Cover letter

Dear Editors,

We are pleased to submit our revised manuscript entitled “RTSEvo v1.0: A Retrogressive Thaw Slump Evolution Model” for your continued consideration.

We sincerely thank you and the reviewers for the constructive and insightful feedback, which has substantially improved the quality, clarity, and rigor of our work.

In this revision, we have carefully addressed all comments raised by the editor and reviewers. The major structural and scientific enhancements are summarized as follows:

(1) In direct response to the editor’s requirements regarding data accessibility, we have created a persistent Zenodo archive corresponding to the exact GitHub release (v1.0) used in this study.

(2) We introduced a new sensitivity experiment utilizing 30m landsat 8 NDVI data to quantitively evaluate the impact of predictor resolution. We also added a dedicated discussion (Section 4.3) showing the robustness of the spatial allocation module against coarse-resolution inputs.

(3) We introduced an independent validation region (the Haiding Nuo’er Subregion) and conducted a new cross-regional experiment to rigorously assess model transferability (Section 3.4).

(4) We expanded our discussion (Section 4.4) to explicitly address the spatial heterogeneity and clarify the regional dependence of the calibrated parameters.

Additionally, please note an update regarding the corresponding authors. Since the initial submission of this work, Shuping Zhao and I have moved to Shanghai Normal University. We have updated our author affiliations and contact information in the revised manuscript to reflect this change.

A point-by-point response is appended below. Thank you again for your time and continued editorial handling of our manuscript.

Sincerely,

Zhuotong Nan,

On behalf of all authors

Response to Review Comment #1

The manuscript presents "RTSEvo v1.0," a dynamic evolution model for Retrogressive Thaw Slumps (RTS). The authors address a critical gap in current research by advancing from static susceptibility mapping to spatiotemporal simulation. I find the methodological framework to be highly innovative. This hybrid approach offers valuable insights into the morphological evolution of RTS and significantly contributes to our understanding of abrupt thaw processes and permafrost degradation mechanisms on the Qinghai-Tibet Plateau. The study is well-structured and tackles a timely issue in permafrost science. However, I have several concerns regarding the spatial scale of the input data, the transferability of the model, and the standardization of the source code that need to be addressed before publication.

Response：We sincerely thank the reviewer for the time dedicated to reviewing our work and for the positive evaluation of our manuscript. Your constructive and insightful comments, particularly regarding model transferability and scale-mismatch, were incredibly valuable. We have conducted new cross-regional experiments and validations that substantially improved the scientific rigor and clarity of this study. We have thoroughly revised both the manuscript and the source code to fully address all your concerns.

1. Scale Mismatch and Resolution Limitations: Retrogressive Thaw Slumps (RTS) are local-scale periglacial landforms. In this study, the selected RTS features have an average area of 2.61 ha, which corresponds to an approximate diameter of 160 m (assuming a square geometry). However, the spatial resolution of several predictor variables used in the model is relatively coarse; for instance, the NDVI data is at 250 m resolution. Even utilizing 30 m

resolution datasets may be insufficient for simulating such small-scale geomorphological processes. Consequently, this "scale gap" represents a significant constraint on the model's precision. I strongly recommend that the authors include a serious discussion regarding this limitation in the manuscript.

Response: We sincerely thank the reviewer for raising this critical point. You are absolutely correct that the inherent scale gap between small scale RTS features and coarse predictors poses a challenge for high-precision modeling. Traditionally, this gap will result in overly smooth and may lead to unrealistic hazard zones.

Your comment inspired us to rigorously evaluate and explicitly discuss how our model handles this scale mismatch. We have conducted a new quantitative sensitivity experiment and added a dedicated discussion section. These additions demonstrate how our framework mitigates this scale gap issues.

- We added a new subsection (Section 4.3: Predictor resolution and the scale-mismatch challenge) to explicitly acknowledge the inherent discrepancy between RTS feature size and coarse predictors.
- We conducted an additional sensitivity experiment focusing on vegetation (NDVI), a critical modulator of ground thermal regimes, to physically test this constraint. We replaced the 250 m MODIS NDVI with a reconstructed, higher-resolution 30 m Landsat 8 NDVI dataset for the 2021 independent simulation.
- The results revealed only marginal performance improvements (FoM increases of 0.53% for LR-EM and 1.16% for RF-EM) when replacing 250m NDVI with 30 NDVI. This confirms that RTSEvo effectively isolates the smoothing effect of coarse predictors by delegating the fine-scale, physical constraints to the high-resolution spatial allocation module.

We put the resulting figures (Figures S4 and S5) to the supplementary information to maintain the flow of the main text and provide evidence for this concern.

The following text has been added to the revised manuscript (Lines 549-574):

"4.3 Predictor resolution and the scale-mismatch challenge

A practical challenge in regional permafrost modeling is the scale mismatch between local geomorphological features and the coarse resolution of continuous satellite-derived environmental predictors. In the Beiluhe Basin, individual RTS features have a mean area of 2.61 ha, corresponding to an approximate diameter of 160 m. When predictors like MODIS NDVI (250m) or reanalysis climate forcing are resampled to 10 m modeling grid, sub-pixel spatial heterogeneity is not explicitly represented. In traditional susceptibility mapping, this scale gap typically results in overly smooth, unrealistic hazard zones that fail to capture the sharp boundaries of actual terrain failure.

RTSEvo is designed to mitigate this scale gap through its two-stage modeling structure, which mathematically isolates the smoothing effect. In the first stage, the probability estimation module uses regional-scale predictors merely to identify permissive bioclimatic areas where RTS expansion is environmentally plausible. In the second stage, the constrained spatial allocation module determines how RTS expansion is actually expressed at high spatial resolution. This allocation process is controlled by neighborhood interactions and RTS expansion rules, allowing fine-scale spatial patterns to emerge even when the background predictors are spatially smooth.

To rigorously validate this structural merit, we focused our sensitivity testing on vegetation (NDVI). Vegetation is a critical modulator of ground thermal regimes and was confirmed as a meaningful predictor in our SHAP analysis (Figure A4). We tested whether its coarse representation was bottlenecking

our simulation. Upgrading the vegetation input from 250m MODIS to a reconstructed 30 m Landsat 8 NDVI dataset for the 2021 independent simulation in the Beiluhe Basin yielded only marginal performance gains (<1.2% FoM increase) (Figures S4, S5). This counterintuitive finding reveals that while finer-resolution vegetation data naturally enhances local contrast, vegetation density only indicates general environmental susceptibility. Because RTSEvo explicitly delegates physical constraints to the high-resolution spatial allocation module, it prevents the smoothing effect of coarse predictors from destabilizing the final simulation.

This finding carries significant operational implications for permafrost geohazard forecasting. High-resolution imagery (e.g., 30 m or finer) on the QTP frequently suffers from severe temporal limitations, including frequent cloud contamination and irregular acquisition intervals. Our results demonstrate that regional simulation efforts do not need to be bottlenecked by the absence of high-resolution datasets. By combining coarse but temporally consistent environmental data with strict, physically based allocation rules, researchers can generate highly realistic, fine-scale evolutionary projections."

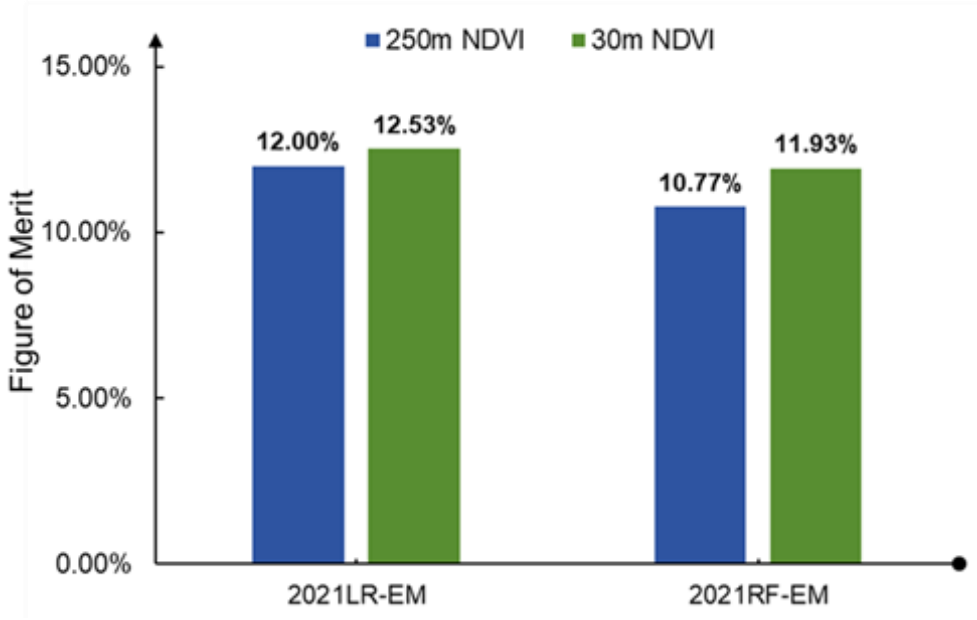

Figure S4. The impact of varying vegetation predictor resolutions on model

predictive performance in the Beiluhe Basin (2021).

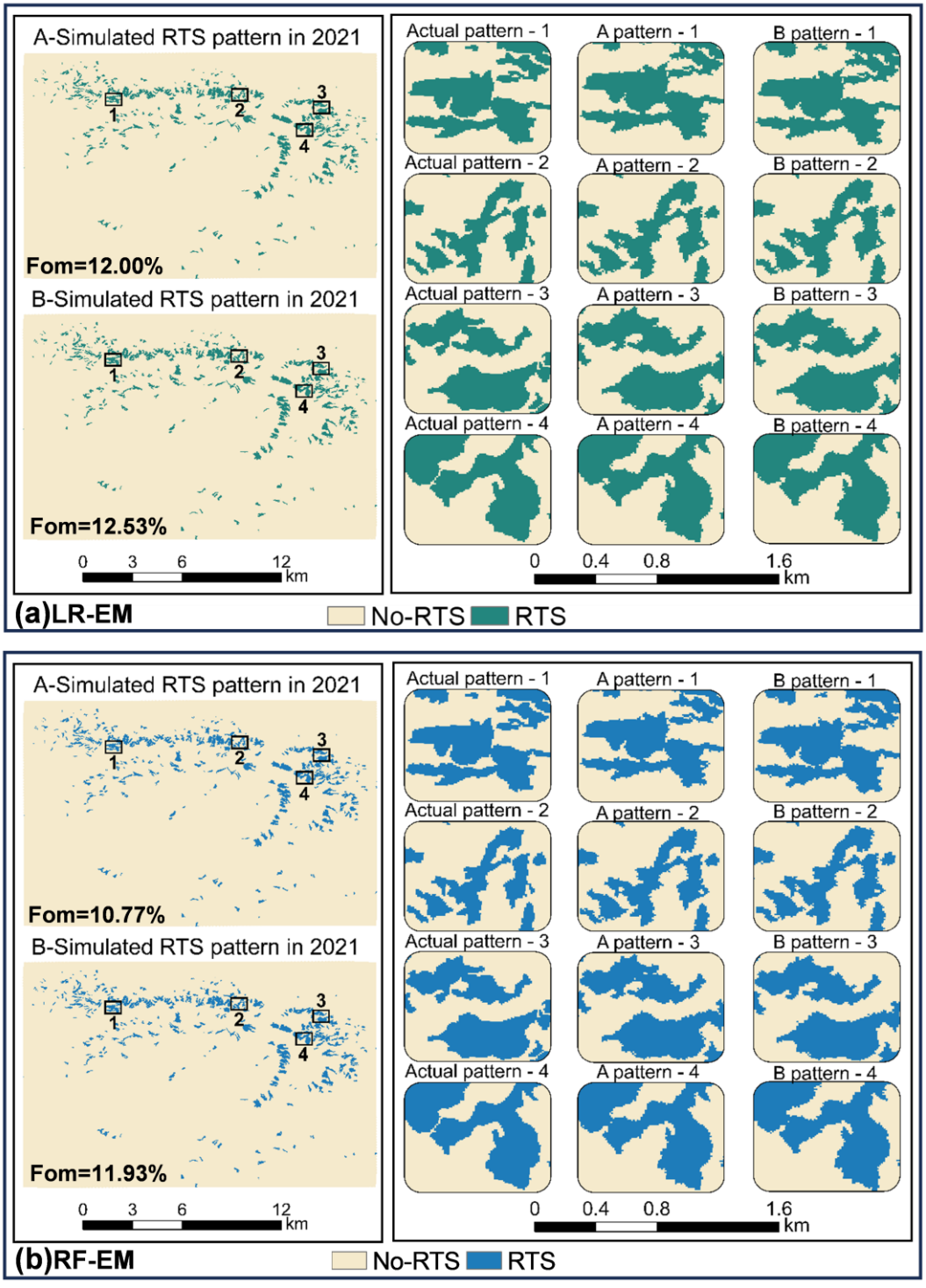

Figure S5. Pixel-level comparison of simulated RTS spatial distributions using different NDVI resolutions in the Beiluhe Basin for the 2021 validation period. (a) The regional and sub-regional results for the LR-EM, and (b) the corresponding results for the RF-EM. "A pattern" refers to the simulated

expansions driven by the 250 m MODIS NDVI, and “B pattern” for the higher-resolution 30m Landsat 8 NDVI. The rectangular insets (numbered 1-4) represent four randomly selected sub-regions of intense RTS expansion.

2. Model Transferability and Generalizability The study constructs the RTSEvo model by combining physical process characterization with machine learning parameter calibration. While the model was validated using over 450 RTS inventory records with positive results, the spatial distribution of these samples is highly concentrated within the Beiluhe Basin. This spatial clustering implies that the climatic conditions, which are key drivers used to predict RTS state and evolution，are highly homogeneous across the training and validation datasets. Therefore, it is questionable whether the parameters calibrated in this specific region are transferable to other permafrost regions with different environmental settings. The authors must strictly clarify the model's transferability and discuss this potential lack of generalizability in the text.

Response: We sincerely thank the reviewer for raising this fundamental and highly insightful point. You are entirely correct that a single Beiluhe Basin naturally raises valid questions regarding the transferability of these specific parameters to other regions. To rigorously address your concern, we have introduced a completely independent test regions into the study and conducted a dedicated cross-region transferability experiment. This allowed us to explicitly assess the model’s performance against different environment controls and directly answer your question. We have detailed these major additions across the manuscript as follows:

- Introduction of a new test region (Section 2.6): We introduced the Haiding Nuo'er Subregion as a secondary test region specifically to independently validate model transferability. To provide a meaningful contrast to the

Beiluhe Basin, this region features distinct environmental conditions, including lower elevations, reduced vegetation greenness, higher silt content, and greater thawing degree days.

- Design of generalization experiments (Section 2.7): We added "Experiment 4: Cross-regional transferability testing" to our methodology. We evaluated the model under two specific scenarios: (1) direct parameter transfer, where parameters optimized in the Beiluhe Basin were applied to the Haiding Nuo'er Subregion without modification; and (2) local recalibration, where parameters were re-optimized using the regional RTS inventory of the Haiding Nuo'er Subregion.
- Results of transferability verification (Section 3.4): While direct parameter transfer predictably yielded lower accuracy baselines (e.g., FoM values of 15.81%, 5.27%, and 6.10% for 2020, 2021, and 2022) compared to a locally tuned model, the physically constrained spatial allocation rules ensured that the simulated expansion patterns remained physically plausible and structurally sound, rather than degrading into random noise. Local recalibration then effectively optimized this baseline performance. This confirms that while the architecture is robust, specific spatial allocation parameters must still be recalibrated to achieve maximum cross-regional accuracy.
- Discussion on model transferability (Section 4.4): We added a discussion emphasizing that this parametric sensitivity is not a structural limitation of the algorithm, but a direct reflection of the spatial heterogeneity inherent to permafrost landscapes. We clarified that while the universal physical rules encoded in the model (such as retrogressive erosion) remain valid across all environments, the statistical weights governing their local spatial expression cannot be universally transferred. We conclude by advocating for a "regionalized" parameterization paradigm to achieve reliable large-scale projections.

The following text has been added to the revised manuscript:

Lines 209-218: "*To independently validate model transferability, the Haiding Nuo'er Subregion (extending from 92.889°E to 93.167°E and from 35.396°N to 35.469°N) is introduced as a secondary test region (Figure 2c). The Haiding Nuo'er Subregion is characterized by alpine meadow and alpine grassland ecosystems, similar in general land cover to the Beiluhe Basin, but it exhibits distinct environmental conditions. The number of mapped RTS features in this region increased from 52 to 92 between 2016 and 2022, indicating active thermokarst development. Compared to the Beiluhe Basin, RTS occurrences in the Haiding Nuo'er Subregion are associated with lower elevations and reduced vegetation greenness, as well as higher silt content in the 30–60 cm soil layer and greater thawing degree days (Figure S1). These differences reflect variations in topographic setting, surface conditions, soil composition, and thermal forcing, which are key controls on permafrost stability and RTS initiation. Such differences provide a valuable opportunity to examine the model's applicability and generalizability under varying environmental controls.*"

Lines 307-314: "*Experiment 4: Cross-regional transferability testing. A key challenge in regional geohazard modeling is whether model parameters calibrated in one geographic setting can be applied to other areas with differing environmental conditions. We designed a cross-regional transferability experiment using the Haiding Nuo'er Subregion as an independent test domain. The experiment was conducted under two scenarios: (1) direct parameter transfer, in which spatial allocation parameters optimized in the Beiluhe Basin were applied without modification to simulate RTS evolution in the Haiding Nuo'er Subregion; and (2) local recalibration, in which parameters were re-optimized using the regional RTS inventory of the Haiding Nuo'er Subregion as the calibration reference. Identical data sources and preprocessing workflows were applied in both scenarios to ensure*

methodological consistency and to isolate the effect of parameter transferability from differences in data preparation."

Lines 448-467: "***3.4 Model transferability and cross-regional generalizability***

Evaluating the model's cross-regional generalizability revealed the physical limitations of direct parameter transfer across heterogeneous permafrost environments (Experiment 4). Baseline parameters derived from the highly concentrated RTS samples in the Beiluhe Basin proved strictly optimized for that region's specific climate forcing. When testing the Holt's linear trend model in the Haiding Nuo'er Subregion, an independent test area, the module achieved strong areal demand forecasting performance (R^2=0.913, MAPE=13.75%). However, this regional forecast required a higher smoothing coefficient (α=0.8401) and a lower trend smoothing coefficient (β^*=0.4330) relative to the Beiluhe Basin. These parameter differences indicate that the temporal dynamics and climatic responsiveness of RTS expansion differ meaningfully between the two regions.

During the spatial allocation phase, the base occurrence probability maps generated by the LR module for the Haiding Nuo'er Subregion (Figure S3) revealed that high probability zones are predominantly distributed in elongated, ribbon-like patterns adjacent to existing RTS features. Applying the direct transfer of Beiluhe Basin spatial allocation parameters to the Haiding Nuo'er Subregion yielded baseline FoM values of 15.81%, 5.27%, and 6.10% for 2020, 2021, and 2022, respectively. Conversely, locally recalibrating the parameters based on the 2020 RTS distribution of the secondary site produced notable performance improvements, increasing the FoM to 18.15% in 2020 and 6.25% in 2021, although a modest decrease to 4.29% was observed in 2022 (Figure 5).

These performance outcomes demonstrate that the RTSEvo architecture remains fully operational and capable of simulating RTS evolution when

deployed in a different permafrost region. At the same time, the consistent accuracy gains achieved through local recalibration prove that spatial allocation parameters exhibit limited cross-regional generalizability. This finding highlights a reality of permafrost geomorphology: local topography, soil composition, and thermal forcing heavily modulate RTS expansion, necessitating region-specific optimization to achieve maximum predictive accuracy.”

Lines 575-591: “***4.4 Model transferability and regional permafrost heterogeneity***

The parametric sensitivity observed during our cross-regional validation is not a structural limitation of the RTSEvo algorithm, but rather a direct reflection of the profound spatial heterogeneity inherent to permafrost landscapes. Because RTS expansion is fundamentally a phase-change driven geomorphic process, regional variations in subsurface thermal regimes, active layer thickness, and ground ice distribution physically alter how a slump propagates. Consequently, a parameter set highly optimized for the specific ice-rich, high-elevation environment of the Beiluhe Basin naturally misaligns with the distinct climatic and topographic boundary conditions of the Haiding Nuo’er Subregion. This confirms that while the universal physical rules encoded in the model (such as retrogressive erosion) remain valid across all environments, the statistical weights governing their local spatial expression cannot be universally transferred.

This physical reality exposes a critical vulnerability in the growing trend of applying monolithic, "one-size-fits-all" machine learning frameworks to pan-Tibetan or pan-Arctic permafrost simulations. Purely data-driven models trained on concentrated regional inventories risk severe spatial uncertainty and catastrophic predictive failure when extrapolated to novel environments because they conflate universal failure mechanisms with local environmental sensitivities. Our results prove that by mathematically separating universal

physical mechanisms (the spatial allocation rules) from local environmental sensitivities (the occurrence probability weights), dynamic models developed a structural resistance to this deterioration and can be successfully ported across regions. However, to achieve reliable continental-scale projections, the permafrost modeling community still have to adopt a "regionalized" parameterization paradigm.”

We have also updated Figure 2 to show the location of the new test region, added Figure S1 to the Supplementary Information to detail the distinct environmental variables, and added Figure 5 to visually display the transferability results.

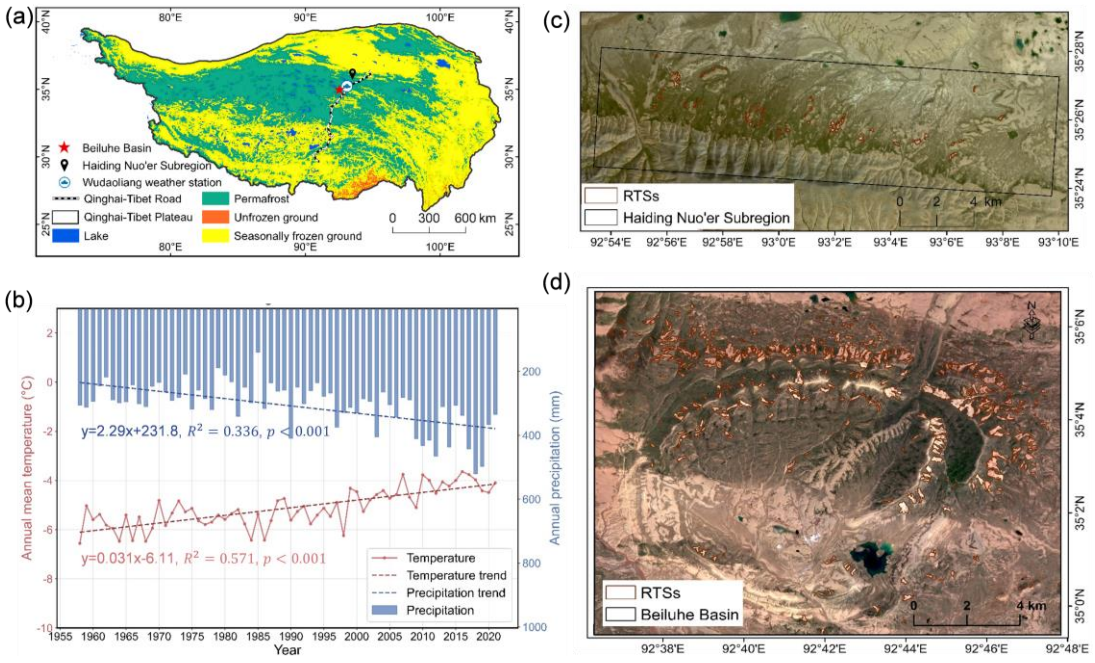

Figure 2. Location of the primary study area (Beiluhe Basin) and the transferability test region (Haiding Nuo'er Subregion), alongside regional climate trends and RTS distribution. (a) Permafrost distribution on the QTP, with the primary study area marked by a red star and the Haiding Nuo’er Subregion marked by a black pin. Permafrost data are from Cao et al. (2023). (b) Air temperature and precipitation records from the nearby Wudaoliang weather station (1955-2021), with trend lines illustrating regional warming and humidification. (c-d) 2022 PlanetScope satellite images (Planet Team, 2025)

of the Haiding Nuo'er Subregion and the Beiluhe Basin, respectively, overlaid with mapped RTS boundaries from Xia et al. (2024a).

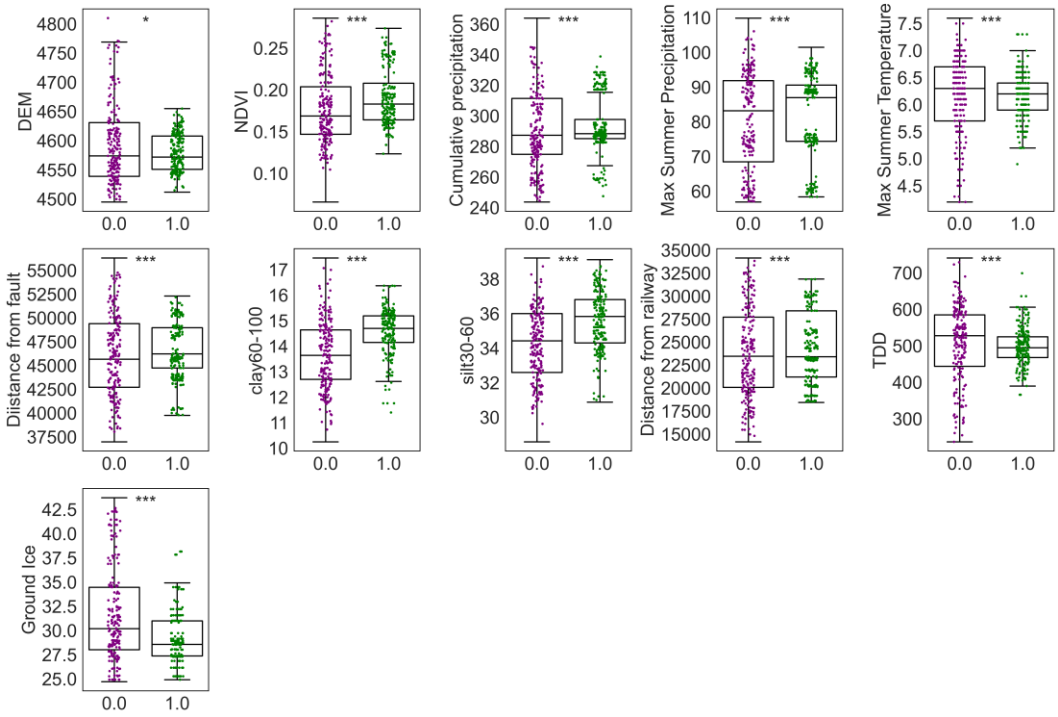

Figure S1. Distribution of numerical predictor variables for RTS versus non-RTS locations within the independent transferability test region (Haiding Nuo'er Subregion). The box-and-whisker plots compare the distributions for zones where RTSs occurred (RTS=1) and where they did not (RTS=0). Significance levels between groups, determined by the Mann-Whitney U test, are indicated as follows: ns, not significant; *p < 0.05; **p < 0.01; ***p < 0.001.

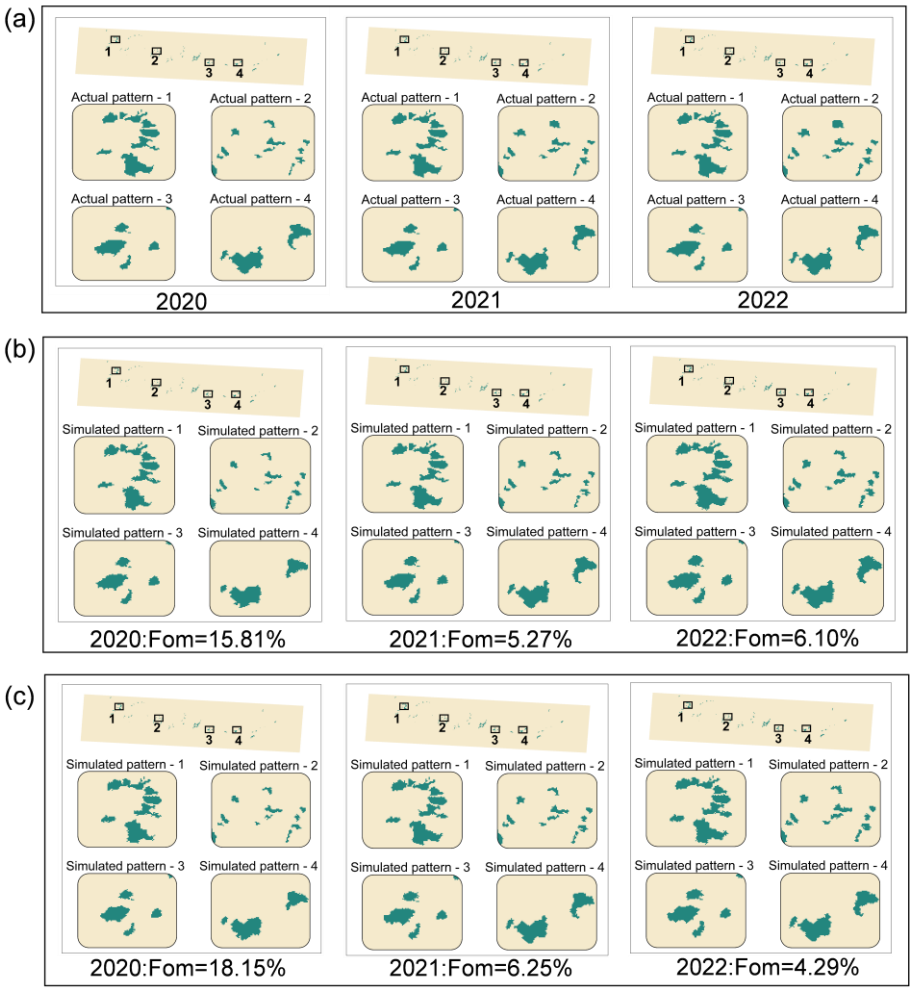

Figure 5. Comparison of simulated and observed RTS spatial distributions in the independent transferability test region (Haiding Nuo'er Subregion) for the 2020-2022 period. (a) Observed RTS spatial distributions based on historical inventories. (b) Simulated RTS spatial distributions using parameters directly transferred from the primary Beiluhe Basin without modification. (c) Simulated RTS spatial distributions using parameters locally recalibrated specifically for the Haiding Nuo'er Subregion.

3. Codes: Geoscientific Model Development (GMD) mandates that code be open-source and sufficiently documented to facilitate reuse by the community (and for the purpose of peer review). Upon reviewing the provided repository. I found that the current documentation is non-standard and lacks rigor. Notably, the code contains comments in languages other than English (e.g., Chinese characters are visible in comments. I strongly suggest the authors

perform a thorough revision of the code and documentation to ensure it meets international standards and undergoes rigorous testing before final publication.

Response: Thank you very much. We have conducted a comprehensive overhaul of the RTSEvo codebase and its accompanying documentation. Specifically, we implemented the following improvements:

- All non-English annotations (including all Chinese characters) have been completely removed. Some redundant or self-explanatory comments were deleted, while all scientifically and algorithmically important annotations were translated into clear technical English.
- We have improved the repository documentation. The revised README.md file now provides a rigorous step by step guide. This includes system requirements, a clear outline of the modular workflow, data preparation instructions, and detailed explanations of all key parameters to facilitate reuse by other researchers.
- The updated codebase was rigorously re-tested across different environments to ensure stability and to verify that the community can accurately reproduce the results presented in this manuscript.

The fully revised source code for the thaw slump evolution model is publicly available on GitHub (https://github.com/nanzt/RTSEvo) and the exact version (v1.0) used to generate the results presented here is permanently archived on Zenodo (https://doi.org/10.5281/zenodo.17850641).

4. Figure A2: Regarding the final plot in Figure A2 (Active Layer Thickness): Why does the distribution exhibit a bimodal pattern with values clustered at two extremes (showing a difference of up to 1 m)? Please clarify the physical or data-processing reason for this distinct separation.

Response：We thank the reviewer for this keen observation. We can confirm that the distinct bimodal distribution of ALT shown in Figure A2 is a real and physical meaningful phenomenon, rather than a data-processing artifact.

This bimodal pattern directly reflects a physical bifurcation between two dominant ground thermal regimes in the Beiluhe Basin. Specifically, undisturbed areas with intact vegetation and thick organic layers tend to maintain relatively shallow active layers due to strong thermal insulation. Meanwhile, disturbed/RTS-affected surfaces experience a drastic reduction in insulation and an enhanced ground heat flux, leading to substantially deeper thaw depths.

We cross-referenced independent field observations from adjacent sites along the Qinghai-Tibet Railway (Zhao et al. 2021). As shown in Figure R1 (provided below for you reference), in-situ measurements demonstrate a sharp contrast in ALT across local sites, with values ranging from roughly 120 cm to 300 cm.

We have clarified this physical mechanism in the revised manuscript. The following text has been added (Lines 243-249):

"In particular, the active layer thickness (ALT) exhibits a clear bimodal distribution (Figure A2). The bimodal distribution reflects the presence of two dominant ground thermal regimes in the Beiluhe Basin. Continuous vegetation and organic layers maintain a shallow active layer through strong thermal insulation, whereas RTS-affected surfaces experience abrupt deepening of the active layer due to vegetation loss, reduced insulation, and enhanced soil heat flux (Wang et al., 2018; Yi et al., 2018). Furthermore, ALT observation data from adjacent sites along the Qinghai-Tibet Railway have also shown significant differences ranging from 120 cm to 300 cm (Zhao et al., 2021), corroborating the physical reality of this distinct separation."

Note, Figure R1 is not included in the revised text and provided here only for reviewer's reference. We have cited the relevant references in the main text.

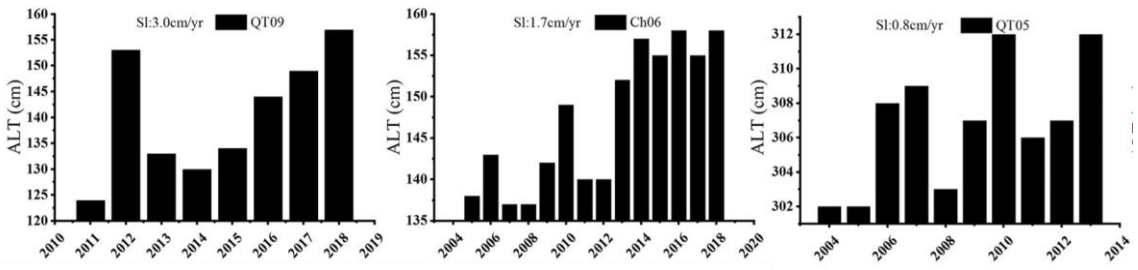

Figure R1. Active layer thickness at different adjacent stations along the Qinghai-Tibet highway (Zhao et al., 2021).

References

Cao, Z., Nan, Z., Hu, J., Chen, Y., and Zhang, Y.: A new 2010 permafrost distribution map over the Qinghai–Tibet Plateau based on subregion survey maps: a benchmark for regional permafrost modeling, Earth Syst. Sci. Data, 15, 3905-3930, https://doi.org/10.5194/essd-15-3905-2023, 2023.

Planet Team: Planet application program interface: In space for life on earth [dataset], Retrieved from https://api.planet.com, 2025.

Wang, Y., Sun, Z., and Sun, Y.: Effects of a thaw slump on active layer in permafrost regions with the comparison of effects of thermokarst lakes on the Qinghai–Tibet Plateau, China, Geoderma, 314, 47-57, https://doi.org/10.1016/j.geoderma.2017.10.046, 2018.

Xia, Z., Liu, L., Mu, C., Peng, X., Zhao, Z., Huang, L., Luo, J., and Fan, C.: Annual inventories of retrogressive thaw slumps across the Qinghai-Tibet Plateau from 2016 to 2022, Zenodo [dataset], https://doi.org/10.5281/zenodo.10928346, 2024.

Yi, Y., Kimball, J. S., Chen, R. H., Moghaddam, M., Reichle, R. H., Mishra, U., Zona, D., and Oechel, W. C.: Characterizing permafrost active layer dynamics and sensitivity to landscape spatial heterogeneity in Alaska, Cryosphere, 12, 145-161, https://doi.org/10.5194/tc-12-145-2018, 2018.

Zhao, L., Zou, D., Hu, G., Wu, T., Du, E., Liu, G., Xiao, Y., Li, R., Pang, Q., Qiao, Y., Wu, X., Sun, Z., Xing, Z., Sheng, Y., Zhao, Y., Shi, J., Xie, C., Wang, L., Wang, C., and Cheng, G.: A synthesis dataset of permafrost thermal state for the Qinghai–Tibet (Xizang) Plateau, China, Earth Syst. Sci. Data, 13, 4207-4218, https://doi.org/10.5194/tc-12-145-2018/10.5194/essd-13-4207-2021, 2021.

Response to Review Comment #2

Xu et al.'s work fills a critical bottleneck in permafrost research by moving from static susceptibility assessment of RTS to dynamic modeling which can be used to predict the future development of RTS. This represents an outstanding contribution to the field. Overall, the proposed framework is well constructed, the modeling strategy is clearly implemented, and the results show reasonable skill in reproducing the observed patterns of RTS development. However, I also have some concerns regarding scale, sampling, and long-term validity require further clarification.

Response：We sincerely thank the reviewer for the highly positive evaluation of our manuscript and for recognizing our work as an "outstanding contribution" to the field. We also thank you very much for your constructive and rigorous feedback regarding upscaling issue, sampling strategies, and the long-term validity of our modeling assumptions. We have carefully addressed all of your concerns in the revised manuscript, and our detailed point-by-point response are provided below.

1. The model operates at a 10 m spatial resolution, while many key driving variables are resampled from much coarse resolutions (e.g., 1km). Please discuss how this resampling affects the precision of the probability estimation and the subsequent spatial allocation module.

Response: We thank the reviewer for raising this profound methodological question. You highlight a fundamental challenge in regional modeling, the inherent uncertainty introduced by interpolating coarse driving variables down to a fine scale modeling grid. We try our best to address your concern as summarized below:

1. Why we run on 10m modeling resolution? Our environmental predictors are relatively coarse, but RTS expansion is a highly localized geomorphic

process (average RTS diameter in our study is ~160m). If we ran the model at 250 m or 1km resolutions to match the predictors, a single pixel would cover entire slumps. It's physically impossible to simulate the geometric boundaries, neighborhood interactions, or specific headwall retreat that defines RTS evolution, at those coarse scales. Therefore, a 10 grid is necessary even if it requires resampling coarser environmental inputs.

2. We acknowledge the resampling 1km or 250m variables to a 10 m grid introduces spatial uncertainty, because sub-pixel heterogeneity is lost at these coarse resolutions. But at a strict 10 m scale, the probability estimation also cannot resolve micro-scale variations. The probability module running at coarse resolution actually is ok to identify broad, permissive bioclimatic areas where expansion is environmentally supported.

3. In our model framework, a spatial allocation module follows the probability module. The allocation module was explicitly designed to compensate for this resampling uncertainty. While the background environmental drivers are smoothed, the spatial allocation rules (such as retrogressive erosion factor, neighborhood state, etc) operate at the 10m scale. Thanks to parameter calibration, this allows the model to capture realistic, high-resolution expansion patterns.

To quantitatively prove that this resampling does not destabilize the spatial allocation module, we conducted a new sensitivity experiment. We replaced the 250 m MODIS NDVI with a reconstructed 30 m Landsat 8 dataset (which still requires resampling to 10 m, but with much less interpolation). The results showed only marginal performance improvements (FoM increases of just 0.53% for LR-EM and 1.16% for RF-EM). This confirms that the model's predictive skill is anchored by the 10 m physical allocation rules, making it robust against the uncertainties of resampled coarse predictors.

We have explicitly discussed the effects of resampling and the scale gap in a new dedicated section (Section 4.3: Predictor resolution and the scale-mismatch challenge)

The following text has been added to the revised manuscript (Lines 549-574):

*"**4.3 Predictor resolution and the scale-mismatch challenge***

A practical challenge in regional permafrost modeling is the scale mismatch between local geomorphological features and the coarse resolution of continuous satellite-derived environmental predictors. In the Beiluhe Basin, individual RTS features have a mean area of 2.61 ha, corresponding to an approximate diameter of 160 m. When predictors like MODIS NDVI (250m) or reanalysis climate forcing are resampled to 10 m modeling grid, sub-pixel spatial heterogeneity is not explicitly represented. In traditional susceptibility mapping, this scale gap typically results in overly smooth, unrealistic hazard zones that fail to capture the sharp boundaries of actual terrain failure.

RTSEvo is designed to mitigate this scale gap through its two-stage modeling structure, which mathematically isolates the smoothing effect. In the first stage, the probability estimation module uses regional-scale predictors merely to identify permissive bioclimatic areas where RTS expansion is environmentally plausible. In the second stage, the constrained spatial allocation module determines how RTS expansion is actually expressed at high spatial resolution. This allocation process is controlled by neighborhood interactions and RTS expansion rules, allowing fine-scale spatial patterns to emerge even when the background predictors are spatially smooth.

To rigorously validate this structural merit, we focused our sensitivity testing on vegetation (NDVI). Vegetation is a critical modulator of ground thermal regimes and was confirmed as a meaningful predictor in our SHAP analysis (Figure A4). We tested whether its coarse representation was bottlenecking

our simulation. Upgrading the vegetation input from 250m MODIS to a reconstructed 30 m Landsat 8 NDVI dataset for the 2021 independent simulation in the Beiluhe Basin yielded only marginal performance gains (<1.2% FoM increase) (Figures S4, S5). This counterintuitive finding reveals that while finer-resolution vegetation data naturally enhances local contrast, vegetation density only indicates general environmental susceptibility. Because RTSEvo explicitly delegates physical constraints to the high-resolution spatial allocation module, it prevents the smoothing effect of coarse predictors from destabilizing the final simulation.

This finding carries significant operational implications for permafrost geohazard forecasting. High-resolution imagery (e.g., 30 m or finer) on the QTP frequently suffers from severe temporal limitations, including frequent cloud contamination and irregular acquisition intervals. Our results demonstrate that regional simulation efforts do not need to be bottlenecked by the absence of high-resolution datasets. By combining coarse but temporally consistent environmental data with strict, physically based allocation rules, researchers can generate highly realistic, fine-scale evolutionary projections."

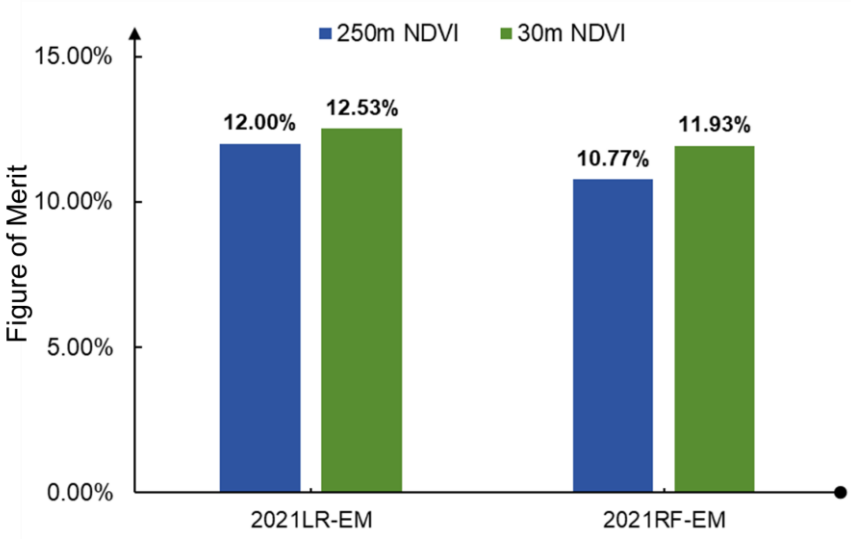

Figure S4. The impact of varying vegetation predictor resolutions on model predictive performance in the Beiluhe Basin (2021).

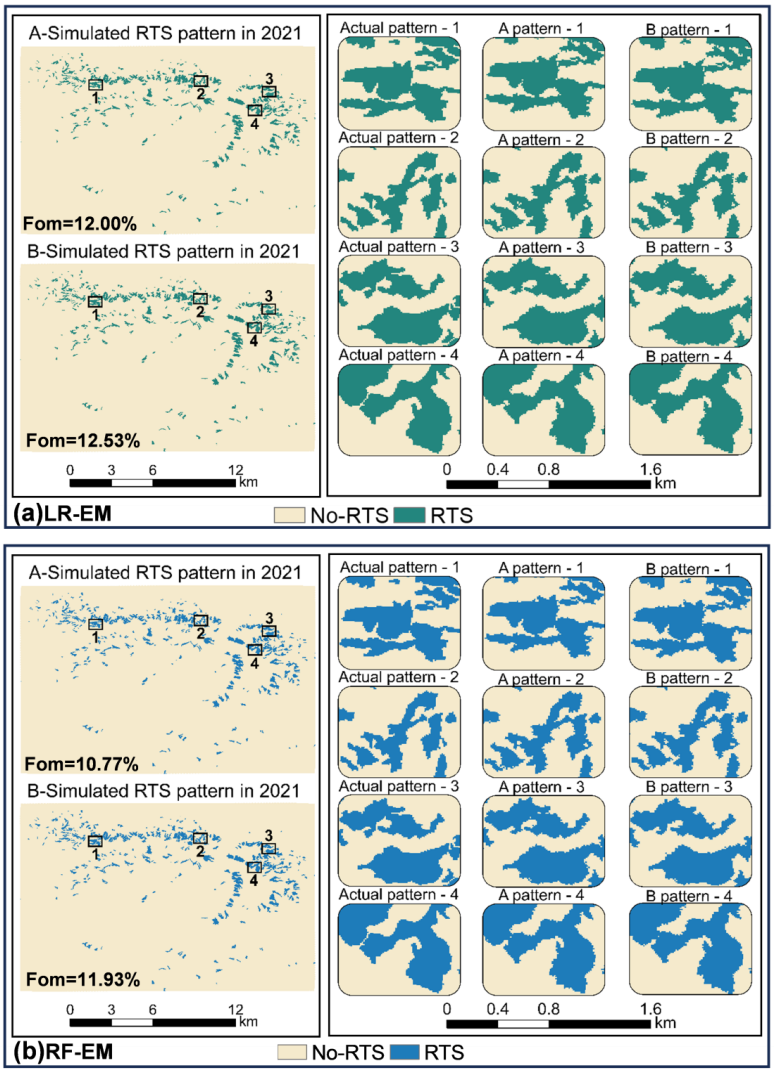

Figure S5. Pixel-level comparison of simulated RTS spatial distributions using different NDVI resolutions in the Beiluhe Basin for the 2021 validation period. (a) The regional and sub-regional results for the LR-EM, and (b) the corresponding results for the RF-EM. "A pattern" refers to the simulated expansions driven by the 250 m MODIS NDVI, and “B pattern” for the higher-resolution 30m Landsat 8 NDVI. The rectangular insets (numbered 1-4) represent four randomly selected sub-regions of intense RTS expansion.

2. The FoM values appear numerically not very high. I know this framework adapted from LUCC modeling, please provide a comparison indicating if these

metrics are consistent with or superior to those reported in related LUCC or other geohazard evolution studies.

Response: Thank you very much. At first glance, FoM values appear low compared to traditional accuracy metrics like Kappa or F1 scores. Standard accuracy metrics are heavily inflated by true negatives, that is, the massive expanses of landscape that remain stable and are easy to prediction. So, FoM evaluates only the intersection of simulated change and observed change, completely excluding correctly predicted stable areas. Because RTS expansion is a highly localized and rare geomorphic processes, achieving a high FoM is very difficult.

Our FoM metric is highly consistent with and often superior to established benchmarks in the literature. In comparable LUCC simulation studies, FoM values typically range between 2% to 25% (Pontius et al. 2008, Figure S7), with the lower end being common for processes like landslides or slumps. Our achieved FoM values up to 20.05% during calibration and 12.00% in independent validation, actually represent highly competitive predictive skill.

Furthermore, we emphasize that the success for RTSEvo is the relative improvement generated by our physical rules. Compared to baseline, probability-only simulations, integrating the retrogressive erosion factor increased the FoM by over 29%. It confirms that the dynamic allocation framework is successfully capturing the physical expansion mechanics.

We added the following text to the revised manuscript (Lines 400-405):

"*In comparable land-use and land-cover change (LUCC) simulation studies, FoM values commonly range between approximately 2% and 25%, particularly for rare or spatially clustered change processes (Pontius et al., 2008). Within this context, the FoM values obtained by RTSEvo are highly competitive and indicate a highly effective capture of the underlying expansion mechanisms. Furthermore, the model demonstrates a substantial relative*

improvement compared to baseline probability-only simulations, with FoM increases exceeding 29% when our dynamic allocation process-based rules are included."

Here, we also provided Figure R2 from Pontius et al. 2008 below for the reviewer's reference, which shows 2-25% FoM range across various established LUCC models. Note, this figure is not included in the revised manuscript.

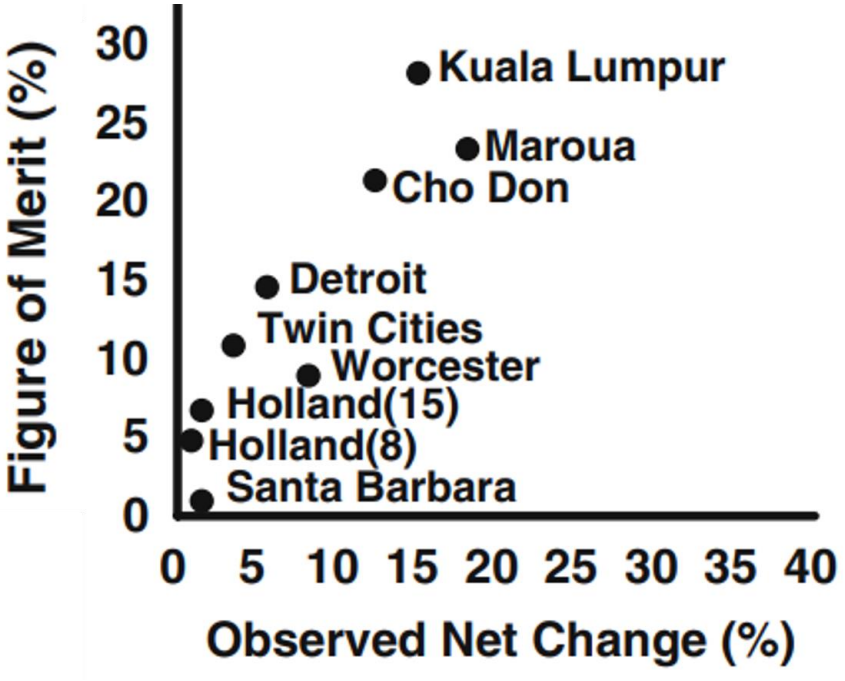

Figure R2. FoM values of different LUCC models in various cities (Pontius et al., 2008).

3. The model assumes that RTS state transitions are irreversible (line 459), which is noted as a limitation for long-term projections. Please clarify the physical justification for maintaining the assumption in v1.0 and discuss if a recovery state could be integrated in future versions. Furthermore, what is the specific temporal horizon (e.g., 10 vs. 50 years) over which the current model remains a reliable predictive tool?

Response: Thanks. We wish to explicitly clarify the intended operational scope of our model. RTSEvo was developed and purposed for short- to

medium-term geohazard forecasting (approx. decade cales, < 20 years). It was not designed for long-term, century-scale evolutionary projections. Because RTSEvo is for near-future simulation, the irreversibility assumption is not just a simplification, but a valid and physically appropriate design choice for this specific timeframe.

Extensive field observations and remote-sensing studies consistently indicate that the complex geomorphic processes of RTS stabilization, revegetation, and full surface recovery generally require multi-decadal timescales (often greatly exceeding 20 years) (Burn and Friele, 1989; Leibman et al., 2014; Nesterova et al., 2024).. Therefore, within our targeted 10-to-20-year simulation window, the irreversibility assumption closely mimics physical reality.

However, we fully agree that if the model were to be applied outside its intended scope, like for long-term temporal horizons (>30-50 years), maintaining this irreversibility assumption would lead to a systematic overestimation of total active RTS extent. To address your point, subsequent versions will resolve this by transitioning from a binary classification to a multi-class state transition framework. We make it clear in the revised manuscript (Lines 609-618):

"*Furthermore, the framework assumes that once a pixel converts to an RTS, it irreversibly remains in that state. This assumption is physically justified over short- to medium-term simulations. Field observations and remote-sensing studies consistently indicate that the processes of RTS stabilization, revegetation, and surface recovery typically operate over multi-decadal timescales, often exceeding 20 years, and are strongly modulated by local climate, sediment availability, and hydrological conditions (Burn and Friele, 1989; Leibman et al., 2014; Nesterova et al., 2024; Krautblatter et al., 2024). However, for extended temporal horizons (e.g., >30-50 years), neglecting stabilization and recovery processes may lead to a systematic overestimation*

of active RTS extent. Future iterations of the model must incorporate multi-class state transitions to simulate the recovery state of the RTS.”

4. The use of 1:1 sampling ratio for RTS vs. non-RTS pixels (line 234) doesn't reflect the natural rarity of RTS in the landscape. Please address how this balanced training approach affects the calibration of the final probability maps and if any threshold adjustments were required.

Response：We thank the reviewer for raising this excellent technical point. In our knowledge, training a machine learning model on the highly imbalanced natural distribution would cause the result biased toward the majority class, which will failing to identify any actual slumps.

However, in this revision, to empirically validate our 1:1 sampling choice, we tested alternative positive-to-negative ratios (1:2 and 1:3). The results demonstrated that our selected models are highly robust to changes in class ratios: the AUC values of the optimized Random Forest (RF) model remained consistently at ~0.99, and the Logistic Regression (LR) model at ~0.87, across all tested ratios. See Figure S2 for our testing results.

Regarding the threshold, in traditional static susceptibility mapping, post-hoc threshold adjustments (for example, a conversion cutoff >0.5) to prevent over-prediction. However, in the RTSEvo framework, this threshold adjustment is not required. This is because the probability map is only stage 1. The actual number of pixels converted is strictly governed by the time-series area demand forecast and the specific locations are selected by the spatial allocation module based on relative probability rankings (not absolute probability) and physical rules.

To clarify this sampling rationale and its validation, we have added the following text to the revised manuscript (Lines 263-266):

"*Sensitivity experiments using alternative RTS to non-RTS sampling ratios (1:2 and 1:3) indicated that model performance is highly robust and insensitive to the choice of class balance (Figure S2). Based on these results, a balanced 1:1 sampling strategy was ultimately adopted, as it ensures stable model learning of the environmental occurrence signatures without degrading predictive accuracy.*"

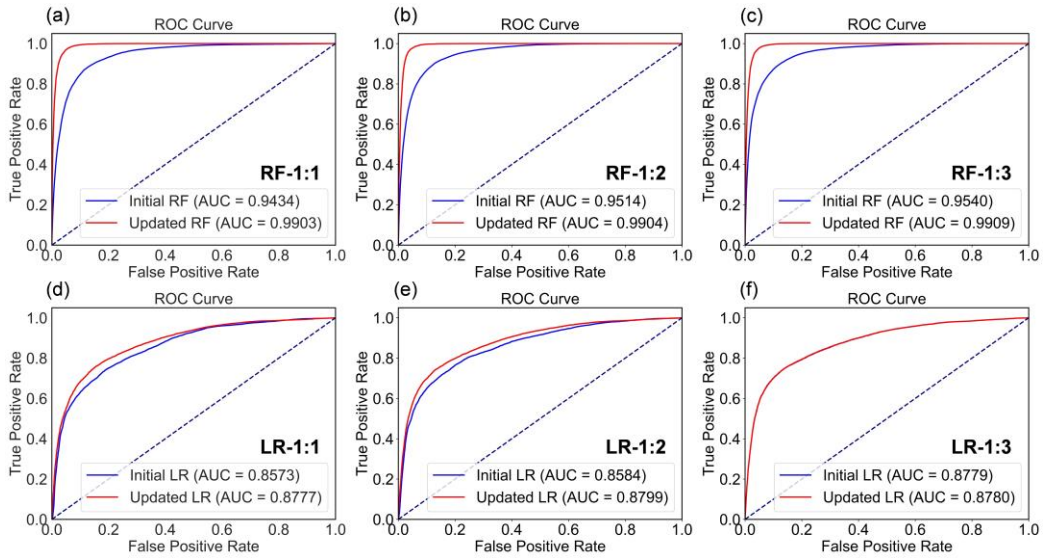

Figure S2. Impact of varying positive-to-negative sample ratios on the predictive accuracy of the Random Forest (RF) and Logistic Regression (LR) models. (a), (b), and (c) show the ROC curves for the RF model under positive-to-negative sample ratios of 1:1, 1:2, and 1:3, respectively (blue lines represent ROC curves before parameter optimization, while red lines indicate those after optimization). (d), (e), and (f) display the corresponding ROC curves for the LR model under the same respective sample ratios.

5. Line 62: "guid the simulation" should be "guide the simulation".

Response：Apologize. It has been corrected.

6. Line 461: "longer temporal horizons" could be more specific (e.g., decadal to centennial scales).

Response：Revised to “extended temporal horizons (e.g., >30-50 years)” for clarity.

7. Line 433: Provide the full name of "SHAP" upon its first occurrence.

Response：Corrected. The full name “SHapley Additive Explanations (SHAP)” has been added in the revision.

References

Burn, C. R. and Friele, P. A.: Geomorphology, vegetation succession, soil characteristics and permafrost in retrogressive thaw slumps near Mayo, Yukon Territory, Arctic, 42(1), 31-40, https://doi.org/10.14430/arctic1637, 1989.

Leibman, M., Khomutov, A., and Kizyakov, A.: Cryogenic Landslides in the Arctic Plains of Russia: Classification, Mechanisms, and Landforms, Landslide Science for a Safer Geoenvironment, Cham, 2014, 493-497,

Nesterova, N., Leibman, M., Kizyakov, A., Lantuit, H., Tarasevich, I., Nitze, I., Veremeeva, A., and Grosse, G.: Review article: Retrogressive thaw slump characteristics and terminology, Cryosphere, 18, 4787-4810, https://doi.org/10.5194/tc-18-4787-2024, 2024.

Pontius, R. G., Boersma, W., Castella, J.-C., Clarke, K., de Nijs, T., Dietzel, C., Duan, Z., Fotsing, E., Goldstein, N., Kok, K., Koomen, E., Lippitt, C. D., McConnell, W., Mohd Sood, A., Pijanowski, B., Pithadia, S., Sweeney, S., Trung, T. N., Veldkamp, A. T., and Verburg, P. H.: Comparing the input, output, and validation maps for several models of land change, Ann. Reg. Sci., 42, 11-37, https://doi.org/10.1007/s00168-007-0138-2, 2008.