Liu Yao (Orcid ID: 0000-0003-2783-3291)
Ogle Kiona (Orcid ID: 0000-0002-0652-8397)
Jackson Stephen T. (Orcid ID: 0000-0002-1487-4652)

Journal: Ecological Monographs
Manuscript type: Article

Title

Estimation of pollen productivity and dispersal: How pollen assemblages in small lakes represent vegetation

Authors

Yao Liu[1,2,*]
Kiona Ogle[3,4,5]
Jeremy W. Lichstein[6]
Stephen T. Jackson[7,8]

Affiliations

1. Department of Geography and Environmental Sciences, Northumbria University, Newcastle upon Tyne, UK
2. Environmental Sciences Division & Climate Change Science Institute, Oak Ridge National Laboratory, Oak Ridge, TN, USA
3. The School of Informatics, Computing, and Cyber Systems, Northern Arizona University, Flagstaff, AZ, USA
4. Center for Ecosystem Science and Society, Northern Arizona University, Flagstaff, AZ, USA
5. Department of Biological Sciences, Northern Arizona University, Flagstaff, AZ, USA
6. Department of Biology, University of Florida, Gainesville, FL, USA.
7. US Geological Survey, Southwest and South Central Climate Adaptation Science Centers, Tucson, AZ, USA
8. Department of Geosciences and School of Natural Resources and Environment, University of Arizona, Tucson, AZ, USA

* Corresponding Author. E-mail: yao2.liu@northumbria.ac.uk

Received 20 July 2021; revised 7 November 2021; accepted 29 December 2021
Handling Editor: Rebecca E. Irwin

Open research

All data used in our analyses are publicly available. Lake pollen and local vegetation survey data are freely available through Jackson (2019) at https://doi.org/10.1002/ecy.2784. Tree abundance in concentric rings from 1 km to 300 km distance from each lake were estimated using data from the publicly available U.S. Forest Inventory and Analysis (FIA) database version 5.1, downloaded in November 2012 from www.fia.fs.fed.us. We used all available field-measured plot samples during the years 2000–2012. These plots were identified as follows: (1) at least one accessible forest condition present (FIA Condition table COND_STATUS_CD = 1); standard production plot (FIA Plot table QA_STATUS = 1 or QA_STATUS = No Data); and (3) plot was measured, as opposed to modeled (FIA Plot table KINDCD not = 4). The Bayesian statistical model was coded in R and is available in Data S1 (pollen_model.r) and also archived on Zenodo at https://doi.org/10.5281/zenodo.5825842 (Liu 2022).

Abstract

Despite ongoing advances, quantitative understanding of vegetation dynamics over timespans beyond a century remains limited. In this regard, pollen-based reconstruction of past vegetation enables unique research opportunities by quantifying changes in plant community compositions over hundreds to thousands of years. Critically, the methodological basis for most reconstruction approaches rests upon estimates of pollen productivity and dispersal. However, previous studies have reached contrasting conclusions concerning these estimates, which may be perceived to challenge the applicability and reliability of pollen-based reconstruction. Here we show that conflicting estimates of pollen production and dispersal are, at least in part, artifacts of fixed assumptions of pollen dispersal and insufficient spatial resolution of vegetation data surrounding the pollen-collecting lake. We implemented a Bayesian statistical model that relates pollen assemblages in surface sediments of 33 small lakes (< 2 ha) in the northeastern United States, with surrounding vegetation ranging from 10^1 to $>10^5$ m from the lake margin. Our analysis reveals three key insights. First, pollen productivity is largely conserved within taxa and across forest types. Second, when local (within 1-km radius) vegetation abundances are not considered, pollen-source areas may be overestimated for a number of common taxa (Cupressaceae, *Pinus*, *Quercus*, and *Tsuga*). Third, pollen dispersal mechanisms may differ between local and regional scales, which is missed by pollen-dispersal models used in previous studies. These findings highlight the complex interactions between vegetation heterogeneity on the landscape and pollen dispersal. We suggest that, when estimating pollen productivity and dispersal, both detailed local and extended regional vegetation must be accounted for. Also, both

deductive (mechanistic models) and inductive (statistical models) approaches are needed to better understand the emergent properties of pollen dispersal in heterogeneous landscapes.

Keywords

Bayesian statistical model, pollen dispersal, pollen productivity, pollen-based vegetation reconstruction, pollen-vegetation relationship

Introduction

Global environmental changes can affect vegetation composition at regional to global scales, including community turnover (Willis et al. 2010) and ecosystem transformations (Nolan et al. 2018, Jackson 2021). These vegetation dynamics can take decades, centuries, or millennia to unfold, and have large consequences for global carbon and energy budgets (Bonan 2008, Bonan and Doney 2018). A primary approach to studying long-term vegetation dynamics is through geohistorical records, using "proxies" such as fossil pollen deposited in lakes, wetlands, and hollows. Interpreting changes in pollen records across space and over time provide insights into the history of vegetation, climate, disturbance, and human impacts on ecosystems (e.g., Marsicek et al. 2018, Nolan et al. 2018, Mottl et al. 2021). While analyses of pollen data have been ongoing for over a century (Edwards et al. 2017), the urgent need to better understand vegetation response to anthropogenic climate change has led to a recent surge in developing and applying a variety of quantitative methods for reconstructing past vegetation from fossil pollen data (Williams et al. 2011, Mazier et al. 2012, Theuerkauf and Couwenberg 2018, Zanon et al. 2018, Dawson et al. 2019, Trachsel et al. 2020).

Fundamental to most of these reconstruction approaches is the robust estimation of pollen productivity and dispersal. Pollen productivity is typically defined as the number of pollen grains produced per-unit relative abundance of a plant taxon. Pollen dispersal describes how far airborne pollen grains travel before being deposited. Although most pollen grains are deposited near their source, a substantial amount of pollen deposited in a lake derives from distant sources owing to the leptokurtic (i.e., fat-tailed) shape of pollen dispersal kernels (Prentice 1985, Jackson and Lyford 1999). Both

pollen productivity and dispersal vary among taxa (pollen taxa are typically defined at the genus- or family-level due to the limitations of pollen identification) but are often assumed to be invariant within taxa. Thus, beyond the poorly studied effects of intraspecific variation, pollen taxa may include multiple species that can differ in pollen productivity and dispersal.

Inconsistent estimates of pollen productivity and dispersal from different studies can cast doubt on the reliability and applicability of quantitative pollen-based vegetation reconstruction. Pollen productivity estimates for a given taxon can differ by one or more orders of magnitude across regions (Prentice and Webb III 1986, Broström et al. 2008, Mazier et al. 2012, Li et al. 2018) or time scales (e.g., interannually or decadally - see Kuoppamaa et al. 2009, Minckley et al. 2012), which raises concerns about the applicability of pollen productivity estimates across space and time (e.g., Broström et al. 2008, Li et al. 2018). Therefore, productivity estimates are often validated and applied only within the same region (Hellman et al. 2008b, 2008a, Sugita et al. 2010). For pollen dispersal, the appropriate spatial scale for relating lake pollen data to the surrounding vegetation remains unclear, with a 100-fold difference among studies in assumed or estimated pollen-source areas. Specifically, some studies identify the "relevant source areas of pollen (RSAP)" (Sugita 1994) for pollen deposited in small and medium sized lakes as ranging from a few hundred meters to a few kilometers (e.g., Bradshaw and Webb 1985, Bunting et al. 2004, Hellman et al. 2009a, Han et al. 2017), whereas recent data-driven studies estimate the primary lake-pollen source area to be on the order of hundreds of kilometers for almost all tree taxa (Kujawa et al. 2016, Dawson et al. 2016).

Apparent contradictions in pollen productivity and dispersal may arise due to poorly specified dispersal kernels or inadequate information on landscape heterogeneity (e.g., the spatial distribution of pollen source taxa) (Liu 2015, Li et al. 2018). Notably, two prominent types of pollen-vegetation models, ERV models (Prentice and Webb III 1986, Sugita 1994, Bunting et al. 2013) and STEPPS models (Paciorek and McLachlan 2009, Dawson et al. 2016), are based on different assumptions. In studies using ERV-type models, the dispersal-functions (weights) applied to actual or simulated vegetation data are constructed *a priori*, using either distance-weighting heuristics or pollen-transport models. These dispersal function assumptions (Jackson and Lyford 1999, Theuerkauf et al. 2013) underlie estimation of the relevant source areas of pollen (RSAP), which are typically hundreds of meters in radius, and determine the areal extent of vegetation surveys (Bunting et al. 2004, Hellman et al. 2009a, 2009b, Li et al. 2018), which are typically on the order of a few kilometers. In contrast, studies based on the STEPPS model (Dawson et al. 2016, 2019, Trachsel et al. 2020) simultaneously estimate pollen productivity and dispersal from a network of vegetation and pollen data, and the estimated dispersal kernels suggest that vegetation from hundreds of kilometers away still strongly influences pollen assemblages. The spatial resolution of vegetation data used in the STEPPS model, however, is much coarser than in ERV models. For example, in the STEPPs model, "local" vegetation surrounding a lake is represented by mean vegetation composition within the corresponding coarse 8 km × 8 km grid cell. This could result in misrepresentation of critical local pollen sources (Paciorek and McLachlan 2009, Liu 2015), and observed pollen may be misattributed to more distant vegetation.

In this study, we aim to rigorously quantify pollen productivity and dispersal by combining the strength of ERV models (a hierarchy of finely resolved vegetation data near the lake) and the STEPPS model (geographically extensive vegetation data, empirically estimated dispersal, and uncertainty quantification) while overcoming their respective limitations. We implemented a Bayesian statistical model to relate pollen assemblages at 33 small lakes (< 2 ha) from three study regions in the northeastern United States to their surrounding vegetation. Our model is informed by detailed forest inventory data resolved at various distances within 1 km from the lakeshore (local), and extends from 1 km to 300 km from the lakeshore (regional). Using these data, the approach simultaneously quantifies the influences of pollen productivity and dispersal. We address the following questions: (Q1) how much do pollen productivity estimates vary within taxa across three study regions of different forest types in the northeastern US? (Q2) What are the pollen dispersal characteristics (i.e., how does the probability of pollen-deposition decline with distance from the pollen source), and how effectively do mechanistic models of pollen dispersal capture the observed patterns? Finally, (Q3) how important is local and regional vegetation for accurately estimating pollen productivity and dispersal? Respectively, these questions correspond to the three major factors influencing how pollen deposited in lakes represents surrounding vegetation: pollen productivity, dispersal, and landscape heterogeneity. Thus, we expect this work to help refine our understanding of pollen productivity and dispersal patterns, and to motivate a new generation of quantitative pollen-based reconstructions to understand vegetation dynamics.

Methods

Study sites and data

The study area is located in the northeastern United States, and includes 33 small lakes (area < 2 ha) from three regions of different forests (Fig. 1). Dense forests around the 14 Fish Creek (FC) sites in the central Adirondack Mountains in New York consist of old-growth hardwood and mixed stands (dominant genera are *Fagus*, *Acer*, *Betula*, *Tsuga*, and *Picea*), selectively logged stands (often dominated by *Fagus*), and second-growth stands (dominated by the aforementioned taxa and *Pinus*). Nine sites from the eastern Adirondack Mountains (EAD), also in New York, are mostly surrounded by second-growth, mixed hardwood/coniferous forests (dominant genera are *Acer*, *Betula*, *Fraxinus*, *Pinus*, *Quercus*, and *Tsuga*), often dominated by *Pinus*. The 10 southern New England (SNE) sites are in *Quercus*-dominated second-growth forests. Vegetation in these regions is described in detail by Jackson (1990, 2019). Using lakes from the three different regions, our model will allow uncertainty estimations across different vegetation compositions as well as within pollen-taxa (i.e., the species of *Betula*, *Pinus*, and *Quercus* differ to some extant among regions).

Due to the varying taxonomic resolution of pollen identification, pollen types grouped by taxa (often genus, sometime species or family) rather than species. Our analysis focused on 13 arboreal taxa (*Abies*, *Acer rubrum*, *A. saccharum*, *Betula*, *Fagus*, *Fraxinus*, *Larix*, *Picea*, *Pinus*, *Populus*, *Quercus*, Cupressaceae, and *Tsuga*), which included all taxa exceeding 5% in pollen assemblages or vegetation proportions at any site.

We operationally define the "local scale" as within 1 km from the lakeshore and the extended "regional scale" as spanning 1-300 km from the lakeshore. This distinction is introduced to distinguish major vegetation-survey datasets and accommodate possible major differences in pollen dispersal between the two scales in our model. And unlike the notion of "background pollen" in ERV models, sites within the same region in our model may have different pollen inputs from regional sources owing to vegetation heterogeneity at scales of 10^1-10^2 km. The pollen sampling procedure and the local vegetation survey protocol follow Jackson (1990, 2019). At each of the 33 small lakes, modern pollen samples were collected from the top 1 cm of sediments at the center of the lake between 1986 and 1990. During the same period, trees growing within 1 km around the lakeshore were surveyed with varying intensity, including exhaustive or inventory-plot measurements within 20 m, and transects of plotless angle-count (Bitterlich) samples. Total basal area (m^2) was calculated for each tree taxon growing within 1-km radius, for each of five distance-intervals from the shore of each small lake: 0-20, 20-50, 50-100, 100-500, and 500-1000 m. Site information is listed in Appendix S1: Table S1. Pollen counts, local basal areas of taxa at each lake, and detailed description of the sampling methods and data are reported in Jackson (2019).

To estimate contemporary vegetation abundances outside the 1-km radius, we obtained data from the Forest Inventory and Analysis (FIA) program for 349,309 individual trees in 7,354 plots between 2000 and 2012 (Fig. 1). Because some plots were re-measured during this time, a total of 10,753 plot-level records were used. The total basal area (m^2) for each tree taxon was calculated for each plot-level record. If a plot location was measured multiple times between 2000 and 2012, the mean basal area of

each taxon was calculated. For each lake site, we calculated the mean vegetation abundances of taxa in ring-shaped bands located at different distances to the lake, and a total of 100 "rings" were considered at each lake site. Each ring is 3-km wide. That is, the smallest/nearest ring has an inner radius of 1 km and an outer radius of 4 km from the lakeshore, whereas the largest/furthest ring has an inner radius of 298 km and an outer radius of 301 km from the lakeshore. We chose the ring width to be 3 km because the coordinates for FIA plots are "fuzzed" within 0.5-1 mile (i.e., 0.8-1.6 km; O'Connell et al. 2017) and some are "swapped" (plot coordinates exchanged between similar private forest plots within the same county); therefore, finer rings in this case will not result in better resolution. Although vegetation within the 1 km radius was sampled in the late 1980s, more than a decade before the FIA plots were sampled, the change in vegetation composition over that decade (and/or difference due to sampling methods) appears to be small (Appendix S1: Fig. S1).

Furthermore, we assessed spatial patterns of vegetation heterogeneity around each lake, using vegetation abundance data from both vegetation surveys at the local scale and forest inventories at the regional scale. To do this, we used squared-chord distance (SCD; Overpeck et al. 1985) to quantify the dissimilarities between vegetation composition (i.e., relative abundances) at different distances from the lakeshore and several "focal" vegetation compositions at 0.02, 0.1, 0.5, 1, 9, 27, and 45 km from the lakeshore.

Bayesian model specification

We developed a Bayesian statistical model that relates deposited pollen assemblages with surrounding vegetation at the 33 small lakes. Recall the "pollen source topography"

analogy: Essentially, our model is informed by pollen counts from the lakes as well as the vegetation distribution maps surrounding the lakes, and estimates taxon-specific pollen productivities (estimated for each study region) and parameters describing dispersal patterns (estimated for local and regional scales respectively).

Our model structure is overall similar to the STEPPS approach (e.g., Dawson et al. 2016) in terms of the Bayesian implementation and the delineation of a focal area around the site and those areas beyond. However, important data choices and model implementations differ. Our analysis has two distinct features that accommodate complex landscape heterogeneity: (i) the finely resolved and extended spatial information of vegetation abundances (the vegetation distribution map), and (ii) the separate, explicit treatment of local versus regional dispersal. Specifically, we decomposed pollen contributions into local sources (originating within 1 km from the lakeshore) and regional sources (originating beyond 1 km from the lakeshore). We estimated taxon-specific local contributions, which are the proportions of pollen contributed by local vegetation relative to all pollen of the same taxa from all distances. Relative vegetation abundances (i.e., vegetation composition) are distance-weighted within their respective scales (local versus regional). At the local scale, the taxon-specific influences of vegetation on pollen are estimated for the five concentric rings (0-20, 20-50, 50-100, 100-500, and 500-1000 m intervals from the lakeshore) at which local vegetation was finely surveyed. At the regional scale, the taxon-specific influences of vegetation on pollen are estimated for increasing radial rings (1-4, 4-7, 7-10, …, and 298-301 km from the lakeshore).

Our modeling approach is as follows. At site s = 1, …, S (S = 33; s = 1, …, 14 belong to the FC region, s = 15, …, 23 belong to the EAD region, and s = 24, …, 33

belong to the SNE region), observations of lake sediment-surface pollen samples, vegetation within 1 km, and FIA plots within 300 km were available as described above. To account for observation error, we modeled pollen counts, $Y_{s,t}$, of taxon $t = 1, \ldots, T$ ($T = 13$) in the sample from site s (where $\mathbf{Y}_s$ is the vector of site-specific pollen counts, of length T) as coming from a multinomial distribution with probability parameters determined by the relative pollen load, $\mathbf{P}_s$, for each taxon at that site ($\mathbf{P}_s$ is also a vector of length T and $\sum_{t=1}^{T} P_{s,t} = 1$):

$$\mathbf{Y}_s \sim \text{Multinomial}(\mathbf{P}_s,\ \mathrm{N}_s) \quad (1)$$

where $N_s = \sum_{t=1}^{T} Y_{s,t}$ is the total number of pollen grains counted in the sample from site s, summing across all taxa.

$P_{s,t}$ at site s is determined by both local and regional sources of pollen and by the taxon-specific pollen productivities, $\Phi_{r,t}$ (scaling factor, per relative vegetation abundances, therefore unitless) for region $r = 1, \ldots, R$ ($R = 3$):

$$P_{s,t} = \Phi_{r(s),t}\left(\gamma_t \cdot VL_{s,t} + \left(1 - \gamma_t\right) \cdot VR_{s,t}\right) \quad (2)$$

where $r(s)$ denotes the region in which site s is location; γ_t is the taxon-specific local weights, representing the proportion of pollen contributed by local trees of taxon t relative to all (local and regional) pollen of that taxon. $VL_{s,t}$ is the distance-weighted relative abundance of each taxa t within 1 km (local vegetation composition) of site s; $VR_{s,t}$ is defined similarly to $VL_{s,t}$, but for trees occurring beyond 1 km and up to ~ 300 km from the lakeshore (regional vegetation composition).

We modeled the taxon- and region-specific pollen productivities ($\mathbf{\Phi}$) hierarchically. We expect $\Phi_{r,t}$ to vary among taxa; in addition, because environmental conditions, phenotypes of trees, and growth forms of trees may vary across regions, $\Phi_{r,t}$

may also vary within taxa across regions. Therefore, for each taxon in region $r = 1, \ldots, R$ ($R = 3$, corresponding to FC, EAD, and SNE), we define the region-level vector of taxon productivities, $\mathbf{\Phi}_r$, as varying around the overall taxon-level productivity, $\mathbf{\Phi}^*$ (both are vectors of length T). Because the productivity of a taxon is only meaningful relative to other taxa (pollen percentages do not measure pollen flux, therefore no absolute pollen productivity can be inferred from pollen data), we imposed a sum-to-one constraint for the relative productivity parameters using a Dirichlet distribution:

$$\mathbf{\Phi}_r \sim \text{Dirichlet}(\alpha \cdot \mathbf{\Phi}^*) \quad (3)$$

where the scalar parameter, α, was given a relatively non-informative, yet realistic uniform prior:

$$\alpha \sim \text{Uniform}(20,\ 1000) \quad (4)$$

Sum-to-one constrains are also employed for the overall taxon-level productivities, to which we assigned a relatively non-informative Dirichlet prior:

$$\mathbf{\Phi}^* \sim \text{Dirichlet}(1,\ \ldots,\ 1) \quad (5)$$

For the local pollen contributions or weights, γ_t, that represent contribution of local pollen (values between 0 and 1) relative to regional pollen, we specified relatively non-informative uniform priors:

$$\gamma_t \sim \text{Uniform}(0,1) \quad (6)$$

for taxon $t = 1, \ldots, T$ ($T = 13$).

To calculate $VL_{s,t}$ and $VR_{s,t}$, the relative abundances (i.e. the compositions) of local and regional vegetation around site s are weighted based on their distances to the lake (i.e., forests closer to the lake get a higher weight than those further away from the lake):

$$VL_{s,t} = \sum_{i=1}^{I} VJCK_{s,i,t} \cdot wl_{s,i,t} \quad (7)$$

$$VR_{s,t} = \sum_{j=1}^{J} VFIA_{s,j,t} \cdot wr_{s,j,t} \quad (8)$$

where $VJCK_{s,i,t}$ (data, from Jackson 2019) denotes the taxa compositions (proportional contribution to the total basal area) within the i^{th} consecutive local ring ($i = 1, \ldots, I$; $I = 5$) of site s, weighted by the local weights $wl_{s,i,t}$. Similarly, $VFIA_{s,j,t}$ denotes the taxa contributions within the j^{th} consecutive regional ring ($j = 1, \ldots, J$; $J = 100$) of site s, weighted by the regional weights $wr_{s,j,t}$.

The outer boundary of the i^{th} consecutive local ring is at DL_i ($DL_i = 20$, 50, 100, and 1000 meters for $i = 1, \ldots, I$, respectively) from the lakeshore. The inner-most boundary of the most proximate regional ring is at 1 km from the lake shore, and the outer boundary of the j^{th} consecutive regional-ring is at DR_j ($DR_i = 4\times10^3$, 7×10^3, 10×10^3, …, and 301×10^3 meters for $j = 1, \ldots, J$, respectively) from the lakeshore. Inspired by Equation (8) in Prentice (1985), the local and regional distance weights are similarly defined (b_t parameters are proportional to pollen fall speeds over wind speed and θ is related to the turbulence parameter) as follows:

$$wl_{s,i,t} = \frac{1}{cl_{s,t}} \left(e^{b_t\left({R_s}^{\theta} - (DL_i + R_s)^{\theta}\right)} - e^{b_t\left({R_s}^{\theta} - (DL_{i+1} + R_s)^{\theta}\right)} \right) \quad (9)$$

$$wr_{s,j,t} = \frac{1}{cr_{s,t}} \left(e^{b_t\left({R_s}^{\theta} - (DR_j + R_s)^{\theta}\right)} - e^{b_t\left({R_s}^{\theta} - (DR_{j+1} + R_s)^{\theta}\right)} \right) \quad (10)$$

where R_s is the radius of the lake at site s, and $cl_{s,t}$ and $cr_{s,t}$ are standardizing constants that ensure that $wl_{s,i,t}$ and $wr_{s,j,t}$ sum to 1 across all i and j, respectively:

$$cl_{s,t} = 1 - e^{b_t\left({R_s}^{\theta} - (DL_I + R_s)^{\theta}\right)} \quad (11)$$

$$cr_{s,t} = 1 - e^{b_t\left({R_s}^{\theta} - (DR_J + R_s)^{\theta}\right)} \quad (12)$$

To facilitate interpretation of model parameters, we also calculate the cumulative influence of vegetation at different distances for a typical, 30-meter radius lake:

$$Fl_{i,t} = (1 - e^{b_t \cdot (30^\theta - (DL_{i+1}+30)^\theta)})/(1 - e^{b_t \cdot (30^\theta - (DL_I+30)^\theta)}) \quad (13)$$

$$Fr_{j,t} = (1 - e^{b_t \cdot (30^\theta - (DR_{j+1}+30)^\theta)})/(1 - e^{b_t \cdot (30^\theta - (DR_J+30)^\theta)}) \quad (14)$$

$$Flr_{j,t} = \gamma_t + (1 - \gamma_t) \cdot Fr_{j,t} \quad (15)$$

where *Fl* is the cumulative contribution of pollen from vegetation within different local distances (*Fl* = 100% at 1 km from the lakeshore), *Fr* is the estimated cumulative contributions of vegetation at different regional distances (*Fr* = 100% at 301 km) when the model is only informed by regional vegetation, but not local vegetation. *Flr* is the actual cumulative contribution of pollen from vegetation within different regional distances (*Flr* = 100% at 301 km), which is calculated from *Fl*, *Fr*, and the local contribution γ.

To compare the local contribution (γ) and the cumulative contribution *Fr* at 1 km predicted by regional pollen dispersal (denoted $\tilde{\gamma}$), we calculate the Localness Index (*LI*):

$$LI_{s,t} = \frac{\gamma_t}{\tilde{\gamma}_t} \quad (16)$$

$$\tilde{\gamma}_{s,t} = \frac{1 - e^{b_t \left({R_s}^\theta - (1000+R_s)^\theta\right)}}{1 - e^{b_t \left({R_s}^\theta - (301\times10^3+R_s)^\theta\right)}} \quad (17)$$

where $\tilde{\gamma}_{s,t}$ is similar to $Fr_{l,t}$, but also takes site-specific lake-radius (R_s) into consideration. A *LI* of 1 (one) would indicate that the regional dispersal pattern can predict local contribution and suggest similarities in dispersal between the local and regional scales. A *LI* higher (lower) than 1 would indicate that local populations are over- (under-) represented by pollen deposited in the lake, after accounting for regional dispersal.

In Equations (9)-(12), we set $\theta = 0.1$ to represent an unstable "atmospheric condition" (temperature decreases with height at a faster rate than the adiabatic lapse rate, therefore representing a more turbulent atmosphere) (Jackson and Lyford 1999). We explored other values, and $\theta = 0.1$ remained the most appropriate (Appendix S1: Section S1, Model Variant 1-2). For each taxon-specific b_t, rather than specifying fixed values based on pollen fall speeds (e.g., Prentice 1985, Sugita et al. 1999, Sugita 2007a, Hellman et al. 2009b), we estimate b_t and gave it a positive-valued, relatively non-informative exponential distribution prior,

$$b_t \sim \text{Exp}(0.1) \quad (18)$$

Similar to the original interpretation (Prentice 1985), the taxon-specific b_t in our analysis can be viewed as a composite parameter representing the overall effect of pollen deposition velocities of each taxa and the atmospheric conditions operating to transport the pollen above the canopy.

To determine the robustness of our model specification related to the distance-weighting of local and regional vegetation, Equations (9) and (10), we tested three additional models that explored different specifications for θ or *wl* (Appendix S1: Fig. S2-S5). The model presented above was the most parsimonious (less risk of over-parameterization compared to other models), and thus, we focus on this model throughout.

Model implementation

The model (Data S1; DOI: 10.5281/zenodo.5825842) was implemented in the Bayesian modeling software, JAGS version 4.2.0, using R version 3.4.2 and the "rjags" and R2jags

(version 0.5-7) packages (Plummer 2003). We ran three parallel MCMC chains for 3,000,000 iterations. For each chain, the starting values were generated from the prior distributions. All chains converged by iteration 2,000,000, and we used the last 1,000,000 iterations, keeping every 200th sample to reduce within chain autocorrelation and to reduce the number of samples stored. Hence, we obtained 5000 independent posterior samples per chain, and from these samples, we calculated the posterior median as a point estimate of each parameter, and the 2.5th and 97.5th percentiles to quantify the uncertainty in each parameter (i.e., 95% Bayesian credible interval [CI]).

Results: Quantification of how pollen represents surrounding vegetation

Posterior estimates from our model are summarized and presented below. Wherever suitable, we also discuss some specific, less-expected results in relation to the model and the general understanding of pollen productivity and dispersal. By clarifying these points and interpretations in the Results, we can focus on addressing the primary research questions in the Discussion.

Model fit

The Bayesian statistical model captured the relationship between vegetation abundances and pollen assemblages, as indicated by the observed versus predicted taxon-specific pollen relative abundances (proportions) at each site (Fig. 2). The observed versus predicted correspondence was best ($R^2 > 0.81$) for the taxa that are abundant on the landscape and that produce a relatively large amount of pollen (*Betula*, *Fagus*, *Pinus*, *Quercus*, and *Tsuga*) comparing to other taxa. In addition, *Larix*, which grows in low

abundance and in highly localized populations, and *Acer saccharum*, which is an abundant species but a low pollen-producer, also were associated with good fits (R^2 = 0.81 and 0.73, respectively). The fit for Cupressaceae (R^2 = 0.49) is dominated by the pollen signal of *Thuja*, since Cupressaceae pollen higher than 0.25% is only found in EAD sites with substantial *Thuja* populations concentrated along some lake margins and no *Chamaecyparis* or *Juniperus*. The remaining taxa (*Abies*, *Acer rubrum*, *Fraxinus*, *Picea*, and *Populus*) display intermediate correspondence between observed versus predicted pollen relative abundances ($0.29 < R^2 < 0.45$, Fig. 2). In general, the observed versus predicted pollen relative abundances varied about the 1:1 line without substantial systematic bias (Fig. 2).

Pollen productivity

The taxon-specific relative pollen productivities at the regional ($\mathbf{\Phi_r}$) and population (overall) ($\mathbf{\Phi^*}$) scales are estimated with high confidence (narrow 95% Bayesian credible intervals [CIs]), distinguishing differences in pollen productivity among taxa (Fig. 3). *Abies*, *Acer rubrum*, and *A. saccharum* are low pollen producers, with posterior medians for relative productivity around 0.02, significantly lower than any other taxa. On the other extreme, *Betula* produces significantly more pollen than all other taxa (posterior median of taxon-level estimate is round 0.28, resulting in roughly 14 times higher productivity than the aforementioned low-producers).

For most taxa, the posterior medians of the regional-level productivities ($\mathbf{\Phi_r}$) are similar within the taxon across the three regions of different forests, and are contained in the 95% CIs of the overall taxon-level productivity ($\mathbf{\Phi^*}$). The posterior estimate of the

scaling parameter, α in Equations (3) and (4), for the regional variability in pollen production is not clustered against the specified lower bound of its prior (median and 95% CI: 91 [69, 126], compared to 20, the lower bound for the prior of α), suggesting similar productivity among regions (higher α indicates more similarity across regions). For example, even though *Betula* populations are composed of *B. alleghaniensis* and *B. papyrifera* in FC and EAD, but dominated by the abundant populations of *B. lenta* and *B. populifolia* in SNE, the productivity estimates are still similar across these regions (Fig. 3).

In several cases, however, regional-level pollen productivity ($\mathbf{\Phi}_r$) varies by a factor of two across regions (Fig. 3). Here, intra-taxon (species- or genus-level) differences in productivity may play an important role. Notably, much lower productivity of *Picea* spp. is found in the SNE region (~ half of the productivity of *Picea* in other regions). In this case, *P. abies* grows in scattered plantations in SNE, whereas *Picea* in the FC and EAD regions is represented by higher regional tree abundances of *P. mariana* and *P. rubens*. Moreover, Cupressaceae pollen productivity is 70% higher in the SNE compared to the EAD and FC regions. This variability may indicate genus-level differences: Cupressaceae is represented by *Chamaecyparis* and *Juniperus* in SNE, but by *Thuja* in EAD and FC. Finally, *Quercus* spp. productivity is approximately two times higher in the SNE, where the dominant *Quercus* species (*Q. alba* and *Q. velutina*) and the climate differ from those of EAD and FC (*Q. rubra*). In contrast, pollen taxa comprising a single species in the study area (*Fagus grandifolia* and *Tsuga canadensis*) have more similar productivities across regions. However, pollen productivity of *Pinus* spp. in the densely forested FC region is significantly lower (approximately half) than in the other

two regions, despite the similarity of dominant *Pinus* species (*P. strobus* and *P. resinosa*) in FC and EAD.

Our relative pollen productivities are subject to the "sum-to-one" constraint, whereas ERV type models calculate productivities as relative to the reference taxa. This difference in methodology can potentially affect productivity estimates. However, we also found that the general patterns of high versus low pollen producers and regional variations in pollen productivity remain robust when relative pollen productivity is calculated in relation to a reference taxon (Appendix S1: Fig. S2-S3), regardless of whether the reference is set to the productivity of *Fagus*, which is present at three regions, or *Tsuga*, which is present at the FC and EAD sites and a few SNE sites.

Local pollen contribution and dispersal

Taxon-specific local pollen contribution (γ_t, Equation 6), which we operationally define as the proportion of pollen deposited in a lake that originated within 1 km from the lakeshore, is summarized in Fig. 4. *Abies*, *Betula*, *Picea*, and *Populus* have the lowest local pollen contribution, with medians < 0.1 and upper CI limits < 0.25. *Acer rubrum*, the pollen of which has both anemophily and entomophily characteristics (Batra 1985), shows the highest local contribution (median = 0.81). The local contributions for *Fraxinus* and *Larix* have comparatively large uncertainties, indicating that given the available data, our model cannot tightly resolve the local contributions for these taxa. However, there are many difficulties in quantifying the influence of taxa that are underrepresented in pollen (Parsons et al. 1983), so the true local influence may be very high for *Larix*, because all sites that have *Larix* pollen present have *Larix* trees near the

lake. Similarly, the true local contribution of *Abies* may also be high. For the remaining taxa (*A. saccharum*, *Fagus*, *Pinus*, *Quercus*, Cupressaceae, and *Tsuga*), the local contribution is estimated with good confidence (width of CI less than the median posterior), and overall, roughly less than half of the pollen originates locally (i.e., within 1 km of the lakeshore).

The local pollen contributions vary with distance from the lakeshore (outer boundaries of concentric rings at 20, 50, 100, 500, and 1000 m from the lakeshore) (Fig. 5). For example, local pollen dispersal can be visualized as the estimated cumulative influences on pollen (*Fl*, Equation 13) from vegetation at different distances (Fig. 5). We compared these estimated cumulative influences with those based on pre-defined functions typically used in distance-weighting (Calcote 1995, Jackson and Kearsley 1998, Gaillard et al. 2008). For the majority of taxa, the estimated cumulative influences (*Fl*) are the most similar to the $1/\text{distance}^2$ weighting (steeper than the unweighted and 1/distance weighting, and generally less steep than weights based on Prentice-Sutton equation). The uncertainty estimate (95% CI widths) for *Fl* is large for *Acer rubrum* and *Larix*, suggesting that it is possible that pollen grains of these two taxa mainly come from populations very close to the lakeshore (i.e., the upper bound of the confidence interval reaches 80% of cumulative influence within 50 m).

We acknowledge that the "true" shape of cumulative influences of local pollen may be more complex than *Fl*. In this regard, despite potential pitfalls associated with the increased number of parameters, empirically estimating these weights (e.g., independent weights to be estimated at each distance for each taxon) allows a more detailed perspective of cumulative influences and how they may differ from those of distance-

weighting functions (Appendix S1: Fig. S4 and Section S1 - Model Variant 3). We found the ultra-local (with 100 m of the lakeshore) populations of Cupressaceae, *Fagus*, and *Pinus* appear to exert significantly greater influences than all distance-weighting methods (CIs higher and do not overlap with those based on distance-weighting). In contrast, *Populus* has significantly smaller influences than all those of distance-weighting approaches at all distances. More generally, the influence of ultra-local populations may be noticeably strong for *Acer rubrum*, Cupressaceae, *Fagus*, and *Pinus*, as more than 50% of the local contribution of these taxa is attributed to populations occurring within 50 m of the lakeshore, which constitutes only ~ 1.6% of the total area within in the 1 km local radius (assuming a 30 m radius of the lake).

Regional pollen dispersal

The importance of regional pollen dispersal is visualized via the cumulative influence (*Flr,* Equation 15) of vegetation at a range of different distances (Fig. 6, cyan lines) and the distance at which the cumulative influence reaches 75% (dashed line; influence at 300 km is 100% by model design). *Betula* appears to have the furthest regional dispersal, reaching the 75% influence at ~ 50 km (the lower CI limit corresponds to ~ 70 km). *Abies*, *Acer saccharum*, and *Fraxinus* also show large pollen input from regional sources, reaching the 75% cumulative influence at 10-20 km. The 75% cumulative influence for Cupressaceae, *Fagus*, *Populus*, *Quercus*, and *Tsuga* is estimated to be only a few kilometers. For *Acer rubrum* and *Larix*, most pollen deposited in the lake comes from within a couple of kilometers of the lakeshore, indicated by the steeply rising cumulative-influence curves, which confirms the common notion that pollen from *Acer rubrum* and

Larix are from nearby sources (Bradshaw and Webb 1985, Jackson 1990). However, the cumulative influence of *Pinus* also increases steeply with distance—a surprising finding because *Pinus* pollen is usually considered to be largely from regional sources (Bradshaw and Webb 1985). Additional exploratory analysis (Appendix S1: Fig. S4-S5 and Section S1 - Model Variant 3) also suggests a strong influence from nearby vegetation: *Pinus* populations within 20 m from the lakeshore may exert strong influence on the pollen abundances retrieved from the lake (Appendix S1: Fig. S4).

A comparison of the regional dispersal curves based on *Flr* (Equation 15, cyan lines in Fig. 6) with those based on *Fr* (Equation 14, red lines in Fig. 6) suggests the potential bias when regional dispersal is estimated solely with regional vegetation data (*VFIA*) and without considering local vegetation (*VJCK*). We found that, when local vegetation data are not used to inform dispersal, regional pollen dispersal of *Acer rubrum*, Cupressaceae, *Larix*, *Pinus*, *Quercus*, and *Tsuga* are notably over-estimated relative to the scenario when such data are considered (Fig. 6). For example, the estimated 75% cumulative influence of *Quercus* is reached at 3 km (local vegetation considered, *Flr*) versus 70 km (local vegetation not considered, *Fr*), respectively. In contrast, and expectedly based on posterior estimates of γ (Fig. 4), the regional dispersal estimates for *Betula* (the taxon with the furthest regional influence), and to a lesser extent, for *Abies* and *Fraxinus*, are not affected by the incorporation of local vegetation data.

Finally, our model estimates that the 75% cumulative influence of *Picea* is reached around 50 km when local vegetation is considered, whereas the same cumulative influence is reached within a few kilometers when local vegetation is not considered (Fig.

6). That is, *Picea* pollen percentage is well explained by vegetation within a few kilometers, but not by local vegetation within 1 km, which is counterintuitive. Exploratory analysis (Appendix S1: Section S1 - Model Variant 3) provides a possible explanation for this seemingly confusing result: *Picea* pollen may be surprisingly insensitive to its population abundance within 50 m from the lakeshore (Appendix S1: Fig. S4), but sensitive to other local populations further away (100-1000 m) and to those within a few kilometers (Appendix S1: Fig. S4-S5). However, this level of flexibility of distance-weighting is not allowed by the pollen dispersal function in our original model, Equations (9) and (10); as a result, our original model may have misattributed observed *Picea* pollen to vegetation further away in order to minimize the influence of vegetation within 50 m from the lakeshore.

Comparing the local and regional dispersal

Localness Index (*LI*) is shown for all taxa besides *Acer rubrum* and *Larix*, which likely lack meaningful regional input (Fig. 7). *LI* is defined as the ratio between the actual (i.e., empirically estimated) local contribution (γ, Fig. 4) and the cumulative contribution at 1 km predicted by *Fr* (Equation 14, red lines in Fig. 6). We found that many taxa have *LI* values significantly higher or lower than 1 (local populations of trees over- or under-represented, respectively), indicating that the dispersal patterns of these taxa differ between the local and regional scales. Notably, *Quercus* and *Tsuga*, which have local relative-abundances typically higher than the regional ones (Appendix S1: Fig. S6, also see descriptions in Jackson 1990, 2019), have *LI* significantly greater than 1. In contrast,

Betula and *Populus*, which have lower abundances by the lakeshores (0-20 m) than further away, have *LI* significantly below 1.

The heterogeneous landscape

The spatial pattern of vegetation heterogeneity was quantified using SCD to reveal how much and how fast vegetation composition changes with distance for several "focal vegetation compositions" (Fig. 8, Appendix S1: Fig. S7). Two zones of high heterogeneity are revealed: The first zone is located at 0-20 m from the lakeshore, where the vegetation composition within this area is substantially different from adjacent vegetation at 20-50 m from lakeshore (especially at FC and EAD) and vegetation further away from the lakeshore (Fig. 8a, SCDs increased sharply starting 20-50m from lakeshore). The second zone is at 1-27 km from the lakeshore. Vegetation compositions within areas located 0.02, 0.1, 0.2, and 0.5 km from lakeshore, respectively, are all dissimilar from adjacent areas (Fig. 8a-d), although on average, vegetation composition within 27 km is similar to that occurring 27-100 km from the lakeshore (Fig. 8f). By uniquely combining the finely resolved vegetation abundance data at the local scale (< 1 km from the lakeshore) and the spatially coarse but extensive FIA information (~ 3 km resolution, extending up to ~ 300 km from the lakeshore), both zones of high heterogeneity are well represented in our study.

Discussion

Our results allow examination of currently contested notions of pollen productivity and dispersal, supporting some previous findings while contradicting others. First, regarding

the intra-taxon variability of pollen productivity (Q1), we found that pollen productivity is largely conserved within taxa across three regions with different forest composition and pattern (Fig. 3). The greatest intra-taxon difference we observed is less than threefold. Second, regarding the overall pattern of pollen dispersal (Q2), the empirically estimated 75% cumulative influence of most taxa is reached within a few kilometers, yet the exceptions are distinct (e.g., *Betula* reaches its 75% cumulative influence around 50 km, Fig. 6). More specifically, although local vegetation generally exerts a strong influence on pollen deposition in lakes (Fig. 4), the dispersal patterns at the local scale may not be fully captured by the commonly used distance-weighting functions or mechanistic model of pollen dispersal (Fig. 5, Appendix S1: Fig. S4). Third, our results also examined the pattern of landscape heterogeneity and demonstrated that both detailed local vegetation and extended regional vegetation are needed to accurately estimate pollen productivity and dispersal (Q3). Based on these findings, in the following sections we suggest best practices for estimating pollen productivity and dispersal, highlight the importance of landscape heterogeneity, and identify key challenges.

Estimating pollen productivity and dispersal: Cautions and suggestions

Our results indicate that pollen productivity, pollen dispersal processes, and the spatial arrangement of vegetation abundance ("landscape heterogeneity") interact to influence how pollen assemblages in lake sediments represent surrounding vegetation. Therefore, inadequate vegetation information can hamper accurate estimation of both pollen productivity and dispersal. Furthermore, erroneous estimation of or inappropriate

assumptions of dispersal processes can lead to inaccurate estimation of pollen productivity.

Knowledge of local vegetation composition is particularly important in estimating pollen productivity and dispersal. Our results indicate that without finely resolved local vegetation abundances, pollen contributions of many taxa (Cupressaceae, *Pinus*, *Quercus*, *Tsuga*) from regional sources would be overestimated (Fig. 6), and consequently, lead to erroneous productivity estimates. When vegetation abundance is only available at coarse resolution (i.e., 3 km intervals in this case), variation in pollen assemblages cannot be effectively attributed to spatial variation in vegetation composition at finer scales. Instead, variation in pollen assemblages is misattributed to differences in vegetation composition further away, biasing pollen productivity and dispersal estimates. This phenomenon may explain the large pollen-source areas, on the order of 10^2 km, estimated by STEPPS model studies using coarser vegetation data (Paciorek and McLachlan 2009, Kujawa et al. 2016, Dawson et al. 2016).

Inaccurate assumptions about pollen dispersal can lead to biased pollen productivity estimates. Our analysis identified some complex and contextual features of pollen dispersal (e.g., the effects of ultra-local and local populations on pollen, Fig. 5, Appendix S1: Fig. S4, Fig. 7), which are not fully represented by the mechanistic models of pollen dispersal currently applied to pollen–vegetation calibration. Various studies have shown that when the effects of dispersal are not properly accounted for, estimated pollen productivity is often entangled with dispersal and landscape heterogeneity (Bradshaw and Webb 1985, Prentice 1985, Jackson 1990, Jackson and Kearsley 1998). The estimates, rather than reflecting the productivity of pollen, only serve as highly

contingent correction-factors (Theuerkauf et al. 2016), and therefore lack predictive power across regions.

In particular, inaccurate dispersal assumptions, when combined with finely resolved local vegetation abundances, may misattribute pollen variation to vegetation variation at the wrong scale (i.e., overemphasizing influence of ultra-local vegetation with 100 m of the lakeshore and overlook the site-specific influence of regional vegetation) and cause errors in pollen productivity estimates. For example, we found that although the relative abundances of *Betula* vary the most across sites and regions at the 0-50 m scale (Appendix S1: Fig. S6), empirically the influence from this scale may be significantly smaller than predicted by widely used distance weightings (Appendix S1: Fig. S4). Meanwhile, *Betula* relative abundance at 100-500 m (often similar to abundance at 500-1000 m within each site, Appendix S1: Fig. S6) explains most of the "local" influence— which amounts to less than 10% of cumulative influence (Fig. 4)—and the majority of *Betula* pollen comes from regional sources, reaching 75% cumulative influence at ~ 50 km (relative to 100% influence at 300 km by model design, Fig. 6). We also found highly similar relative productivities of *Betula* across the three study regions (medians at 0.30, 0.26, 0.23, respectively; for comparison, productivity of *Tsuga* is around 0.08), and strong correspondence ($R^2 = 0.85$) between predicted and observed *Betula* pollen relative abundances. Although many region-specific factors may be at play, our result contrasts with the large variation in *Betula*-pollen productivity in Europe found using ERV models—those from lake sites in Germany (Matthias et al. 2012) are approximately 4 times higher than those obtained from the Swiss Plateau (Soepboer et al.

2007) and Estonia (Poska et al. 2011). The difference between assumed and actual dispersal pattern may explain the anomaly.

To overcome these issues, we offer two general suggestions for estimating pollen productivity and dispersal from empirical data. First, the spatial resolution and areal extent of the vegetation survey must be adequate. For studies using small lakes, the survey resolution at the local scale should be adequate to distinguish populations in the immediate vicinity of the lake (< 50 m) and those beyond (50 m to 1 km); This is not only because of the strong influence of ultra-local vegetation (Sugita 1993), but also because vegetation composition and abundance at this scale is often notably variable within regions (Fig. 8). This emphasis in resolution may be less important for larger lakes due to the non-linear relationship between circumference and area. The vegetation survey area, regardless of lake size, should extend to span the other high-heterogeneity zones (Fig. 8, see result section "The heterogeneous landscape"). For sites in this study, this heterogenous zone spans to ~ 30 km. Future studies should consider the extent of this zone in developing models. Second, the effects of pollen productivity, dispersal, and landscape heterogeneity must be considered simultaneously. In particular, because current models may inadequately represent some important features of pollen dispersal (Jackson and Lyford 1999), it may be desirable to estimate local and regional pollen dispersal using statistical models fitted to empirical data (Klein et al. 2003) rather than purely mechanistic pollen-transport models.

Complexity in the heterogenous landscape

Our findings point to the important role of vegetation heterogeneity surrounding a lake in determining pollen-assemblage composition (Fig. 8, Appendix S1: Fig. S7), which provides a critical addition to previous studies. Here, landscape heterogeneity should not be confused with simple patchiness, in which different patches of vegetation are randomly distributed with constant probability or form a monotonic gradient with distance to the lake (Sugita 1994, Hellman et al. 2009b, 2009a). Instead, we use landscape heterogeneity to refer to the observation that local and even regional populations of trees are, in general, distributed non-randomly with respect to proximity to lakes (Appendix S1: Fig. S6), owing to at least two factors (Jackson 1990, 1994, 2012). First, depositional basins are often associated with microclimates and microenvironments unrepresentative of the broader landscape (e.g., cold-air drainage, wet margins, fire breaks). Second, lakes themselves are not randomly distributed on the landscape, but are often restricted to specific, and sometimes idiosyncratic, topographic, edaphic, hydrological, and lithological settings that may or may not be representative of the broader region. It is worth noting that these biasing factors do not necessarily affect studies using moss polsters, in which sites can be selected randomly (Bunting et al. 2013, Li et al. 2017).

The non-random distribution of vegetation interacts with pollen transport processes to create complex patterns of pollen dispersal. First, we found that depending on the taxon, ultra-local populations (e.g., within 50 m of the lakeshore, Appendix S1: Section S1 - Model Variant 3) may exert influences significantly smaller or greater than those predicted by heuristic distance-weighting or by Prentice's function (Appendix S1: Fig. S4). In particular, considerable pollen contributions from populations of *Acer*

rubrum, Fagus, and *Pinus* were estimated within 20 m, 50 m, and 20 m of the lakeshore, respectively. This phenomenon could be due to "gravity deposition" (i.e., anthers, microsporangia, and pollen falling from branches overhanging the lake surface (Tauber 1965, Jacobson and Bradshaw 1981)), deposition of pollen clumps rather than individual grains (e.g., a tetrad of *Pinus* grains will fall 4 times faster than a single grain), or other physical processes of pollen transport that are not fully considered in current models (e.g., Theuerkauf et al. 2016, also see discussions in Jackson and Lyford 1999, Theuerkauf et al. 2013). In contrast to the strong influence of ultra-local populations for some taxa, we found that *Picea* pollen is virtually unaltered by populations within 50 m of the lake, which is counterintuitive given its apparent high production, large grain size, and evidence from other regions (Jackson and Smith 1994). This may be due to the local growth form of *Picea* on the lake margin at our sites, which often comprised stunted individuals on waterlogged histosols. Second, many taxa have a Localness Index (*LI*) significantly different from 1 (Fig. 7), indicating that local populations of many taxa are over- or under- represented relative to regional populations, after accounting for regional pollen dispersal. In general, taxa (e.g., *Quercus* and *Tsuga*) that are more abundant near the lakes than across the region (Appendix S1: Fig. S6, also see descriptions in Jackson 1990, 2019) tend to have $LI > 1$ (local population over-represented in pollen), which likely allowed additional pollen grains or pollen clumps to be transported to the lake. In contrast, *Betula* and *Populus* have $LI < 1$ (local population under-represented in pollen), which may be driven by the low abundances (and lack of pollen input from ultra-local sources) of these taxa by the lakeshores (0-20 m) than further away. Together, these complexities challenge our current understanding and representation of pollen dispersal

processes: Although pollen dispersal has been represented by a single function (known or to be estimated) in most previous studies, the actual pattern may be subject to multiple distinct processes and require more complex representations.

The complex patterns of vegetation distribution and dispersal also present implications and challenges for pollen-based vegetation reconstruction. Our productivity estimates may apply reliably to a range of lake sizes and the spatial extent of reconstruction, because unlike many other attempts, they are not entangled with factors such as dispersal and landscape heterogeneity. These productivity estimates can also serve as priors for multi-scale reconstruction such as REVEALS/LOVE (Sugita 2007a, 2007b) or large-scale efforts such as those using the STEPPS models (Dawson et al. 2016). In contrast, the empirically estimated dispersal is specific to lake size because it accounts for multiple pathways of pollen transport, ranging over scales from 1 m to 100 km, and the relative influence of these pathways may change with the size of the depositional basin. For example, gravity deposition is governed by trees along the lakeshore. As the radius of a lake increases, lake area increases more steeply than the length of its shoreline, and hence the influence of gravity deposition per unit lake area decreases rapidly. To gain predictive understanding of how pollen dispersal scales with lake sizes for reconstruction, empirically estimated dispersal across different lake sizes and regions can be compared to develop qualitative insights that can be further tested by process-based models.

Addressing long-standing challenges

Our study revisits several long-standing, interlinked issues in the quantitative application of pollen percentages in vegetational inference: the intra-taxon variability of pollen productivity, the physical processes governing pollen transport, and the role of landscape heterogeneity for understanding links between vegetation and pollen (Webb and McAndrews 1976, Jacobson and Bradshaw 1981, Bradshaw and Webb 1985, Prentice 1985, Jackson 1994, Davis 2000). Numerous solutions have been proposed to address these issues, and understanding of pollen–vegetation relationships have advanced considerably over the past several decades. However, as our results have shown, the solutions developed to date are incomplete, and are subject to ancillary assumptions (Oreskes et al. 1994). As efforts at reconstruction of past vegetation continue, it is important to be clear about limitations of these partial solutions. Otherwise, there is an increasing danger that untested assumptions embedded in these partial solutions over time are likely to be taken for granted, i.e., "ignorance creep" (Jackson 2012).

Much remains unknown concerning the variability of pollen productivity and the nature of pollen dispersal. With regard to productivity, our analysis suggests that it may be much more conserved within pollen taxa than previously suggested, including in recent syntheses (Mazier et al. 2012, Li et al. 2018), and intra-taxon variation (less than three-fold in our analysis) may be explained by differences in species and environmental conditions (Fig. 3). To determine the influence of phylogeny versus environmental factors, future studies could integrate species characteristics and environmental information into the estimation of pollen productivity. For example, Equation (2) in our framework could be modified to estimate species-level productivity (as nested within taxon-level productivity) and the effect of environmental covariates such as temperature

and precipitation. Comparative studies of pollen productivity similar to our analysis here need to be carried out across regions and vegetation types.

With regard to pollen dispersal, our results suggest that the widely used mechanistic pollen-transport models with *a priori* parameter values (e.g., the Prentice-Sutton model, Prentice 1985) and distance-weighting function (Calcote 1995, Jackson and Kearsley 1998, Gaillard et al. 2008) may capture the general pattern of local dispersal (Fig. 5), but with potential caveats suggested by Model Variant 3 (Appendix S1: Fig. S4 and Section S1). Also, pollen dispersal patterns differ between local and regional scales (Fig. 7, a localness index value of 1 indicates the same dispersal pattern between local and regional scales). Advancing the understanding of pollen dispersal may require dual application of deductive (i.e., mechanistic pollen-transport models) and inductive (i.e., semi-mechanistic and phenomenological models in which dispersal patterns are statistically estimated using empirical data) approaches. Most dispersal studies to date are based on the Prentice-Sutton model (Prentice 1985b), which assumes that pollen grains are transported in the atmosphere similar to ground-level Gaussian plumes. However, it has been suggested that a Lagrangian stochastic model may better capture the outcome of long-distance pollen dispersal (Kuparinen et al. 2007, Theuerkauf et al. 2013, 2016, though not everywhere Wan et al. 2020); more tests on different regions and models are needed. Because multiple physical processes govern pollen transport, a single mechanistic model or dispersal kernel may ultimately be elusive (Jackson and Lyford 1999). It is therefore important to develop semi-mechanistic (e.g., this study and Klein et al. 2003) and phenomenological models (Kujawa et al. 2016, Dawson et al. 2016) of pollen dispersal. In addition, statistically fitted semi-mechanistic and phenomenological

models such as the STEPPS (Dawson et al. 2016) can provide formal and coherent quantification of uncertainties.

Understanding and quantifying how pollen represents vegetation is at the heart of pollen analysis. Much has been learned since the initial proposals for quantitatively linking pollen and vegetation percentages (Davis 1963), but some key questions may be better answered with alternative and emerging techniques. Manipulative experiments (e.g., isotope labeling of pollen grains (Colwell 1951); experimental pollen release and monitoring (Raynor et al. 1974, 1975)), which have become uncommon in today's studies, may effectively disentangle the relative importance of multiple pollen-transport processes operating at varying spatial scales. Although radioisotopic labeling is no longer feasible, genetic markers can be useful in tracing dispersal of pollen (Dawson et al. 1997). Remote sensing techniques (e.g., using drones) could be leveraged to facilitate or even replace laborious vegetation survey (e.g., Williams et al. 2009, 2011), whereas machine learning approaches may be powerful for extracting vegetation information from remotely sensed images (Zanon et al. 2018). These approaches, together with iterative improvements to statistical and simulation models of pollen productivity and dispersal, will build toward a rigorous inferential basis, maximizing the unique potential of fossil pollen records in understanding and addressing global ecological and environmental challenges.

Acknowledgments. This manuscript has been authored in part by UT-Battelle, LLC, under contract DE-AC05-00OR22725 with the US Department of Energy (DOE). The US government retains and the publisher, by accepting the article for publication, acknowledges that the US government retains a nonexclusive, paid-up, irrevocable, worldwide license to publish or reproduce the published form of this manuscript, or allow others to do so, for US government purposes. DOE will provide public access to these results of federally sponsored research in accordance with the DOE Public Access Plan (http://energy.gov/downloads/doepublic-access-plan). Any use of trade, firm, or product names is for descriptive purposes only and does not imply endorsement by the U.S. Government. This work is funded by National Science Foundation grant EAR-1003848. Y.L. is partly supported by UT-Battelle, LLC, under contract DE-AC05-00OR22725 with the US Department of Energy. We thank Andria Dawson, Chris Paciorek, Dunbar Carpenter, Jack Williams, Marie-José Gaillard, John Calder, Simon Goring, and Colleen Iversen for discussion and comment. Major computational tasks were run on Northern Arizona University's Monsoon computing cluster, funded by Arizona's Technology and Research Initiative Fund.

Author contributions. Y.L. and S.T.J. designed the study; S.T.J. and J.W.L. provided data and advice; Y.L. performed the data analysis with advice from K.O. and S.T.J.; Y.L. wrote the paper with contributions from S.T.J., K.O., and J.W.L.

References

Batra, S W T. 1985. 'Red Maple (*Acer Rubrum L.*), an Important Early Spring Food Source for Honey Bees and Other Insects'. *Journal of the Kansas Entomological Society* 58 (1): 169–72.

Bonan, Gordon B. 2008. 'Forests and Climate Change: Forcings, Feedbacks, and the Climate Benefits of Forests'. *Science* 320 (5882): 1444–49. https://doi.org/10.1126/science.1155121.

Bonan, Gordon B., and Scott C. Doney. 2018. 'Climate, Ecosystems, and Planetary Futures: The Challenge to Predict Life in Earth System Models'. *Science* 359 (6375). https://doi.org/10.1126/science.aam8328.

Bradshaw, R. H.W., and T. Webb. 1985. 'Relationships between Contemporary Pollen and Vegetation Data from Wisconsin and Michigan, USA.' *Ecology* 66 (3): 721–37. https://doi.org/10.2307/1940533.

Broström, Anna, Anne Birgitte Nielsen, Marie José Gaillard, Kari Hjelle, Florence Mazier, Heather Binney, Jane Bunting, et al. 2008. 'Pollen Productivity Estimates of Key European Plant Taxa for Quantitative Reconstruction of Past Vegetation: A Review'. In *Vegetation History and Archaeobotany*, 17:461–78. https://doi.org/10.1007/s00334-008-0148-8.

Bunting, M J, A Brostrom, S Sugita, and R Middleton. 2004. 'Vegetation Structure and Pollen Source Area'. *Holocene* 5: 651–60.

Bunting, M. Jane, J. Edward Schofield, and Kevin J. Edwards. 2013. 'Estimates of Relative Pollen Productivity (RPP) for Selected Taxa from Southern Greenland:

A Pragmatic Solution'. *Review of Palaeobotany and Palynology* 190 (March): 66–74. https://doi.org/10.1016/j.revpalbo.2012.11.003.

Calcote, Randy. 1995. 'Pollen Source Area and Pollen Productivity: Evidence from Forest Hollows'. *Journal of Ecology* 83 (4): 591–602. https://doi.org/10.2307/2261627.

Colwell, Robert N. 1951. 'The Use of Radioactive Isotopes in Determining Spore Distribution Patterns'. *American Journal of Botany* 38 (7): 511–23.

Davis, Margaret B. 1963. 'On the Theory of Pollen Analysis'. *American Journal of Science* 261 (10): 897–912. https://doi.org/10.1126/science.5.127.888.

Dawson, Andria, Christopher J. Paciorek, Simon J. Goring, Stephen T. Jackson, Jason S. McLachlan, and John W. Williams. 2019. 'Quantifying Trends and Uncertainty in Prehistoric Forest Composition in the Upper Midwestern United States'. *Ecology* 100 (12). https://doi.org/10.1002/ecy.2856.

Dawson, Andria, Christopher J. Paciorek, Jason S. McLachlan, Simon Goring, John W. Williams, and Stephen T. Jackson. 2016. 'Quantifying Pollen-Vegetation Relationships to Reconstruct Ancient Forests Using 19th-Century Forest Composition and Pollen Data'. *Quaternary Science Reviews* 137 (April): 156–75. https://doi.org/10.1016/j.quascirev.2016.01.012.

Dawson, I. K., R. Waugh, A. J. Simons, and W. Powell. 1997. 'Simple Sequence Repeats Provide a Direct Estimate of Pollen-Mediated Gene Dispersal in the Tropical Tree Gliricidia Sepium'. *Molecular Ecology* 6 (2): 179–83. https://doi.org/10.1046/j.1365-294X.1997.00163.x.

Edwards, Kevin J., Ralph M. Fyfe, and Stephen T. Jackson. 2017. 'The First 100 Years of Pollen Analysis'. *Nature Plants 2017 3:2*.

Gaillard, M J, S Sugita, M J Bunting, R Middleton, A Brostrom, C Caseldine, T Giesecke, et al. 2008. 'The Use of Modelling and Simulation Approach in Reconstructing Past Landscapes from Fossil Pollen Data: A Review and Results from the POLLANDCAL Network'. *Vegetation History and Archaeobotany* 17 (5): 419–43. https://doi.org/10.1007/s00334-008-0169-3.

Han, Yue, Hongyan Liu, Qian Hao, Xu Liu, Weichao Guo, and Huailiang Shangguan. 2017. 'More Reliable Pollen Productivity Estimates and Relative Source Area of Pollen in a Forest-Steppe Ecotone with Improved Vegetation Survey'. *Holocene* 27 (10): 1567–77. https://doi.org/10.1177/0959683617702234.

Hellman, S., M. J. Bunting, and M. -J. Gaillard. 2009. 'Relevant Source Area of Pollen in Patchy Cultural Landscapes and Signals of Anthropogenic Landscape Disturbance in the Pollen Record: A Simulation Approach'. *Review of Palaeobotany and Palynology* 153 (3): 245–58. https://doi.org/10.1016/j.revpalbo.2008.08.006.

Hellman, Sofie E. V., Marie-josé Gaillard, Anna Broström, and Shinya Sugita. 2008. 'Effects of the Sampling Design and Selection of Parameter Values on Pollen-Based Quantitative Reconstructions of Regional Vegetation: A Case Study in Southern Sweden Using the REVEALS Model'. *Vegetation History and Archaeobotany* 17 (5): 445–59. https://doi.org/10.1007/s00334-008-0149-7.

Hellman, Sofie, Marie-José Gaillard, Anna Broström, and Shinya Sugita. 2008. 'The REVEALS Model, a New Tool to Estimate Past Regional Plant Abundance from

Pollen Data in Large Lakes: Validation in Southern Sweden'. *Journal of Quaternary Science* 23 (1): 21–42. https://doi.org/10.1002/jqs.1126.

Hellman, Sofie, Marie-José Gaillard, Jane Mairi Bunting, and Florence Mazier. 2009. 'Estimating the Relevant Source Area of Pollen in the Past Cultural Landscapes of Southern Sweden — A Forward Modelling Approach'. *Review of Palaeobotany and Palynology* 153 (3–4): 259–71. https://doi.org/10.1016/j.revpalbo.2008.08.008.

Jackson, Stephen T. 1990. 'Pollen Source Area and Representation in Small Lakes of the Northeastern United States'. *Review of Palaeobotany and Palynology* 63 (1–2): 53–76. https://doi.org/10.1016/0034-6667(90)90006-5.

———. 1994. 'Pollen and Spores in Quaternary Lake Sediments as Sensors of Vegetation Composition: Theoretical Models and Empirical Evidence'. In *Sedimentation of Organic Particles*, 253–86. Cambridge University Press, Cambridge.

———. 2012. 'Representation of Flora and Vegetation in Quaternary Fossil Assemblages: Known and Unknown Knowns and Unknowns'. *Quaternary Science Reviews* 49: 1–15. https://doi.org/10.1016/j.quascirev.2012.05.020.

———. 2019. 'Modern Pollen-Assemblage Data from Small Lakes Paired with Local Forest-Composition Data in Northeastern United States'. *Ecology* 100 (10): 2784. https://doi.org/10.1002/ecy.2784.

Jackson, Stephen T., and Jennifer B. Kearsley. 1998. 'Quantitative Representation of Local Forest Composition in Forest-Floor Pollen Assemblages'. *Journal of Ecology* 86 (3): 474–90. https://doi.org/10.1046/j.1365-2745.1998.00277.x.

Jackson, Stephen T., and Mark E. Lyford. 1999. 'Pollen Dispersal Models in Quaternary Plant Ecology: Assumptions, Parameters, and Prescriptions'. *Botanical Review* 65 (1): 39–75. https://doi.org/10.1007/BF02856557.

Jackson, STEPHEN T., and SUSAN J. Smith. 1994. 'Pollen Dispersal and Representation on an Isolated, Forested Plateau'. *New Phytologist* 128 (1): 181–93. https://doi.org/10.1111/j.1469-8137.1994.tb04001.x.

Klein, Etienne K, Claire Lavigne, Xavier Foueillassar, Pierre-henri Gouyon, Source Ecological Monographs, No Feb, and Catherine Laredo. 2003. 'Corn Pollen Dispersal : Quasi-Mechanistic Models and Field Experiments'. *Ecological Monographs* 73 (1): 131–50.

Kujawa, Ellen Ruth, Simon Goring, Andria Dawson, Randy Calcote, Eric C. Grimm, Sara C. Hotchkiss, Stephen T. Jackson, et al. 2016. 'The Effects of Anthropogenic Land Cover Change on Pollen-Vegetation Relationships in the American Midwest'. *Anthropocene* 15: 60–71. https://doi.org/10.1016/j.ancene.2016.09.005.

Kuoppamaa, Mari, Antti Huusko, and Sheila Hicks. 2009. 'Pinus and Betula Pollen Accumulation Rates from the Northern Boreal Forest as a Record of Interannual Variation in July Temperature'. *Journal of Quaternary Science* 24 (5): 513–21. https://doi.org/10.1002/jqs.1276.

Kuparinen, Anna, Tiina Markkanen, Hermanni Riikonen, and Timo Vesala. 2007. 'Modeling Air-Mediated Dispersal of Spores, Pollen and Seeds in Forested Areas'. *Ecological Modelling* 208 (2–4): 177–88. https://doi.org/10.1016/j.ecolmodel.2007.05.023.

Li, Furong, Marie-José Gaillard, Shinya Sugita, Florence Mazier, Qinghai Xu, Zhongze Zhou, Yuyun Zhang, Yuecong Li, and Dominique Laffly. 2017. 'Relative Pollen Productivity Estimates for Major Plant Taxa of Cultural Landscapes in Central Eastern China'. *Vegetation History and Archaeobotany* 26 (6): 587–605. https://doi.org/10.1007/s00334-017-0636-9.

Li, Furong, Marie-José Gaillard, Qinghai Xu, Mairi J. Bunting, Yuecong Li, Jie Li, Huishuang Mu, et al. 2018. 'A Review of Relative Pollen Productivity Estimates From Temperate China for Pollen-Based Quantitative Reconstruction of Past Plant Cover'. *Frontiers in Plant Science* 9 (September): 1214. https://doi.org/10.3389/fpls.2018.01214.

Liu, Yao. 2015. 'Integrating Fossil Pollen Data into Ecological Modeling to Study Long-Term Ecology: From Sampling to Inference to Hypothesis Testing'. PhD Thesis, University of Wyoming.

Liu, Yao. 2022. yliu11/bayesian-vegetation-pollen: Pollen-vegetation model (Version v1). Zenodo. https://doi.org/10.5281/zenodo.5825842

Matthias, Isabelle, Anne Birgitte Nielsen, and Thomas Giesecke. 2012. 'Evaluating the Effect of Flowering Age and Forest Structure on Pollen Productivity Estimates'. *Vegetation History and Archaeobotany* 21 (6): 471–84.

Mazier, F., M.-J. Gaillard, P. Kuneš, S. Sugita, A.-K. Trondman, and A. Broström. 2012. 'Testing the Effect of Site Selection and Parameter Setting on REVEALS-Model Estimates of Plant Abundance Using the Czech Quaternary Palynological Database'. *Review of Palaeobotany and Palynology* 187 (November): 38–49. https://doi.org/10.1016/j.revpalbo.2012.07.017.

Nolan, Connor, Jonathan T. Overpeck, Judy R. M. Allen, Patricia M. Anderson, Julio L. Betancourt, Heather A. Binney, Simon Brewer, et al. 2018. 'Past and Future Global Transformation of Terrestrial Ecosystems under Climate Change'. *Science* 361 (6405): 920–23. https://doi.org/10.1126/science.aan5360.

O'Connell, B.M., B.L. Conkling, A.M. Wilson, E.A. Burrill, J.A. Turner, S.A. Pugh, G. Christensen, T. Ridley, and J. Menlove. 2017. 'The Forest Inventory and Analysis Database: Databased Description and User Guide for Phase 2 (Version 7.0)'. *User Guide*. Vol. 7.

Oreskes, N, K Shrader-Frechette, and K Belitz. 1994. 'Verification, Validation, and Confirmation of Numerical Models in the Earth Sciences'. *Science* 263 (5147): 641.

Overpeck, J. T., T. Webb, and I. C. Prentice. 1985. 'Quantitative Interpretation of Fossil Pollen Spectra: Dissimilarity Coefficients and the Method of Modern Analogs'. *Quaternary Research* 23 (1): 87–108. https://doi.org/10.1016/0033-5894(85)90074-2.

Paciorek, Christopher J., and Jason S. McLachlan. 2009. 'Mapping Ancient Forests: Bayesian Inference for Spatio-Temporal Trends in Forest Composition Using the Fossil Pollen Proxy Record'. *Journal of the American Statistical Association* 104 (486): 608–22. https://doi.org/10.1198/jasa.2009.0026.

Parsons, R. W., A. D. Gordon, and I. C. Prentice. 1983. 'Statistical Uncertainty in Forest Composition Estimates Obtained from Fossil Pollen Spectra via the R-Value Model'. *Review of Palaeobotany and Palynology* 40 (3): 177–89. https://doi.org/10.1016/0034-6667(83)90035-0.

Plummer, M. 2003. 'JAGS: A Program for Analysis of Bayesian Graphical Models Using Gibbs Sampling'. *Proceedings of the 3rd International Workshop on Distributed Statistical Computing (DSC 2003)*, 20–22. https://doi.org/10.1.1.13.3406.

Poska, Anneli, Vivika Meltsov, Shinya Sugita, and Jüri Vassiljev. 2011. 'Relative Pollen Productivity Estimates of Major Anemophilous Taxa and Relevant Source Area of Pollen in a Cultural Landscape of the Hemi-Boreal Forest Zone (Estonia)'. *Review of Palaeobotany and Palynology* 167 (1–2): 30–39.

Prentice, I C, and Thompson Webb III. 1986. 'Pollen Percentages, Tree Abundances and the Fagerlind Effect'. *Journal of Quaternary Science* 1 (1): 35–43.

Prentice, I. Colin. 1985. 'Pollen Representation, Source Area, and Basin Size: Toward a Unified Theory of Pollen Analysis'. *Quaternary Research* 23 (1): 76–86. https://doi.org/10.1016/0033-5894(85)90073-0.

Raynor, Gilbert S, V Janet Hayes, and Eugene C Ogden. 1974. 'Particulate Dispersion into and within a Forest'. *Boundary-Layer Meteorology* 7 (4): 429–56.

———. 1975. 'Particulate Dispersion from Sources within a Forest'. *Boundary-Layer Meteorology* 9 (3): 257–77.

Soepboer, Welmoed, Shinya Sugita, André F. Lotter, Jacqueline F.N. Van Leeuwen, and Willem O. Van Der Knaap. 2007. 'Pollen Productivity Estimates for Quantitative Reconstruction of Vegetation Cover on the Swiss Plateau'. *Holocene* 17 (1): 65–77. https://doi.org/10.1177/0959683607073279.

Sugita, S., M. J. Gaillard, and A. Broström. 1999. 'Landscape Openness and Pollen Records: A Simulation Approach'. *Holocene* 9 (4): 409–21. https://doi.org/10.1191/095968399666429937.

Sugita, Shinya. 1993. 'A Model of Pollen Source Area for an Entire Lake Surface'. *Quaternary Research* 39 (2): 239–44. https://doi.org/10.1006/qres.1993.1027.

———. 1994. 'Pollen Representation of Vegetation in Quaternary Sediments: Theory and Method in Patchy Vegetation'. *The Journal of Ecology* 82 (4): 881. https://doi.org/10.2307/2261452.

———. 2007a. 'Theory of Quantitative Reconstruction of Vegetation I: Pollen from Large Sites REVEALS Regional Vegetation Composition'. *Holocene* 17 (2): 229–41. https://doi.org/10.1177/0959683607075837.

———. 2007b. 'Theory of Quantitative Reconstruction of Vegetation II: All You Need Is LOVE'. *Holocene* 17 (2): 243–57. https://doi.org/10.1177/0959683607075838.

Sugita, Shinya, Tim Parshall, Randy Calcote, and Karen Walker. 2010. 'Testing the Landscape Reconstruction Algorithm for Spatially Explicit Reconstruction of Vegetation in Northern Michigan and Wisconsin'. *Quaternary Research* 74 (2): 289–300. https://doi.org/10.1016/j.yqres.2010.07.008.

Theuerkauf, Martin, John Couwenberg, Anna Kuparinen, and Volkmar Liebscher. 2016. 'A Matter of Dispersal: REVEALSinR Introduces State-of-the-Art Dispersal Models to Quantitative Vegetation Reconstruction'. *Vegetation History and Archaeobotany* 25 (6): 541–53. https://doi.org/10.1007/s00334-016-0572-0.

Theuerkauf, Martin, Anna Kuparinen, and Hans Joosten. 2013. 'Pollen Productivity Estimates Strongly Depend on Assumed Pollen Dispersal'. *Holocene* 23 (1): 14–24. https://doi.org/10.1177/0959683612450194.

Trachsel, Mathias, Andria Dawson, Christopher J. Paciorek, John W. Williams, Jason S. McLachlan, Charles V. Cogbill, David R. Foster, et al. 2020. 'Comparison of

Settlement-Era Vegetation Reconstructions for STEPPS and REVEALS Pollen–Vegetation Models in the Northeastern United States'. *Quaternary Research* 95 (May): 23–42. https://doi.org/10.1017/qua.2019.81.

Wan, Qiuchi, Yaze Zhang, Kangyou Huang, Qianwen Sun, Xiao Zhang, Marie-José Gaillard, Qinghai Xu, Furong Li, and Zhuo Zheng. 2020. 'Evaluating Quantitative Pollen Representation of Vegetation in the Tropics: A Case Study on the Hainan Island, Tropical China'. *Ecological Indicators* 114 (July): 106297. https://doi.org/10.1016/j.ecolind.2020.106297.

Williams, John W., Bryan Shuman, and Patrick J. Bartlein. 2009. 'Rapid Responses of the Prairie-Forest Ecotone to Early Holocene Aridity in Mid-Continental North America'. *Global and Planetary Change* 66 (3–4): 195–207. https://doi.org/10.1016/j.gloplacha.2008.10.012.

Williams, John W., Pavel Tarasov, Simon Brewer, and Michael Notaro. 2011. 'Late Quaternary Variations in Tree Cover at the Northern Forest-Tundra Ecotone'. *Journal of Geophysical Research* 116 (G1): G01017. https://doi.org/10.1029/2010JG001458.

Willis, Kathy J, Keith D Bennett, Shonil A Bhagwat, and H John B Birks. 2010. '4° C and beyond: What Did This Mean for Biodiversity in the Past?' *Systematics and Biodiversity* 8 (1): 3–9.

Zanon, Marco, Basil A. S. Davis, Laurent Marquer, Simon Brewer, and Jed O. Kaplan. 2018. 'European Forest Cover During the Past 12,000 Years: A Palynological Reconstruction Based on Modern Analogs and Remote Sensing'. *Frontiers in Plant Science* 9. https://doi.org/10.3389/fpls.2018.00253.

Figure Legends

Figure 1. The study sites include 33 small lakes (area < 2 ha) located in three regions in the northeastern United States: 14 lakes in Fish Creek, New York (FC, blue squares), nine (9) in the eastern Adirondack Mountains, New York (EAD, yellow spades), and 10 in the southern New England including Massachusetts, Connecticut, and Rhode Island (SNE, green triangles). Because some sites are closely clustered and their symbols largely overlap on the map, the positions of the symbols are slightly jittered for clearer presentation. Correspondence between letters and sites is given in Appendix S1: Table S1. Locations of Forest Inventory and Analysis (FIA) plots measured between 2000 and 2012 are represented by the small, dense, gray points.

Figure 2. Observed versus predicted pollen relative abundances (proportions), where the predicted values are represented by the posterior means for $P_{s,t}$ (see Equations (1) and (2)). Taxon-specific coefficients of determination (R^2) are based on a linear regression (solid line) of observed versus predicted values, overlaid with the 1:1 line (dashed line). Site symbols (letters) are described in Appendix S1: Table S1 and the region color-scheme is defined in Fig. 1.

Figure 3. Posterior estimates (median and 95% credible interval [CI]) of the relative pollen productivity. Taxa are alphabetically ordered along the x-axis, and region-level productivity ($\mathbf{\Phi}_r$, Equations (2) and (3); shown in blue, yellow, and green symbols) are shown in relation to the overall productivity (population-level, across all sites, $\mathbf{\Phi}^*$,

Equations (3) and (5); shown in black diamonds). The region-level productivities within each region and the population-level productivities all sum to one, respectively. For productivity scaled to reference taxa, see Appendix S1: Fig. S2-S3.

Figure 4. Posterior estimates (median and 95% credible interval [CI]) of the local contribution from within 1 km radius, γ_t (see Equations (2) and (6)), of vegetation to pollen relative abundances. Taxon-specific γ_t is assumed to be the same across all sites.

Figure 5. Posterior estimates for the cumulative influence of local vegetation at different distances from a lakeshore, Fl (Equation (13)), within 1 km of a typical, 30 m radius lake. Estimates from the model (orange points [median] and orange shaded area [95% CI]) are compared with cumulative influences based other widely-used distance weighting. The weighting using Prentice-Sutton equation (thick solid line) is based on pollen fall speeds compiled in Jackson and Lyford (1999), with parameters corresponding to neutral atmospheric conditions). A cumulative influence of 1 (100%) represents the total influence of all pollen that originated locally, where the total local influence relative to local and regional combined is given by γ_t in Fig. 4.

Figure 6. Posterior estimates for the cumulative influence of regional vegetation at different distances from a lakeshore, *Flr* (Equation (15)), from >1 km to 300 km, for a typical, 30 m radius lake. Estimates (medians [solid lines] and the 95% CIs [colored shading]) are shown for regional distances from 3 km and up to 300 km, based on assuming no local vegetation contribution (*Fr*, Equation (14); red lines) and based on

including local vegetation contribution (*Flr*, Equation (15); cyan lines) (Figs. 4 and 5). Dashed horizontal line indicates 75% cumulative influence. A cumulative influence of 1 (100%) represents the total influence of all pollen that originated regionally, where the total regional influence relative to local and regional combined is given by ($1-\gamma_t$) in Fig. 4.

Figure 7. Localness Index (*LI*, Equation (16)) is defined and calculated as the ratio between the actual local contribution (γ_t, Fig. 4) and the cumulative influence at 1 km from the lakeshore predicted by regional pollen dispersal ($\tilde{\gamma}_{s,t}$, Equation (17)). Horizontal dashed line indicates $LI = 1$; Posterior medians (symbols) and 95% CIs for *LI* are overlaid for all sites within a region (14, 9, and 10 sites respectively for FC, EAD, and SNE). CIs that significant higher (lower) than 1 would indicate that local populations are over-(under-) represented by pollen deposited in the lake, after accounting for regional dispersal. Because *Acer rubrum* and *Larix* likely lack meaningful regional input (see result section "Local pollen contribution and dispersal" and discussion section "Comparing the local and regional dispersal"), their *LI* values cannot be accurately estimated and therefore were masked in grey.

Figure 8. Spatial pattern of heterogeneous landscapes calculated from local and regional vegetation data. The dissimilarities, measured by the squared-chord distance (SCD), are shown for "focal vegetation" within (a) 20 m, (b) 100 m, (c) 500 m, (d) 1000 m, (e) 9 km, (f) 27 km, and (g) 45 km of a lakeshore. For focal vegetation within each of the aforementioned concentric ring, the dissimilarity between the its composition and the vegetation composition at a given distance from the lakeshore (x-axis) is calculate. Pollen

records from different forest types typically have SCDs > 0.2 (dashed lines). For clarity, we show the regional mean here; site-level dissimilarities can be found in Appendix S1: Fig. S7. Lines are colored by region.

Figure 1.

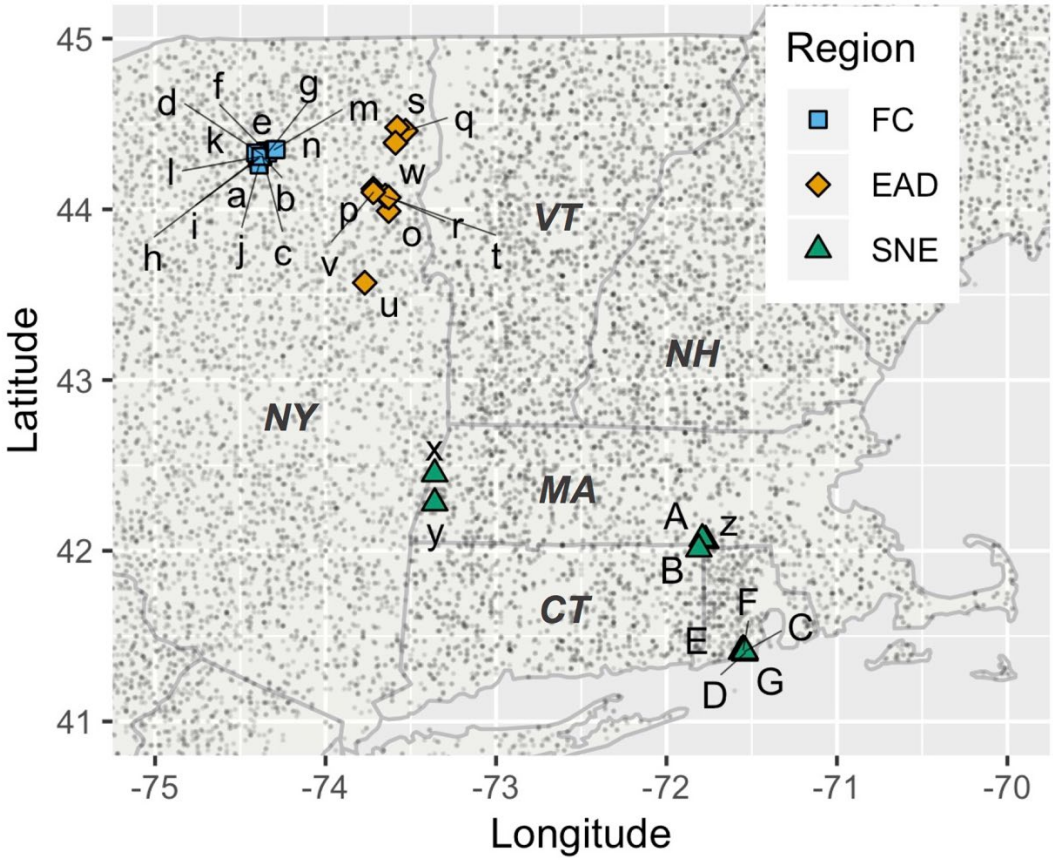

Figure 2.

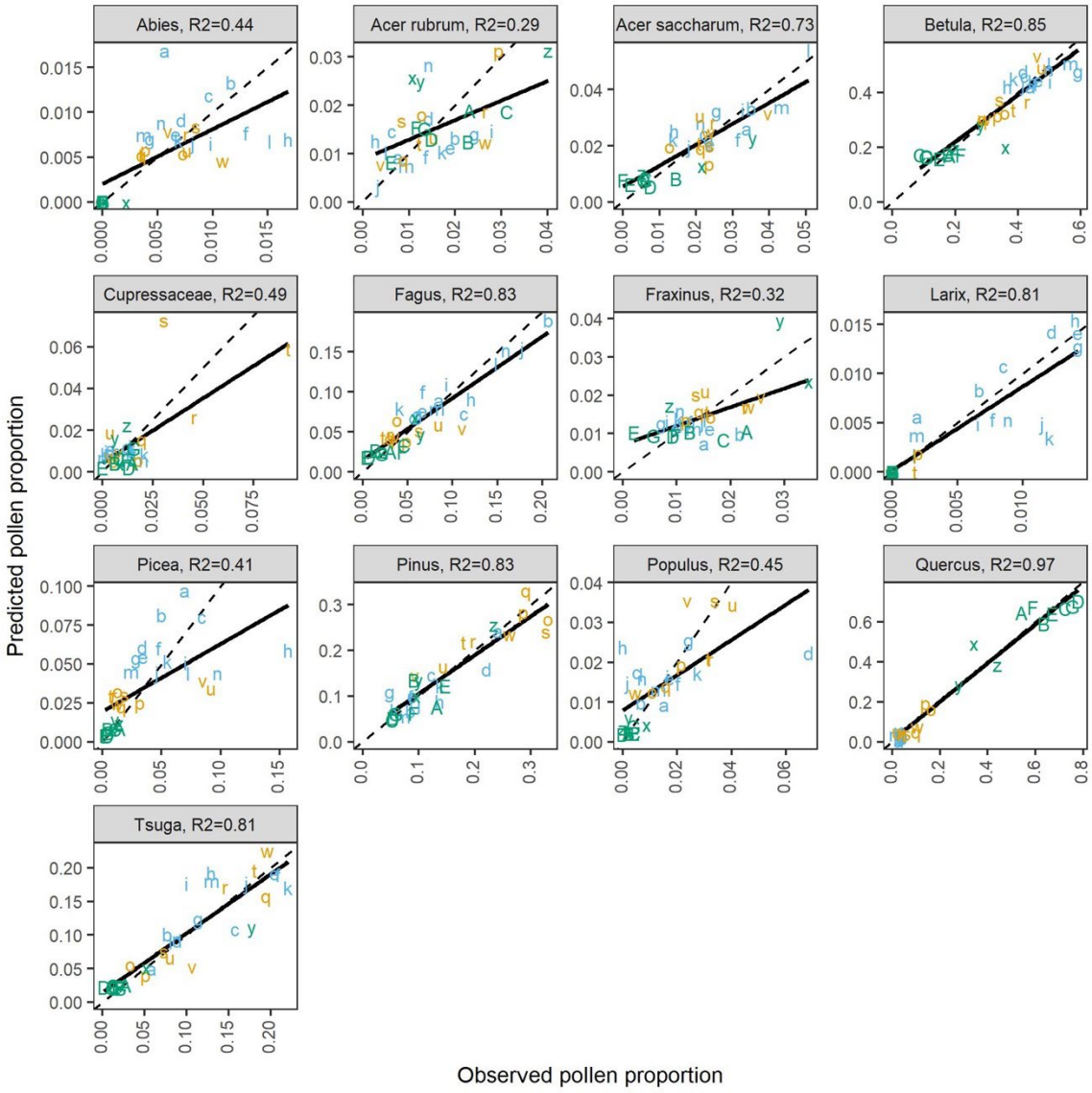

Figure 3.

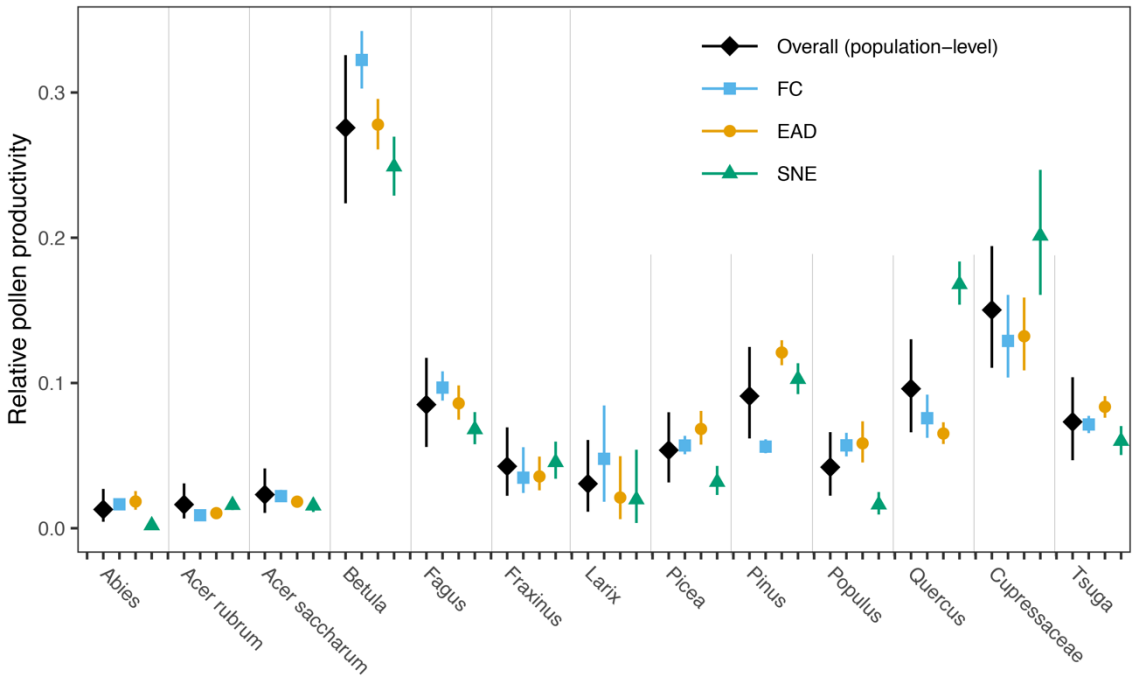

Figure 4.

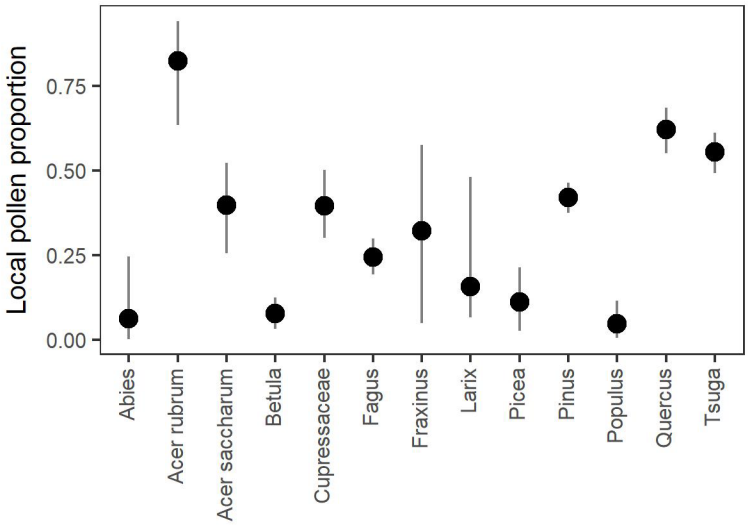

Figure 5.

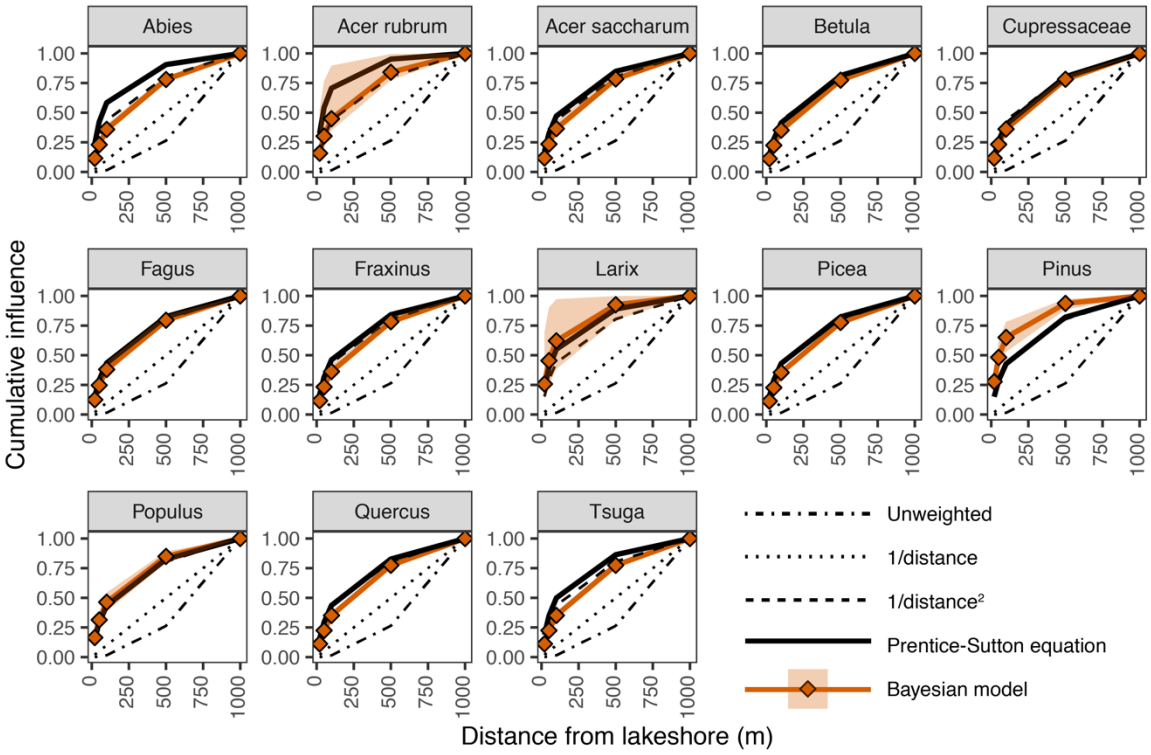

Figure 6.

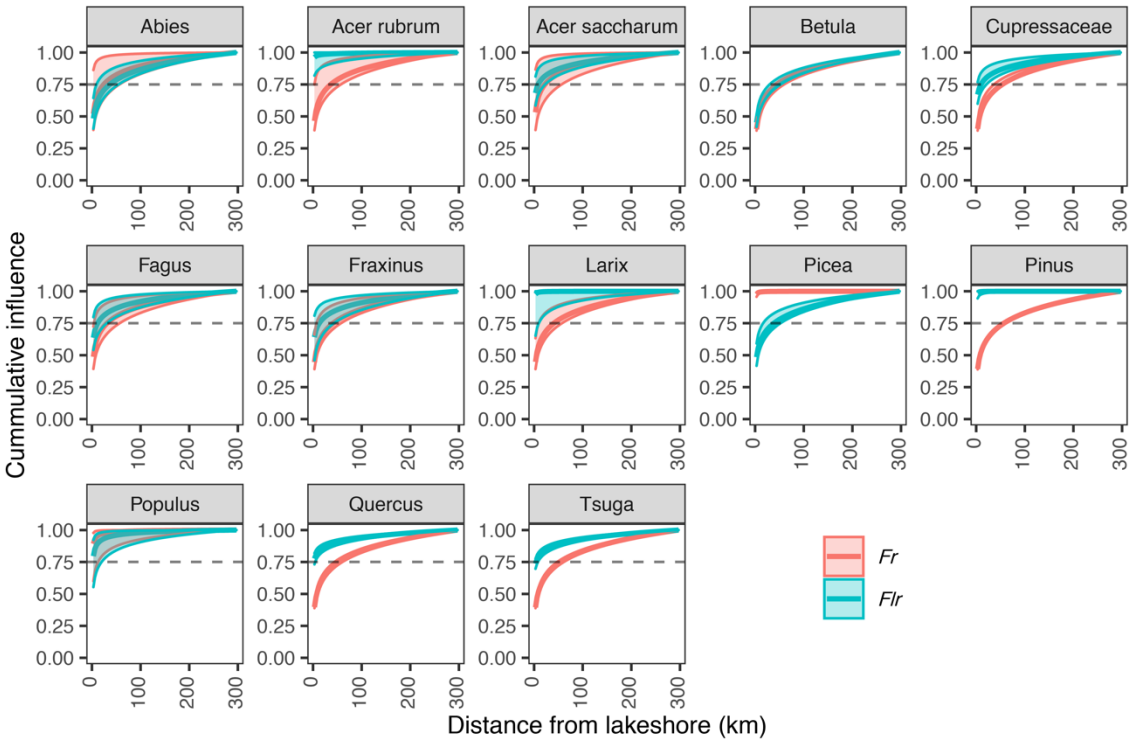

Figure 7.

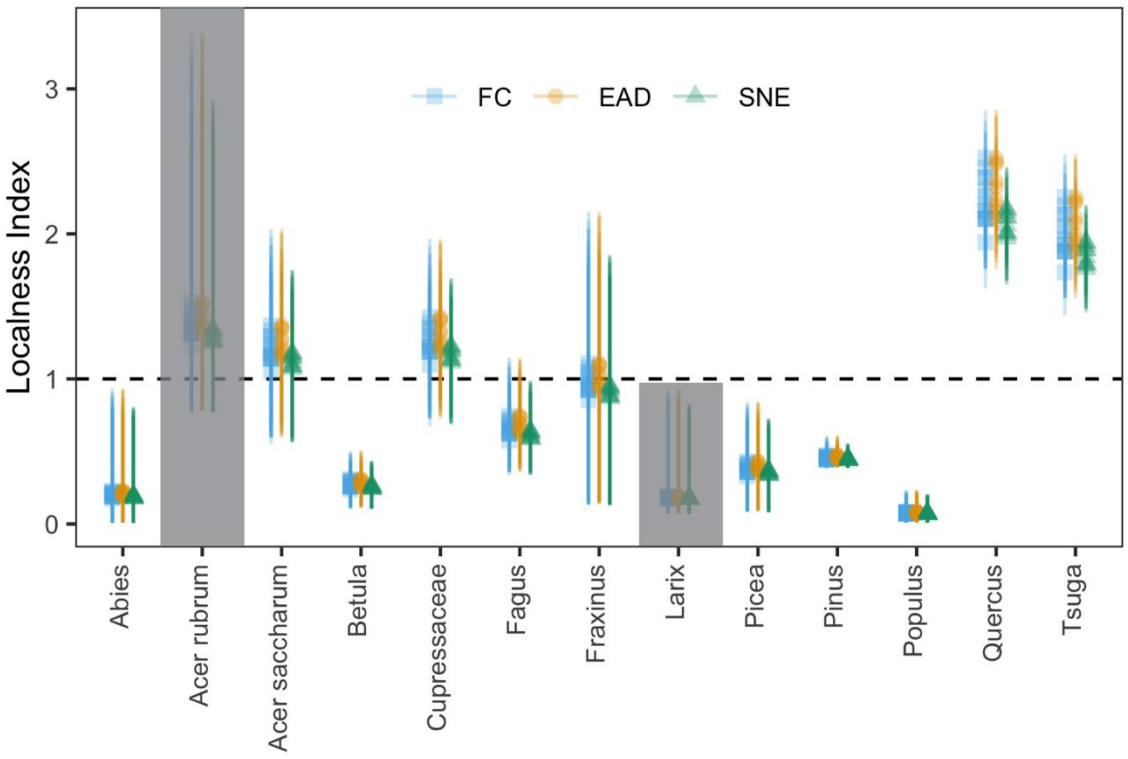

Figure 8.