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：Thank you very much for the constructive and thoughtful comments, which helped us improve the quality and clarity of the manuscript. We prepared revisions to address all of these comments.

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: Thanks. Resampling primarily affects the spatial smoothness of the estimated occurrence probability. Pixels falling within the same coarse-resolution source cell therefore share similar baseline probabilities.

The influence of this smoothing on model precision is mitigated by the two-stage structure of RTSEvo. The probability estimation module defines where RTS expansion is environmentally plausible at a regional scale, while the subsequent constrained spatial allocation module governs how RTS growth is realized at fine spatial scales. Local scale heterogeneity in RTS expansion is thus driven mainly by allocation constraints, rather than by downscaled predictor variability alone.

To further assess the impact of predictor resolution on probability estimation and spatial allocation, we conducted a supplementary sensitivity experiment in which MODIS NDVI (250 m) was replaced by a reconstructed Landsat 8 NDVI (30m), while keeping all other components of the RTSEvo workflow

unchanged. Results show that higher-resolution NDVI produces modest improvements in spatial allocation accuracy, but the overall gain in predictive performance remains limited (FoM improvement < 1.2%) (Figure S1). This indicates that while finer-resolution predictors enhance local contrast, coarse-resolution inputs do not substantially bias probability estimation nor destabilize the spatial allocation module.

We therefore conclude that resampling coarse predictors to 10 m introduces a known smoothing effect on probability surfaces, but this effect is explicitly controlled by the model architecture. We have clarified this point in the revised Discussion (Section 4.3), and we explicitly acknowledge this as an inherent limitation of regional scale RTS modeling.

Figures S1 is added to the supplementary file in the revision.

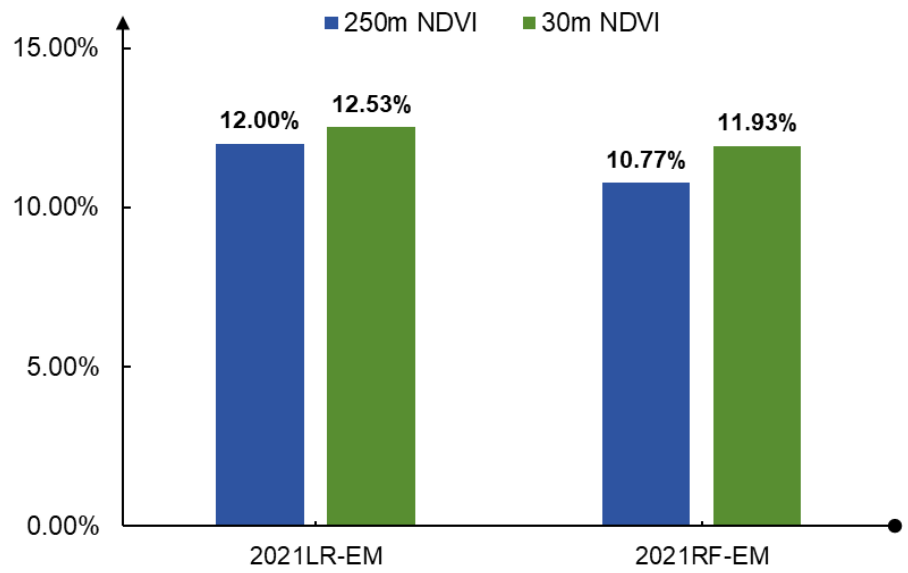

Figure S1. The impact of different NDVI resolutions on model performance.

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: FoM is a stringent metric that evaluates only correctly predicted changes, excluding stable areas, and therefore typically yields lower numerical values than overall accuracy metrics such as Kappa or F1.

In LUCC simulation studies, FoM values commonly range between approximately 2-25% (Figure S2), particularly for rare or spatially clustered change processes. Within this context, the FoM values obtained by RTSEvo (up to 20.05% during calibration and 12.00% in independent validation) are comparable to, and in many cases exceed, those reported in related LUCC applications. Moreover, RTS expansion is a highly localized and rare process, which further constrains achievable FoM values. Importantly, our model demonstrates substantial relative improvement compared to baseline probability-only simulations, with FoM increases exceeding 29% when process-based rules are included. This improvement highlights the added value of the dynamic allocation framework rather than the absolute FoM magnitude alone.

The following text is added to the revised manuscript:

“In LUCC simulation studies, FoM values commonly range between approximately 2-25%, particularly for rare or spatially clustered change processes (Pontius et al., 2008).”

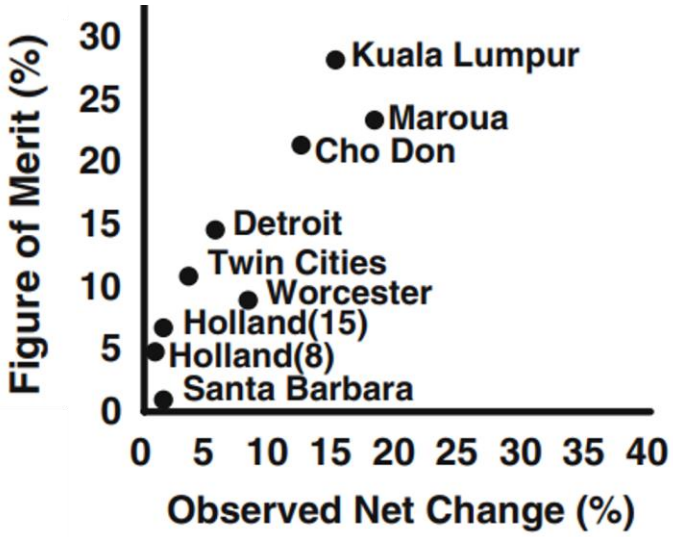

Figure S2. FoM values of different LUCC models in various cities (Pontius et al., 2008).

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. This assumption is physically justified for short to medium term simulations (approximately decadal scales), which is the intended application range of the current model. Field observations and remote-sensing studies indicate that RTS stabilization, revegetation, and surface recovery generally occur over several decades, often exceeding 20 years, depending on local climate, sediment supply, and hydrological conditions (Burn and Friele, 1989; Leibman et al., 2014; Nesterova et al., 2024).

In theory, it is possible to integrate a state recovery in future versions. However, this requires extensive long-term RTS monitoring to determine which locations have returned to a recovery state. A recovery state could be explicitly introduced by extending the base occurrence probability module from a binary to a multi-class framework, distinguishing among non-RTS, active RTS, and recovered RTS states.

The following text is added to the revised manuscript:

“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 longer-term projections (e.g., >30-50 years), neglecting stabilization and recovery processes may lead to overestimation of persistent RTS extent. Future work needs to incorporate simulation of 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：Thank you. To assess the effect of this choice, we tested alternative ratios (1:2 and 1:3) on model performance (Figure S3). The results show that the logistic regression (LR) and random forest (RF) models are

robust to changes in class ratios. With sample ratios of 1:1, 1:2, and 1:3, the AUC values of the optimized RF model are all around 0.99, while the AUC values of the optimized LR model are all around 0.87, indicating that changes in sample ratios have little impact on the accuracy of the models.

Since the sample ratio has little impact on model performance, and considering the relatively small proportion of RTS in the study area, a 1:1 balanced sampling strategy between RTS and non-RTS pixels was ultimately adopted to effectively prevent the classifier from biasing towards the majority class.

The following text is added to the revised manuscript:

“Sensitivity experiments using alternative RTS:non-RTS sampling ratios (1:2 and 1:3) indicate that model performance is insensitive to the choice of class balance. Based on these results, a balanced 1:1 sampling strategy between RTS and non-RTS pixels was adopted, as it ensures robust model learning and effectively prevents the classifier from being biased toward the majority class.”

Figures S3 is added to the supplementary file in the revision.

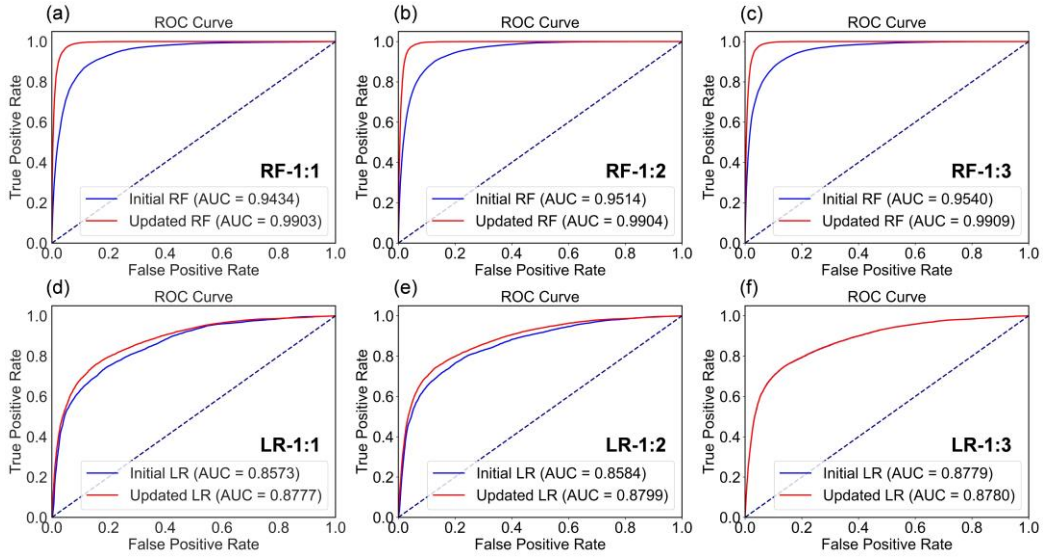

Figure S3. Impact of different positive-to-negative sample ratios on the accuracy of Random Forest and Logistic Regression models. (a), (b), and (c) show the ROC curves for the Random Forest model under positive-to-negative sample ratios of 1:1, 1:2, and 1:3, respectively (blue lines represent ROC curves before parameter optimization, while red lines indicate those after optimization). (d), (e), and (f) display the ROC curves for the Logistic

Regression model under positive-to-negative sample ratios of 1:1, 1:2, and 1:3, respectively.

5. Line 62: “guid the simulation” should be “guide the simulation”.

Response：Corrected.

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

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

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

Response： 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.