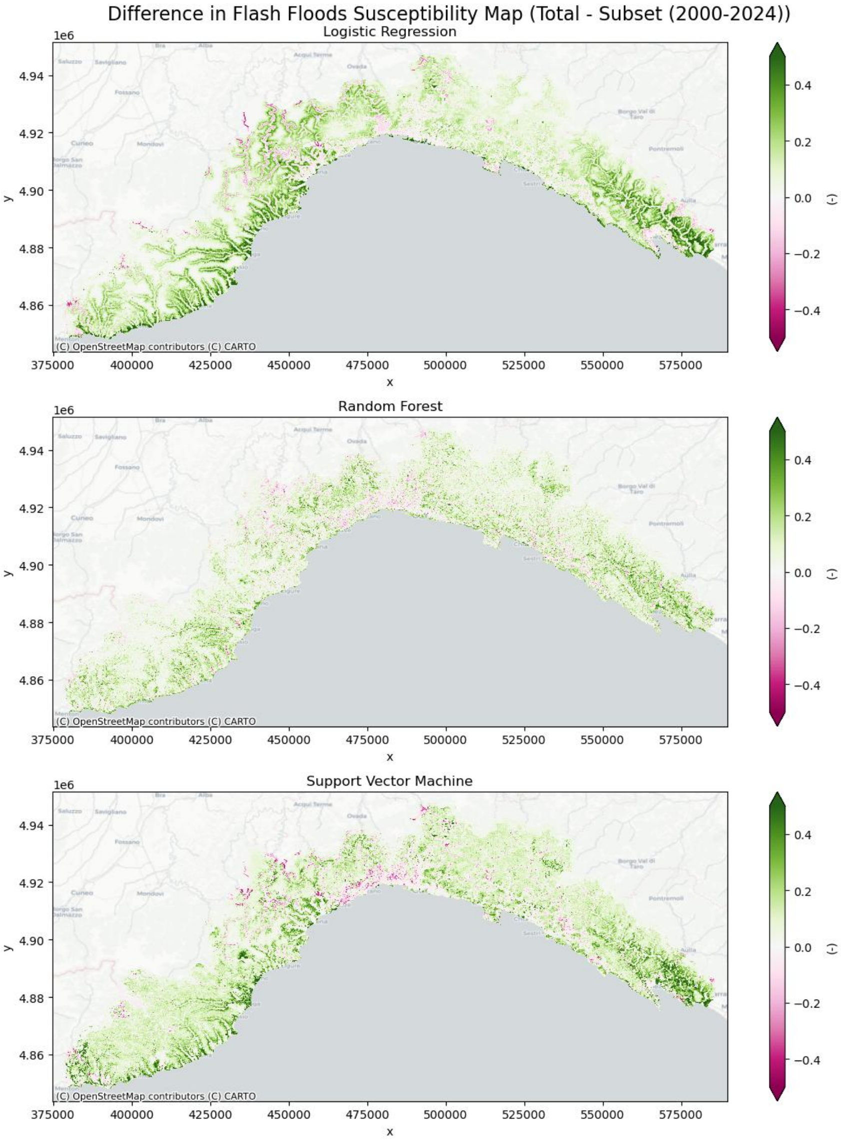
Referee comments 2

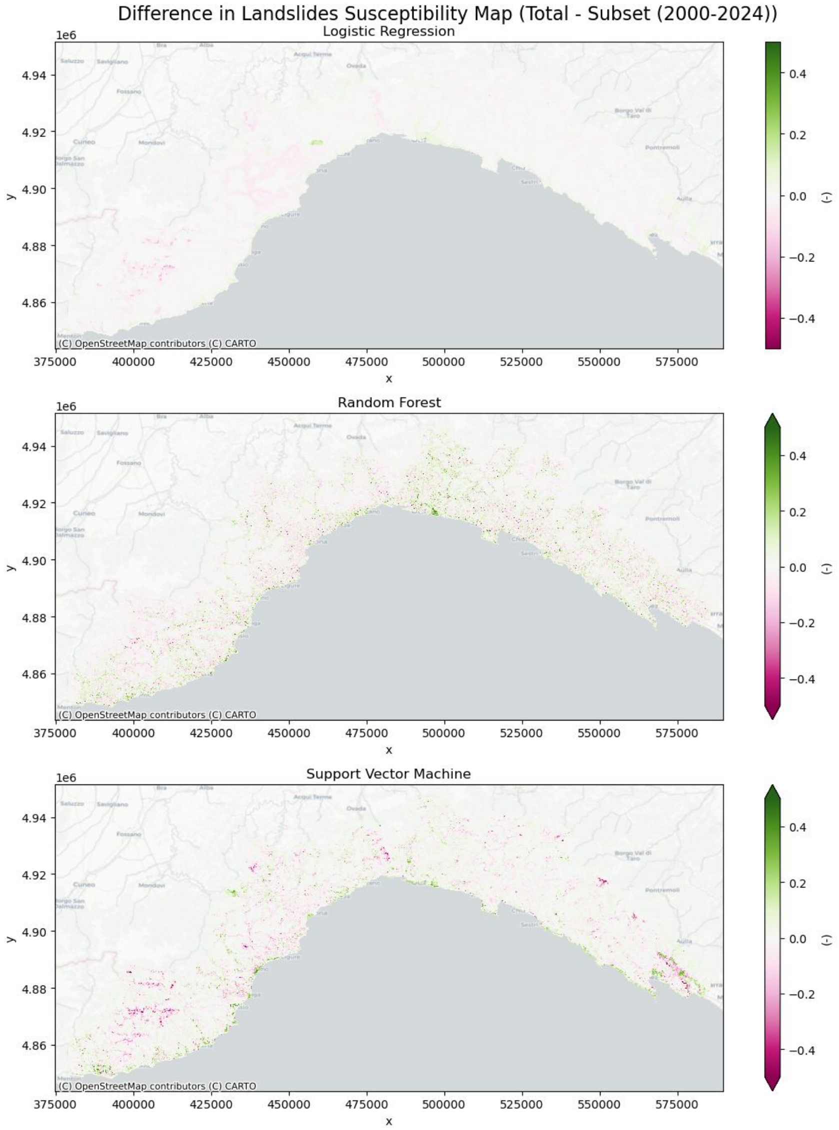
This manuscript investigates flash flood and landslide susceptibility using a unified machine-learning framework applied to the Liguria region in Italy. The study addresses an important topic in the context of multi-hazard risk assessment and benefits from the use of multiple event inventories, a consistent modelling framework across hazard types, and a careful discussion of several limitations. Overall, the paper is technically sound and clearly written, and it has the potential to make a valuable contribution to regional-scale multi-hazard susceptibility analysis. The comments below are intended to help strengthen the methodological clarity, interpretability, and positioning of the contribution.

Comment 1:
The flash flood inventory spans a very long time period (approximately 1950 to present), whereas the susceptibility modelling relies on present-day static conditioning factors (e.g. land cover, river network, road networks, …). Over this period, substantial changes have occurred in Liguria, particularly related to urbanisation, river modification, and land-use change, which may violate the assumption of stationarity between past events and current susceptibility conditions. While the long inventory improves statistical robustness, it may also introduce temporal inconsistencies that affect the interpretation of the resulting susceptibility maps, especially in urbanised valleys where many early events predate current conditions. This issue is particularly relevant for flash floods, which are highly sensitive to anthropogenic hydrological modifications. The authors are encouraged to (i) acknowledge this limitation more explicitly, and/or (ii) demonstrate that it does not substantially affect the results. Possible approaches could include comparing susceptibility maps derived from a more recent subset of events (e.g. post-1990) with the full inventory-based map, or applying temporal weighting to events based on their distance from present-day conditions.

Response 1:
This is a valid point. We can indeed relate high hazard susceptibility to urbanized areas (Section 4.3) and the level urbanization increased over time, especially for the area of Genoa (Faccini et al., 2015 and Lanza, 2004). We have added a comment to section 5.6 Role of inventories and data completeness in framework transferability. "*In addition, the framework has been applied and trained on the full period of data availability at once (only excluding scattered events from before 1940), while especially land use changes such as urbanization may have affected the landslide susceptibility in the region over time. To verify that the long inventory does not introduce temporal inconsistencies that may affect our conclusion, we derived new susceptibility maps for both hazards using only the more recent events (2000-2024) (refer to Figure 3). We found no significant changes in the susceptibility maps with more limited changes for landslides than for flash floods (figures not shown).*"

We found no significant changes in the susceptibility maps especially for landslides. As for flash floods there are some more differences but least pronounced for Random Forest. Figures will not be included in the manuscript for conciseness. Overall, the maps based on only the period 2000-2024 shows a somewhat lower susceptibility, which is most likely a result of the lower number of events included in the training.


Difference in Flash Floods Susceptibility Map (Total - Subset (2000-2024))
Logistic Regression
Random Forest
Support Vector Machine
1e6
4.94
4.92
4.90
4.88
4.86
y
375000
400000
425000
450000
475000
500000
525000
550000
575000
x
0.4
0.2
0.0
−0.2
−0.4
(-)
(C) OpenStreetMap contributors (C) CARTO



Comment 2:
The Introduction states that this study transfers the framework of Tehrani et al. (2021) to flash flood susceptibility and applies it within a unified multi-hazard context. However, the precise methodological and conceptual advances beyond the original framework are not sufficiently articulated. The authors are encouraged to clarify more explicitly what aspects of the framework are novel (e.g. unified hazard treatment, inventory construction, comparative analysis), and how this work differs from or extends Tehrani et al. (2021) beyond a change in hazard type and study area. Strengthening this framing would help readers better appreciate the innovative potential of this study.

Response 2:
We have clarified the text in the introduction: "*Our starting point is Tehrani et al. 2021 which originally developed a landslide detection ML model, and was subsequently adapted it to a landslide susceptibility mapping framework. The latter has been extended in this study with (i) the inventory construction, (ii) the systematic inclusion of input layers relevant to flash floods and landslides, (iii) the assessment of multicollinearity between the layers, and (iv) the study of behaviour similarity between both hazards resulting in a unified susceptibility framework.*"

Comment 3:
The title emphasises a Mediterranean-scale application, while the analysis is conducted at a local/regional scale focused on Liguria. While Liguria is representative of Mediterranean hydro-geomorphic settings, the current title may overstate the spatial scope of the application. The authors may wish to consider revising the title to better reflect the local-scale case study, while still acknowledging Mediterranean relevance.

Response 3:
We've adjusted the title:

A single framework for assessing flash flood and landslide susceptibility: an application to the Mediterranean Liguria region, Italy

Comment 4:
In the Methodology section, the manuscript states that hyperparameters were calibrated using grid search and cross-validation, but the specific parameter ranges and final selections for each algorithm (LR, RF, SVM) are not described in detail. For transparency and reproducibility, the authors are encouraged to provide: the tested hyperparameter ranges, and the final selected values for each model, possibly in an appendix or supplementary material. This would strengthen the methodological clarity of the paper.

Response 4:
The parameter ranges and the final selected values for each model and hazard will be added in the Appendix B.

Comment 5:
Class imbalance: the flash flood and landslide inventories differ substantially in size, and the binary classification problem is inherently unbalanced (event vs non-event pixels). While model performance metrics (AUC, accuracy, confusion matrices) are reported, the role of class imbalance in influencing false positives and false negatives is not explicitly discussed. The authors are encouraged to either: discuss how class imbalance may affect the reported performance metrics, or justify the decision not to apply balancing techniques (e.g. class weighting, SMOTE). Explicitly addressing this point would improve confidence in the interpretation of model performance.

Response 5:
Since the susceptibility mapping for the two hazards is done independently, the impact of the imbalance between inventories will be limited. However, we do recognize that the imbalance in the number of events in the inventories over time can have affected the accuracy of the model. A comment on this has been added to section 5.6 of the discussion and a figure with events over time in the inventory section.

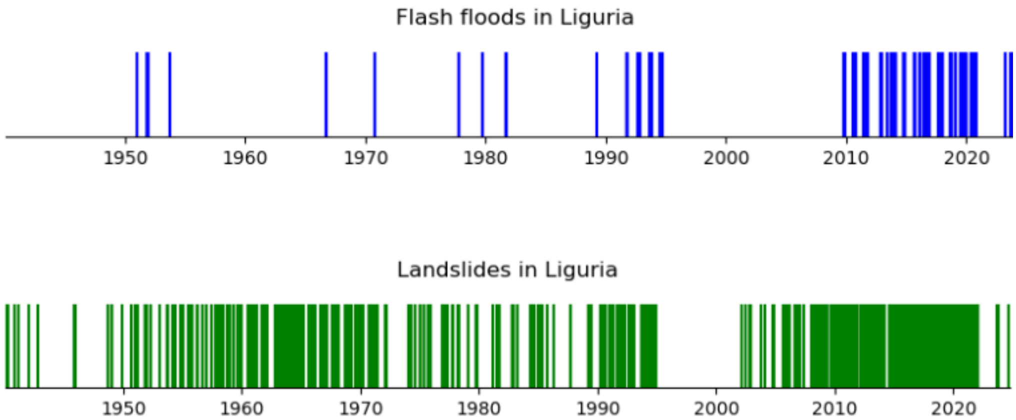


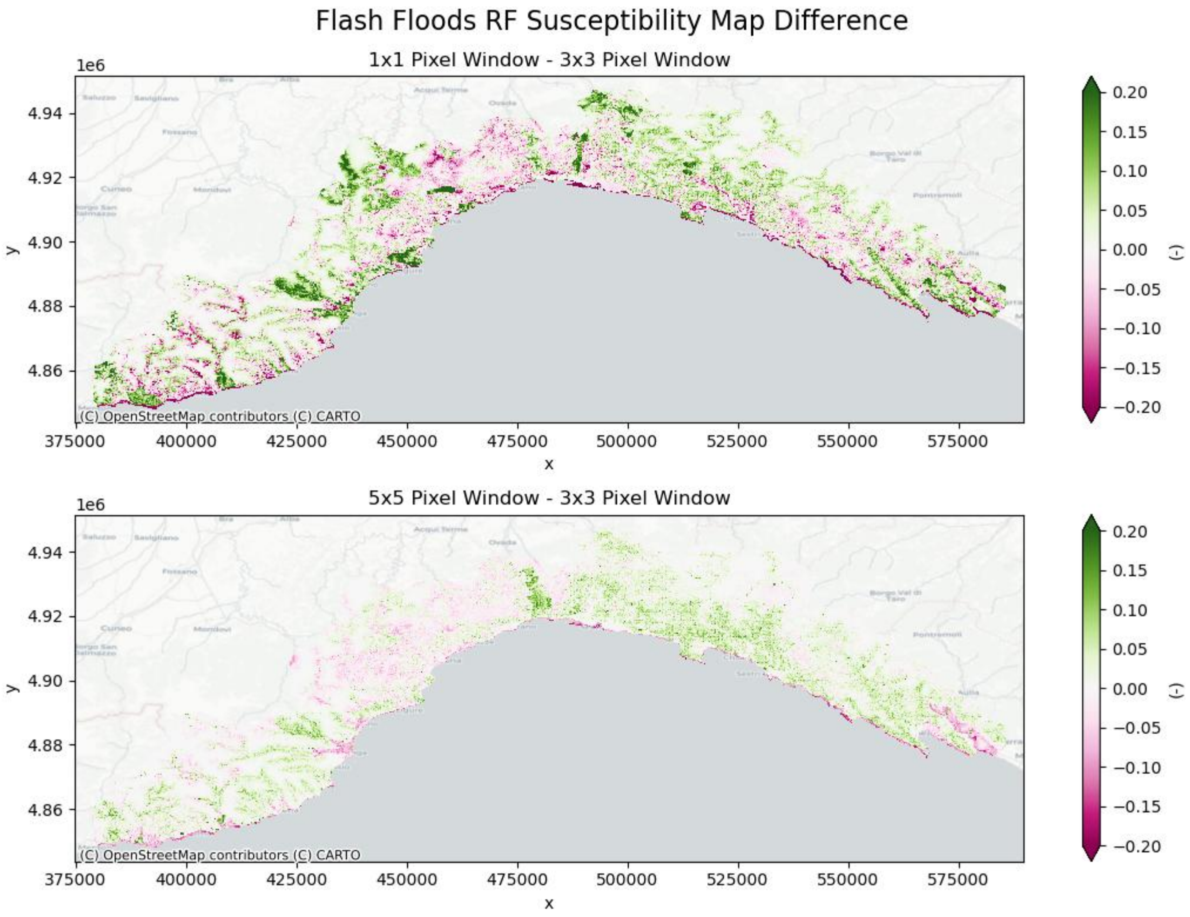
Figure 3: Flash flood (top) and landslide (bottom) events distribution from 1940-2024

We have added a comment on the inherently balanced binary classification in section 5.8: "*The fewer occurrences of events –whether landslides or flash floods– with respect to non-events lead to a class imbalance which can cause bias in the ML model by predicting the event (majority class) more often. One approach to handle this imbalance is to under sample the majority class and/or oversample the minority class, the latter is the more common in hazard mapping. A popular method is the synthetic minority oversampling technique (SMOTE) which for each event finds the k nearest neighbours and generates a random event at the chosen neighbour creating new events along the segment between the original event and the chosen neighbour. Han and Semnani (2025), who confirm the challenges of addressing class imbalance in landslides, found the best overall results for gridded hyperspace even sampling with a variant of SMOTE. However, this method can potentially create physically unrealistic events. Another approach is to apply class weights or cost sensitive weighting (Chen, C. et al., 2004) in which the model is penalized for misclassifying the minority class. This means that by balancing the classes, a trade-off occurs: an increase of TP to the detriment of creating more FP. As previously indicated, our confusion matrix results are quite positive, limiting but not removing the need of applying a trade-off which could be studied in future work.*"

Comment 6:
To account for location uncertainty in point-based inventories, a 3x3 pixel window is used around each event. While this is a reasonable approach, it may introduce spatial smoothing that blurs the influence of certain conditioning factors, particularly for landslide susceptibility in steep terrain. As already partially acknowledged by the authors, this may contribute to unexpected results (e.g. landslides associated with gentler slopes). The authors may consider discussing alternative or complementary strategies, such as: exclusion or buffer zones around event pixels to reduce spatial autocorrelation, or sensitivity tests on window size. If considered too technical for implementation, a clearer discussion of the implications of this choice would still be valuable.

Response 6:

We have added in the methods a sentence on this and in the results. We reran the ML models for both 1x1 and 5x5 pixel window sizes. There are only minor differences, though not significant, for the flash floods susceptibility maps. There are no differences in the susceptibility maps for landslides. The figures will not be included for conciseness.



Comment 7:

The framework is described as globally applicable; however, its performance depends strongly on the availability and quality of local hazard inventories. The authors are encouraged to expand the discussion on spatial and temporal transferability, particularly regarding:

- data-rich versus data-scarce regions;
- which conditioning variables may be problematic when transferring the model (e.g. the authors mention aspect -southward facing slopes- dependence, how would that change in regions with different dominant exposure?);
- the potential role of high-resolution dynamic inputs (e.g. convection-permitting reanalysis and future projections) in improving applicability across regions and time periods.

Response 7:

We will extend the discussion on this in section 5.6 that specifically links the data completeness to transferability of the framework:

Previous studies (Free et al., 2022; Modrick and Georgakakos, 2015; Uwihirwe et al., 2022) have shown that in many other, especially remote, regions of the world the number of recorded events is substantially lower. Global inventories can fill some gaps but are often coarse, incomplete or inconsistent in spatial and temporal coverage. This will affect the training of the ML algorithm and its accuracy, limiting the transfer of the framework to other data rich regions, or the transfer of the trained framework to regions with similar terrain, meteorological conditions, land cover, soil types, and land use including degree of urbanization. To overcome the latter, transfer *learning for landslides susceptibility modelling in dissimilar areas was applied by Wang et al. (2022) by using Domain Adaptation (DA) in which a latent feature space is defined where the source and target areas have the same distribution. In particular, we expect physics-informed predictors like TWI, SPI, slope, curvature to be more easily transferable (after aligning their distributions) to other regions. On the other hand, elevation, proximity to roads, proximity to rivers, NDVI, land cover, and lithology are region dependent and could be transferred to similar regions using e.g. Case-Based Reasoning (CBR) (Wang et al., 2022). Alternatively, they may be harmonized to represent properties rather than classes e.g. grain size instead of lithology classes. Aspect on its own is not an informative variable but could be rederived into Windward-Leeward Index (WLI) that incorporates the influence of orographic precipitation. Still, there may be other aspects like wetness, seasonality and variability of precipitation that are not similar between regions affecting the accuracy of the transferred framework.*

The potential role of high-resolution dynamic inputs will be included in 5.7 Framework applications:

In this study all conditioning factors were static. Including these dynamic parameters would enhance the capacity of the models to capture transient conditions and could potentially even allow the framework to update susceptibility in near real time. Moreover the framework holds the potential to be applied in early warning applications, as we previously explored in Uwihirwe et al. (2022), using high-resolution forecasted, or even now-casted precipitation (short-term). The same holds for future climate change assessments using high resolution precipitation projections from convection-permitting climate models (Zander et al., 2022). The framework would anyhow benefit from higher resolution precipitation products as these capture better the effect of the high spatial variability in heterogeneous terrain (Lee et al., 2022).