samples of built-up areas are used, the Non-Negative Positive-Unlabeled (NNPU) risk estimator proposed by Kiryo et al. (2017) can be applied:

Comment 06:

Equations (2) and (6): The right square bracket is missing for the second term (the expected loss, E_u) on the right-hand side of the equation.

Answer

The missing right square bracket in the second term (the expected loss, E_u) of Equations (2) and (6) has been corrected in the revised manuscript.

Changes in manuscript:

Equations (2) and (6):

$$\tilde{R}_{pu}^{\mathrm{NNPU}}(g) = \pi_p \mathbb{E}_p[l(B_{ij}, +1)] + \max\left\{0, \mathbb{E}_u[l(B_{ij}, -1)] - \pi_p \mathbb{E}_p[l(B_{ij}, -1)]\right\}, \quad (2)$$

$$\tilde{R}_{pu}^{\mathrm{NNNU}}(g) = \pi_n \mathbb{E}_n[l(B_{ij}, -1)] + \max\left\{0, \mathbb{E}_u[l(B_{ij}, +1)] - \pi_n \mathbb{E}_n[l(B_{ij}, +1)]\right\}, \quad (6)$$

Comment 07:

Line 118: Maybe a typo: pi_m should be pi_n?

Answer

The reviewer is right, it was a typo. It has now been corrected.

Changes in manuscript:

Line 130:

125

$$\tilde{R}_{pu}^{\mathrm{NNNU}}(g) = \pi_n \mathbb{E}_n[l(B_{ij}, -1)] + \max\left\{0, \mathbb{E}_u[l(B_{ij}, +1)] - \pi_n \mathbb{E}_n[l(B_{ij}, +1)]\right\}, \quad (6)$$

where $\pi_n = 1 - \pi_p$ is the prior probability of non-urban samples. This complementary formulation enhances model robustness by explicitly incorporating reliable non-urban information into the semi-supervised training process. A combined risk function

130 is then proposed:

Comment 08:

Section 3.1: What is the spatial resolution of the images for deep learning modeling? Would this affect the model performance since the resolution of WSF dataset is 10 m?

Answer

The spatial resolution of the images used for deep learning modeling is 10 m, corresponding to Sentinel-2 imagery. Therefore, the input data have the same spatial resolution as the World Settlement Footprint (WSF) dataset (10 m), and no spatial resampling or resolution harmonization was required.

Because both datasets share the same spatial resolution, model performance is not affected by resolution discrepancies between the reference data and the input imagery.

In the revised manuscript, we have clarified the spatial resolution explicitly in the Dataset section to avoid ambiguity.

Changes in manuscript:

Line 154:

3.1 Dataset

Sentinel-2 satellite imagery with a spatial resolution of 10 m was used as input to the model. One image per year from 2016
155 to 2025 was selected (Table 1). For each year, the image corresponds to the period between January and March, when cloud

Comment 09:

Line 170 and Figure 6: It is typically expected that the model performance in validation is poorer than that in training, but this figure shows that the loss values of the two stages almost overlap with each other. It is suggested to randomly split the dataset into training and validation to guarantee the model's robustness. In addition, which set of model weights among the 200 epochs were chosen for further model comparison?

Answer

We thank the reviewer for this important observation.

The reviewer is correct that, in most cases, validation loss is expected to be higher than training loss. In our study, however, the training and validation datasets were generated from randomly extracted image patches within a single Sentinel-2 tile covering Metropolitan Lima. Because only about 432 non-overlapping patches of size 512×512 pixels can cover the study area, overlapping patches were allowed in order to obtain a sufficiently large number of samples (12,000 patches). As a result, some validation patches partially overlap with training patches, which explains the similar behavior of the training and validation loss curves shown in Figure 6.

To reduce possible confusion, we clarify that the dataset was randomly split into 70% training samples and 30% validation samples prior to the iterative calibration process. This information has now been explicitly stated in the manuscript.

In addition, the risk function shown in Figure 6 is computed using partially labeled samples within the semi-supervised learning framework and therefore does not fully represent the model performance over independent reference data. For this reason, an additional evaluation was performed using manually interpreted validation samples from Sentinel-2 imagery acquired in 2019, which were not used during model calibration. The corresponding comparison with the WSF dataset is presented in Figure 7 and Table 2.

Regarding the selection of model weights, the weights obtained at the final training epoch were used because both training and validation loss curves reached stable convergence without signs of overfitting. This clarification has also been added to the manuscript.

Changes in manuscript:

Lines 191-194:

trained iteratively. Approximately 600 polygons were annotated through this human-in-the-loop process, adding about 490,000 labeled pixels (0.5% of the total training data). The dataset was randomly divided into 70% training samples and 30% valida-

tion samples prior to the iterative calibration procedure. Because the patches were extracted from a single Sentinel-2 tile and overlapping patches were allowed to increase the sample size, some validation samples partially overlap with training samples, which explains the similar behavior of the training and validation loss curves shown in Figure 6.

To assess the performance of our model, we compared our results with the World Settlement Footprint (WSF) 2019 (German

Comment 10:

Line 193: It stated that “This result is expected since WSF effectively represents consolidated urban zones worldwide”. If that is the case, both the precision and recall evaluation metrics of WSF should be higher than those of the proposed framework.

Answer:

The reviewer is correct that if one method consistently represents consolidated urban areas better than another, both precision and recall would typically be higher. In our results, however, WSF shows slightly higher recall, while our model shows slightly higher precision in consolidated areas (Table 2). The statement in the manuscript was intended to refer to the overall performance measured by the F1 score, for which WSF exceeds our model by approximately 1%.

To avoid possible ambiguity, the sentence has been revised to clarify that WSF performs slightly better overall in consolidated areas in terms of the F1 score, rather than in both precision and recall individually.

Changes in manuscript:

Lines 216-217

The accuracy scores of our model and the WSF urban map are summarized in Table 2. In consolidated urban areas, when considering average scores, the WSF outperforms our model in recall by 2% and in F1 by 1%, but underperforms in precision by 2%. This result is consistent with the slightly higher F1 score obtained by WSF in consolidated areas, where large and homogeneous built-up structures are typically well represented in global settlement products. In peripheral areas, our model outperforms WSF in recall by 12% and in F1 by 2%, while underperforming in precision by 4%. In remote areas, the improvement is more significant: our model exceeds WSF in recall by 37% and in F1 by 12%, with a 4% decrease in precision.

Comment 11:

Lines 197-199: What are the possible reasons why the performance difference is relatively large in these cases?

Answer

The larger performance differences arise from two main factors related to the training data and model design.

First, in consolidated urban areas, the proposed approach relies predominantly on positive (urban) samples, while the available non-urban samples are limited and mainly correspond to parks. In contrast, the WSF model is trained in a fully supervised manner using both urban and non-urban samples, which leads to a more complete representation of non-urban classes. This explains the higher non-urban recall achieved by WSF in consolidated areas.

Second, in remote areas, the proposed model benefits from the closed-boundary strategy, which provides a large number of non-urban samples surrounding the urban fringe. This improves the model's ability to distinguish emerging built-up areas from natural terrain. In addition, the model is calibrated using local data from Lima, allowing it to better capture the specific spatial patterns of informal and fragmented urban growth. In contrast, WSF is trained on global datasets, which may underrepresent these localized patterns. This explains the higher urban recall achieved by our model in remote areas.

Changes in manuscript:

Lines 223-227

urban recall. These differences can be explained by the characteristics of the training data and model design. In consolidated areas, the proposed approach relies predominantly on urban samples, with limited non-urban representation, which reduces non-urban recall compared to WSF. In contrast, in remote areas, the closed-boundary strategy provides abundant non-urban samples and improves the detection of emerging built-up areas. Additionally, the use of local training data allows the model to better capture the spatial patterns of informal urban growth, which are often underrepresented in global datasets such as WSF.

Comment 12:

Figure 11(a): It should be noted that the uncertainty in the flood modeling process is not negligible. Thus, it is suggested to employ the probabilistic flood inundation maps instead of the deterministic maps for the further exposure analysis if possible. Please refer to the paper below.

Reference:

"Uncertainty analysis and quantification in flood insurance rate maps using Bayesian model averaging and hierarchical BMA" (https://doi.org/10.1061/JHYEFF.HEENG-58)

Answer

We agree that tsunami inundation modeling involves significant uncertainty, and that probabilistic approaches provide a more comprehensive representation of this uncertainty.

In this study, however, we did not perform tsunami simulations. Instead, we used the official inundation maps produced by the Directorate of Hydrography and Navigation of the Peruvian Navy, which are the standard reference used by national and local authorities for disaster risk management and evacuation planning. These maps are generated through numerical simulations (e.g., TUNAMI model) that incorporate seismic source scenarios, bathymetry, topography, and coastal characteristics.

According to the available technical documentation, the methodology for estimating maximum inundation and run-up follows established procedures consistent with IOC/ITSU guidelines, including post-tsunami field survey practices. However, detailed information regarding uncertainty quantification in these official products is limited.

We acknowledge that deterministic inundation maps do not capture the full range of possible tsunami scenarios. To address this point, we have added a discussion in the revised manuscript.

Changes in manuscript:

lines 331-338:

located near the shoreline should also be aware that, following a major earthquake, evacuation should begin immediately without waiting for an official warning. It is important to note that the tsunami inundation map used in this study corresponds to

a deterministic scenario derived from numerical simulations (e.g., TUNAMI model (Imamura, 1995)) based on specific seismic source assumptions and coastal conditions. Such maps are used for planning purposes by Peruvian governmental agencies, but they do not explicitly account for uncertainties in source characterization, wave propagation, and coastal interactions. Recent studies have highlighted the importance of probabilistic tsunami hazard assessments (Huang and Merwade, 2023; Davila et al., 2025). Therefore, the exposure estimates presented in this study should be interpreted within the context of the selected deterministic scenario. Future work could explore the integration of probabilistic inundation maps to better capture the uncertainty in tsunami hazard. Past earthquakes have shown that reclaimed land is highly susceptible to liquefaction (Kramer, 1996; Konagai et al., 2013); furthermore, previous reports indicate that the District of Callao experienced liquefaction during past seismic events (Alva-Hurtado and Ortiz-Salas, 2020). Therefore, detailed geotechnical assessments are needed for the

Comment 13:

Figure 12: How are the clusters defined and what is “Ha” in the horizontal axis? It is also suggested to change the label of the vertical axis in Figures 12(b)-12(d) to the accumulated area for the corresponding hazard.

Answer

We thank the reviewer for these helpful comments. The clusters shown in Figure 12 were defined based on spatial grouping of contiguous affected areas identified in the exposure analysis. Specifically, pixels classified as exposed were aggregated into spatially connected components, and each connected component was considered a cluster. This approach allows us to analyze the spatial concentration and extent of affected zones.
Regarding the horizontal axis, “Ha” refers to hectares, which represent the area of each cluster. We acknowledge that this abbreviation may not have been sufficiently clear.
Following the reviewer’s suggestion, we have revised the figure to (i) clarify the definition of clusters in the text, (ii) explicitly indicate that “Ha” corresponds to hectares, and (iii) modify the vertical axis labels in Figures 12(b)–12(d) to reflect the accumulated area for the corresponding hazard, thereby improving clarity and interpretability.

Changes in manuscript:

Figure 12

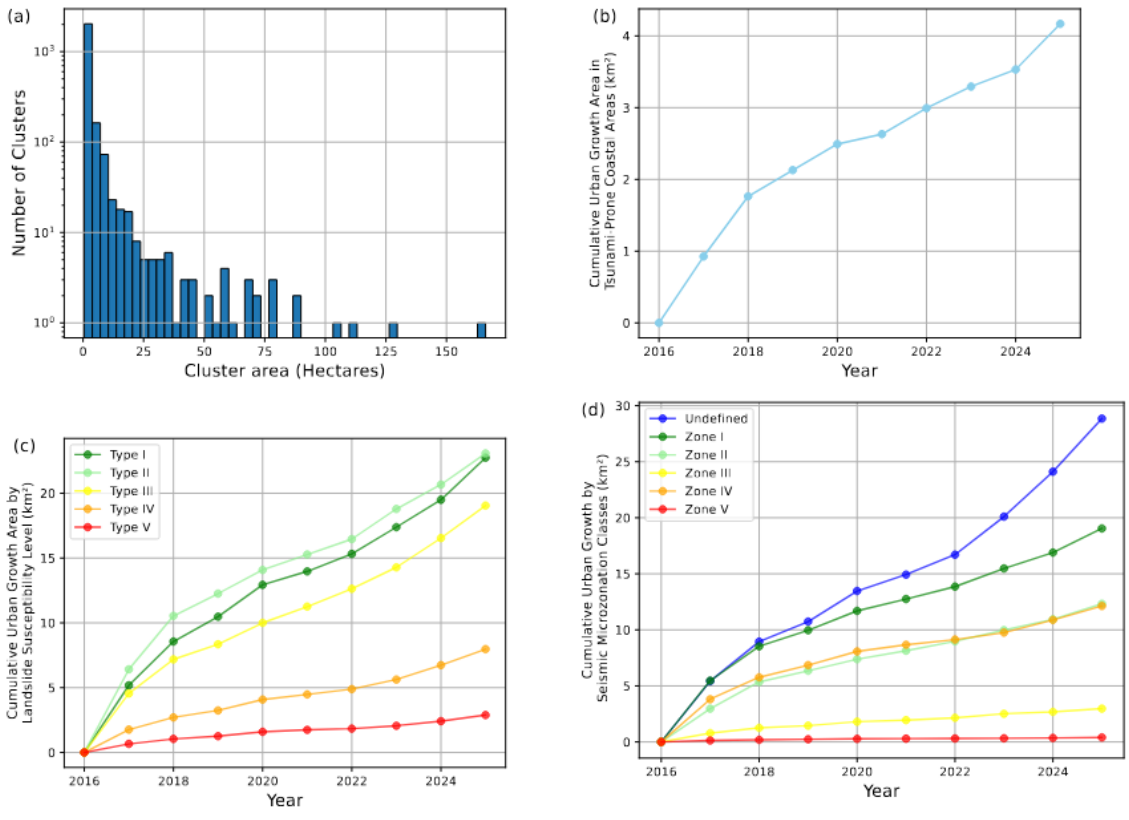


Figure 12. (a) Distribution of urban growth between 2016 and 2025 by clusters. (b) Cumulative urban growth by year in tsunami-prone coastal areas. (c) Cumulative urban growth by year across different landslide susceptibility levels. (d) Cumulative urban growth by year across different seismic microzonation classes.

Comment 14:

Lines 339-340: The statement that “the improved recall in peripheral and remote areas” may be true only for the urban area according to Table 2.

Answer

We appreciate this observation. The original statement referring to “improved recall in peripheral and remote areas” was not sufficiently precise. As shown in Table 2, the improvement is not consistent across all recall metrics.

What we intended to emphasize is the improvement in the average recall across both urban and non-urban areas, as well as the increase in F1-score specifically in peripheral and remote regions. This indicates a better overall balance between precision and recall rather than a uniform gain in recall alone.

To address this, we have revised the manuscript to clarify that the improvement pertains primarily to the average F1-score, ensuring consistency with the results presented in Table 2.

Changes in manuscript:

Lines 374-375:

375 Methodologically, the improved average F1-score in peripheral and remote areas, where the WSF underperforms, demonstrates the model’s ability to more effectively capture small and fragmented structures typical of informal settlement The iterative human-in-the-loop refinement effectively enhanced local consistency without requiring exhaustive manual labeling. Despite these advantages, the approach has several limitations. The 10 m resolution of Sentinel-2 imagery constrains the de-