Report #1

We thank the reviewer for the rigorous review of the revised manuscript and appreciate the constructive feedback.

- Reviewer original comment (RC in blue)
- Authors' answer to reviewer (AA in black) – Line numbers (L) refer to the previously submitted manuscript; *text in italics refers to passages from the manuscript*.

Below we respond to each of the reviewer comments:

#1 RC: The revised manuscript has improved in clarity and framing, and the authors have responded constructively to the previous comments. The study presents a valuable and well-structured data-driven framework for dynamic, spatially explicit modelling of mass movement impacts, with clear relevance to impact-based early warning.

AA: We thank the reviewer for this positive assessment and for acknowledging the improvements made. We are encouraged that the overall structure, clarity, and relevance of the proposed framework are recognized. The following RC (#2) covers several interconnected aspects that will be addressed through substantial revisions. Accordingly, we present a comprehensive response below, instead of reiterating the same arguments multiple times.

#2 RC: However, I still recommend major revision before the manuscript can be accepted. The main concern remains the spatial transferability of the proposed models. The training data are limited to Austria and South Tyrol, whereas the model outputs are extrapolated to the entire Alpine Space. Although the authors now acknowledge this limitation, the manuscript still needs to more explicitly distinguish between model calibration, spatial extrapolation, and operational applicability. The current results should not be interpreted as validated Alpine-wide impact predictions. The absence of independent impact data outside the training region is a substantial limitation, and the uncertainty associated with extrapolated predictions should be more clearly reflected in the Results, Discussion, figures, captions, and Conclusions. In particular, statements suggesting broad operational usefulness should be further softened. In addition, the reliability of the landslide inventory and derived spatial patterns requires more rigorous clarification. The manuscript relies heavily on impact-based inventories (e.g., GEORIOS, IFFI), which inherently emphasize damage-causing events and underrepresent non-damaging or remote landslides. This may introduce systematic biases in both spatial distribution and model learning. The authors should explicitly quantify or at least discuss the completeness, temporal consistency, and spatial representativeness of the landslide data, and clarify how these limitations affect the derived landslide distribution maps and subsequent modelling results. It is also recommended to provide additional supporting materials (e.g., inventory density maps, temporal coverage plots, or uncertainty descriptions) to demonstrate the reliability and robustness of the input data.

AA: We thank the reviewer for the careful and constructive assessment. We agree that the distinctions between model calibration, spatial extrapolation, and operational applicability, as well as issues related to inventory incompleteness and their implications for modelling results, require further clarification and strengthening throughout the manuscript.

In summary, we will address these concerns through a comprehensive revision that further refines the overall framing of the study, while also incorporating additional clarifications and discussions, as well as updated figures and maps. In the revised manuscript we will strictly distinguish between the model calibration/validation area ("core study area": Austria and South Tyrol) and the area used for spatial extrapolation beyond this training domain. This distinction will be consistently reflected across all major sections and associated maps, with a clear focus on interpretations (and map zoom-ins) related to the validated core area, while avoiding interpretations that imply validated Alpine-wide impact predictions. Furthermore, several of the issues addressed by the reviewer will be tackled in a new dedicated discussion section entitled *"6.5 A word of caution: Interpreting probability scores within and beyond the training domain."*

Also, aspects related to the operational applicability of the models will be presented more cautiously, alongside a more explicit discussion of inventory incompleteness and its implications.

Hereafter, we highlight major revisions that will be made chapter by chapter (for minor textual modifications/clarifications: see document with tracked changes):

Abstract:

We will revise the description of the study scope to clearly distinguish between the calibration domain and the extrapolation domain:

L19f: *"The framework integrates (…) data to estimate daily impact potential across the Alpine region (450,000 km²)."* **Revision:** *"It integrates (…) data to estimate daily impact potential for an Alpine core study area covering more than 91,000 km² (Austria and South Tyrol), and subsequently explores the transferability of these relationships to the wider Alpine region, subject to important limitations."*

L25f: *"Results highlight the strong operational potential of slide- and flow-type models, while the fall-type model exhibits limited usability for early warning, due to its low sensitivity to short-term weather conditions."* **Revision:** *"Results indicate promising performance of slide- and flow-type models for the core study area, suggesting potential for operational application, while the fall-type model shows limited applicability for early warning, likely due to its lower sensitivity to short-term weather conditions. The primary contribution of this research lies in demonstrating a transferable and generalizable modelling framework, rather than delivering an operational regional warning system."*

Introduction:

L115f: "This study addresses these gaps by developing and evaluating interpretable data-driven models to dynamically assess the impact potential of precipitation-induced mass movement processes across the Alpine region at daily scale." **Revision:** *"This study addresses these gaps by developing and evaluating interpretable, data-driven models to dynamically assess the daily impact potential of precipitation-induced mass movement processes for a large area covering Austria and South Tyrol (91,000 km²), while treating non-validated spatial extrapolation to the wider Alpine region (450,000 km²) as a demonstrator to explore the potential applicability of the framework to very large terrain."*

Study area:

L133: "This study focuses on the Alpine Space as delineated by the Interreg Alpine Space Programme of the European Union (Fig. 1), an area that encompasses approximately 450,000 km² across seven countries: Austria, France, Germany, Italy, Liechtenstein, Slovenia and Switzerland." **Revision:** *"This study focuses on an Alpine core study area covering Austria and the Italian province of South Tyrol (91,000 km²), which forms the basis for model development and model evaluation. The broader Alpine Space, as delineated by the Interreg Alpine Space Programme of the European Union, encompasses approximately 450,000 km² across seven countries (Austria, France, Germany, Italy, Liechtenstein, Slovenia, and Switzerland) and is considered for demonstrator-based spatial extrapolation."*

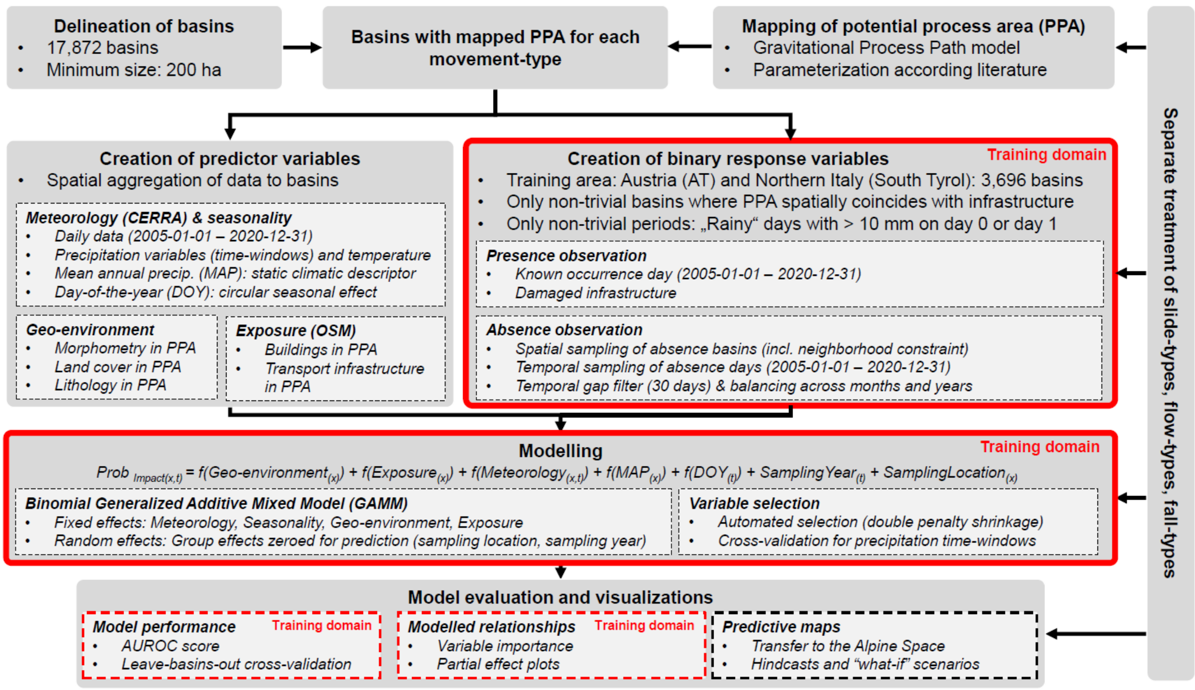
Accordingly, also the Figure 1 caption will be revised to *"Location and topography of the study area. Austria (red) and South Tyrol (blue) indicate the core study area for model training and quantitative validation, while the wider Alpine Space represents the domain used for spatial extrapolation."*

Data:

194f: "The datasets cover the entire Alpine Space, while mass movement event data for model training were available for Austria and the Italian province South Tyrol (Fig. 4), a region considered reasonably representative of the environmental variability across the Alpine Space." **Revision:** *"Model training and quantitative validation were conducted within the core Alpine study area covering Austria and the Italian province of South Tyrol (Fig. 4), while predictor datasets were created for the entire Alpine Space, technically enabling spatial extrapolation beyond the training domain."*

Then (formally L198), we will add an explicit statement highlighting the limitations of generalized predictor variables and their implications for spatial transferability, with a clear reference to the Discussion: *"The Alpine-wide generalization of input variables, including the reclassification of lithology and land cover (cf. Section 4.2.3), introduces inherent limitations and constrains the ability of the model to estimate impact potential beyond the training domain (cf. Section 6, Discussion)."*

We will also revise Figure 2 and its caption to explicitly distinguish which steps are conducted exclusively within the training domain (highlighted in red) and which are applied across the entire Alpine Space:



New Figure 2 caption *"Methodical framework of this study. Aspects referring exclusively to the training domain (Austria, South Tyrol) are highlighted in red, while processing steps applied to the entire Alpine Space are shown without red borders."*

Methods:

We will revise the methods section by adapting section titles and corresponding descriptions to explicitly reflect the respective focus areas. Revised section titles:

- **"4.3 Modelling within the core study area"**
- **"4.4.1 Assessment of model performance within the core study area"**
- **"4.4.2 Assessment of modelled relationships within the core study area"**
- **"4.4.3 Generation of predictive maps within and beyond the core study area"**

The revised methods Chapter 4.4.3 will now explicitly clarify that model predictions were extended to the entire Alpine Space beyond the training domain and will direct the reader to sections where the associated limitations are discussed in greater detail: *"A series of predictive maps and animations were produced by applying the modelled relationships from the three models within the validated core study area, as well as across the other, non-validated regions of the Alpine Space. (...) It should be noted that the applications beyond the validated training domain serve as a technical demonstrator and are subject to important limitations. Predictions outside the core area should therefore be interpreted with caution, as they represent a non-validated similarity score indicating whether generalized local conditions resemble or differ from those associated with damage-causing landslides in the spatially distant training domain (cf. Section 6.4 and 6.5 for guidance on interpretation and relevance for operational implementation)."*

Results:

The Results section will be revised to focus on describing the results within the calibration area, avoiding interpretations outside the training domain. Also, Section titles will be adapted to enhance clarity:

- **"5.1 Model performance within the core study area"**
- **"5.2 Modelled relationships within the core study area"**
- **"5.3 Maps and visualizations within and beyond the core study area"**

Example of the revised results section (full details will be provided in the tracked-changes document): *"Slide-type models generally indicate the largest spatial extent of potentially impacted areas, encompassing not only the high-relief regions of western Austria and South Tyrol but also extensive parts of the Austrian Alpine foreland in the east, where relief energy is comparatively moderate (Fig. 10d). In contrast, impacts from flow-type processes are less widespread and more strongly confined to steeper valley settings (Fig. 10i). Potential impact areas associated with fall-type movements are the most spatially restricted, with elevated values concentrated in high-relief zones of western Austria and South Tyrol, where buildings and transport infrastructure are frequently located within reach of steep rock faces (Fig. 10n)."*

Accordingly, the figures will also be revised to explicitly highlight the core study area, with zooms confined to locations within the training domain in order to avoid interpretations beyond the calibration area. Figure 10 will be revised as follows: instead of presenting the full Alpine Space, the figure will focus exclusively on the model calibration areas (red and blue), including a detailed zoom into the training domain.

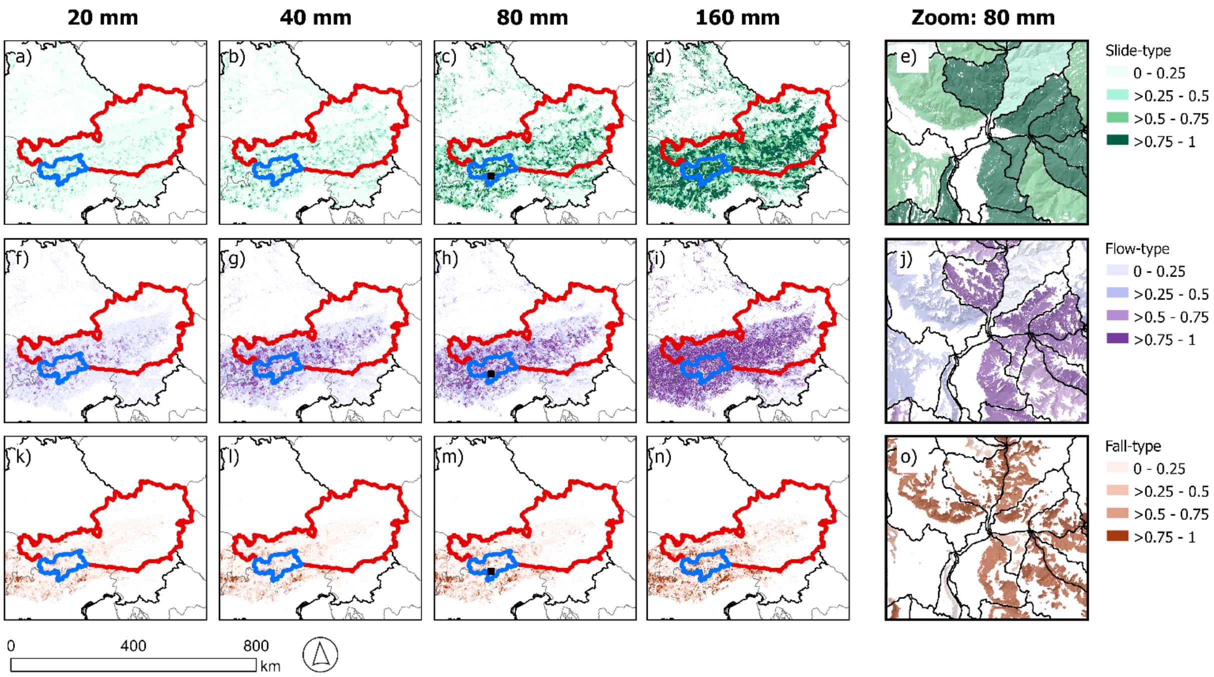


Figure 1: Visualization of "what-if" scenarios for slide-type models (a–e), flow-type models (f–j) and fall-type models (k–o) under spatially uniform increases in short-term precipitation. The figure depicts the core study area (red: Austria; blue: South Tyrol) and a portion of the surrounding Alpine Space. Shown are (…)

For the revised Fig. 11, we will explicitly depict the calibration area using blue and red polygons. However, we propose to retain a map showing the entire Alpine Space in this context, as the event affected several regions outside the calibration area (see also cited literature in the text). To avoid misinterpretation, the figure caption will explain the meaning of predictions outside the core study area once more: *"Note that the model was calibrated and validated only within the core study area (red and blue polygons). Predictions outside this area reflect the similarity of generalized local conditions to those causing damage-inducing landslides in the core area (values close to 1 indicate higher similarity, values close to 0 lower similarity)."*

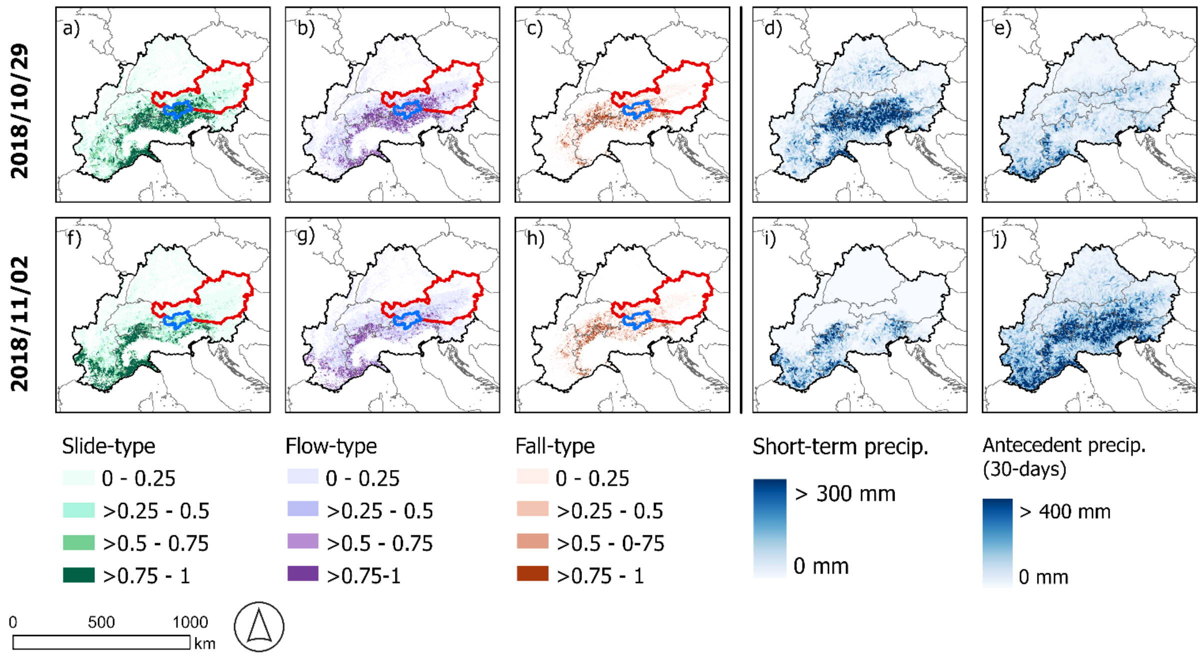


Figure 2: Hindcast example of Storm Vaia for the core study area (red: Austria; blue: South Tyrol) and the surrounding Alpine Space (regions outside the red and blue polygons). (…) Note that the model was calibrated and validated only within the core study area (red and blue polygons). Predictions outside this area reflect the similarity of generalized local conditions to those causing damage-inducing landslides in the core area (values close to 1 indicate higher similarity, values close to 0 lower similarity).

In the revised Fig. 12, we will now avoid showing the full Alpine Space and instead focus on a zoomed view within the validated core area. The revised figure caption and accompanying text will only emphasize patterns observed within the validated calibration area.

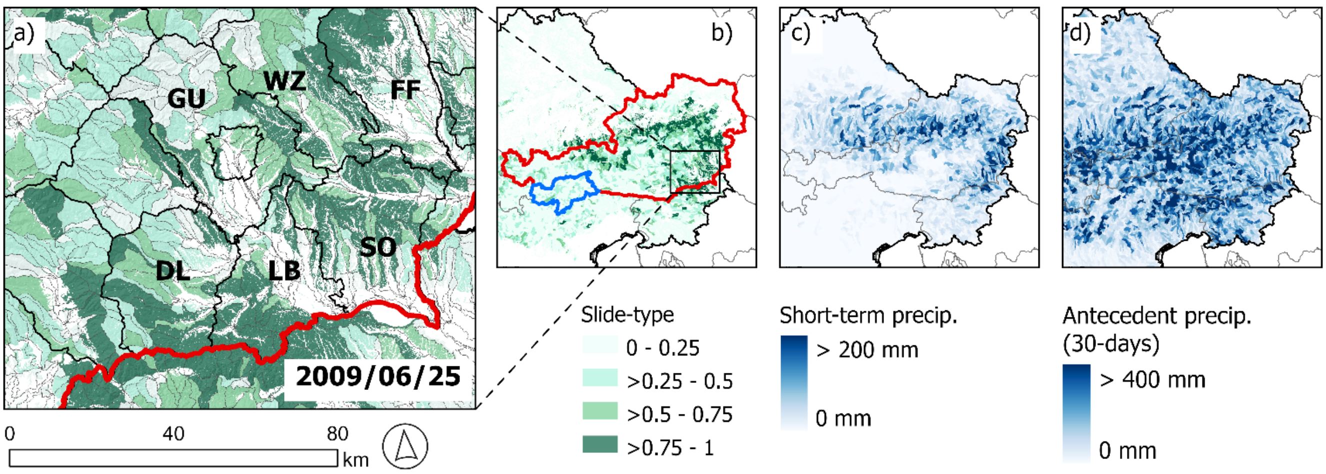


Figure 3: Hindcast example of a severe event in Styria, Austria. Model predictions from the slide-type model for June 25, 2009, are shown for the focus area of the event (a) and the calibration area of Austria (red in b) and South Tyrol (blue in b) (…).

Discussion:

Also, here we will avoid interpretation outside the core study area and revise the text accordingly.

Most importantly, we will expand the discussion on the interpretation of predictor effects and inventory-related issues, and added a new discussion section that explicitly addresses transferability issues.

New discussion section:

"6.5 A word of caution: Interpreting probability scores within and beyond the training domain

Another important consideration when interpreting the presented model predictions concerns the spatial domain in which the models were trained and validated versus the regions to which they are applied (Fig. 1). The models were calibrated and quantitatively validated exclusively within the 91,000 km² core study area (Austria and South Tyrol). Consequently, predictions generated outside this domain constitute non-validated spatial extrapolations. In these regions, model outputs should not be interpreted as reliable impact estimates, but rather as a basic indicator of the similarity between generalized local conditions and those associated with observed impact-generating events within the distant training domain. The lack of consistent inventory data for large areas (cf. Section 6.1) becomes even more critical beyond the widely applied landslide susceptibility context, particularly in impact-based spatio-temporal modelling, where not only the location of past events is required, but also their timing and, crucially, associated damage information. Accordingly, future improvements and a more robust transfer of modelling approaches to the Alpine-wide scale would benefit from the development and harmonization of time-stamped, impact-oriented landslide inventories across the Alpine arc.

Given that even compiling consistent process inventories at such scales is associated with substantial uncertainties (Caleca et al., 2025), this is likely to remain a challenging and resource-intensive task. Therefore, we emphasize promising methodological advances that explicitly address the transfer of model-derived information from well-investigated domains to data-sparse domains. In particular, transfer learning techniques, combined with spatially explicit model tuning and validation strategies, offer valuable pathways to improve model transferability and enhance the robustness of predictions under persistent and difficult-to-overcome data limitations (Wang et al., 2022; Schratz et al., 2024; Brenning and Suesse, 2026).

A further important aspect concerns the interpretation of the predicted probability scores, also within the core study area where the model was trained and validated. These values do not represent absolute probabilities of impact occurrence but conditional probabilities influenced by the class balance of the training data. Given that absence observations substantially outnumber presence observations in our dataset, the predicted probabilities are generally biased toward lower values. Moreover, the raw probability scores produced by the different model variants are not directly comparable, as they are derived from datasets with differing presence-to-absence ratios. Misinterpretation of these scores may hinder user uptake and reduce trust. For instance, a value of 0.7 should not be interpreted as a 70% likelihood of a mass movement impact occurring in a given basin on a specific day. Instead, higher values indicate greater similarity to conditions associated with observed impact events within the model training domain, whereas lower values reflect greater similarity to non-impact conditions. For practical application in impact-based warning systems, these continuous similarity scores would need to be transformed into discrete warning levels (e.g., red, orange, yellow). While objective thresholding approaches based on performance metrics such as true positive and false alarm rates exist (Steger et al., 2024), their effective implementation requires careful calibration against independent data and, ideally, co-development with end-users. Incorporating local knowledge and context-specific requirements is essential to ensure that derived warning thresholds are both interpretable and actionable (Budimir et al., 2025). Overall, while the presented modelling framework shows clear promise for data-driven, impact-based warning, the operational usefulness of the resulting models remains limited at the current stage of development. This is particularly the case beyond the training domain."

Newly added references:

Brenning, A. and Suesse, T.: Aligning Validation with Deployment in Spatial Prediction: Target-Weighted Cross-Validation, https://doi.org/10.48550/arXiv.2603.29981, 2026.

Schratz, P., Becker, M., Lang, M., and Brenning, A.: Mlr3spatiotempcv: Spatiotemporal resampling methods for machine learning in R, Journal of Statistical Software, 111, 1–36, https://doi.org/10.18637/jss.v111.i07, 2024.

Wang, Z., Goetz, J., and Brenning, A.: Transfer learning for landslide susceptibility modeling using domain adaptation and case-based reasoning, Geoscientific Model Development, 15, 8765–8784, https://doi.org/10.5194/gmd-15-8765-2022, 2022.

#3 RC: "A further concern is the interpretation of predictor effects, especially temperature-related variables. While the revised explanation is more cautious, potential confounding among temperature, elevation, seasonality, inventory bias, and exposure remains insufficiently resolved. The manuscript should avoid over-interpreting these effects as direct physical controls unless supported by additional analyses.

We agree that the interpretation of predictor effects requires greater care and should be considered within the appropriate "uncertainty context". Accordingly, we will expand the associated discussion section (Section 6.1), emphasizing a more cautious interpretation ("correlation does not imply causation"), outlining key underlying reasons (e.g. inventory biases), and highlighting potential future solution strategies (e.g. transfer learning).

Newly added arguments to the model interpretation section (6.1):

"Although we aimed to produce plausible results through a tailored and innovative landscape representation (Section 6.2), a carefully designed sampling strategy (Section 6.3), and enhanced model interpretability (Section 6.4), the above process-oriented interpretations should nevertheless be approached with caution. In data-driven landslide modelling, particularly under data bias conditions, correlation does not necessarily imply geomorphic causation (Steger et al., 2021). An underlying reason for this limitation is the frequent lack of consistent and comprehensive landslide inventory data used for model training. This issue becomes particularly critical for very large-area assessments (e.g. at the European scale), where the integration of heterogeneous datasets from different sources poses additional challenges (Caleca et al., 2025).

Even within comparatively well-studied regions, such as the training domain comprising Austria and South Tyrol, mass movement inventories exhibit considerable inconsistencies and pronounced biases that can influence modelling results (Lima et al., 2021; Steger et al., 2021). In particular, the inventories used in this study are affected by positional inaccuracies, which were addressed through a generalized landscape representation (Section 6.2), as well as by a systematic underrepresentation of slope instabilities that did not result in damage. This limitation substantially reduces their suitability for purely process-oriented mass movement analyses, as events occurring in remote areas are typically not systematically recorded (Trigila et al., 2010; Heiser et al., 2019; Steger et al., 2021). In line with previous modelling approaches focusing on damage-causing landslides (e.g., Steger et al., 2021), the present study is assumed to be less adversely affected by such spatial bias, as it explicitly targets mass movement impacts, which are generally mapped more consistently in the datasets. This shift in focus is expected to enable a more targeted identification of areas where landslides are likely to cause damage. In other words, by redefining the modelling objective, from general mass movement occurrence to areas and time periods likely affected by damaging events, the approach leverages characteristics of the underlying data that would otherwise be considered detrimental biases. Thus, in this modelling context, the focus of the inventories on damage-causing events becomes advantageous rather than limiting, as it enhances, together with a targeted landscape representation, the ability of the approach to systematically identify areas with impact potential."

#4 RC: Moreover, issues such as class imbalance, sampling strategy, and the focus on infrastructure impacts should be more critically evaluated in terms of their influence on model performance and generalization.

The revised manuscript clarifies the sampling design (Section 6.3), including the rationale for starting with a 1:1 ratio at the basin level, ensuring an equal spatial representation of presence and absence locations. While all time periods were initially equally represented, the application of the rainfall filter (deliberately introduced to focus on relevant rainfall days) resulted in a purposeful temporal imbalance. This design choice reflects the distribution of rainy days throughout the year and allows for a meaningful representation of seasonal (day-of-year) effects. These methodological aspects are described in detail in our previous work (Steger et al., 2023), and we therefore prefer to avoid repeating them extensively here, instead referring the reader to that publication. Nevertheless, the revised manuscript will include an expanded discussion of class imbalance and sampling strategy, and their implications for model predictions.

For transparency, we report the underlying class ratios:

- Slide: 2,077 presences vs. 5,823 absences (~36% presences)
- Flow: 1,386 vs. 4,077 (~34%)
- Fall: 647 vs. 3,422 (~19%)

The baseline sample prevalence across our data categories ranges from approximately 19% to 36%. These data represent a moderate class imbalance and fall within ranges commonly encountered in data-driven modelling. They do not fall within the low baseline prevalence conditions (<10%) where balancing datasets has been shown to improve such models (e.g., Steen et al., 2021). However, we explicitly acknowledge in the manuscript that class imbalance affects the interpretation of predicted probabilities, which reflect the prevalence in the training data and are therefore are shifted to lower values in our case.

Revised Methods Section 4.4.3 now reads as follows: *" (…) probability scores do not represent absolute occurrence probabilities, but conditional probabilities influenced by the class balance in the training data (i.e., presence to absence ratio). Higher scores denote (…)"*

Revised section 6.4: *"The lower number of presence observations, together with the more pronounced class imbalance between presences and absences (647 vs. 3,422; 19%), may contribute to the reduced performance of the fall-type model. However, the prevalence of presence observations still remains within a range (>10%) commonly considered sufficient for model fitting (Steen and Tingley, 2021)."*

Newly added reference:

Steen, V. A., Tingley, M. W., Paton, P. W. C., and Elphick, C. S.: Spatial thinning and class balancing: Key choices lead to variation in the performance of species distribution models with citizen science data, Methods in Ecology and Evolution, 12, 216–226, https://doi.org/10.1111/2041-210X.13525, 2021.

More important implications of this class imbalance will be discussed in the new Section 6.5:

"These values do not represent absolute probabilities of impact occurrence, but conditional probabilities influenced by the class balance of the training data. Given that absence observations substantially outnumber presence observations in our dataset, the predicted probabilities are generally biased toward lower values. Moreover, the raw probability scores produced by the different model variants are not directly comparable, as they are derived from datasets with differing presence-to-absence ratios. (…) For practical application in impact-based warning systems, these continuous similarity scores would need to be transformed…"

Finally, the expanded discussion in Section 6.1 (cf. see answer to #3 RC) will highlight implications and advantages of focusing on infrastructure damage rather than general mass movement occurrence, notably by enabling a more consistent use of data that directly represent damaging events, as further discussed in the already cited reference Steger et al. (2021).

#5 RC: Finally, the conclusions should be rewritten to emphasize that the primary contribution lies in a transferable methodological framework rather than a fully validated or operational early warning system. Overall, the manuscript has clear scientific merit, but further revision is required to ensure that the claims are fully consistent with the data quality, uncertainty, and validation constraints.

We agree. The thoroughly revised conclusion will read as follows:

"This study presents a comprehensive, dynamic, and spatially explicit modelling framework for the impact-based assessment of mass movements across large areas. By addressing key limitations of traditional static susceptibility approaches, the framework integrates geo-environmental, exposure, and dynamic meteorological data into a unified modelling strategy. Process-specific models for slide-, flow-, and fall-type movements capture distinct spatio-temporal patterns and provide interpretable outputs that can support risk-informed decision-making within the training domain. The primary contribution of this work lies in the development of a transferable methodological framework rather than an operational early warning system. While the slide- and flow-type models demonstrate encouraging potential for application in an early warning context within the training domain, their operational use remains constrained and requires further development. The fall-type model, in particular, shows

limited sensitivity to meteorological forcing and therefore bearing more the character of an almost static spatial assessment. To our knowledge, this study presents the first implementations of an impact-based mass movement modelling framework that is explicitly process-specific and integrates geo-environmental, exposure, and weather information across scales. Its adaptability to other hazards, compatibility with climate data, and applicability across regions highlight the versatility of the framework and its broader potential for disaster risk reduction and climate impact assessment."