Review 1

We thank the reviewer for their careful assessment of the manuscript and their detailed and constructive feedback. Their comments highlight the need for stronger spatial validation, systematic parameter optimisation and sensitivity testing, dynamic glacial constraints, and a clearer distinction between the resolution of the output grid and the spatial detail supported by the proxy evidence.
We have already implemented spatial block partitions for training, validation, and testing, as well as Gaussian blending of overlapping prediction tiles, and have begun implementing a systematic hyperparameter optimisation protocol. We are also working on incorporating glacial masks into the source area reconstruction and U-Net interpolation. In addition, we will document the temporal resolution of the original pollen records, assess the influence of interpolated records and ecological smoothing, and clarify the construction and interpretation of the uncertainty components.
Some concerns can already be addressed through clarification of the existing approach: the original pollen records frequently have a temporal resolution finer than the 500-year reconstruction interval, with interpolation required only for a limited number of missing bins, and the individual uncertainty components are already presented separately, although this needs to be made clearer. We will also clarify that the forest elevation example describes changes in the elevational distribution of reconstructed forest cover and does not independently establish treeline migration. More generally, we agree that the discussion should place greater emphasis on methodological constraints and will strengthen the presentation of the pipeline as a proof of concept, including the limitations of its spatial and temporal interpretation.
Below, we address the reviewer's general comments and major concerns individually and outline the changes already implemented and those planned for revision.

General comments:

The manuscript outlines a comprehensive three-step framework combining Elastic Net regression, the ordination-based "Pixel Flow" spatial method, and an autoregressive U-Net to reconstruct 20,000 years of continuous land cover across Alaska and western Canada at 300 m resolution. Transforming sparse pollen data into detailed spatial grids tackles a key challenge in paleoecology, and the authors thoughtfully maintain data provenance by including empirical source-area maps with the modeled grids.

However, the pipeline relies heavily on heuristic hyperparameters, scaling factors, and loss weights, with no systematic sensitivity testing, obscuring whether reconstructed patterns reflect genuine ecological dynamics or algorithmic priors. Evaluating accuracy primarily on training pixels conflates internal fitting with spatial predictive skill, leaving model performance in proxy-void regions unvalidated. This is compounded by a fundamental scale mismatch: downscaling regional pollen signals (which integrate over tens to hundreds of kilometers) to a

300 m grid risks implying unsupported micro-scale precision. Furthermore, omitting dynamic glacial constraints allows vegetation to be predicted beneath the Laurentide and Cordilleran Ice Sheets during the Last Glacial Maximum (LGM), which is physically untenable.
Establishing this framework as a defensible paleoecological benchmark requires rigorous spatial out-of-sample validation, systematic ablation and sensitivity analyses, dynamic ice masking, and a candid discussion of proxy resolution limits. The discussion should focus on these foundational methodological constraints rather than speculative downstream applications.

Our general reply

We agree that (hyper)-parameters should be systematically chosen and have already implemented a two-step sweep using Optuna for our revisions. We outline the specifics in a reply to one of the major concerns below. In addition, we were already able to implement a spatial block split for training, testing and validation and thank the reviewer for this suggestion.

While the difference in input and output scale can be interpreted as a mismatch, it is precisely the aim of any downscaling method and common in for example climate sciences (e.g. Karger et al., 2017; Thrasher et al., 2022). By supplementing the knowledge of pollen and land cover frequencies for each time scale with spatially higher resolved climatic information, we are able to find the most probable distribution of those land cover classes in space. We agree that accuracy of the reconstruction results cannot be guaranteed and will deepen emphasis on the manuscript's character as a proof of concept and the methodological constraints that come with the output.

Even though ice covers the study area only before 12 ka, we agree that glacial masking should be implemented for those time steps and are working on adding this to steps 2 and 3 of the reconstruction pipeline.

We additionally agree that larger sections of the Discussion should focus on methodological constraints, while keeping an outlook on application, and will edit the chapter accordingly. We do, however, not think that an ablation study with different reconstruction methods would be in the scope of this manuscript.

Major Scientific Concerns:

We add our replies after the respective comments:

- Evaluating model performance across non-overlapping tiles yields overly optimistic metrics (U-Net model accuracy ranging from 0.75-0.90 and 0.80-0.90 in Sections 3.2 and 3.3) because strong spatial autocorrelation in climatic and topographic predictors persists across tile boundaries. To determine genuine predictive skill across proxy-void regions, the framework requires a spatial block or leave-one-site-out cross-validation scheme where geographic clusters of pollen records and their surrounding areas are systematically withheld during training.

We agree and implement a spatial block validation scheme (Fig. 1). Validation blocks make up 15% of the entire data and include all land cover classes. Test blocks make up 15% as well. We implement three folds, i.e. three different validation blocks to get a better accuracy estimate per timeslice.

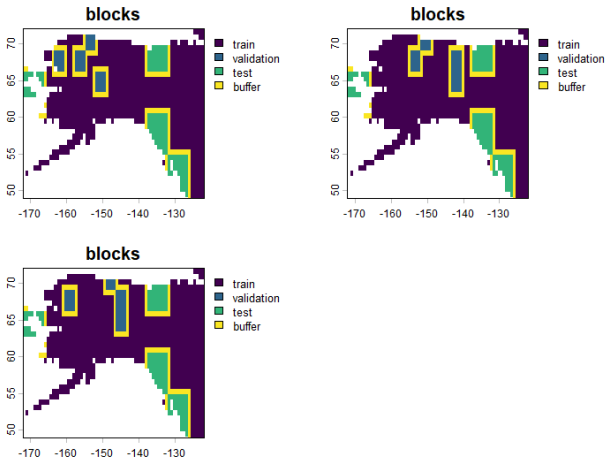


Fig. 1: Three fold block set-up for test-train-validation splits in the study area.

- The architectural complexity of a U-Net requires quantitative justification against standard spatial interpolation baselines, such as spatial Random Forests, Gaussian Process Regression, or Inverse Distance Weighting. The authors must also include an ablation study that systematically removes the modern land-cover prior (13th Channel), the autoregressive predecessor (14th Channel), and the temporal consistency loss (λ!"#$%&'() to confirm that the model captures authentic paleoecological dynamics rather than relying on temporal inertia and modern spatial patterns.

  We believe that an ablation study comparing all listed methods would be outside of the scope for this manuscript. UNet models have been used extensively for the spatial interpolation of climatic variables and land cover, which we believe warrants our application (e.g. Appel, 2024; Huber et al., 2024; Putra and Hsu, 2024).We do, however, agree that temporal input needs to be handled carefully and include a sweep of the strength of their inclusion during our hyperparameter sweeps.

- Omitting dynamic ice-sheet boundaries allows the model to predict vegetation beneath the Laurentide and Cordilleran Ice Sheets during the LGM and the late-glacial transition (~20–12 ka BP), rendering early time slices physically and ecologically invalid. Rather than treating this in Sections 4.1 and 4.4 as an optional future improvement, the authors must apply an established time-transgressive glacial mask (e.g., ICE-6G, ICE-7G, or GLAC-1D) to exclude ice-covered areas before presenting the late-glacial reconstructions as valid scientific products.

  The reconstructions produced between 0 and 12 ka are largely not impacted by this restraint at all (Fig. 2). For time slices previous to those we agree with the reviewer and are working on implementing glacial masks in the source area reconstruction as well as the UNet interpolation.

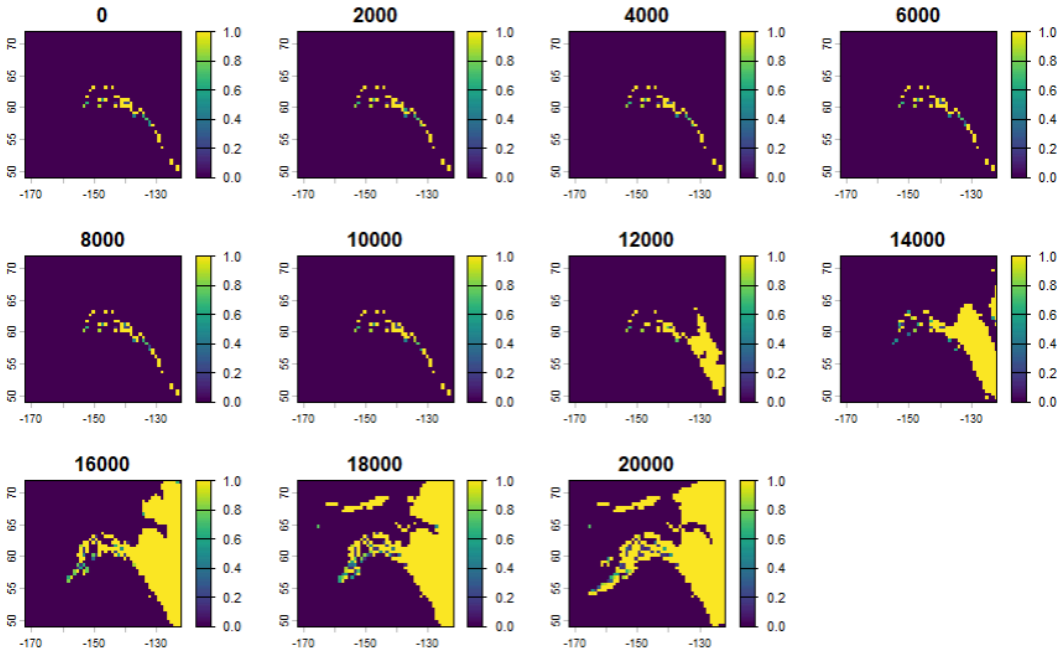


Fig. 1: Ice mask in the study area for selected timeslices (glac1d).

- Linearly interpolating raw pollen records with millennial-scale sampling (often at coarse temporal resolution) into fixed 500-year slices creates synthetic time steps that are statistically collinear with adjacent points. Training sequential U-Nets on these slices optimizes the network on interpolated pseudoreplicates rather than on empirical observations. The authors should report the raw temporal-resolution distribution across the 119 time slices and, if possible, assess how temporal autocorrelation in the training labels affects the reported accuracy.

  We can include the raw temporal resolution of the pollen records. However, the original temporal resolution within records is often higher than the 500-years used in our reconstruction (see Fig. 3). A binning step is usually applied to pollen records, to account for age uncertainty due to sedimentary processes and age-depth modelling. Only few require interpolation to fill missing bins. And while this of course introduces temporal autocorrelation within the pollen compositions, this is a very valid assumption. Vegetation records are high in temporal autocorrelation anyways.

  We still plan on training some UNet timeslices without interpolated records to investigate how their presence impacts models and final reconstructions compared to those with interpolated and non-interpolated records. The number of training pixels will be kept equal for both tests.

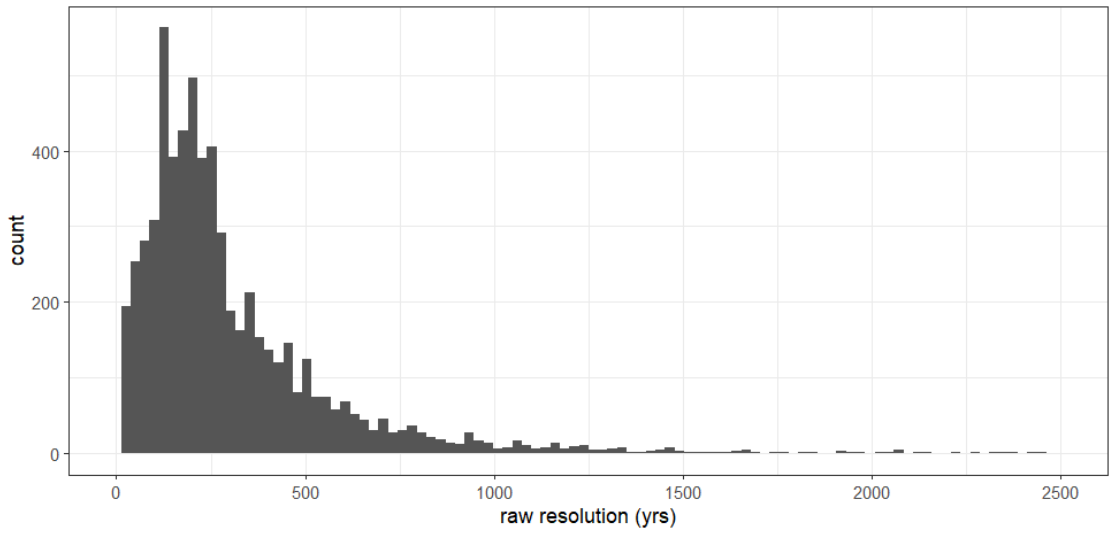


Fig. 3: Raw temporal resolution of the pollen records included.

- Averaging five mathematically disparate metrics (UReg, URDA, Utemp, UError, Uunc) into an unweighted arithmetic mean in Equation 10 obscures the physical interpretation of the resulting uncertainty surface. The framework should instead provide disaggregated raster layers that separate input calibration error, spatial allocation ambiguity, and model epistemic and aleatoric uncertainty. Any single composite index must be supported by formal variance propagation or by statistical weighting.

  We already show the uncertainty components separately but can do so more clearly. While we understand the risks of having one aggregated uncertainty index, we believe it useful as a quick overview for areas that might be problematic especially with users that are not acquainted well with the used methods.

- Converting discrete pollen counts into soft probability targets via softmax($-D/\tau$) in Equation 3 risks causing class collapse, where distinct transitional classes (e.g., lichen and moss) are smoothed into dominant forest classes before training. The manuscript must provide the mathematical formulation, symmetry constraints, and normalization of matrix D, clarifying how it penalizes structural versus compositional dissimilarity.

  We will clarify that $D_{ij} = 1 - O_{ij}$, where $O_{ij} \in [0, 1]$ measures overlap between land-cover classes in environmental space within the study area. The matrix is symmetrized, assigned a zero diagonal, and normalized by its maximum off-diagonal value. It therefore represents environmental dissimilarity; structural and compositional differences are captured only insofar as they correspond to environmental separation.

  The row-normalized kernel $softmax(-D/\tau)$ preserves total vote mass and contains no class-frequency prior favoring forest. However, this does not rule out excessive smoothing of environmentally similar classes. We will assess this risk

by examining redistribution between lichen/moss and forest and comparing class-specific performance with a baseline without ecological smoothing.

- Key hyperparameters, including the ecological smoothing temperature (τ), Dirichlet concentration parameter (α)), discrete channel weights (0.35), temporal loss weight (λ!"#$%&'(= 0.18), and the confidence gate threshold (0.60), are presented without tuning criteria. A sensitivity analysis of the start equation tau and lambda sub temporal, alongside a documented optimization protocol, is essential for methodological reproducibility.

  We will document a staged optimization protocol using fixed spatial training and validation partitions, with mean validation Macro F1 across development time slices as the selection criterion. First, we will tune ecological smoothing temperature τ jointly with focal gamma, learning rate, and weight decay. Second, we will fix those settings and tune temporal loss weight λ_temporal and the modern and previous land-cover channel scales. A joint refinement will then assess interactions between these parameters, followed by separate sensitivity analyses for the confidence gate and Dirichlet parameter α. We will report the search ranges, trial budgets, selected values, and sensitivity results, including a λ_temporal = 0 baseline. The final configuration will be assessed on held-out test data after parameter selection.

- Inferring treeline shifts from the 90th percentile of forest elevation without controlling for basin hypsometry (like the case study in Section 2.3 and Section 3.4) confounds true ecotone migration with geometric area constraints and patch-size fluctuations. The reconstructed elevation distributions must be normalized against the underlying digital elevation model's hypsometric curve and validated against independent local proxy evidence, such as macrofossils or regional lapse-rate constraints.

  We agree that the 90th percentile of forest elevation is sensitive to the distribution of forest area across elevations and cannot uniquely identify treeline migration. Our intention was to illustrate changes in the elevational distribution of reconstructed forest cover as an application of the product. We will revise the text and figure caption to clarify this distinction and explicitly acknowledge the influence of basin hypsometry and changes in forest area at different elevations. We do not interpret this statistic alone as a reconstruction of treeline position.

- Applying a hard, winner-take-all selection based on edge-weighted confidence to stitch overlapping 256 × 256 tiles risks introducing linear seam artifacts across continuous spatial domains. The inference pipeline should compute a distance-weighted blend (e.g., linear or Gaussian feathering) across overlapping probability or logit tensors before argmax classification.

  We agree and have already implemented smoothed inference at the tile overlap with softmax class probabilities from each tile and gaussian weights. Each pixel is finally assigned the class with the highest combined probability.