Supplementary materials for this article are available at https://doi.org/10.1007/s13253-023-00582-x.

Integrating Different Data Sources Using a Bayesian Hierarchical Model to Unveil Glacial Refugia

Mauricio CAMPOS, Bo LI, Guillaume de LAFONTAINE, Joseph NAPIER, and Feng Sheng HU

Rapid anthropogenic climate change has elevated the interest in studying the biotic responses of species during the Last Glacial Maximum. During this period, species retreated to highly spatially restricted geographic regions where survival was possible, known as glacial micro-refugia, from which they migrated and expanded when conditions became more suitable. Several distinct sources of evidence have contributed to developing a new understanding of how these regions might have impacted the sustainability of the natural populations of many species. Pollen records in Eastern Beringia have been used to explore the possibility that the region harbored glacial refugia for several plants from the arctic tundra and/or the boreal forest biomes common to the region. Our study focuses on *Alnus viridis* and *Picea glauca*, two predominant species of arcto-boreal vegetation. We propose to integrate genomic, SDM, and existing fossil data in a hierarchical Bayesian modeling (HBM) framework to determine whether multiple refugia existed in isolated geographic areas. This study demonstrates how the flexibility of HBMs makes the formal synthesis of such disparate data sources feasible. Our results highlight the regions of plausible refugia that can guide future investigations into studying the role of glacial refugia during climate change.

Supplementary materials accompanying this paper appear online.

Key Words: Data integration; Genetic; INLA; Pollen fossil; Spatial process; Species distribution models.

M. Campos (✉) · B. Li, Department of Statistics, University of Illinois, Urbana-Champaign, USA. (E-mail: *mcampos3@illinois.edu*) (E-mail: *libo@illinois.edu*).
G. de Lafontaine, Département de biologie, chimie et géographie, Université du Québec, Rimouski, Canada. (E-mail: *Guillaume_deLafontaine@uqar.ca*).
J. Napier, Department of Integrative Biology, University of Texas, Austin, USA. (E-mail: *joseph.napier@austin.utexas.edu*).
F. Sheng, Department of Biology, Washington University, St. Louis, USA. (E-mail: *deanhu@wustl.edu*).
F. Sheng, Department of Earth and Planetary Sciences, Washington University, St. Louis, USA.

© 2023 International Biometric Society
Journal of Agricultural, Biological, and Environmental Statistics
https://doi.org/10.1007/s13253-023-00582-x

Published online: 15 November 2023

1. INTRODUCTION

Anthropogenic climate change has become a major concern for the sustainability of the natural populations of many species. This has renewed interest in understanding the biotic responses to climate variations in the paleorecord, because such understanding will be essential in anticipating future changes in biodiversity and informing ecosystem management (e.g., Dawson et al. 2011). To shed light on this issue, we particularly study the species range shifts within the Quaternary from the Pleistocene during the Last Glacial Maximum (LGM) (colloquially referred to as the *Ice Age*) to the current-day Holocene. During this period, the varying climates had a major impact on altering the biodiversity patterns of the region, and thus, understanding shifts in species distribution during this period offers much evidence of the species response to climate change (Davis and Shaw 2001; de Lafontaine et al. 2018; Napier et al. 2020a).

During periods of atypical regional climate, species retreated to geographic regions where survival was possible, known as glacial refugia, from which they migrated and expanded when conditions became more suitable (Hampe and Jump 2011; Keppel et al. 2012; Gavin et al. 2014). Recent genetic studies (e.g., Anderson et al. 2006; Parducci et al. 2012; De Lafontaine et al. 2013; Hao et al. 2018; Napier et al. 2019, 2020b) have demonstrated the possibility that many arcto-boreal plants survived the LGM in small disjunct populations that later expanded in the post-glacial period. These "cryptic refugia", usually undetected using the fossil record (Provan and Bennett 2008), challenge the traditional understanding regarding the role of low-latitude refugia in the post-glacial vegetation development (e.g., Petit et al. 2003; McLachlan et al. 2005; Magri et al. 2006; Stewart et al. 2010; Mosblech et al. 2011). It was previously believed that most of the post-glacial colonization came from refugia located in warmer lower latitudes, but now high-latitude refugia offer another insight into the process of how species flourished into the Holocene (Feurdean et al. 2013). This is of particular importance since the existence of small refugial populations might contribute to explaining the "Quartenary conundrum"—there being little evidence of species extinction during the dramatic climate shifts of the Quaternary, as opposed to the massive extinctions predicted by our current climate change (Botkin et al. 2007)—thus creating forecasts that lessen the overestimation of extinction likelihood (Luoto and Heikkinen 2008; Randin et al. 2009; Mosblech et al. 2011).

Evidence of the existence of cryptic refugia has risen from many regions in the northern hemisphere but our study will focus mostly on Eastern Beringia (Alaska and adjacent Canada), which has been featured extensively in the literature and recognized as a site of possible refugia (e.g., Shafer et al. 2010). The dense network of fossil pollen records recovered from lake sediments captured over several decades in the region has been used to examine the possibility that it harbored glacial refugia for arcto-boreal taxa (Hopkins et al. 1981; Bigelow et al. 2003; Brubaker et al. 2005). Additionally, phylogeographic surveys have also contributed to these studies by analyzing DNA markers of extant populations (e.g., Abbott and Brochmann 2003; Anderson et al. 2006, 2011; de Lafontaine et al. 2010; Napier et al. 2019, 2020b). Altogether, the evidence seems to indicate that several arcto-boreal species managed to persist through the LGM in Eastern Beringia. However, details about the whereabouts of such refugial populations remain unknown.

Much of the evidence used in uncovering the refugia comes from three data sources: pollen fossil records, phylogeographic surveys, and species distribution models (SDMs). Fossil pollen is recovered from lake-sediment cores. If enough pollen that dates back to the LGM is found in the cores, it would be direct proof of past presence in the vicinity of the coring site. As such it is likely the most robust line of evidence, but recovering this information is a resource-intensive procedure; thus, we only have limited data collected over various decades. Phylogeography relies on analyzing the geographical pattern of DNA diversity from modern-day samples to infer the past evolutionary scenarios that generated the observed modern-day genetic lineages. Since it relies on sampling present-day individuals, genetic information is easier to obtain than pollen fossil records, at the expense that the inferences about past refugia are less direct. Finally, SDM is the association between known modern-day occurrence and climate variables that is projected on past climate reconstructions to obtain probabilities of suitable climate for a given species over the landscape. This provides insight into regions where climate conditions might have been suitable for the species to be present but provides no direct evidence of past presence.

A review of the literature in paleoecology has revealed that many different statistical techniques have been employed to recover refugia from each data source (Gavin et al. 2014). For example, analysis of fossil data typically consists of comparing modern-day pollen assemblages with those observed in the past to infer the composition and location of ancient forests.

Phylogeography relies on analyzing geographical patterns of genetic diversity and structure from natural populations to infer the historical evolutionary processes that lead the distribution of past geographical genealogical lineages to the present distributions. Refugia locations can usually be identified due to their lower intrapopulation genetic diversity and higher interpopulation diversity, resulting in stronger genetic differentiation but lower spatial genetic structure, than biological populations located in recolonized areas (Hewitt 2000; De Lafontaine et al. 2013). Population genetics have employed different approaches to studying genetic variation used for phylogeographic inferences. For instance, techniques such as AMOVA (Analysis of Molecular Variance; Excoffier and Smouse 1994) have borrowed statistical methods to provide objective historical inferences. AMOVA aims to estimate population differences similar to the statistical analysis of variance (ANOVA) (Meirmans and Liu 2018). The total genetic variance is decomposed into three covariance components: between-population, between-individuals within a population, and within-individuals, which are then used to construct the test statistics similar to F-statistics (Meirmans and Liu 2018).

Lemmon and Lemmon (2008) used likelihood methods to both test a prior conjecture regarding refugia as well as estimate the phylogeographic history of a gene in the absence of such conjecture. A Bayesian alternative to the estimation methods has also been employed to model the locations of taxa along each branch in the phylogeny (e.g., Lemey et al. 2009, 2010; Manolopoulou and Emerson 2012; Marske et al. 2012). Due to the increasing complexity of the likelihood models for estimating ancestral refugia, it is common in the field to fit Bayesian models using Approximate Bayesian Computation methods (Gao et al. 2012; Li et al. 2013; Budde et al. 2013; Tsuda et al. 2016; Wang et al. 2016; Cornejo-Romero et al. 2017; Ren et al. 2017; Aoki et al. 2019).

Each line of evidence provides its own set of strengths and weaknesses. Different data sources also seem to capture different information regarding refugia and postglacial expansion. For example, evidence from fossil pollen records (e.g., Anderson and Brubaker 1994) implies that taxa, such as spruce, resided in one general area and expanded in a single direction during the postglacial. However, genetic analyses suggested the existence of multiple microrefugia in Eastern Beringia (Napier et al. 2019), consistent with the prevailing pattern that has also emerged in other regions around the globe (Hao et al. 2018). It is imperative to develop an integrative method that can jointly glean information from all lines of evidence. Several attempts have been made and most of them emphasized the integration of genetic data and SDM information to obtain better estimates of refugia location (see Section IV of Gavin et al. 2014). These methods typically use SDM as a filter for identifying plausible refugia locations over which multiple genetic scenarios are then simulated and compared to the observed genetic data with statistical tests (e.g., MANOVA/ANOVA) to determine which ones provide the most likely locations (e.g., Knowles and Alvarado-Serrano 2010; Brown and Knowles 2012; Espíndola et al. 2012; Aoki et al. 2019; Napier et al. 2019). Bayesian hierarchical models (BHMs) have also been used for this purpose as a foundation of dynamic geographical range models. These models combine abundance information (usually obtained from SDMs) with environmental data and demographic rates to estimate niches and range dynamics, which in turn inform the presence of refugia (Marion et al. 2012; Pagel and Schurr 2012; Schurr et al. 2012).

BHMs have been a popular approach for data fusion due to their advantage of enabling joint modeling while being flexible to take into account the unique characteristics of each data type (Clark 2005). In addition, the posteriors of BHMs naturally provide uncertainty quantification for the estimates of unknown variables. BHMs have shown great promise in paleoclimate and paleoecological studies (e.g., Li et al. 2010; Urban et al. 2013). Advances in computation power as well as alternatives to MCMC, such as INLA (Rue et al. 2009), have made it possible for the estimation to be timely and efficient. To our knowledge, there is no systematic and rigorous method to combine all three major data sources to infer refugia locations. We propose to integrate species distribution, genomic, and existing fossil data in a BHM framework to elucidate glacial refugia of green alder (*Alnus viridis*) and white spruce (*Picea glauca*) in Eastern Beringia. Our method allows for the strengths of one data source to compensate for the weaknesses of others. We hope that the uniqueness and strength of this proposed method make it a useful tool for paleoecology and enlighten new follow-up studies.

The rest of the paper is organized as follows: Sect. 2 reviews the three distinct lines of evidence that are commonly used to locate the most possible arcto-boreal refugia in Eastern Beringia. Section 3 introduces the BHM that integrates all three lines. Section 3.2 contains a brief explanation of how the estimation procedure is implemented using Integrated Nested Laplace Approximation (INLA). A small simulation study verifying our method is shown in Sect. 4. Finally, Sect. 5 presents the results for both arcto-boreal species under study.

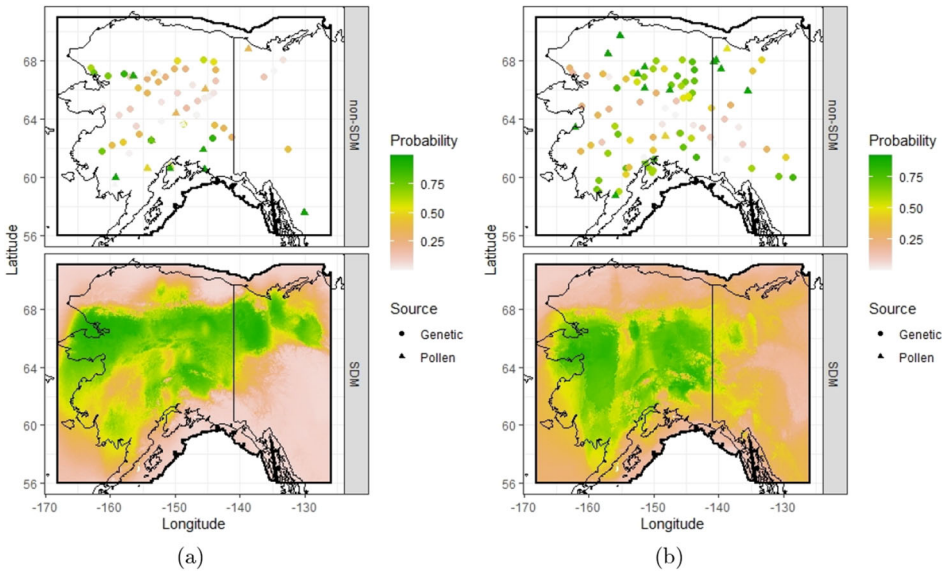


Figure 1. Observed data from all sources for **a** green alder and **b** white spruce. The upper panel in **a** and **b** shows the genetic and pollen data while the lower panel shows the species distribution model (SDM) data. The pollen and genetic data shown here are already processed with the interpolations discussed in Sects. 2.2 and 2.3. The shaded region corresponds to Eastern Beringia during the Last Glacial Maximum, with modern-day Alaska superimposed for reference.

2. DATA

Our data come from three different sources: niche models, genetic lineages, and pollen fossil records. All three types of data are acquired for green alder (*Alnus viridis*) and white spruce (*Picea glauca*) and are shown in Fig. 1. The particular type of niche modeling used in this paper is species distribution models (SDMs). The pollen and genetic data are much more sparse than the SDM.

2.1. SPECIES DISTRIBUTION MODEL

SDMs determine the probability of suitable climates using environmental variables (Franklin 2010) and have been widely applied due to their simplicity and growing accessibility (e.g., Thuiller et al. 2009). SDMs were developed based on available modern species occurrence and climate data and then applied to climate simulations to hindcast probabilities of species past occurrence. SDM for green alder is taken from Napier et al. (2019) whereas for white spruce it was generated using the same approach. The improved availability of paleoclimate simulations has led to the increasing application of SDMs to paleoecology (Nogués-Bravo 2009; Svenning et al. 2011), including the study of historical refugia. As the output of numerical models, SDM can be obtained at a very fine resolution, with around 334,000 sites in our study region.

However, several assumptions and uncertainties, such as the assumed static species-climate relationships despite changes in the environment, unaccounted putative dispersal

limitations, and biotic interactions, limit the utility of SDMs and complicate their interpretation (Guisan and Thuiller 2005).

Thus SDMs are best viewed as a tool for preliminary analysis regarding the past locations (e.g., Porto et al. 2013) and dynamics (e.g., Graham et al. 2010) of species that needs cross-validation with independent evidence such as genomic and other paleoecological data. Following Allouche et al. (2006), we consider SDM probabilities greater than a model-specified threshold τ_m (maximizing the accuracy of the model based on the True Skill Statistics) as an indicator for favorable conditions for refugia and otherwise unfavorable. For regions where SDM probabilities are below τ_m, it is safe to assume that those regions are not refugia. However, where SDM probability is above τ_m, we need complementary lines of evidence to better locate the exact refugia area. This characteristic of SDM makes SDM more appropriate for playing the role of classifying whether refugia are present or absent. Therefore, we transform SDM into binary data.

Let $P_m(\boldsymbol{s})$ represent the SDM probability that site $\boldsymbol{s}$ is a refugia location. We define $Y_m(\boldsymbol{s}) = I\{P_m(\boldsymbol{s}) \geq \tau_m\}$, where $I(A)$ is an indication function with $I(A) = 1$ if A is true and 0 otherwise. The threshold τ_m was chosen to be the True Skill Statistic (TSS) defined in Allouche et al. (2006). TSS is defined as TSS = sensitivity + specificity-1, where the sensitivity and specificity are obtained by comparing the SDM predictions with a set of validation sites. Napier et al. (2019) suggest using 0.54 and 0.506 thresholds for green alder and white spruce, respectively.

2.2. POLLEN DATA

The fossil data come from pollen records that have been collected in Beringia since the early 1980s. The information used represents the effort over multiple decades of several research teams to uncover evidence of refugia and yet only a few of them can be used to posit our species of interest during the LGM. Coring sites that actually date back to the LGM are scarce, and thus, the spatial resolution of this database is coarse. The observed pollen data measure the proportion of pollen fossil records belonging to a specific species at a given depth of a sediment core. The greater this proportion, the stronger the evidence of the site being refugia. Despite this continuous association, the pollen data were mainly used as a binary indicator by thresholding the records.

We wish to utilize pollen data to a greater extent than as a binary variable indicating presence/absence. We will still respect that usually a site $\boldsymbol{s}$ is considered to be refugia of a species with probably $P_p(\boldsymbol{s}) \geq \gamma_p$ for a large γ_p, if its composition proportion $c(\boldsymbol{s}) \geq \tau_p$ for a species-specific threshold τ_p. Also, since there are many pollen types in the sediment samples, the composition percentages are usually small. Due to the small nature of these percentages, we consider $P_p(\boldsymbol{s}) = 1$ if $c(\boldsymbol{s}) = 0.5$. In the observed data, no composition percentage reaches this threshold. To use the pollen data properly, we propose to transform the composition proportions $c(\boldsymbol{s})$ into probabilities $P_p(\boldsymbol{s})$ subject to the above considerations:

$$p_p(\boldsymbol{s}) = \begin{cases} \frac{c(\boldsymbol{s})}{\tau_p}\gamma_p & \text{if } c(\boldsymbol{s}) \leq \tau_p \\ \{2c(\boldsymbol{s})\}^{\log(\gamma_p)/\log(2\tau_p)} & \text{if } c(\boldsymbol{s}) > \tau_p. \end{cases} \tag{1}$$

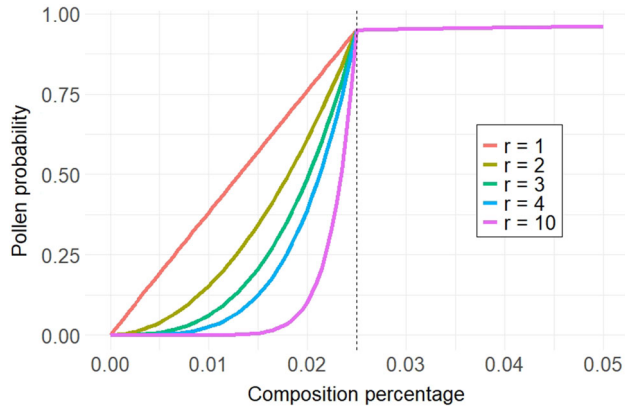


Figure 2. Different polynomial interpolations for pollen data. The $r = 1$ curve corresponds to the choice used in Eq. (1). The dashed line represents τ_p and how all interpolations have the same tail afterward.

The transformation (1) features a linear interpolation of probabilities when $c(\boldsymbol{s}) \leq \tau_p$, and approaching probability 1 in polynomial when $c(\boldsymbol{s}) > \tau_p$, as shown in Fig. 2. The coefficient 2 and the polynomial power are determined by the conditions $P_p(\boldsymbol{s}) = \gamma_p$ when $c(\boldsymbol{s}) = \tau_p$ and $P_p(\boldsymbol{s}) = 1$ when $c(\boldsymbol{s}) = 0.5$. There is no established literature with regard to the interpolations formula, and our proposed transformations were simply constructed for being sound choices and conforming to our prior knowledge. There could be other choices. For example, the linear interpolation for the range $c(\boldsymbol{s}) \leq \tau_p$ can be generalized to a polynomial interpolation with power r, i.e., $p_p(\boldsymbol{s}) = \left\{\frac{c(\boldsymbol{s})}{\tau_p}\right\}^r \gamma_p$ for $c(\boldsymbol{s}) \leq \tau_p$. Curves for different r are shown in Fig. 2. We choose $r = 1$ for its simplicity and for representing more reasonable probabilities for small $c(\boldsymbol{s})$ than $r > 1$. We found our results are insensitive to other reasonable transformations.

The threshold τ_p depends on species: green alder uses $\tau_p = 2.5\%$, whereas white spruce uses 1% (Napier et al. 2019; Warren et al. 2016). Likewise, γ_p differs for both species. These were chosen as $\gamma_p = 0.95$ for alder and $\gamma_p = 0.90$ for spruce as sensible choices that represent 'high' probabilities.

2.3. GENETIC EVIDENCE

We obtain genetic data from genetic surveys that report separate lineages (see Napier et al. 2019 for more detail). Genetic data for green alder consists of evidence from only two lineages, while white spruce has five different lineages. For a particular site $\boldsymbol{s}_i$, the ancestry coefficient $a_k(\boldsymbol{s}_i)$ is defined as the proportion of site i's genome that originated from lineage k (Pritchard et al. 2000). This implies that $\sum_k a_k(\boldsymbol{s}_i) = 1$, where the summation is taken over all lineages represented in the study for a particular species. Similar to pollen data, genetic information was traditionally used as a binary source of evidence. To use genetic data more efficiently, we likewise transform each genetic assemblage to a probability of being refugia.

The transformation is based on the relative percentages of different lineages for each species. Only the dominance of one single lineage indicates the higher chance of this location

being a refugium. The transformation differs depending on species as each one has a different number of lineages; nevertheless, the underlying principle remains the same.

For both species, let $\tilde{A}(\boldsymbol{s}) = \max_k a_k(s)$ for site $\boldsymbol{s}$ and let $P_g(\boldsymbol{s})$ represent the probability of site $\boldsymbol{s}$ being refugia according to genetic data. Since green alder only has two lineages, the transformation should interpret a site with $\tilde{A}(\boldsymbol{s})$ closer to 0.5 corresponding to a smaller $P_g(\boldsymbol{s})$, and the probability should grow larger as $\tilde{A}(\boldsymbol{s})$ increases. To meet those requirements, we propose a transformation as a polynomial of power r:

$$P_g(\boldsymbol{s}) = \left\{2\left(\tilde{A}(\boldsymbol{s}) - 0.5\right)\right\}^r. \tag{2}$$

Using this method we obtain a "soft" threshold for the data, where we retain all information in the data but only a few sites receive high probabilities while the rest are much lower, reflecting the higher uncertainty of being refugia if the sample at that site is more mixed (See Fig. 3a). The value of r can be chosen according to how well the transformed data conforms with expert knowledge or prior information. A sensitivity analysis for different r values shows that overall as r increases the higher $P_g(\boldsymbol{s})$ remain similar, although the bulk of $P_g(\boldsymbol{s})$ decreases. We choose $r = 3$ because this value seems to reach a better balance between keeping the sites with high $\tilde{A}(\boldsymbol{s})$ as high $P_g(\boldsymbol{s})$ and decreasing the rest to more conservative levels, according to expert knowledge.

White spruce has five different lineages. We first identify the dominant lineage for each site by finding which lineage corresponds to $\tilde{A}(\boldsymbol{s})$. Denote m_j as the total number of sites that have the j-th lineage as the dominant. Let $\tilde{A}_j(\boldsymbol{s}_i), i \in \{1, 2, 3, ..., m_j\}$ represent the ancestry coefficient at the i-th location that is dominated by the j-th lineage. Let $\xi_j = \max_i \tilde{A}_j(\boldsymbol{s}_i)$ and $\delta_j = \min_i \tilde{A}_j(\boldsymbol{s}_i)$ be the maximum and minimum ancestry coefficient for lineage j, respectively. Also, let $A_{\max}$ and $A_{\min}$ represent the maximum and minimum ancestry coefficients observed among all sites and all lineages, which for our white spruce data are 0.789 and 0.0001, respectively. We define the probability of refugia for the genetic data:

$$P_g(\boldsymbol{s}) = \frac{A_{\min}[\xi_j - \tilde{A}_j(\boldsymbol{s})] + A_{\max}[\tilde{A}_j(\boldsymbol{s}) - \delta_j]}{\xi_j - \delta_j}. \tag{3}$$

With this definition, the lowest ancestry coefficient for each lineage will be assigned $A_{\min}$ as its probability of being refugia, whereas the largest lineage-specific coefficient will be assigned $A_{\max}$ as its corresponding probability (see Fig. 3b), meaning that all lineages will have the same interpolated probability range. Note that the probability $P_g(\boldsymbol{s})$ can be interpreted as the weighted average of $A_{\min}$ and $A_{\max}$, weighing by the distances from $\tilde{A}_j(\boldsymbol{s})$ to the extremes ξ_j and δ_j. In other words, the closer $\tilde{A}_j(\boldsymbol{s})$ is from its lineage's highest possible ancestry coefficient (i.e., ξ_j), the more the probability will approach $A_{\max}$. Likewise, the closer $\tilde{A}_j(\boldsymbol{s})$ is from its lineage's lowest possible ancestry coefficient (i.e., δ_j), the more the probability will approach $A_{\min}$. Since the values used in this interpolation method are the lineage-specific ancestry coefficients, they are naturally bounded by $A_{\min}$ and $A_{\max}$, which represent the highest and lowest possible values of $P_g(\boldsymbol{s})$, respectively. This guarantees that none of the interpolated probabilities will be less than 0 or greater than 1.

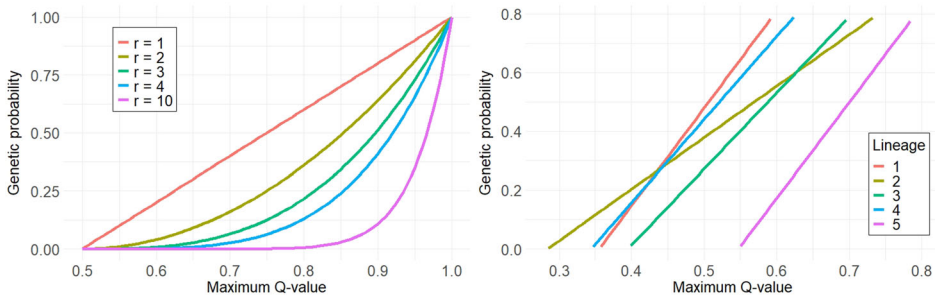


Figure 3. Genetic data interpolations: **a** different polynomial curves are shown for green alder whereas **b** linear interpolations are shown for each genetic lineage of white spruce.

Like pollen fossil data, the interpolations presented here are not unique and there is no established formula to follow. Other transformations could be considered, but we merely wish to translate the raw lineage information into more interpretable probabilities that can then be used in our model. The linear interpolation was chosen for its simplicity in interpretation and reasonable performance.

3. BAYESIAN HIERARCHICAL MODEL

All three lines of evidence contain useful information in unveiling the refugia, though they each have their strength and weakness. We aim to integrate these complementary data sources to identify the possible locations of refugia, which is expected to be more efficient and powerful than using a single line of evidence. As discussed earlier, we have transformed the three data sources into $Y_m(\boldsymbol{s})$, $P_p(\boldsymbol{s})$, and $P_g(\boldsymbol{s})$, according to their characteristics. Our model is constructed based on the transformed data.

3.1. MODEL SPECIFICATION

Let $P(\boldsymbol{s})$ denote the probability of location $\boldsymbol{s}$ being a refugium. We attempt to obtain coherent estimates of $P(\boldsymbol{s})$, given the binary $Y_m(\boldsymbol{s})$ derived from the SDM and the probabilities $P_p(\boldsymbol{s})$ and $P_g(\boldsymbol{s})$ derived from pollen and genetic respectively. To accomplish this, we need to carefully model the relationship between $P(\boldsymbol{s})$ and the three data sources based on the characteristics of each data. Since all three data are observations from different perspectives given the true refugia, it is natural to establish the forward model of each data given a common underlying $P(\boldsymbol{s})$. The SDM has been mainly used as a preliminary screening tool through the binary $Y_m(\boldsymbol{s})$, hence we will employ a logistic model for $Y_m(\boldsymbol{s})$. Since the information from pollen and genetics is more quantitatively related to the probability of refugia, we build a model to reflect this feature. The forward models should also recognize that, when compared with temporally variable and spatially inconsistent pollen data, genetic data is often easier to obtain from comprehensive spatial grids.

Let S_m, S_g, $and S_p$ denote the collection of sites for SDM, genetic, and pollen data, with sizes n_m, n_g, and n_p, respectively. The total sample size is given by $n = |S| = |S_m \cup S_g \cup S_p|$ where $|S|$ denotes the cardinality of S. In our application, all data subsets are disjoint so that $n = n_m + n_g + n_p$. There is a strong unbalance in the sample sizes, such that $n_m \gg n_g + n_p$. Furthermore, we define two subregions for the SDM data: S_{m0} and S_{m1}, where S_{m1} is defined as the collection of SDM sites where SDM is greater than the threshold, i.e., $Y_m(\boldsymbol{s}) = 1$, and S_{m0} the rest. To lift the constraint of modeling probabilities, we first perform an inverse probability integral transform on $P(\boldsymbol{s})$, $P_p(\boldsymbol{s})$, and $P_g(\boldsymbol{s})$ using a standard normal cumulative distribution function $\Phi(\cdot)$ to turn probabilities into Gaussian random variables. Specifically, we have $\mu + X(\boldsymbol{s}) = \Phi^{-1}(P(\boldsymbol{s}))$, where $X(\boldsymbol{s})$ is assumed to be a mean zero Gaussian random variable, $Y_p(\boldsymbol{s}) = \Phi^{-1}(P_p(\boldsymbol{s}))$ and $Y_g(\boldsymbol{s}) = \Phi^{-1}(P_g(\boldsymbol{s}))$. Then we propose the following forward models as the first level of our BHM:

FIRST LEVEL: DATA MODELS

$$\begin{aligned} \text{logit}\left(P\left(Y_m(s)=1\right)\right) &= \alpha_m + \beta_m\left\{\mu + X(\boldsymbol{s})\right\} + Z(\boldsymbol{s}), & \mathbf{Z} &\sim GP\left(\mathbf{0}, \boldsymbol{\Sigma}(\sigma_m^2(\boldsymbol{s}), \rho_m)\right), \\ Y_p(\boldsymbol{s}) &= \alpha_p + \beta_p\left\{\mu + X(\boldsymbol{s})\right\} + \epsilon_p, & \epsilon_p &\sim N(0, \sigma_p^2), \\ Y_g(\boldsymbol{s}) &= \mu + X(\boldsymbol{s}) + \epsilon_g, & \epsilon_g &\sim N(0, \sigma_g^2), \end{aligned} \tag{4}$$

where $\mathbf{Z}$ is the vector consisting of all $Z(\boldsymbol{s})$ for $\boldsymbol{s} \in S_m$.

Our model respects the fact that all three data are trying to capture the true probability of refugia in different manners, albeit with uncertainties. Additionally, the model assumes that $Y_m(s)$, $Y_p(\boldsymbol{s})$, and $Y_g(\boldsymbol{s})$ are conditionally independent given the latent process $X(\boldsymbol{s})$. Since genetics is considered the more spatially comprehensive quantitative data source, we model the genetic data as an unbiased source of the true refugia probability. SDM and pollen data are taken as deviations with both additive and multiplicative biases, in addition to Gaussian errors on the models. This model specification also ensures the identifiability of unknown parameters.

The Gaussian error process $Z(\boldsymbol{s})$ models the extra uncertainty in SDM data beyond what a Bernoulli distribution can capture. Considering different levels of credibility of SDM in showing whether there are refugia, we employ a non-stationary covariance function with unknown spatially varying variance, $\sigma_m^2(\boldsymbol{s})$, and an invariant range parameter, ρ_m for $Z(\boldsymbol{s})$. The covariance between two sites can be expressed as

$$\text{cov}(Z(\boldsymbol{s}_i), Z(\boldsymbol{s}_j)) = \sigma_m(\boldsymbol{s}_i)\sigma_m(\boldsymbol{s}_j)C(||\boldsymbol{s}_i - \boldsymbol{s}_j||),$$

where $C(||\boldsymbol{s}_i - \boldsymbol{s}_j||)$ can be any valid correlation function. We choose the Matérn correlation function for $Z(\boldsymbol{s})$. Let $d = ||\boldsymbol{s}_i - \boldsymbol{s}_j||$, a Matérn correlation function is defined as

$$C(d) = \frac{2^{1-\nu}}{\Gamma(\nu)}\left(\frac{\sqrt{8\nu}}{\rho}d\right)^{\nu} K_\nu\left(\frac{\sqrt{8\nu}}{\rho}d\right), \tag{5}$$

where ν is the smoothness parameter, ρ represents the range and K_ν is the modified Bessel function of the second kind. The range parameter measures how quickly the spatial correlation decays with spatial distance and the smoothness parameter determines how smooth

the random process is in terms of mean square differentiability (Stein 1999). We fix the smoothness parameter to $\nu = 1$ in our model due to the limitation of INLA computing algorithm (Bakka et al. 2018). Currently, R-INLA only allows values of $\nu \in (-1, 1]$ for spatial applications in two dimensions, where fields with negative values lack a point-wise interpretation (Lindgren and Rue 2015). Nevertheless, $\nu = 1$ is a reasonable choice for environmental processes and Whittle (1954) has argued that $\nu = 1$ is a more natural choice for two-dimensional processes than the exponential $\nu = 1/2$ alternative.

The spatially varying variance parameter, $\sigma_m(\boldsymbol{s})$, for $Z(\boldsymbol{s})$ switches between two regions: S_{m0} and S_{m1}. It is believed that evidence in S_{m0} is often more certain to show that this land was not suitable as refugia (e.g., evidence of ice sheets), compared to evidence in favor of S_{m1}. This is also why SDM data is often used as a classifier for refugia. Thus, we model SDM in the region S_{m0} with relatively smaller variance compared to S_{m1} as follows:

$$\log(\sigma_m(\boldsymbol{s})) = \theta_1 + \theta_2 \cdot I\{\boldsymbol{s} \in S_{m1}\},$$

where

$$I\{\boldsymbol{s} \in S_{m1}\} = \begin{cases} 1 & \text{if } \boldsymbol{s} \in S_{m1}, \\ 0 & \text{otherwise.} \end{cases}$$

In the second level of BHM, we model the unknown latent process $X(\boldsymbol{s})$ as a spatially correlated Gaussian process, and then we specify the priors in the third level to close the hierarchy.

SECOND LEVEL: LATENT SPATIAL PROCESS

$$X(\boldsymbol{s}) \sim GP\left(\mathbf{0}, \boldsymbol{\Sigma}(\sigma_x^2, \rho_x)\right). \tag{6}$$

To model the spatial correlation in $X(\boldsymbol{s})$, we again use a Matérn correlation function as defined in (5), with a range parameter ρ_x. Since there is no evidence to support that the variance of $X(\boldsymbol{s})$ should be spatially varying, we assume $X(\boldsymbol{s})$ is a stationary random process with a constant variance σ_x^2. We still fix the smoothness parameter to $\nu = 1$ for the reasons elaborated earlier and treat the variance and range parameters, σ_x^2 and ρ_x, as unknown.

THIRD LEVEL: PRIORS

$$\begin{aligned} \mu, \alpha_m, \alpha_p &\sim N(0, 1000), \\ \beta_m, \beta_p &\sim N(1, 1000), \\ \log(1/\sigma_p^2) &\sim \text{LogGamma}(1, 0.00005), \\ \sigma_g &\sim \text{PC Prior}, \\ (\sigma_x^2, \rho_x)^T &\sim \text{PC Prior}, \\ (\theta_1, \theta_2, \log \rho_m)^T &\sim N\left((0, 1, 0)^T, \boldsymbol{I}_3\right). \end{aligned}$$

The *penalized complexity* (PC) prior was introduced in Simpson et al. (2017) to create a weakly informative prior that penalizes the complexity of a hierarchical model structure. These priors assign a nonzero mass to the simplest possible model, thus allowing the data to manifest itself freely when considering the necessity of including more parameters. Additionally, the PC priors have the following nice properties: invariant to reparameterizations, having a natural connection to Jeffreys' priors, supporting Occam's Razor, and are robust. They are also easily defined by the user, who only has to specify the tail probability of the prior, giving them more straightforward interpretability. The PC prior for σ_g is an exponential with rate determined by specifying the tail probability $P(\sigma_g > 1) = 0.01$. This prior form is determined by penalizing the distance (in terms of Kullback–Leibler divergence) from a simple model with no nugget to that of a more complex model that includes one. This prior further assists in respecting genetic data as a more spatially consistent data source than pollen by giving σ_g^2 a smaller value a priori.

For the Matérn covariance parameters, the joint PC prior is specified by the following marginal tail probabilities: $P(\sigma_x > 3) = 0.01$ and $P(\rho_x < 1) = 0.01$. These tail probabilities are chosen to reflect unlikely events so that the prior can be considered weakly informative. Additionally, it penalizes complexity by shrinking the range toward infinity and the marginal variance toward zero (Fuglstad et al. 2019). In our particular scenario, the joint prior can be expressed as the product of a marginal Inverse Weibull density for the range and another Exponential for its standard deviation.

After we obtain posterior samples of $X(\boldsymbol{s})$ and μ, we apply the probability integral transform to derive the probability of refugia, $P(\boldsymbol{s}) = \Phi(\mu + X(\boldsymbol{s}))$. Then posterior inference is made on the samples of $P(\boldsymbol{s})$.

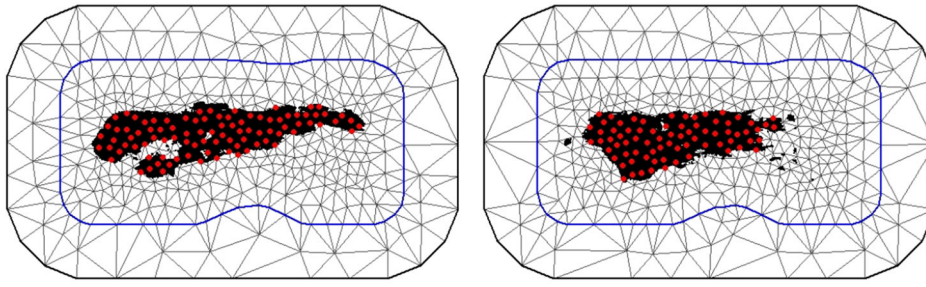
3.2. ESTIMATION USING INLA

A well-known bottleneck for large spatial data analysis is the computation of its likelihood. The number of sites in our data and of our interest makes computation a serious issue. This restricts the usage of traditional Markov chain Monte Carlo (MCMC) sampling methods for our BHM.

To bypass the computational challenge, we resort to the Integrated Nested Laplace Approximation (INLA) (Rue et al. 2009) to derive the posterior densities. INLA has been a popular strategy for Bayesian estimation for large spatial random fields, by employing approximate Bayesian inference for latent Gaussian models controlled by a small number of hyperparameters. Using integrated nested Laplace approximations, INLA can obtain fast and accurate posterior estimates compared to MCMC (Rue et al. 2009; Lindgren and Rue 2015).

INLA assumes that the latent field follows a Gaussian Markov random field (GMRF) with a sparse precision matrix, which allows for faster computations of the approximations and integrals. However, this becomes a limitation when modeling continuously indexed spatial fields.

Nevertheless, Lindgren et al. (2011) showed that an approximate stochastic weak solution to a linear stochastic partial differential equation (SPDE) will provide a Gaussian random field (GRF) with a Matérn covariance function, defined by the parameters of the SPDE. This



(a) (b)

Figure 4. Constrained refined Delaunay triangulation mesh for **a** green alder and **b** white spruce. The black regions represent S_{m0} and the red dots represent the node points that correspond to this region (Color figure online).

means that modeling can be done in continuous space using GRFs, but the inference will gain the computational speed obtainable from working with sparse precision matrices on GMRFs formulated on a triangulation of the spatial domain.

R-INLA employs a Delaunay triangulation mesh (see Fig. 4) where each vertex corresponds to a point where the GMRF is fitted. Following the approach from Lindgren et al. (2011), the solution to the SPDE is approximated by a finite sum of basis functions, giving a continuously indexed approximation of the Gaussian random field. Any point in a triangle is approximated by a linear interpolation of the basis functions used at each node. To improve the mesh construction, the border of the study region is used to delineate small and big triangles, with the smaller, more regular ones inside. This increases variability near the boundaries which helps mitigate the boundary effect of the estimations (Lindgren and Rue 2015; Bakka et al. 2018).

Having to define the spatial model in the discrete field means that the special regions for SDM, S_{m0} and S_{m1}, must also be defined in the triangular mesh. For that purpose, we must identify the nodes of the mesh that are contained in said regions. This is done by defining a radius around each node and counting the proportion of SDM sites that belong to S_{m1}. If the proportion is above a certain threshold value, then we count the node as being part of S_{m1}. Both the radius and threshold are tailored by the user to achieve reasonable results. For both species, a radius of 0.7 and a threshold proportion of 0.25 were used to define the S_{m1} mesh nodes (see the red dots in Fig. 4).

The R-INLA package, obtained from www.r-inla.org, was used to run the INLA method for the Bayesian inference (Lindgren and Rue 2015).

4. SIMULATION STUDY

We conduct a small simulation study to verify that our method recovers the underlying refugia probability $P(\boldsymbol{s})$ if the data follow the models outlined in (4). We also evaluate the sensitivity of our estimates to the number of SDM sites as we will use only a subset of the very dense SDM data in our real data analysis.

Table 1. Different sampling scenarios for SDM (n_m), pollen (n_p), and genetic (n_g) data

Sampling scenario	n_m	n_p	n_g
S1	500	15	60
S2	1000	15	60
S3	1000	1000	1000

4.1. SETUP

To mimic the real data, we adopt Eastern Beringia as our spatial domain and randomly select 60 locations for genetic data and 15 locations for pollen data. We randomly choose 500 SDM locations which are much sparser than the SDM data we have, for ease of computation. We also consider the scenario of 1000 SDM locations to evaluate the sensitivity of our estimation to the amount of SDM data used in the model. Additionally, we consider a third scenario with 1000 locations for each of the genetic, pollen, and SDM data, to study the effect of the sparsity of genetic and pollen data on the refugia probability estimation. These three scenarios are summarized in Table 1. In addition to the sites with observations, we also randomly sample 200 locations that do not overlap with the data sites and will be used for evaluating the probability estimation. For each scenario, we run the simulation 50 times.

We first simulate the $X(\boldsymbol{s})$ process by following the latent process model (6), and then generate $Y_p(\boldsymbol{s})$ and $Y_g(\boldsymbol{s})$ following their corresponding models in (4). All parameters in the models are assigned values that represent the conditions of the real data and are deferred to the Supplement. The SDM data generation requires simulating the spatial error $Z(\boldsymbol{s})$ that has different variances depending on the region. To accomplish that, we first simulate a $Z(\boldsymbol{s})$ process with fixed variance $\sigma_m^2(S_{m0})$ for all n_m sites, based on which we calculate a preliminary $P(Y_m(s) = 1)$ using the logit model in (4). For a given τ_m, all sites such that $P(Y_m(s) = 1) > \tau_m$ have their $Z(\boldsymbol{s})$ multiplied by a correction factor to switch their variance to $\sigma_m^2(S_{m1})$. We recalculate $P(Y_m(s) = 1)$ and finally, the $Y_m(\boldsymbol{s})$ process is generated based on these updated probabilities and the threshold τ_m.

4.2. COMPARING $\hat{P}(\boldsymbol{s})$ AND $P(\boldsymbol{s})$

To obtain estimates of $P(s)$, for a given site $\boldsymbol{s}$, we generate posterior samples of $P(\boldsymbol{s}) = \Phi(\mu + X(\boldsymbol{s}))$ and take the posterior mean as our estimate $\hat{P}(\boldsymbol{s})$. We estimate $P(\boldsymbol{s})$ at all sites with observations and the extra 200 data-absent locations. We calculate the typical mean squared error (MSE) of $\hat{P}(\boldsymbol{s})$ as one measure of the estimation performance. Since in our data application, identification of refugia is based on the relative size of the $\hat{P}(\boldsymbol{s})$, capturing the spatially varying pattern is the key to correctly differentiating refugia and non-refugia areas. For this reason, we also use the Pearson correlation between $\hat{P}(\boldsymbol{s})$ and $P(\boldsymbol{s})$ to measure the performance of $\hat{P}(\boldsymbol{s})$.

Figure 5 presents the simulation results summarized over all locations. Scenario 1 (S1) and 2 (S2) are comparable in their correlation and MSE, with S2 performing slightly worse in terms of MSE but slightly better in correlation. This indicates that the probability estimation

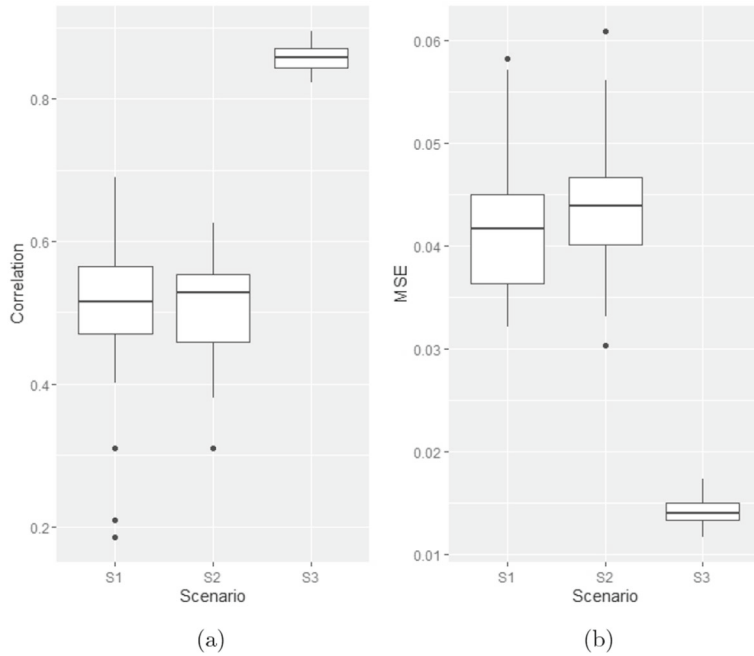


Figure 5. Simulation results showing **a** correlation and **b** MSE between the estimated probabilities, $\hat{P}(\boldsymbol{s})$, and true probabilities, $P(\boldsymbol{s})$, for the three distinct sampling scenarios, S1, S2, and S3 .

is insensitive to the number of SDM sites, due to the binary feature of SDM data. This supports our choice of using a random subset of all available SDM sites in the real data analysis. However, since genetic and pollen data are more informative, the abundance of these data can have a significant impact on the estimation. We observe low correlations between the estimated and true probabilities, around 0.5, for Scenarios S1 and S2, indicating the struggle of estimates with scarce genetic and pollen data. Results for S3 show that with dense genetic and pollen data, the capacity of our model for estimating the refugia probability is much improved.

We also evaluate the estimation performance at genetic, pollen, and SDM sites as well as data-absent locations separately. The pattern of three scenarios for each category remains similar as Fig. 5 shows. There is not much difference between the different types of sites, with those without any data being only slightly less accurate than the rest. The full extent of these results are deferred to the Supplement.

Figure 6 compares the estimated and the true refugia probability of one particular simulation run for both S2 and S3. The particular run was chosen to correspond to the median correlation between $\hat{P}(\boldsymbol{s})$ and $P(\boldsymbol{s})$ for their respective scenario. With fewer genetic and pollen data in S2, the estimation show a blurry version of the true probability. Nevertheless, the estimation still captures a rough pattern of high and low probabilities. With the ideal situation of S3 that has dense genetic and pollen data, the estimated refugia probabilities recover considerable amount of details of the true probabilities. Although it is unclear how to exactly interpret the empirical coverage of credible intervals, we report the coverage for

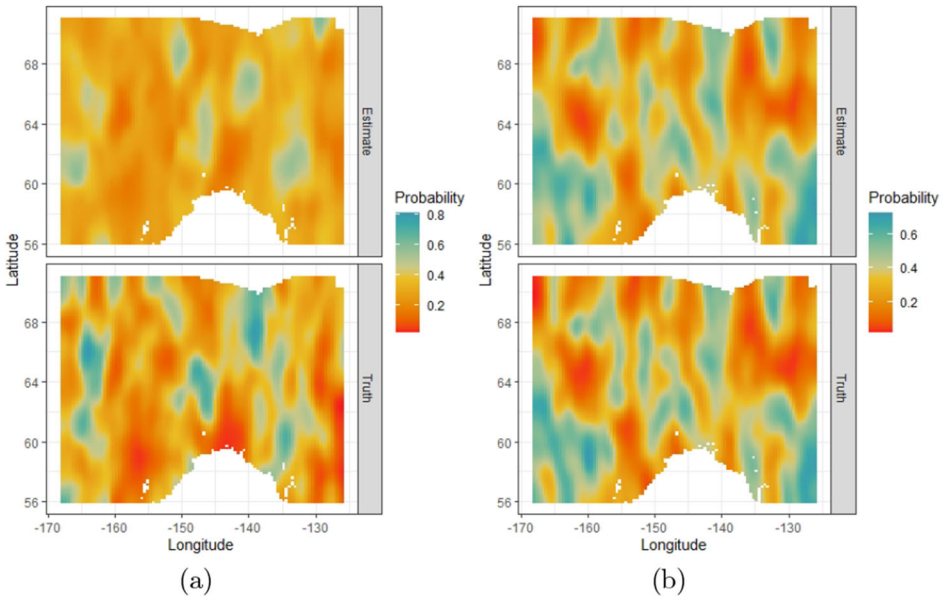


Figure 6. $\hat{P}(\boldsymbol{s})$ and $P(\boldsymbol{s})$ of the simulation run for which the correlation between $\hat{P}(\boldsymbol{s})$ and $P(\boldsymbol{s})$ is the median for **a** S2 and **b** S3.

$P(\boldsymbol{s})$ in the Supplement. The empirical coverage for S2, which resembles the real data, is lower than 95% by 10%. This might imply our uncertainty estimate for the real data could be somewhat lower than it should be.

In addition, we found the variance parameter estimates for pollen, genetic and SDM seem to suffer from a bias though the parameter of primary interest, σ_x, can be estimated with no bias, as shown in Fig. 7. In practice, it is often challenging to accurately estimate all the parameters, particularly those in the variance-covariance matrix, of a complex system, especially when data is sparse. It is also often difficult to diagnose issues for a complex system. To investigate the cause of bias, we consider a simple spatial model as if only genetic data is observed in the first level:

$$Y_g(\boldsymbol{s}) = \mu + X(\boldsymbol{s}) + \epsilon_g(\boldsymbol{s}),$$

where $X(\boldsymbol{s})$ is a Gaussian process with Matérn correlation function and variance σ_x^2, and $\epsilon(\boldsymbol{s})$ are white noise with variance σ_g^2. We treat the Matérn correlation structure of X as known so we can focus on only estimating σ_x and σ_g. We generate data on 1000 locations and then fit the model using INLA and maximum likelihood. Both methods offered an unbiased estimate for σ_x, whereas the INLA estimate for σ_g^2 remains biased while the maximum likelihood estimate is unbiased. We repeat the experiment using different prior specifications for σ_g^2 but observe the same pattern, even when centering the prior around the true value of σ_g^2. It is unclear what leads to the difference. It might be due to the approximation employed by INLA or we may need to explore alternative mesh structures for INLA. A thorough investigation of the computation approach is currently underway but beyond the scope of this project.

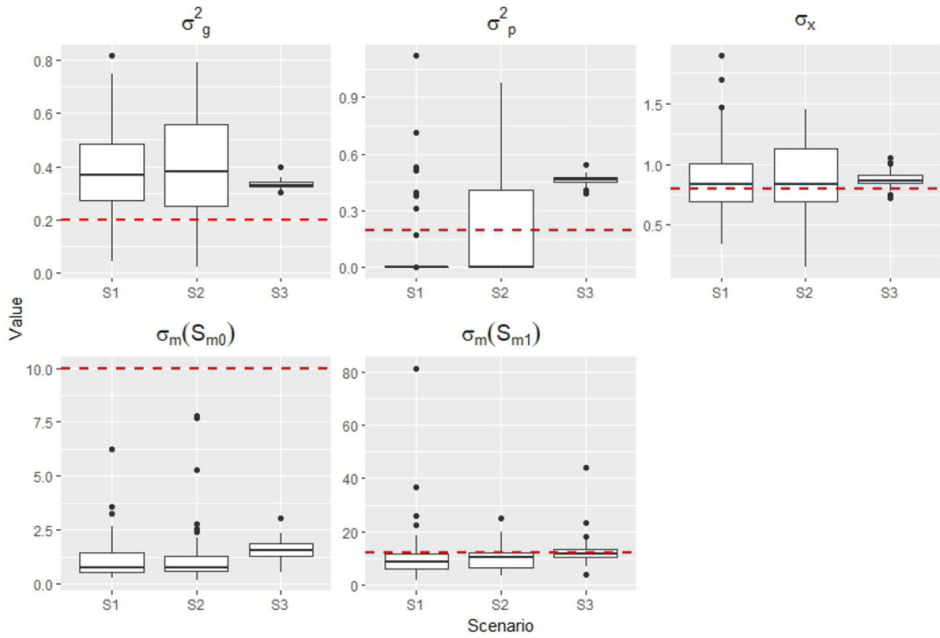


Figure 7. Posterior means of five variability parameters for the three different sampling scenarios (S1, S2, and S3) based on 50 simulations. The red dashed lines represent the respective true value used in simulating the data.

5. REFUGIA FOR GREEN ALDER AND WHITE SPRUCE

We apply our BHM to the green alder and white spruce data to unveil the refugia of these two species in Eastern Beringia during the Last Glacial Maximum (LGM, 23-19 thousand years ago) period.

5.1. GREEN ALDER

SDM output for green alder is composed of 334,000 sites that cover the Alaskan peninsula and parts of adjacent Canada. To ease the computation for this very large and dense data set, a random sample of 50,000 sites was drawn and used in the analysis. For the other two data sources, there were 47 genetic and 18 pollen observations, as shown in Fig. 1a.

The posterior estimates of model parameters and hyperparameters are reported in Table 2. The negative estimate for μ suggests that on average the true probabilities are smaller than 0.5. This is mostly due to the overall SDM values being zero, but also to a lesser extent to using $r = 3$ for genetic interpolation which ensures most genetic probabilities are low. Additionally, all sources of evidence seem to agree among themselves since they all have positive β, i.e., if an area has high observed probabilities among data sources then the common probability of that area being refugia is also expected to be high. The posterior probability for $\theta_2 > 0$ was calculated and was found to be almost 1. This seems to signal that there is some tangible difference in uncertainty between S_{m0} and S_{m1}, with S_{m1} having a greater variability.

Table 2. Posterior estimates of *Alnus viridis* model

Parameter	Mean	St. Dev	2.5% Quantile	97.5% Quantile
μ	−0.5599	0.307	−1.1433	0.0805
α_m	−29.5723	8.0242	−48.8235	−17.0337
α_p	−0.2138	1.0189	−1.9813	2.0104
β_m	0.661	1.2104	−1.8335	2.912
β_p	1.2099	1.1049	−0.8111	3.5188
$\sigma_m(S_{m0})$	19.6507	4.5121	14.5895	27.0541
$\sigma_m(S_{m1})$	23.6677	4.7163	17.8428	31.3969
σ_p^2	8.1744	3.0616	3.8316	15.7032
σ_g^2	0.4164	0.3083	0.0897	1.2436
σ_x	1.4068	0.3866	0.7915	2.2974
ρ_m	11.5636	2.9265	8.4688	16.2158
ρ_x	2.564	0.9674	1.1734	4.9291

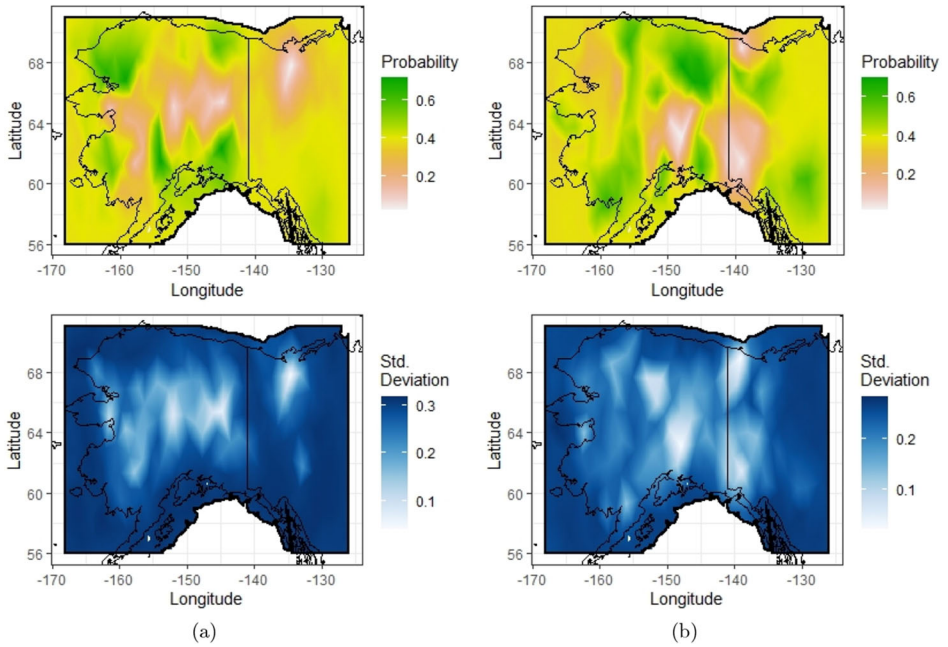


Figure 8. Posterior mean and standard deviation of the true probability of being refugia, $P(\boldsymbol{s})$, of **a** green alder and **b** white spruce. The shaded region corresponds to Eastern Beringia during the Last Glacial Maximum, with modern-day Alaska superimposed for reference (Color figure online) .

Figure 8a shows the posterior mean and standard deviation of $P(\boldsymbol{s})$ on all sites within the study region. We can see that the areas of low variability coincide with the regions where we have pollen or genetic data, whereas high uncertainty occurs in places where neither of these two data sources is present. Reduced variability at the co-occurrence of disparate data suggests that the different lines of evidence provide unique, complementary information about the true source locations of past populations (Gavin et al. 2014). Our results broadly

Table 3. Posterior estimates of *Picea glauca* model

Parameter	Mean	St. Dev	2.5% Quantile	97.5% Quantile
μ	−0.4612	0.247	−0.9454	0.0263
α_m	−25.6744	6.7762	−39.3970	−12.9561
α_p	1.5932	0.1800	1.2677	1.9845
β_m	1.2203	0.2027	0.7864	1.6075
β_p	1.021	0.192	0.6602	1.4474
$\sigma_m(S_{m0})$	15.0940	1.6044	11.4026	17.5576
$\sigma_m(S_{m1})$	16.4105	1.7294	13.0457	18.7793
σ_p^2	0.0002	0.0003	0	0.0009
σ_g^2	0.6045	0.1185	0.4148	0.8778
σ_x	1.0954	0.0888	0.9424	1.3586
ρ_m	12.2107	1.4162	9.0695	14.8249
ρ_x	3.6014	0.5473	2.5171	4.967

match the findings of Napier et al. (2019) who analyzed the same SDM and genetic dataset but in a framework that did not integrate those two data sources.

5.2. WHITE SPRUCE

For white spruce, there are also 334,000 sites of SDM information covering the same study region as green alder. Likewise, we randomly sampled 50,000 sites for our analysis. Additionally, there were 79 genetic and 14 pollen observations, as shown in Fig. 1b.

The posterior estimates for all model parameters and hyperparameters are shown in Table 3. Notice that the estimate for σ_p^2 is very small. This is most likely due to the pollen data for white spruce being almost invariable. Even though this hyperparameter represents the uncertainty of pollen data, which is expected to be larger than that of genetic information, the small variation of the pollen data caps the magnitude of the values σ_p^2 can take. Similar to green alder, the overall mean is also negative for the white spruce estimation, thus suggesting that indeed the locations of refugia are sparse. Furthermore, the variance for the z process has a posterior probability of 0.896 of being larger in S_{m1} than in S_{m0}.

Figure 8b shows the mean and standard deviation of the posterior density of $P(\boldsymbol{s})$ on all sites within the study region. Same as with green alder, the regions where there are genetic and pollen data have lower variability, while the regions lacking those two types of data have increased uncertainty. This is again evidence of how sites with multiple sources of evidence coincide in the estimation of $P(\boldsymbol{s})$. Since our results for white spruce are the first attempt to find the exact locations of refugia, there is no existing literature to verify our results yet. However, the resulting highlighted regions, such as a possible refugium in Alaska, seem reasonable from a scientific point of view.

6. CONCLUSIONS

We propose an innovative Bayesian hierarchical model to utilize the diverse pieces of evidence collected in paleoecology to uncover cryptic refugia. Specifically, we integrate SDMs,

pollen fossil records, and genetic surveys as three complementary sources of evidence to produce a single unified map showing the possible refugia locations in Eastern Beringia for the *Alnus viridis* and *Picea glauca*, respectively. The simplicity of the model plus the computational convenience offered by INLA allow for researchers to quickly and efficiently implement the model with their data sources. The flexibility given by the Bayesian hierarchical modeling also allows researchers to model their specific data sources in any way they consider best. Furthermore, we hope that the method disclosed in this paper can turn into a powerful and useful tool for further investigations in paleoecology for many other species. With the insight gained from such studies, we can better prepare for the coming challenges brought by rapid climate change.

Our models are constructed based on the perceived data quality and where each data source exhibits credibility. According to their reliability, pollen and genetic data were modeled using linear biases, whereas SDM used a nonlinear bias with a logit transformation. This allows the results to resemble more accurate data sources while receiving due help from other available evidence. The high uncertainty of the probability estimates in regions far from pollen and genetic data suggests that SDM provides only a little information to identify high-plausible refugia. However, SDM helps identify the areas where refugia can be safely discarded, as evidenced by the reduced variance of S_{m0} in both species. Pollen and genetic data are then responsible for bringing to light some specific regions of high probability.

Due to the intricate relationship between the raw data and the true probability of being refugia, we transform the data beforehand to enable a more straightforward relationship to the true probabilities based on empirical experience and prior knowledge. Alternatively, we could consider incorporating the data pre-processing process into the hierarchical model and learning all parameters from the data. However, there are no widely accepted pre-possessing models for us to borrow at this point. We thus only focus on integrating different lines of evidence rather than extensively exploring pre-processing approaches. Although we consider those transformations sound choices, we acknowledge that further investigation is needed.

Even though we try to rigorously integrate three different data sources to provide the refugia estimates, our method is still an indirect approach to detecting refugia. It would be overly optimistic to conclude that these highlighted regions are the sites of glacial refugia. The results obtained through our analysis should be treated with caution. To formally confirm some locations to be refugia of a species, we need to find in situ macrofossils that can prove this species' past presence. Searching for such macrofossils is an extremely difficult task without prior knowledge of where to target (de Lafontaine et al. 2014). The plausible refugia we identified here, however, can be safely used to guide the search for actual proof of the species' past presence and inform future field expeditions for the further study of cryptic refugia in Eastern Beringia.

Finally, we notice our estimation of some nuisance parameters in the model carries bias. We suspect that the Markov chain Monte Carlo estimation may alleviate this issue, though the computation can become an obstacle due to the large number of spatial locations. Nonetheless, many methods have been developed to ease computation for large spatial data such as lattice kriging (Nychka et al. 2015), fixed-rank kriging (Cressie and Johannesson

2008), multi-resolution approximations (Katzfuss 2017), and nearest neighbors processes (Datta et al. 2016), among others.

ACKNOWLEDGEMENTS

The authors would like to thank the anonymous referees, an Associate Editor, and the Editor for their constructive comments that improved the quality of this paper.

Funding This research is supported by NSF-1418339, NSF-2124576 and NSF-2118329.

Declarations

Conflicts of interest The authors declare no conflicts of interest.

[Received July 2022. Revised October 2023. Accepted October 2023.]

REFERENCES

Abbott RJ, Brochmann C (2003) History and evolution of the arctic flora: in the footsteps of Eric hultén. Mol Ecol 12(2):299–313

Allouche O, Tsoar A, Kadmon R (2006) Assessing the accuracy of species distribution models: prevalence, kappa and the true skill statistic (TSS). J Appl Ecol 43(6):1223–1232

Anderson PM, Brubaker LB (1994) Vegetation history of northcentral alaska: a mapped summary of late-quaternary pollen data. Quatern Sci Rev 13(1):71–92

Anderson LL, Hu FS, Nelson DM, Petit RJ, Paige KN (2006) Ice-age endurance: Dna evidence of a white spruce refugium in alaska. Proc Natl Acad Sci 103(33):12447–12450

Anderson LL, Hu FS, Paige KN (2011) Phylogeographic history of white spruce during the last glacial maximum: uncovering cryptic refugia. J Hered 102(2):207–216

Aoki K, Tamaki I, Nakao K, Ueno S, Kamijo T, Setoguchi H, Murakami N, Kato M, Tsumura Y (2019) Approximate Bayesian computation analysis of est-associated microsatellites indicates that the broadleaved evergreen tree *castanopsis sieboldii* survived the last glacial maximum in multiple refugia in japan. Heredity 122(3):326–340

Bakka H, Rue H, Fuglstad G-A, Riebler A, Bolin D, Illian J, Krainski E, Simpson D, Lindgren F (2018) Spatial modeling with r-inla: a review. WIREs Comput Stat 10(6):e1443

Bigelow NH, Brubaker LB, Edwards ME, Harrison SP, Prentice IC, Anderson PM, Andreev A A, Bartlein PJ, Christensen TR, Cramer W, Kaplan JO, Lozhkin AV, Matveyeva NV, Murray, DF, McGuire,AD, Razzhivin VY, Ritchie JC, Smith B, Walker DA, Gajewski K, Wolf V, Holmqvist BH, Igarashi Y, Kremenetskii K, Paus A, isaric MFJ, Volkova VS (2003) Climate change and arctic ecosystems: 1. vegetation changes north of 55°n between the last glacial maximum, mid-holocene, and present. J Geophys Res Atmosp 108(D19)

Botkin DB, Saxe H, Araújo MB, Betts R, Bradshaw RHW, Cedhagen T, Chesson P, Dawson TP, Etterson JR, Faith DP, Ferrier S, Guisan A, Hansen AS, Hilbert DW, Loehle C, Margules C, New M, Sobel MJ, Stockwell DRB (2007) Forecasting the effects of global warming on biodiversity. Bioscience 57(3):227–236

Brown JL, Knowles LL (2012) Spatially explicit models of dynamic histories: examination of the genetic consequences of pleistocene glaciation and recent climate change on the American Pika. Mol Ecol 21(15):3757–3775

Brubaker LB, Anderson PM, Edwards ME, Lozhkin AV (2005) Beringia as a glacial refugium for boreal trees and shrubs: new perspectives from mapped pollen data. J Biogeogr 32(5):833–848

Budde K, González-Martínez SC, Hardy OJ, Heuertz M (2013) The ancient tropical rainforest tree symphonia globulifera l. f. (clusiaceae) was not restricted to postulated Pleistocene refugia in Atlantic equatorial Africa. Heredity 111(1):66–76

Clark JS (2005) Why environmental scientists are becoming bayesians. Ecol Lett 8(1):2–14

Cornejo-Romero A, Vargas-Mendoza CF, Aguilar-Martínez GF, Medina-Sánchez J, Rendón-Aguilar B, Valverde PL, Zavala-Hurtado JA, Serrato A, Rivas-Arancibia S, Pérez-Hernández MA et al (2017) Alternative glacial-interglacial refugia demographic hypotheses tested on cephalocereus columna-trajani (cactaceae) in the intertropical Mexican drylands. PLoS ONE 12(4):e0175905

Cressie N, Johannesson G (2008) Fixed rank kriging for very large spatial data sets. J R Stat Soc Ser B Stat Methodol 70(1):209–226

Datta A, Banerjee S, Finley AO, Gelfand AE (2016) Hierarchical nearest-neighbor gaussian process models for large geostatistical datasets. J Am Stat Assoc 111(514):800–812

Davis MB, Shaw RG (2001) Range shifts and adaptive responses to quaternary climate change. Science 292(5517):673–679

Dawson TP, Jackson ST, House JI, Prentice IC, Mace GM (2011) Beyond predictions: biodiversity conservation in a changing climate. Science 332(6025):53–58

De Lafontaine G, Ducousso A, Lefèvre S, Magnanou E, Petit RJ (2013) Stronger spatial genetic structure in recolonized areas than in refugia in the European beech. Mol Ecol 22(17):4397–4412

de Lafontaine G, Amasifuen Guerra CA, Ducousso A, Sanchez-Goni M-F, Petit RJ (2014) Beyond skepticism: uncovering cryptic refugia using multiple lines of evidence. New Phytol 204(3):450–454

de Lafontaine G, Turgeon J, Payette S (2010) Phylogeography of white spruce (*Picea glauca*) in eastern north America reveals contrasting ecological trajectories. J Biogeogr 37(4):741–751

de Lafontaine G, Napier JD, Petit RJ, Hu FS (2018) Invoking adaptation to decipher the genetic legacy of past climate change. Ecology 99(7):1530–1546

Espíndola A, Pellissier L, Maiorano L, Hordijk W, Guisan A, Alvarez N (2012) Predicting present and future intra-specific genetic structure through niche hindcasting across 24 millennia. Ecol Lett 15(7):649–657

Excoffier L, Smouse PE (1994) Using allele frequencies and geographic subdivision to reconstruct gene trees within a species: molecular variance parsimony. Genetics 136(1):343–359

Feurdean A, Bhagwat SA, Willis KJ, Birks HJB, Lischke H, Hickler T (2013) Tree migration-rates: narrowing the gap between inferred post-glacial rates and projected rates. PLoS ONE 8(8):e71797

Franklin J (2010) Mapping species distributions: spatial inference and prediction. Cambridge University Press, Cambridge

Fuglstad G-A, Simpson D, Lindgren F, Rue H (2019) Constructing priors that penalize the complexity of gaussian random fields. J Am Stat Assoc 114(525):445–452

Gao J, Wang B, Mao JF, Ingvarsson P, Zeng QY, Wang XR (2012) Demography and speciation history of the homoploid hybrid pine Pinus densata on the Tibetan plateau. Mol Ecol 21(19):4811–4827

Gavin DG, Fitzpatrick MC, Gugger PF, Heath KD, Rodríguez-Sánchez F, Dobrowski SZ, Hampe A, Hu FS, Ashcroft MB, Bartlein PJ et al (2014) Climate refugia: joint inference from fossil records, species distribution models and phylogeography. New Phytol 204(1):37–54

Graham CH, VanDerWal J, Phillips SJ, Moritz C, Williams SE (2010) Dynamic refugia and species persistence: tracking spatial shifts in habitat through time. Ecography 33(6):1062–1069

Guisan A, Thuiller W (2005) Predicting species distribution: offering more than simple habitat models. Ecol Lett 8(9):993–1009

Hampe A, Jump AS (2011) Climate relicts: past, present, future. Annu Rev Ecol Evol Syst 42:313–333

Hao Q, de Lafontaine G, Guo D, Gu H, Hu FS, Han Y, Song Z, Liu H (2018) The critical role of local refugia in postglacial colonization of Chinese pine: joint inferences from dna analyses, pollen records, and species distribution modeling. Ecography 41(4):592–606

Hewitt G (2000) The genetic legacy of the quaternary ice ages. Nature 405(6789):907–913

Hopkins D, Smith P, Matthews J (1981) Dated wood from Alaska and the Yukon: implications for forest refugia in Beringia. Quatern Res 15(3):217–249

Katzfuss M (2017) A multi-resolution approximation for massive spatial datasets. J Am Stat Assoc 112(517):201–214

Keppel G, Van Niel KP, Wardell-Johnson GW, Yates CJ, Byrne M, Mucina L, Schut AG, Hopper SD, Franklin SE (2012) Refugia: identifying and understanding safe havens for biodiversity under climate change. Glob Ecol Biogeogr 21(4):393–404

Knowles LL, Alvarado-Serrano DF (2010) Exploring the population genetic consequences of the colonization process with spatio-temporally explicit models: insights from coupled ecological, demographic and genetic models in montane grasshoppers. Mol Ecol 19(17):3727–3745

Lemey P, Rambaut A, Drummond AJ, Suchard MA (2009) Bayesian phylogeography finds its roots. PLoS Comput Biol 5(9):e1000520

Lemey P, Rambaut A, Welch JJ, Suchard MA (2010) Phylogeography takes a relaxed random walk in continuous space and time. Mol Biol Evol 27(8):1877–1885

Lemmon AR, Lemmon EM (2008) A likelihood framework for estimating phylogeographic history on a continuous landscape. Syst Biol 57(4):544–561

Li B, Nychka DW, Ammann CM (2010) The value of multiproxy reconstruction of past climate. J Am Stat Assoc 105(491):883–895

Li L, Abbott RJ, Liu B, Sun Y, Li L, Zou J, Wang X, Miehe G, Liu J (2013) Pliocene intraspecific divergence and plio-pleistocene range expansions within *Picea likiangensis* (Lijiang spruce), a dominant forest tree of the Ginghai-tibet plateau. Mol Ecol 22(20):5237–5255

Lindgren F, Rue H (2015) Bayesian spatial modelling with R-INLA. J Stat Softw 63(19)

Lindgren F, Rue H, Lindström J (2011) An explicit link between gaussian fields and Gaussian Markov random fields: the stochastic partial differential equation approach. J Roy Stat Soc Ser B (Stat Methodol) 73(4):423–498

Luoto M, Heikkinen RK (2008) Disregarding topographical heterogeneity biases species turnover assessments based on bioclimatic models. Glob Change Biol 14(3):483–494

Magri D, Vendramin GG, Comps B, Dupanloup I, Geburek T, Gömöry D, Lataıowa M, Litt T, Paule L, Roure JM, Tantau I, Van Der Knaap WO, Petit RJ, De Beaulieu J-L (2006) A new scenario for the quaternary history of European beech populations: palaeobotanical evidence and genetic consequences. New Phytol 171(1):199–221

Manolopoulou I, Emerson BC (2012) Phylogeographic ancestral inference using the coalescent model on haplotype trees. J Comput Biol 19(6):745–755

Marion G, McInerny GJ, Pagel J, Catterall S, Cook AR, Hartig F, O'Hara RB (2012) Parameter and uncertainty estimation for process-oriented population and distribution models: data, statistics and the niche. J Biogeogr 39(12):2225–2239

Marske KA, Leschen RA, Buckley TR (2012) Concerted versus independent evolution and the search for multiple refugia: comparative phylogeography of four forest beetles. Evolut Int J Organ Evolut 66(6):1862–1877

McLachlan JS, Clark JS, Manos PS (2005) Molecular indicators of tree migration capacity under rapid climate change. Ecology 86(8):2088–2098

Meirmans PG, Liu S (2018) Analysis of molecular variance (Amova) for autopolyploids. Front Ecol Evol 6:66

Mosblech NAS, Bush MB, van Woesik R (2011) On metapopulations and microrefugia: palaeoecological insights. J Biogeogr 38(3):419–429

Napier JD, de Lafontaine G, Heath KD, Hu FS (2019) Rethinking long-term vegetation dynamics: multiple glacial refugia and local expansion of a species complex. Ecography 42(5):1056–1067

Napier JD, de Lafontaine G, Chipman ML (2020a) The evolution of paleoecology. Trends Ecol Evolut 35(4):293–295

Napier JD, Fernandez MC, de Lafontaine G, Hu FS (2020b) Ice-age persistence and genetic isolation of the disjunct distribution of larch in Alaska. Ecol Evol 10(3):1692–1702

Nogués-Bravo D (2009) Predicting the past distribution of species climatic niches. Glob Ecol Biogeogr 18(5):521–531

Nychka D, Bandyopadhyay S, Hammerling D, Lindgren F, Sain S (2015) A multiresolution gaussian process model for the analysis of large spatial datasets. J Comput Graph Stat 24(2):579–599

Pagel J, Schurr FM (2012) Forecasting species ranges by statistical estimation of ecological niches and spatial population dynamics. Glob Ecol Biogeogr 21(2):293–304

Parducci L, Jørgensen T, Tollefsrud MM, Elverland E, Alm T, Fontana SL, Bennett KD, Haile J, Matetovici I, Suyama Y, Edwards ME, Andersen K, Rasmussen M, Boessenkool S, Coissac E, Brochmann C, Taberlet P, Houmark-Nielsen M, Larsen NK, Orlando L, Gilbert MTP, Kjær KH, Alsos IG, Willerslev E (2012) Glacial survival of boreal trees in northern Scandinavia. Science 335(6072):1083–1086

Petit RJ, Aguinagalde I, de Beaulieu J-L, Bittkau C, Brewer S, Cheddadi R, Ennos R, Fineschi S, Grivet D, Lascoux M, Mohanty A, Müller-Starck G, Demesure-Musch B, Palmé A, Martín JP, Rendell S, Vendramin GG (2003) Glacial refugia: hotspots but not melting pots of genetic diversity. Science 300(5625):1563–1565

Porto TJ, Carnaval AC, da Rocha PLB (2013) Evaluating forest refugial models using species distribution models, model filling and inclusion: a case study with 14 Brazilian species. Divers Distrib 19(3):330–340

Pritchard JK, Stephens M, Donnelly P (2000) Inference of population structure using multilocus genotype data. Genetics 155(2):945–959

Provan J, Bennett K (2008) Phylogeographic insights into cryptic glacial refugia. Trends Ecol Evolut 23(10):564–571

Randin CF, Engler R, Normand S, Zappa M, Zimmermann NE, Pearman PB, Vittoz P, Thuiller W, Guisan A (2009) Climate change and plant distribution: local models predict high-elevation persistence. Glob Change Biol 15(6):1557–1569

Ren G, Mateo RG, Liu J, Suchan T, Alvarez N, Guisan A, Conti E, Salamin N (2017) Genetic consequences of quaternary climatic oscillations in the Himalayas: *Primula tibetica* as a case study based on restriction site-associated dna sequencing. New Phytol 213(3):1500–1512

Rue H, Martino S, Chopin N (2009) Approximate Bayesian inference for latent gaussian models by using integrated nested Laplace approximations. J Roy Stat Soc Ser B (Stat Methodol) 71(2):319–392

Schurr FM, Pagel J, Cabral JS, Groeneveld J, Bykova O, O'Hara RB, Hartig F, Kissling WD, Linder HP, Midgley GF et al (2012) How to understand species' niches and range dynamics: a demographic research agenda for biogeography. J Biogeogr 39(12):2146–2162

Shafer AB, Cullingham CI, Cote SD, Coltman DW (2010) Of glaciers and refugia: a decade of study sheds new light on the phylogeography of northwestern North America. Mol Ecol 19(21):4589–4621

Simpson D, Rue H, Riebler A, Martins TG, Sørbye SH (2017) Penalising model component complexity: a principled, practical approach to constructing priors. Stat Sci 32(1):1–28

Stein ML (1999) Interpolation of spatial data: some theory for kriging. Springer, Berlin

Stewart JR, Lister AM, Barnes I, Dalén L (2010) Refugia revisited: individualistic responses of species in space and time. Proc Roy Soc B Biol Sci 277(1682):661–671

Svenning J-C, Fløjgaard C, Marske KA, Nógues-Bravo D, Normand S (2011) Applications of species distribution modeling to paleobiology. Quatern Sci Rev 30(21–22):2930–2947

Thuiller W, Lafourcade B, Engler R, Araújo MB (2009) Biomod-a platform for ensemble forecasting of species distributions. Ecography 32(3):369–373

Tsuda Y, Chen J, Stocks M, Källman T, Sønstebø JH, Parducci L, Semerikov V, Sperisen C, Politov D, Ronkainen T et al (2016) The extent and meaning of hybridization and introgression between Siberian spruce (*Picea obovata*) and Norway spruce (*Picea abies*): cryptic refugia as stepping stones to the west? Mol Ecol 25(12):2773–2789

Urban MA, Nelson DM, Kelly R, Ibrahim T, Dietze M, Pearson A, Hu FS (2013) A hierarchical Bayesian approach to the classification of c3 and c4 grass pollen based on spiral δ13c data. Geochim Cosmochim Acta 121:168–176

Wang Q, Liu J, Allen GA, Ma Y, Yue W, Marr KL, Abbott RJ (2016) Arctic plant origins and early formation of circumarctic distributions: a case study of the mountain sorrel, *Oxyria digyna*. New Phytol 209(1):343–353

Warren E, de Lafontaine G, Gérardi S, Senneville S, Beaulieu J, Perron M, Jaramillo-Correa JP, Bousquet J (2016) Joint inferences from cytoplasmic dna and fossil data provide evidence for glacial vicariance and contrasted post-glacial dynamics in tamarack, a transcontinental conifer. J Biogeogr 43(6):1227–1241

Whittle P (1954) On stationary processes in the plane. Biometrika 41(3/4):434–449

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Springer Nature or its licensor (e.g. a society or other partner) holds exclusive rights to this article under a publishing agreement with the author(s) or other rightsholder(s); author self-archiving of the accepted manuscript version of this article is solely governed by the terms of such publishing agreement and applicable law.