Submitted to the Annals of Applied Statistics

A BAYESIAN RECORD LINKAGE APPROACH TO APPLICATIONS IN TREE DEMOGRAPHY USING OVERLAPPING LIDAR SCANS

BY LANE DREW[1,a], ANDEE KAPLAN[1,b] AND IAN BRECKHEIMER[2,c]

[1]*Department of Statistics, Colorado State University,* [a]*lane.drew@colostate.edu;* [b]*andee.kaplan@colostate.edu*

[2]*Rocky Mountain Biological Laboratory,* [c]*ikb@rmbl.org*

In the information age, it has become increasingly common for data containing records about overlapping individuals to be distributed across multiple sources, making it necessary to identify which records refer to the same individual. The goal of record linkage is to estimate this unknown structure in the absence of a unique identifiable attribute. We introduce a Bayesian hierarchical record linkage model for spatial location data motivated by the estimation of individual specific growth-size curves for conifer species using data derived from overlapping LiDAR scans. Annual tree growth may be estimated dependent upon correctly identifying unique individuals across scans in the presence of noise. We formalize a two-stage modeling framework, connecting the record linkage model and a flexible downstream individual tree growth model, that provides robust uncertainty quantification and propagation through both stages of the modeling pipeline via an extension of the linkage-averaging approach of Sadinle (2018). In this paper, we discuss the two-stage model formulation, outline the computational strategies required to achieve scalability, assess the model performance on simulated data, and fit the model to a bi-temporal dataset derived from LiDAR scans of the Upper Gunnison Watershed provided by the Rocky Mountain Biological Laboratory to assess the impact of key topographic covariates on the growth behavior of conifer species in the Southern Rocky Mountains (USA).

arXiv:2501.13285v2 [stat.ME] 5 Jun 2025

1. Introduction. The characterization and quantification of forest dynamics has been an area of interest for ecologists for more than a century and has become an increasingly important metric for understanding the effects of climate change (Hyyppä et al. (2008)). Historical investigations of forest dynamics have relied on field surveys over limited spatial domains, which are generally time consuming and potentially difficult to perform (Saatchi et al. (2011); Wensel, Meerschaert and Biging (1987)). The advent and ongoing refinement of aerial laser scanning (ALS) technology has ushered in a new age of data collection in terms of scalability. The use of ALS data in the modeling of forest structure has become a standard approach, as it enables researchers to examine the health and behavior of forests at larger scales than has previously been possible by field survey (Babcock et al. (2016); Dalponte and Coomes (2016)). The obvious extension of these efforts is to functions that rely on repeat measurements over time. Despite improvements in the accuracy of ALS technology, there remains inherent uncertainty in both the scanning mechanism and subsequent post-processing of the data as discussed by Huo and Lindberg (2020). In the existing literature, the mechanisms employed for identifying the unique individuals from scans across multiple time points have been largely heuristic and rely upon manual verification, as in Ma et al. (2018), and employ a two-stage modeling schema which fails to incorporate the uncertainty in the segmentation and matching procedures into the downstream task. To address these issues, we present an alternative two-stage framework utilizing a record linkage approach for

Keywords and phrases: Record Linkage, Entity Resolution, Bayesian Hierarchical Model, Bi-temporal LiDAR, Tree Demography.

spatial location data that is capable of efficiently identifying unique individuals across larger spatial domains while providing robust uncertainty quantification for the linkage that may then be propagated into the downstream modeling objective.

As our ability to collect and store data has exploded, so too has our need to engage in record linkage (also called deduplication or entity resolution). At its core, the field of record linkage is concerned with the resolution of unique records across overlapping files in the absence of a unique identifier. In this paper we use the term record linkage to encompass the process of identifying coreferent records both between and within files. Historically, files have represented repeated surveys over time (Steorts (2015)), or non-temporally linked overlapping databases such as patient records across different providers in the healthcare system (Padmanabhan et al. (2019)). The earliest approaches to probabilistic record linkage, as formalized by Fellegi and Sunter (1969), performed probabilistic matching between pairs of records according to a decision theoretic framework. The field has developed consistently since its inception, and advances in Bayesian computational methods have given rise to a new class of probabilistic modeling approaches over the last 20 years as discussed by Liseo and Tancredi (2011a). While methods that perform matching between pairs of records directly have remained useful (Sadinle (2017)), models built upon latent clustering structures have become increasingly popular alongside methods addressing alternative types of data (Steorts, Hall and Fienberg (2016); Liseo and Tancredi (2011b)). Recently, record linkage has been successfully used to improve wildlife population inference from a series of sequential aerial photographs (Lu et al. (2022)). In a similar vein, we introduce a record linkage model for multi-temporal spatial location data derived from light detection and ranging (LiDAR) scans intended to improve tree demography inference.

While record linkage is an interesting and challenging endeavor on its own, it generally functions as the first step in the sequence of a statistical pipeline as discussed by Kaplan, Betancourt and Steorts (2022). We implement our spatial record linkage model in a two-stage framework in which the record linkage and downstream task are performed sequentially. Our proposed approach performs the downstream modeling task using a randomly sampled subset of iterations from the posterior linkage structure, which propagates the uncertainty from the linkage into the second stage of the pipeline according to the linkage-averaging (LA) approach of Sadinle (2018). Crucially, in the LA framework, the linkage is not informed by the downstream task. Consequently, the output from the first stage record linkage model may be used as the input for a variety of downstream models, offering researchers a high degree of flexibility when adopting this modeling framework.

In our application, we pair the spatial record linkage model with an individual tree growth model, where we define growth as the change in canopy volume between surveys on an annual scale. The empirical data is comprised of LiDAR scans of Gunnison National Forest from 2015 and 2019 provided by the Rocky Mountain Biological Laboratory (RMBL). The individual tree crown polygons and associated attributes were obtained from a 1/3 m resolution canopy height model using the `ITCsegment` (Dalponte and Coomes (2016)) algorithm in the `lidR` (version 3.4, Roussel et al. (2020)) package in R. We note the record linkage and growth models are both designed to account for various sources of biological variation and measurement error in the data collection and post-processing procedures.

The remaining structure of the paper is as follows. In Section 2, we highlight the ecological hypotheses and empirical data that motivate our novel modeling approach. In Section 3, we introduce the relevant notation for the spatial record linkage model and downstream growth model. Additionally, we outline the computational strategies used to fit the model. In Section 4, we provide a discussion of the theoretical justification for the LA approach in a general auxiliary data task setting. In Section 5, we provide an analysis of the empirical data and in Section 6, we examine the performance of the proposed modeling approach in a series of numerical experiments on simulated data. We conclude with a discussion and directions for future work in Section 7.

TABLE 1
LiDAR data collection attributes for the 2015 and 2019 scans provided by RMBL.

Scan Attribute	2015 Scan	2019 Scan
Acquisition Dates	August 7, 2015 and August 10, 2015	August 21 – September 24, 2019
Aircraft Used	Piper Navajo	Cessna Caravan
Sensor	Reigl (Leica) Q1560	Riegl (Leica) VQ1560i
Maximum Returns / Pulse	5	15
Target Pulse Density	Average 8 pulses m^2	Average 2 pulses m^2
Realized Point Density	9.4 pts / m^2	9.4 pts / m^2
Survey Altitude (AGL)	550 m	1159 m
Field of View	58°	58.5°

2. LiDAR Derived Individual Tree Characteristics from Multi-temporal Scans of Snodgrass Mountain. We apply our two-stage modeling approach to identify unique trees across time points and to estimate spatial patterns of tree growth as related to certain environmental drivers in a spruce-fir forest site located in the Southern Rocky Mountains (USA). The study site is a two square kilometer forested domain located on Snodgrass Mountain near the site of RMBL in the vicinity of Crested Butte, Colorado. The domain spans montane to lower subalpine mountain slopes at elevations from 2891-3395m and experiences a cold continental climate with persistent seasonal snowpack accounting for the majority of annual precipitation (Carroll, Gochis and Williams (2020)). Evergreen forests in the domain are dominated by Engelmann spruce (*Picea engelmannii*) and subalpine fir (*Abies lasiocarpa*), which account for more than 80% of the tree canopy. These forests also contain scattered lodgepole pine (*Pinus contorta*) and Rocky Mountain Douglas-fir (*Pseudotsuga menziesii subsp. glauca*). Deciduous quaking aspen (*Populus tremuloides*) forms large single-species stands on lower slopes of the study area, but these areas were excluded from the analysis due to the difficulty of assessing the growth of this species.

Forest structural data was collected for the study area via LiDAR in two intervals during 2015 and 2019. Both scans were collected from an airplane-based sensor in late summer before the drop of deciduous leaves. The laser scanner records discrete peaks of reflected energy at near-infrared wavelengths and uses the integrated Real-time Kinematic (RTK) sensor position and estimated time-of-flight of laser pulses to locate laser reflections ("returns") in geographic space. The scanning process yields a dense (8-16 pts / m^2) cloud of 3-dimensionally located returns representing reflections from the ground, tree canopies, and other reflective surfaces. Details of each LiDAR dataset are provided in Table 1.

The LiDAR-derived point clouds were segmented and summarized to yield estimates of per-tree structural characteristics including tree top locations, maximum heights, and canopy volumes using functions in the R package `lidR` (version 3.4, Roussel et al. (2020)). Although numerous approaches exist to segment individual trees in LiDAR data (see Aubry-Kientz et al. (2019) for a recent comparison), for this analysis we adopted the commonly-used `ITCsegment` algorithm (Dalponte and Coomes (2016)). `ITCsegment` is a region-growing approach which iteratively incorporates points into candidate tree canopies starting at a set of seed locations. Seed locations (putative tree tops) were selected using a local maximum filter, identifying laser returns with high heights relative to a height-dependent local neighborhood. Canopy volumes were calculated by summing canopy heights for each segment using a 1 m resolution canopy surface model generated using the `pit-free` algorithm implemented in `lidR`. The segmentation and canopy surface model generation steps generate imperfect representations of individual tree locations and crown geometries. Errors in sensor geo-positioning and ranging measurements can lead to systematic spatial shifts in datasets collected at different times. Moreover, the location and trajectory of individual laser pulses

TABLE 2
Designations and details for the topographic covariates included in the analysis.

Growth Constraint	Covariate	Data Source	Resolution (m)
Energy	Folded Aspect	Goulden et al. (2020)	1
Energy	Growing Degree Days	Breckheimer (2023)	30
Water	HAS Wetness Index	Goulden et al. (2020); Nobre et al. (2011)	1
Water	Snowpack Persistence	Breckheimer (2023)	30

differs between scans, leading to small amounts of variability in estimated maximum heights and crown locations as discussed by Poorazimy et al. (2022).

In Western North American conifer forests, tree growth is thought to be constrained by the (potentially interacting) availability of water and energy (Buechling, Martin and Canham (2017); Heilman et al. (2022)). To investigate these constraints, we assembled estimates of key water and energy proxies across the domain from diverse remote sensing datasets. A 1 m resolution LiDAR-derived Digital Elevation Model (Goulden et al. (2020)) was used to compute topographic aspect "folded" about the north-south axis to distinguish high solar-radiation south-facing slopes from low solar-radiation north-facing slopes. We also computed a topographic wetness index (TWI) (Nobre et al. (2011)) as a water availability proxy. We augmented these topographic proxies with gridded climate data interpolated from weather station and microclimate sensors as well as satellite-derived maps of the persistence of seasonal snowpack (Breckheimer (2023)). Data for snowpack persistence and growing degree days was aggregated annually from 2015 to 2019 and the median observed values were used during modeling to capture the relative impact of these covariates over the period between LiDAR scans. Details pertaining to the covariates may be found in Table 2 along with a visualization of the derived tree geometries and raster images in Figure 1.

3. Models and Notation. In this section we detail the spatial record linkage and growth models employed in our two-stage LA approach. We first define the necessary notation and present the proposed spatial record linkage model before developing the downstream individual tree growth model. We note that throughout this section, distributions identified with subscripts refer to truncated distributions over the specified bounds. For example a truncated Inverse-Gamma distribution with parameters c and d over the range $[0, b]$ is denoted as Inverse-Gamma$_{[0,b]}(c, d)$. We finish this section with a discussion of the computational strategies employed to facilitate scalability of the record linkage model to spatial domain sizes that are of practical interest.

3.1. *Record Linkage Model.* We begin by presenting the spatial record linkage model as a standalone component to introduce the model structure and to establish a baseline for inclusion in modeling pipelines with alternative downstream tasks. We provide a general model capable of handling two files, which is also capable of performing deduplication within files. We employ a Bayesian hierarchical structure based on latent matching, as discussed by Steorts, Hall and Fienberg (2016) and Liseo and Tancredi (2011b), such that records are linked to unobserved latent entities with true field values instead of being probabilistically matched to other records directly through comparison vectors, as in the work of Fellegi and Sunter (1969) and Sadinle (2017). In the spatial record linkage model, the value associated with the latent entity is the true unobserved location of the individual. We treat the observed data (i.e. location) as a noisy observation of the latent location and include provisions for different potential sources of noise. We consider error introduced as a function of translation and rotation in the data collection process, post-processing of the data, and due to biological mechanisms. We specify the data model and relevant notation as follows.

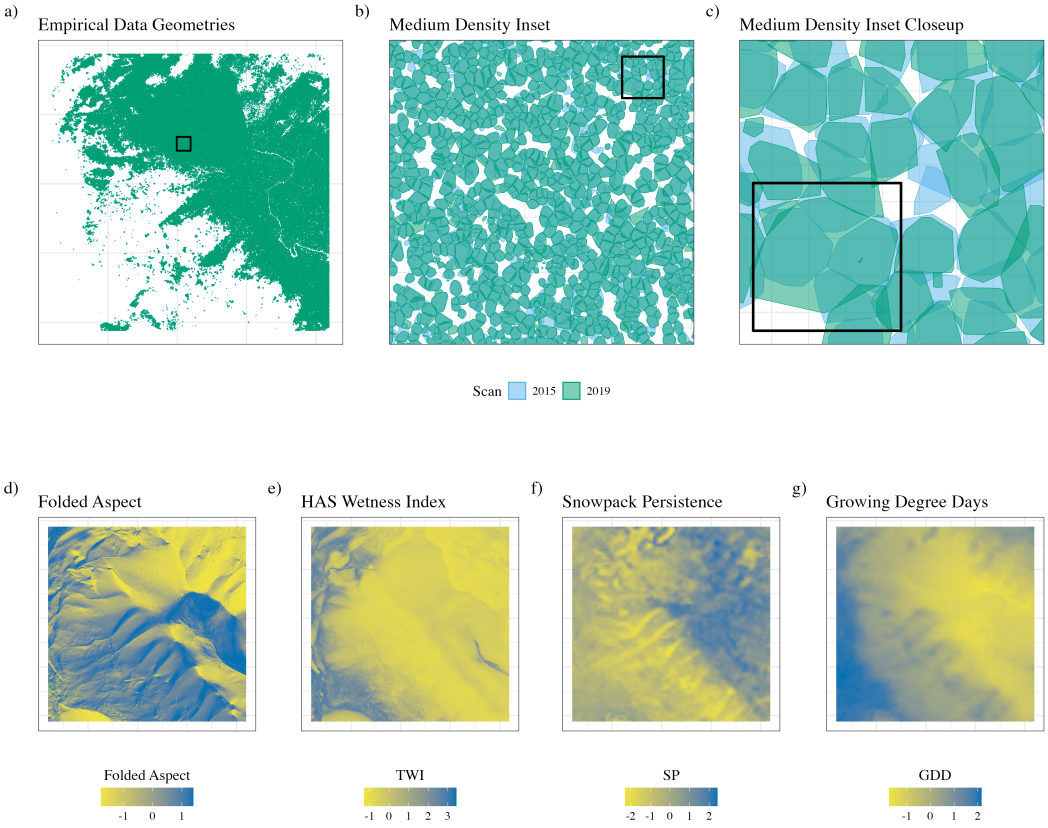

FIG 1. *Plots a) and b) show the derived crown geometries from the 2015 and 2019 LiDAR scans performed by RMBL for the full datasets and a medium density inset (outlined in plot a)). Plot c) shows a closeup from the inset in b), highlighting an instance in which multiple trees in the first file overlap with a single tree in the second file. Plots d) - g) show the scaled raster images for the topographic covariates of interest over the study domain.*

The model is constructed to handle two files, where the files are indexed by $i = 1, 2$ with relative size n_i for each file. The records within files are indexed from $j = 1, \ldots, n_i$ such that the total number of observed records is $n = \sum_{i=1}^{2} n_i$. We denote the observed location data for the jth record in file i as $\boldsymbol{y}_{ij}$, where $\boldsymbol{y}_{ij}$ is a numerical vector of length 2, i.e. $\boldsymbol{y}_{ij} = (x, y)_{ij}$, corresponding to the spatial coordinates of the record in the (x, y)-plane. We note that in our application the term file is synonymous with a LiDAR scan and record with an individual identified tree such that n is the total number of individual trees detected across all scans.

We define the latent location vector as $\boldsymbol{s}_{j'}$, where $j' = 1, \ldots, N$ such that N is the maximum number of unique latent individuals in the population under consideration. The observed locations $\boldsymbol{y}_{ij}$ are modeled as noisy versions of the latent locations $\boldsymbol{s}_{j'}$. We assume that the record set $\boldsymbol{y}_{1j}$ exists in the same space as the latent locations, while the record set $\boldsymbol{y}_{2j}$ is modeled as a transformed version of the associated $\boldsymbol{s}_{j'}$, where we restrict the possible transformations considered to rotation and translation.

The linkage structure, which identifies the relationship between the observed records and the latents, is a vector of length n denoted as $\boldsymbol{\Lambda} = \{\lambda_{ij} : i = 1, 2;\ j = 1, \ldots, n_i\}$, where λ_{ij} is an integer indicating which $\boldsymbol{s}_{j'}$ the jth record in file i refers to. The linkage is implicitly dependent on the maximum latent population size N, as $\lambda_{ij} \in \{1, \ldots, N\}$. The specified linkage structure naturally defines a set of N clusters denoted $\mathcal{C}(\boldsymbol{\Lambda})$, which specify the records that

are linked to each $\boldsymbol{s}_{j'}$ such that $\mathcal{C}(\boldsymbol{\Lambda}) = \{\mathcal{C}_1, \ldots, \mathcal{C}_N\}$. The individual clusters are defined as the sets $\mathcal{C}_{j'} = \{(i,j) : \lambda_{ij} = j'\}$ for $j' = 1, \ldots, N$, and we note that the clusters may be empty or may contain records from the same file in addition to records across files highlighting the capacity of the model for performing record linkage and deduplication simultaneously. In the context of our application, duplicates within files potentially occur during the LiDAR processing due to the segmentation algorithm, such that a single individual may be erroneously split into multiple entities as seen in panel c) of Figure 1.

Previous record linkage modeling approaches are known to be sensitive to the specification of N, which functions as a hyperparameter in the model and quantifies our belief regarding the upper bound on the number of unique entities across all files (Steorts, Hall and Fienberg (2016)). We specify $N = q \times \max(n_i)$ where the scale factor q is chosen to reflect the assumed degree of overlap between the files and such that $N \leq n$. Alternatively, a practitioner could specify $N = n$ to avoid making any a priori assumptions about the number of unique individuals across files. Although not explicitly a parameter in the model, we do effectively obtain an estimate of the number of unique individuals across files which is often of interest in studies investigating species abundance and provides some sense of the effective sample size for estimation of the downstream model parameters. Additional discussion regarding the specification of N is provided in the Supplementary Materials in Appendix A (Drew, Kaplan and Breckheimer (2025)).

The data model, which describes the relationship between the observed point patterns and $\boldsymbol{s}$, allows us to model the variation produced by the underlying biological process (assumed to be tree growth in this application) separately from the error introduced in the LiDAR scanning and post-processing procedures by incorporating the image alignment framework introduced by Green and Mardia (2006). The model is specified as follows,

$$\boldsymbol{y}_{ij} \mid \boldsymbol{s}_{\lambda_{ij}}, \sigma^2, \boldsymbol{t}_i, \theta_i, D \sim \text{Normal}_{2,[D]}\left(\boldsymbol{R}(\theta_i)\left(\boldsymbol{s}_{\lambda_{ij}} - \boldsymbol{\mu}_D\right) + \boldsymbol{t}_i + \boldsymbol{\mu}_D, \sigma^2\boldsymbol{I}\right),$$

for $i = 1, 2$. The rotation, $\boldsymbol{R}(\theta_i)$, is the standard counterclockwise rotation matrix given by

$$\boldsymbol{R}(\theta_i) = \begin{bmatrix} \cos(\theta_i) & -\sin(\theta_i) \\ \sin(\theta_i) & \cos(\theta_i) \end{bmatrix}$$

and $\boldsymbol{t}_i$ is the two dimensional translation vector. We allow the rotation angle and translation to vary across files (i.e. scans). The rotation for each file is around the midpoint, denoted $\boldsymbol{\mu}_D$, of the spatial domain of interest D. We also note that the record set $\boldsymbol{y}_{1j}$ can be expressed in terms of the rotation and translation framework with fixed $\theta_1 = 0$ and $\boldsymbol{t}_1 = \begin{bmatrix} 0 & 0 \end{bmatrix}^\mathsf{T}$, assuming that the records in the first file exist in the same space as the latent locations, $\boldsymbol{s}$, to reduce the effective number of parameters in the model. We adopt this expression to simplify notation going forward. The full spatial record linkage model is specified as follows

$$\begin{aligned}
\boldsymbol{y}_{ij} \mid \boldsymbol{s}_{\lambda_{ij}}, \sigma^2, \theta_i, \boldsymbol{t}_i, D &\overset{ind}{\sim} \text{Normal}_{2,[D]}\left(\boldsymbol{R}(\theta_i)\left(\boldsymbol{s}_{\lambda_{ij}} - \boldsymbol{\mu}_D\right) + \boldsymbol{t}_i + \boldsymbol{\mu}_D, \sigma^2\boldsymbol{I}\right) \\
\boldsymbol{s}_{j'} \mid N &\overset{iid}{\sim} \text{Uniform}(D^*) \\
\sigma^2 &\sim \text{Inverse-Gamma}_{[0,b_\sigma]}(c_\sigma, d_\sigma) \\
\lambda_{ij} \mid N &\overset{iid}{\sim} \text{Uniform}\{1, \ldots, N\} \\
\theta_i &\propto \exp(\kappa\cos(\nu)\cos(\theta_i) + \kappa\sin(\nu)\sin(\theta_i))\, I\{|\theta_i| < b_\theta\} \\
\boldsymbol{t}_i &\sim \text{Normal}_2\left(\boldsymbol{0}, \sigma_t^2\boldsymbol{I}\right),
\end{aligned}$$

where the prior for θ_i, the rotation parameter for file i, is the kernel of a truncated von Mises distribution, as discussed by Green and Mardia (2006).

As mentioned above, we model the observed locations as noisy transformations of the unobserved true $\boldsymbol{s}_{j'}$ according to a Gaussian noise process. We specify the underlying latent point process, $\boldsymbol{s}$, to follow a uniform distribution over a slightly expanded spatial domain D^*, such that $D \subseteq D^*$, to allow for the possibility that the true location of an individual is outside of the observed spatial domain (i.e. the tree base is located outside of D, but the observed tree crowns are inside). We assume the simplest and least informative prior specification for $\boldsymbol{s}$, corresponding to complete spatial randomness with a fixed number of points. This prior allows the observed locations to provide the bulk of the information in determining the distribution of the $\boldsymbol{s}_{j'}$ and enables this distribution to vary adequately over space. In practice, users could specify more complicated latent process models, which may fit the observed data more precisely but with additional assumptions, computational cost, and complexity (see Leininger (2014) for an in-depth discussion of point process models in a Bayesian hierarchical framework). We provide the derivations of the joint posterior and full conditional distributions for the model alongside the algorithm for fitting the model in Appendix A of the Supplementary Materials (Drew, Kaplan and Breckheimer (2025)).

Many latent record linkage modeling approaches employ a hit-miss mechanism for whether the observed records are a noisy distortion of the true latent value in a given field as in Steorts, Hall and Fienberg (2016). In contrast, our model treats every observed record as a noisy copy of the latent as we are dealing with spatial locations over a continuous domain. In the context of our motivating application, the observed locations are the tree crowns while the latent location is the tree base. Thus, the assumption that every observed location is a noisy observation of the truth has a clear physical interpretation in this case as well. We place a conjugate truncated Inverse-Gamma distribution prior on the measurement error parameter σ^2, where the upper bound corresponds to the maximum displacement that would be considered plausible based on the biological mechanisms of tree growth and the calibration of the LiDAR scanning equipment.

We note that our record linkage inspired modeling approach could be interpreted as a microclustering approach such that cluster sizes remain small even as the number of records grows. Betancourt, Zanella and Steorts (2022) introduced a class of random partition models with microclustering behavior to achieve a similar goal, i.e. guaranteeing that clusters remain small by assuming exchangeable sequences of clusters instead of exchangeable data points. In contrast to this approach, we obtain microclustering behavior in our model through the specification of N and the prior on σ^2 by guaranteeing a maximum number of unique latent locations and distance that observed locations may be from their associated true latent location, despite the fact that the prior for the linkage does not explicitly guarantee this property.

3.2. *Downstream Growth Model.* We now turn to the individual tree growth model we employ in this application. The model leverages the known allometric relationship between size and growth by using a flexible nonlinear function of the generalized Michaelis-Menten type to describe the annual individual growth-size curve while allowing for measurement error in the observed growth (measured by the change in canopy volume). Michaelis-Menten functions have been applied broadly across biological and ecological growth models, as described by López et al. (2000) and Bolker (2008). The generalized Michaelis-Menten function can take on a range of shapes from a logistic to a sigmoidal curve depending on the parameterization, where our specification may be seen in Eq. (1) below. The relative simplicity and flexibility of the function class combined with the clear biological interpretations of the parameters make this model a compelling choice. Michaelis-Menten growth models have been found to be ideal for describing the relationship between diameter at breast height (DBH) and height for trees (Barbosa et al. (2019); Brahma et al. (2017)), which is analogous to the size-growth relationship between canopy volume and tree growth in our application.

Although literature employing a Michaelis-Menten function in a measurement error model is scarce, the extension is natural and straight forward.

Our specification of the growth function incorporates topographically derived covariates, discussed in Section 2, allowing us to better understand the impact environmental drivers have on the growth of conifer species. We introduce relevant notation for the downstream growth model and clarify the relationship with the spatial record linkage model as follows.

We previously defined the set of clusters $\mathcal{C}(\mathbf{\Lambda})$ derived from the linkage structure of the spatial record linkage model, and we further restrict this set to the clusters for which growth is observed. We define the set of growth clusters $\mathcal{C}^G(\mathbf{\Lambda})$, with respect to several ecological conditions, which correspond to the individual trees identified by the linkage model for which a plausible change in canopy volume has occurred between the two time points. For notational clarity, we introduce the functions $\min^f()$ and $\max^f()$ which return the minimum and maximum file index, respectively, for a given cluster. We note the implicit ordering in the file indices relative to time, such that the first file is the oldest and the second file is the most recent. We denote the observed canopy volume of the jth record in file i as v_{ij}, measured in cubic meters, where the file index is synonymous with the data collection time point associated with the file. The set of growth clusters may be defined accordingly as $\mathcal{C}^G(\mathbf{\Lambda}) = \{\mathcal{C}_{j'} : \max^f(\mathcal{C}_{j'}) \neq \min^f(\mathcal{C}_{j'}) \ \& \ r_1 \cdot v^*_{\min^f(\mathcal{C}_{j'})} < v^*_{\max^f(\mathcal{C}_{j'})} < r_2 \cdot v^*_{\min^f(\mathcal{C}_{j'})}\}$ such that $\mathcal{C}^G(\mathbf{\Lambda}) \subseteq \mathcal{C}(\mathbf{\Lambda})$. Where $v^*_{\min^f(\mathcal{C}_{j'})}$ and $v^*_{\max^f(\mathcal{C}_{j'})}$ denote the summed volumes of the records associated with the minimum and maximum file indices in the cluster. This implicitly defines a procedure that merges the volumes of linked records within files as a result of deduplication from the linkage model. The hyperparameters r_1 and r_2 control the lower and upper bounds for the change in canopy volume for a growth cluster relative to the typical and biologically feasible growth behavior for the time interval between the observed records and such that $0 < r_1 < r_2$. For example, if we specify $r_1 = .9$ and $r_2 = 1.6$, then we would restrict our set of growth clusters to those that saw between a 10% loss and a 60% increase in canopy volume over the interval between measurements. We exclude the clusters which do not satisfy the specified growth rate constraints from the set of growth clusters used to estimate the growth model parameters. We note that the excluded clusters from the linkage model correspond to changes in canopy volume due to abiotic factors or obvious errors in the linkage resulting in biologically implausible growth rates. While it is possible that a tree may experience a decline in canopy volume over time (e.g. during the mortality process), we have chosen to emphasize a method that focuses on the growth of healthy trees with the recognition that this restricts our understanding of the growth relationship conditional on the fact that a tree grew, or experienced a small enough decline in canopy volume that the change could be attributed to errors in the LiDAR scanning and post processing.

The row vector $\boldsymbol{x}_{s_c}$, of length $p+1$, contains p observed covariates at the latent location $\boldsymbol{s}_c$ for the growth cluster $\mathcal{C}^G_c$ with first element corresponding to the baseline growth rate asymptote. We note that topographic covariates (Folded Aspect, Growing Degree Days, HAS Wetness Index, and Snowpack Persistence in this application) are assumed to be centered and scaled across the entire surface of the domain of interest D prior to inclusion in the model.

For each growth cluster $\mathcal{C}^G_c \in \mathcal{C}^G(\mathbf{\Lambda})$, g_c is the observed annual growth for cluster c and is defined as a function of the first and last cumulative volume measurements for the linked record set $\mathcal{C}^G_c$ such that

$$g_c = \frac{v^*_{\max^f(c)} - v^*_{\min^f(c)}}{t\left(\max^f(c)\right) - t\left(\min^f(c)\right)}.$$

The function $t()$ returns the year associated with the file index so that the difference in observed canopy volumes is scaled by the length of the interval between measurements to place

the observed growth on an annual scale. We model the observed annual growth as a function of the true growth, which we specify as a Michaelis-Menten type function dependent upon the initial observed canopy volume and the environmental covariates associated with the record's latent location $\boldsymbol{s}_c$, while allowing for measurement error, such that

$$g_c \mid \gamma, \boldsymbol{\beta}, \tau, \boldsymbol{\Lambda}, \boldsymbol{x}_{\boldsymbol{s}_c}, \boldsymbol{v}^* \sim \text{Skewed t}\left(\mu_c, \tau, \delta, \omega\right), \text{ for } \mu_c = \frac{\left(\boldsymbol{x}_{\boldsymbol{s}_c}\boldsymbol{\beta}\right) {v^*_{\min^f(c)}}^{\alpha}}{\gamma^{\alpha} + {v^*_{\min^f(c)}}^{\alpha}}. \tag{1}$$

As mentioned above, a notable advantage of the generalized Michaelis-Menten style growth function is that the parameters of the true growth function have clear biological interpretations. The linear component of the function adjusts the maximum growth asymptote as a function of the covariates at a given location. The parameter γ controls the size at which the growth rate saturates due to size scaling, establishing the inflection point of the growth curve. The parameter α controls the curvature of the growth function, where values of $\alpha > 1$ result in a sigmoidal curve and values of $\alpha \leq 1$ result in a shape more akin to a logistic curve. While the true growth is generally assumed to be non-negative, our model allows for the observed growth to be negative as a function of measurement error.

The full growth model is defined as follows

$$\begin{aligned}
g_c \mid \gamma, \boldsymbol{\beta}, \tau, \boldsymbol{\Lambda}, \boldsymbol{x}_{\boldsymbol{s}_c}, \boldsymbol{v}^* &\overset{ind}{\sim} \text{Skewed t}\left(\mu_c, \tau, \delta, \omega\right) \\
\tau &\sim \text{Uniform}\left(0, b_\tau\right) \\
\delta &\sim \text{Normal}_{[-1,1]}\left(0, \sigma^2_\delta\right) \\
\omega &\sim \text{Gamma}\left(2, b_\omega\right) \\
\gamma &\sim \text{Uniform}\left(a_\gamma, b_\gamma\right) \\
\alpha &\sim \text{Beta}_{[c_\alpha, d_\alpha]}\left(a_\alpha, b_\alpha\right) \\
\beta_0 &\sim \text{Normal}\left(\mu_0, \sigma^2_0\right) \\
\beta_k &\overset{ind}{\sim} \text{Normal}\left(\mu_{\beta_k}, \sigma^2_{\beta_k}\right), \text{ for } k = 1, \ldots, p,
\end{aligned}$$

where the specification of the hyperparameters is informed by the advice of domain science experts, while being adequately diffuse where appropriate. We model the observed growth, annual change in canopy volume, as a nonlinear function of the initial size and a set of environmental covariates at the latent location $\boldsymbol{s}_c$ of the individual with skewed measurement error derived from the LiDAR processing algorithm described in Section 2. We place a weak Uniform prior on γ with the lower and upper bounds specified as the minimum reasonable growth saturation value as a function of size and the maximum size observed in the first file, as the parameter corresponds to the size at which the growth rate reaches its half maximum. We specify a shifted and scaled Beta prior for the shape parameter α, which controls the curvature of the growth function, where the specified bounds limit the degree of possible curvature. A non-informative version of this prior takes $a_\alpha = b_\alpha = 1$ corresponding to a Uniform distribution over the specified range. We assume the individual β_k coefficients are independent a priori and assign appropriately diffuse Normal priors with the understanding that all covariates have been centered and scaled prior to inclusion in the model. We place a Gamma prior on ω, where ω controls the kurtosis of the distribution. Finally, we specify a truncated Normal prior for the skewness parameter δ, where the truncation bounds follow the support of the parameter. In this model formulation we adopt a nonlinear regression skew-t error model, as presented by De la Cruz and Branco (2009), to account for the observed structure of our empirical data. For our specific skew-t density, we follow the formulation

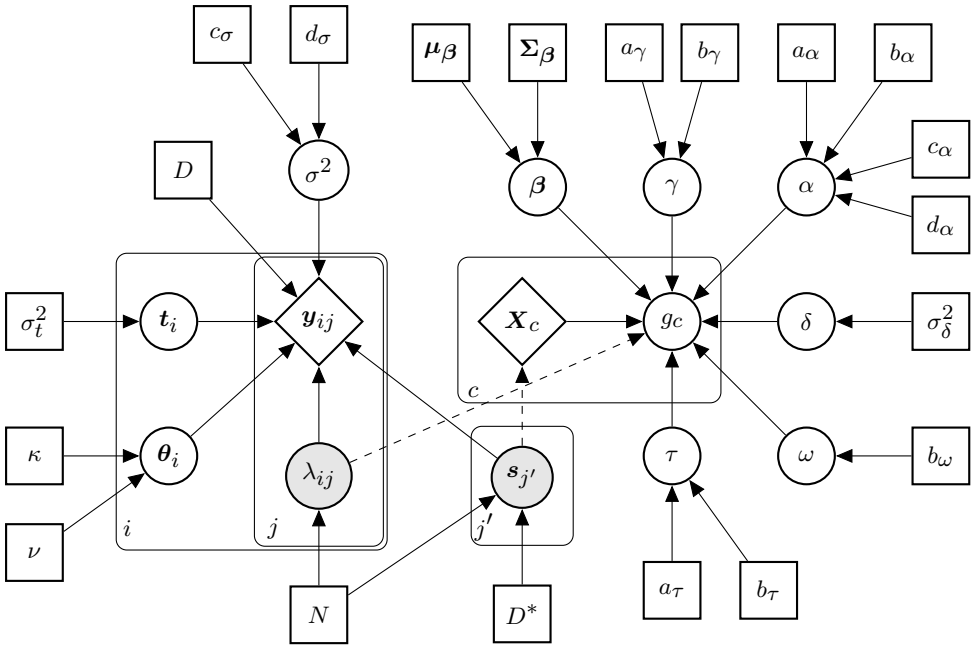

FIG 2. *Plate diagram for the two-stage record linkage and downstream growth model. Where $i = 1, 2$ denotes the file index, $j = 1, \ldots, n_i$ denotes the record index within file i, $j' = 1, \ldots, N$ denotes the latent location index, and $c = 1, \ldots, |\mathcal{C}^G(\boldsymbol{\Lambda})|$ denotes the growth cluster index. The round nodes indicate parameters, while square nodes indicate hyperparameters. We note that solid arrows denote stochastic relationships, while dashed arrows identify the inputs from the record linkage model to the downstream growth model. This framework provides the structure for the LA approach discussed in detail in Section 4.*

of Hansen (1994) and perform the appropriate location and scale adjustments such that μ_c and τ are the mean and variance of the distribution respectively when fitting the model. We do note however, that this modeling framework may be applied more generally with alternative assumed error processes dependent upon the requirements of a given application. For example, we consider a normal error process in the simulation study that we present in Section 6.

Combining the spatial record linkage and downstream growth models, as seen in the plate diagram in Figure 2, we obtain the structure for the two-stage modeling approach. We would like to emphasize the distinction between the two models and the assumptions that are made in each. The linkage model is designed to be as flexible as possible to identify the relationship between the observed records and the latent locations, while the growth model is designed to estimate the growth of the trees. The models are specified to be used together, but the assumptions made in the linkage model are not necessarily identical to those made in the growth model. This two-stage approach allows the linkage to be used for a variety of downstream modeling objectives without the need to rerun the linkage model for each task. Following the LA procedure outlined by Sadinle (2018), we propose using a random sample of iterations from the marginal posterior of the linkage structure $\boldsymbol{\Lambda}$ and the latent spatial point process $\boldsymbol{s}$, obtained from the linkage stage, as inputs for the downstream model. The LA approach effectively marginalizes out the uncertainty from the linkage and the latent locations and provides equivalent inference for the growth model parameters compared to the marginal inference that would be obtained from a joint model under certain conditions. We discuss the requirements and justification for this approach in greater depth in Section 4.

3.3. *Computational Strategies.* The two-stage Bayesian hierarchical model framework that we propose has many strengths including interpretability, the ability to incorporate relevant prior knowledge, and robust uncertainty quantification across the entire modeling pipeline. However, Bayesian record linkage modeling approaches are known to carry additional computational overhead that can be prohibitive to the use of these models in practice when dealing with large datasets (see Steorts et al. (2014) and Marchant et al. (2021) for more complete discussions). We alleviate some of the computational expense associated with the use of a Markov chain Monte Carlo Gibbs sampling algorithm for our model through a few key mechanisms discussed below.

One of the most common approaches for improving the scalability of record linkage models is to exact some form of deterministic blocking for records to reduce the number of comparisons necessary. This mechanism is commonly employed across both Bayesian and frequentist record linkage implementations as a preprocessing step that invalidates certain linkages that are deemed to be implausible (Steorts et al. (2014); Murray (2015)). While certain deterministic blocking schemes may impact the accuracy of the linkage and fail to adequately quantify the uncertainty associated with the procedure, we are able to take advantage of the spatial structure of our data and the biological limitations that invalidate certain links between observed records and $\boldsymbol{s}_{j'}$ as a function of euclidean distance. As an alternative to blocking, we implement a sampling scheme for $\boldsymbol{\Lambda}$ in our Gibbs sampling algorithm that allows us to approximate the posterior linkage structure under the assumption that the observed location for an individual must be within a maximum distance of the true latent location $\boldsymbol{s}_{j'}$ that it is associated with. In contrast to blocking schemes which invalidate links as a function of comparisons between records, our approach limits the linkage structure directly. Absent the use of a blocking or approximation scheme, the time required to sample from the true posterior distribution of the linkage structure $\boldsymbol{\Lambda}$ increases quadratically with the number of records. Instead of considering the full set $\boldsymbol{s}$ of possible latent locations for each record when sampling the latent matching structure, we consider only $\boldsymbol{s}_{j'}$ within a bounding box around the observed record. Additionally, we impose the restriction that there must be at least two candidate $\boldsymbol{s}_{j'}$ within the bounding box otherwise we increase the size of the box iteratively until this condition is met to ensure a reasonable approximation of the cluster assignment probabilities. We note that the spatial bounding approach yields samples from an approximate posterior distribution, however the $\boldsymbol{s}_{j'}$ removed from consideration have near zero probability associated with them as possible matches and their removal allows us to maintain a consistent computational cost in the sampling of each individual λ_{ij}. In Figure 3, we consider the correlation between posterior similarity scores, which measure how often records are estimated to be coreferent, for bounding boxes of varying sizes. We see that the correlations between posterior similarity scores are close to 1 across bounding box sizes, demonstrating the accuracy of the spatial bounding box approach for approximating the true posterior distribution of the linkage structure for moderately dense subsets of the empirical data.

Functionally, this scheme improves both the speed and efficiency of the record linkage model, as seen in Figure 4, allowing the model to scale to much larger domains of interest, which is a clear limitation of alternative modeling approaches. As the size of the bounding box increases, we observe a near exponential increase in the average time required per iteration of the sampler run on a dense 300 m^2 subset of our empirical dataset. Over this domain, we observe a 97.3 times speedup per iteration on average when using a bounding box of 3 m with the resulting linkage having a posterior similarity score correlation of .9984 compared to not using a bounding box. However, we note that the speed improvement for decreasing bounding box sizes is not universal as at some point we are required to expand the size of the box iteratively to meet the conditions established for guaranteeing a reasonable approximation of the cluster assignment probabilities.

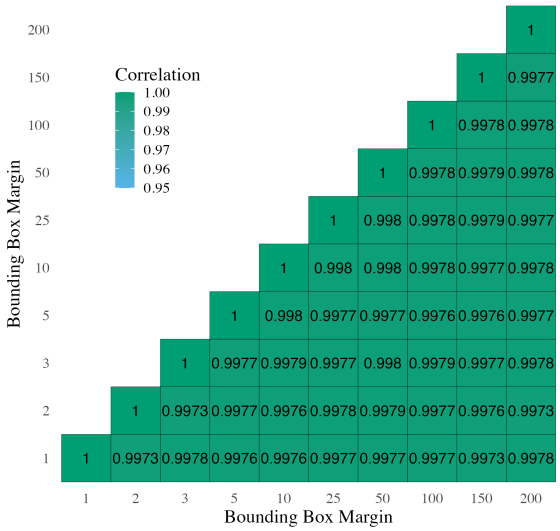

FIG 3. *Correlation heatmap for the posterior similarity cluster scores for varying bounding box margins for an area of size* 200 m^2. *The bounding box margin value specifies the distance to the boundary from the observed point. For example, a margin of* 2 m *corresponds to a* 16 m^2 *bounding box centered at the the observed location of the individual.*

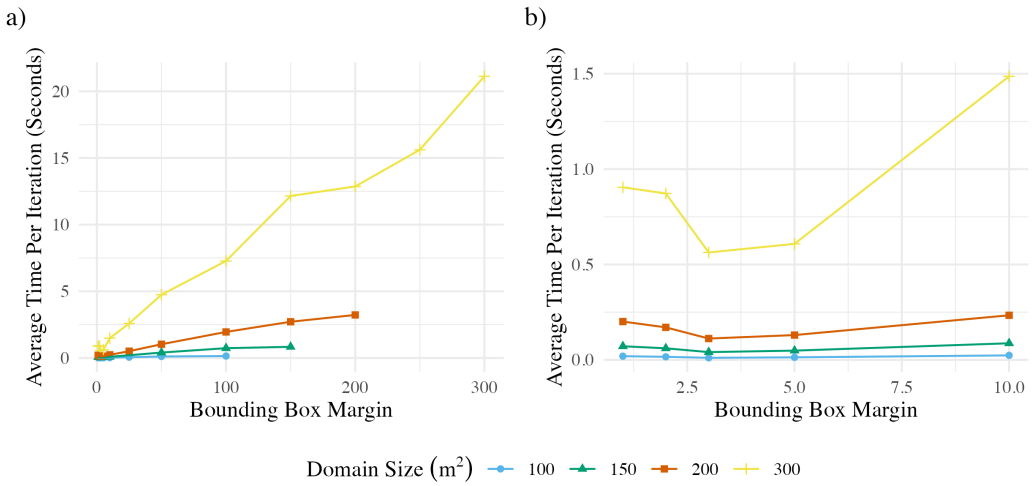

FIG 4. *Record linkage model Gibbs sampler timing results per iteration for varying bounding box margins around each sampled point. We consider the timing for areas of size* 100 m^2, 150 m^2, 200 m^2, *and* 300 m^2. *Plot a) shows the full timing results, while plot b) shows an inset for smaller bounding box margins.*

To optimize the raw computation speed, the MCMC sampler is written in R and C++ using `Rcpp` (Eddelbuettel et al. (2023a)) and `RcppArmadillo` (Eddelbuettel et al. (2023b)) to improve scalability over a base R implementation. We implement the downstream growth model in `rstan` to take advantage of the speed and flexibility of the NUTS algorithm (Stan Development Team (2023)). Additionally, the optimized parallel computation available in `rstan` reduces the time required to fit the downstream growth model with minimal additional architecture required. The details of our Gibbs sampling algorithm for the spatial

record linkage model may be found in the Supplementary Materials in Appendix A (Drew, Kaplan and Breckheimer (2025)).

4. Linkage-Averaging for Parameters from Auxiliary Data Models. We present a discussion of the theoretical justification for the LA approach, introduced by Sadinle (2018) for population size estimation, for a general downstream task with auxiliary data, i.e. regression, when paired with a record linkage model that models the observed records as noisy versions of a set of true latent field values as in Steorts, Hall and Fienberg (2016) and Liseo and Tancredi (2011b) and as defined in Section 3.1. We demonstrate that under the following two mild conditions, this LA approach may be reframed to provide proper Bayesian inference for the parameters of a more general downstream task.

CONDITION 1. Our beliefs regarding the linkage structure $\boldsymbol{\Lambda}$ and the true latent field values $\boldsymbol{s}$ are quantified by the joint posterior distribution $p_{\text{LRL}}(\boldsymbol{\Lambda}, \boldsymbol{s} \mid \boldsymbol{y})$, arising from a record linkage model employing a latent matching structure, where the posterior is proportional to the product of the likelihood $\mathcal{L}_{\text{LRL}}(\boldsymbol{\Lambda}, \boldsymbol{s} \mid \boldsymbol{y})$ and the joint prior $p(\boldsymbol{\Lambda}, \boldsymbol{s})$.

We note that this condition depends on the use of a record linkage model that employs a latent matching structure in which the fields are modeled directly, though it would be straight forward to adapt for an alternative linkage model construction.

CONDITION 2. If the true linkage structure $\boldsymbol{\Lambda}$ and latent field values $\boldsymbol{s}$ were known, the posterior $p_{\text{AD}}(\boldsymbol{\Theta} \mid \mathcal{C}(\boldsymbol{\Lambda}), \boldsymbol{X}(\boldsymbol{s}))$ arising from an auxiliary data model with likelihood $\mathcal{L}_{\text{AD}}(\boldsymbol{\Theta} \mid \mathcal{C}(\boldsymbol{\Lambda}), \boldsymbol{X}(\boldsymbol{s}))$ and joint prior $p(\boldsymbol{\Theta})$, which may be further decomposed depending on the structure of the model, would encapsulate our beliefs regarding the downstream model parameters.

The second condition describes the inferential process for $\boldsymbol{\Theta}$, the vector of parameters from the auxiliary data model, under the assumption that the true linkage structure and latent field values are known. The combination of these two conditions provide the basis for our underlying argument, such that if these conditions hold, the following relationship

$$p_{\text{LA}}(\boldsymbol{\Theta}) = \mathrm{E}_{\boldsymbol{\Lambda}, \boldsymbol{s} \mid \boldsymbol{y}}\left[p_{\text{AD}}(\boldsymbol{\Theta} \mid \mathcal{C}(\boldsymbol{\Lambda}), \boldsymbol{X}(\boldsymbol{s}))\right] = \sum_{\boldsymbol{\Lambda}} \sum_{\boldsymbol{s}} p_{\text{AD}}(\boldsymbol{\Theta} \mid \mathcal{C}(\boldsymbol{\Lambda}), \boldsymbol{X}(\boldsymbol{s})) \mathcal{L}_{\text{LRL}}(\boldsymbol{\Lambda}, \boldsymbol{s} \mid \boldsymbol{y}),$$

is obvious to consider given its clear interpretation. We also demonstrate that $p_{\text{LA}}(\boldsymbol{\Theta})$ is a proper posterior distribution. In order to perform inference on $\boldsymbol{\Theta}$ and $(\boldsymbol{\Lambda}, \boldsymbol{s})$ given $\boldsymbol{y}$, we require a joint prior for $(\boldsymbol{\Theta}, \boldsymbol{\Lambda}, \boldsymbol{s})$ such that

$$p(\boldsymbol{\Theta}, \boldsymbol{\Lambda}, \boldsymbol{s}) = p_{\text{AD}}(\boldsymbol{\Theta} \mid \mathcal{C}(\boldsymbol{\Lambda}), \boldsymbol{X}(\boldsymbol{s})) p(\boldsymbol{\Lambda}) p(\boldsymbol{s}),$$

which follows naturally from Conditions 1 and 2 above.

THEOREM 4.1 (Bayesian validity of linkage-averaged auxiliary data model parameters joint posterior). *The marginal posterior of $\boldsymbol{\Theta}$ under the likelihood $\mathcal{L}_{LRL}(\boldsymbol{\Lambda}, \boldsymbol{s} \mid \boldsymbol{y})$ of the latent record linkage model and joint prior $p_{AD}(\boldsymbol{\Theta} \mid \mathcal{C}(\boldsymbol{\Lambda}), \boldsymbol{X}(s)) p(\boldsymbol{\Lambda}) p(\boldsymbol{s})$ is $p_{LA}(\boldsymbol{\Theta})$.*

Theorem 4.1 establishes $p_{\text{LA}}(\boldsymbol{\Theta})$ as a valid posterior distribution. We provide the proof for 4.1 in Appendix B of the Supplementary Materials (Drew, Kaplan and Breckheimer (2025)), as the details are similar to the proof of Sadinle (2018). We note that the proof

of Sadinle (2018) holds specifically for population size estimation, in comparison to the result for a general downstream task with auxiliary data which we have established. In practice, we approximate the linkage-averaged posterior of $\boldsymbol{\Theta}$, $p_{\text{LA}}(\boldsymbol{\Theta})$, with a random sample $\boldsymbol{\Theta}^{(1,t)}, \ldots, \boldsymbol{\Theta}^{(l,t)} \sim p_{\text{AD}}\left(\boldsymbol{\Theta} \mid \mathcal{C}(\boldsymbol{\Lambda})^{(t)}, \boldsymbol{X}(\boldsymbol{s})^{(t)}\right)$, for each $t = 1, \ldots, k$, such that

$$p_{\text{LA}}(\boldsymbol{\Theta}) \approx \frac{1}{kl} \sum_{t=1}^{k} \sum_{u=1}^{l} I\left(\boldsymbol{\Theta} = \boldsymbol{\Theta}^{(u,t)}\right),$$

where k and l are chosen to be sufficiently large to provide a reasonable approximation to the true posterior.

The spatial record linkage and downstream growth models detailed in Sections 3.1 and 3.2 respectively, clearly satisfy the construction discussed above with $\boldsymbol{\Theta} = (\alpha, \gamma, \boldsymbol{\beta}, \tau, \delta, \omega)$ and where $\boldsymbol{X}(\boldsymbol{s})$ represents the auxiliary data component of the model. We note that an alternative to the LA approach is to model the record linkage and downstream task jointly, which allows the downstream task to inform the file linkage procedure. For example, Gutman, Afendulis and Zaslavsky (2013) discuss a joint modeling approach based on multiple imputation that iteratively samples the unknown linking partition and the downstream model parameters. In their framework, the unknown links are treated as missing data and imputed. While the joint modeling approach may potentially improve the linkage, it is often accompanied by a substantially increased computational burden and the performance is sensitive to model misspecification for the downstream model. In contrast, the LA framework that we present provides equivalent marginal inference for the downstream model parameters as that obtained from a joint model, under the assumptions of Conditions 1 and 2, and allows more flexibility for the researcher to recycle the linkage for multiple downstream tasks of interest that may be implemented in parallel in a straight forward and efficient fashion.

5. Estimation of Annual Growth Curves for Rocky Mountain Conifer Forests. In this section, we return to the empirical data and related hypotheses regarding the annual growth behavior of Southern Rocky Mountain conifer forests presented in Section 2. We employ the two-stage LA approach for estimating the downstream growth model parameters using $k = 100$ randomly sampled iterations from the joint posterior distribution of the linkage structure $\boldsymbol{\Lambda}$ and latent locations $\boldsymbol{s}$ as the input for the growth model. For each pair $\left(\boldsymbol{\Lambda}^{(k)}, \boldsymbol{s}^{(k)}\right)$, we derive the set of growth clusters, $\mathcal{C}^G\left(\boldsymbol{\Lambda}^{(k)}\right)$, and the set of location dependent covariates, $\boldsymbol{X}\left(\boldsymbol{s}^{(k)}_{\mathcal{C}^G(\boldsymbol{\Lambda}^{(k)})}\right)$, which are then used to fit the growth model defined in Section 3.2. This procedure allows us to obtain estimates of the marginal posterior distributions of the growth model parameters of interest that are equivalent to the marginals obtained from a joint model for the linkage structure, $\boldsymbol{s}_{j'}$, and the growth model parameters, as discussed in Section 4. For this analysis, we specify $r_1 = .9$ and $r_2 = 1.6$ such that we consider primarily positive growth with a maximum increase of 60% of the initial observed canopy volume over the 4 year study period. These cutoffs reflect typical growth behavior and disqualify implausible clusters arising from errors in the linkage and due to environmental mechanisms like damage from extreme wind or lightning that do not reflect the biological mechanisms of tree growth.

In addition to the topographic covariates discussed in Section 2, we also include three inter-tree competition metrics. Fagerberg et al. (2022) highlight the importance of including competition indices in individual tree growth models for conifer species. For our application, we consider relative spacing index (RSI), the ratio of the nearest neighbor distance to the average neighbor distance, larger neighbor volume (LNV), the summed canopy volumes of a tree's larger neighbors, and neighborhood density (ND), the density of individuals within the

neighborhood of the individual. All competition metrics are calculated using the observed locations from the first scan (2015) within a 15 m neighborhood around each point such that all three are considered semi-distance dependent competition indices. Ma et al. (2018) calculate LiDAR derived tree competition indices using a 15 m neighborhood, and we adopt the same neighborhood size for our analysis. To ensure that these metrics are accurate for all points considered, the downstream growth model is only fit to growth clusters located more than 15 m from the boundary of the study domain. We note that RSI and ND are measures of symmetric competition while LNV captures asymmetric competition among individuals to account for the variation possible across the range of competitive effects. An in depth discussion of competition indices and their construction may be found in Pommerening and Sánchez Meador (2018) and Contreras, Affleck and Chung (2011).

Given the size of the study domain (~2 km^2), we fix the rotation parameter for the second scan, θ_2, to be zero as even a very small degree of rotation can have a large effect on points near the boundary of the domain. We allow for the possibility of scan-wide translation in this analysis, and note that the choices of which image alignment components to include and the strength of their constituent priors will likely depend on the application. We select noninformative and weak priors, where appropriate, for the record linkage and downstream growth model parameters according to the outline in Section 3. We specify $q = 1.25$ in determining the maximum number of unique latent individuals N to provide flexibility across the range of point densities observed in the study area. The convergence of the record linkage model and downstream model variants were assessed by examining traceplots and Gelman Rubin statistics (Gelman and Rubin (1992)) for each approach. The full model specification details and select convergence diagnostics may be found in Appendix C of the Supplementary Materials (Drew, Kaplan and Breckheimer (2025)).

In concert with the two-stage LA approach, we consider two alternative heuristic strategies for linking trees across scans which we term nearest distance matching (NDM) and polygon overlap matching (POM). The NDM algorithm matches each tree crown location from the first scan with the closest point from the second scan. The POM approach uses the derived crown geometries from the LiDAR scans and traces the crown polygons from the 2015 scan forward and considers the overlap with the polygons from the 2019 scan and the change in canopy volume is calculated as the difference between the estimated volumes. While these methods do not perform deduplication and fail to provide uncertainty quantification for the linkage, they represent simple and easy to implement strategies for identifying unique individuals across scans and obtaining growth estimates that are analogous to methods being used in practice. For example, Ma et al. (2018) use a more sophisticated heuristic matching algorithm coupled with manual review of marginal matches. We apply the same growth cluster restrictions for these methods as for the LA approach, so the set of derived growths is characteristically equivalent across the three linkage procedures. During model fitting, we considered 4 candidate growth models, 3 with the non-linear Michaelis-Menten mean function, for each linkage strategy (Skewed t, Skew Normal, Normal, and Multiple Linear Regression with Normal Errors) and evaluated the model fit using the Continuous Ranked Probability Score (CRPS) as suggested by De la Cruz and Branco (2009) following the discussion of Gneiting and Raftery (2007). We used the scaled CRPS (sCRPS) presented by Bolin and Wallin (2023) which has been shown to be locally scale invariant and an improvement over the standard CRPS for model selection. The Skewed t model discussed in Section 3 was identified by the sCRPS metric as the top performing model for all 3 of the linkage schemes (additional details may be found in Appendix C of the Supplementary Materials (Drew, Kaplan and Breckheimer (2025))). Figure 5 displays the 90% credible intervals for the covariate coefficients obtained from the three linkage approaches for the Skewed t model.

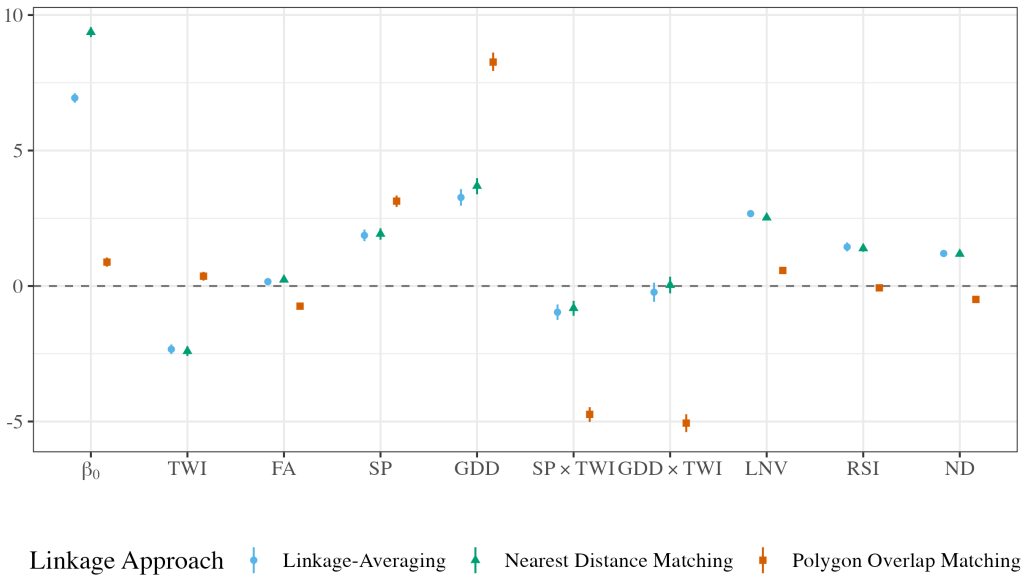

FIG 5. *Comparison of the 90% credible intervals for the growth asymptote and the topographic and competition metric covariate coefficients obtained from the downstream model fit using data derived from the three different linkage approaches (LA, NDM, and POM).*

We see from the coverage plot that the POM approach results in drastically different estimates for the coefficients in terms of magnitude, and in some instances sign, coupled with high degrees of certainty in the estimates. In contrast, the NDM and LA approaches produce more similar estimates for the coefficients given that they are both distance based linkage approaches, although the NDM is deterministic and so does not marginalize out the uncertainty from the linkage procedure when fitting the downstream task like the LA approach. We also note that the estimates for the growth asymptote β_0 are notably different across the three linkage approaches, even for models with similar estimates of the covariate coefficients. In Figure 6, we provide a comparison of the estimated size dependent growth curves arising from the LA model under high and low growth scenarios as a function of the topographic covariates Snowpack Persistence and Growing Degree Days, which are measures of water and energy availability, respectively, as discussed in Section 2. The annual growth curve, μ_c (Equation 1), is a function of size, where the growth asymptote is adjusted by the covariate values at the location of the tree. We consider the 20th and 80th quantiles of the empirical distribution for these covariates while holding all other covariates at their median values to highlight the marginal impact on growth for these individual covariates with 90% credible bands. In panel c), we examine the growth curves for both covariates simultaneously to demonstrate the combined impact of the covariates on growth behavior in both sub-optimal and optimal growth conditions.

Our analysis suggests the importance of including environmental variables related to growth conditions in addition to competition indices in modeling size-dependent individual tree growth over large spatial domains (Ford et al. (2017); Maes et al. (2019)). While the growth behavior is primarily constrained by size in our analysis, these additional metrics are influential in determining the growth behavior of forests across varied terrains and localized densities. Our work reinforces other studies on environmental constraints to conifer growth in the region which emphasize that both available energy and water from snowpack are important growth constraints (Berkelhammer et al. (2020); Carroll, Gochis and Williams (2020)).

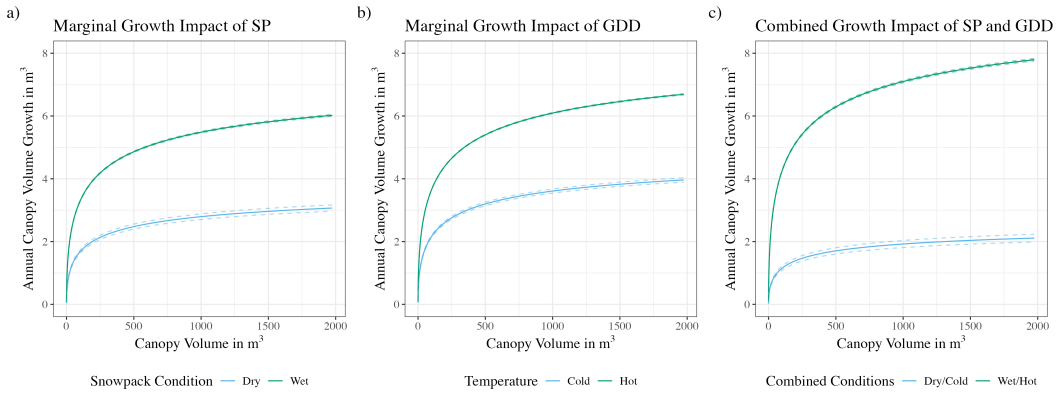

FIG 6. *Comparison of the estimated growth curves for varying quantiles of covariates of interest where plot a) is the growth curve for Snowpack Persistance, plot b) is the growth curve for Growing Degree Days, and plot c) is the growth curve for Snowpack Persistance and Growing Degree Days simultaneously. The plots demonstrate the change in growth behavior for low and high quantiles of the covariates while holding all other covariates at their median values, where the designations Dry/Cold and Wet/Hot correspond to the 20th and 80th quantiles respectively across the plots.*

Somewhat surprisingly, we observed negative effects of a soil moisture proxy (HAS Wetness Index, Figure 1 panel d) on tree growth, indicating that growth of some trees in our study domain may be limited in the wettest soils (Marks et al. (2020)). We also observe that the interaction of topographic proxies for energy and water availability may have an impact on growth when accounting for some collection of symmetric and asymmetric competition metrics. This inference for the downstream model is sensitive to the choice of linkage approach, particularly when considering the estimated growth asymptote.

6. Simulation Results. In this section, we perform a sequence of simulation studies to examine the efficacy of our modeling framework. In Section 6.1, we introduce the data generation algorithm for producing biologically realistic simulated data sets. Subsequently, in Section 6.2, we assess the performance of the LA model as applied to a collection of simulated datasets under various scenarios modeled after the empirical data discussed in Section 2.

6.1. *Data Simulation.* Historically, ecological surveys of forest growth dynamics have relied extensively on field measurement data for validating models using ALS data, as in Ma et al. (2018). These field surveys are often time consuming and expensive to perform, and provide a limited characterization of the model performance across a wide range of scenarios. Due to the scale and complexity of the study area we are considering, a validation dataset is unavailable. Instead, we gauge the efficacy of our model by considering the performance on simulated data across a variety of possible conditions as motivated by our empirical data. The majority of the existing simulation frameworks for marked point processes assume independence between the spatial point process and the mark distribution as noted by Yongtao and Afshartous (2007), however our application necessitates location dependent marks which are specified to be canopy volumes.

We address the disconnect between the available off the shelf methods and the requirements of our application through the use of a data simulation algorithm constructed to approximate the underlying marked spatial point process and the relevant biological mechanisms of forest populations, such as growth and recruitment, using three subjectively selected subsets

of the empirical data with varying point densities as a basis. We initialize the procedure by simulating the latent point process $\boldsymbol{s}$ with a modification of the modeling scheme of Møller, Ghorbani and Rubak (2016) for marked point processes, in order to include topographically derived covariates in the simulation of the mark distribution. We consider three point densities motivated by the range of densities observed in the RMBL dataset. The point densities are .04, .06, and .08 individuals per square meter, which we describe as low, medium, and high, respectively, throughout the remainder of this section. We note that the simulated data is constructed to generate data from two files corresponding to the empirical data in our application.

One of the key innovations of the Møller, Ghorbani and Rubak (2016) approach is to equate a marked spatial point process with a spatio-temporal point process by ordering the marks and treating them as arrival times in a spatio-temporal process. For each selected density, we use the observed sizes from the 2015 dataset to generate the approximate arrival times of the points and then predict the mark associated with each point as a function of time, neighborhood characteristics, and topographic covariates in an iterative fashion until we obtain a point realization matching the intensity of the empirical reference pattern. We employ an embedded gradient boosted tree model, built using the `xgboost` package (Chen et al. (2023)) in R, to predict the marks given the set of derived features. The empirical data demonstrate a notable pattern of inhibition, or regularity, at the $100\,\mathrm{m}^2$ scale, so we include provisions for interpoint interaction as a function of size in the data generation procedure to capture this behavior. The interaction function for point patterns defined by regularity, such as a Strauss process, are often specified with a hard core radius such that points in the process cannot be within a certain radius of each other (Leininger (2014)). In our process we assume a soft core interaction radius such that we allow points to violate the hard core interaction radius with low probability. All of the simulated data is generated over $130\,\mathrm{m}^2$ areas, and then restricted to the center $100\,\mathrm{m}^2$ area to account for possible edge effects in the point patterns. We use the raster images of the topographic covariates, provided by RMBL, to draw the location specific covariate values in order to simulate data that approximates the real data as closely as possible.

To accurately reflect the biological mechanism of juvenile recruitment (i.e. the seeding of offspring trees) over time, we generate the number and locations of potential recruits according to the realized parent point process. Each parent point is assigned a number of recruits based on its size and the locations of recruits are modeled as arising from a $\mathrm{t}(1)$ distribution centered at the parent point. The marks for recruits are simulated from a heavily right-skewed scaled Beta distribution bounded by the minimum size observed in the distribution of the parent points in order to capture the fact that a relatively small number of recruits survive long enough to be identified. We note the LiDAR process used to collect the empirical data has a height detection threshold of approximately 2 m, and so empirical data for the size distribution of smaller trees was unavailable. However, the mechanics of recruitment are well studied, and we incorporated relevant domain knowledge in the construction of our mechanism allowing larger individuals to seed more potential recruits using a sampling mechanism incorporating the individual's proportion of the total biomass contribution as the sampling weight. Johnson et al. (2021) provide a thorough discussion of the mechanisms involved in recruitment processes of conifer species which we leverage in our data generating process.

We proceed in generating the observed data from two time points by applying a simplified growth model to appropriate transformations of the latent point configuration $\boldsymbol{s}$. For purposes of illustration, we consider a growth model with the same mean structure as the model presented in Section 3.2. As noted before, our modeling approach may be adapted for different error processes, and so for simplicity we assume a standard Gaussian error process for this simulation study instead of the Skewed t process discussed previously. We introduce noise

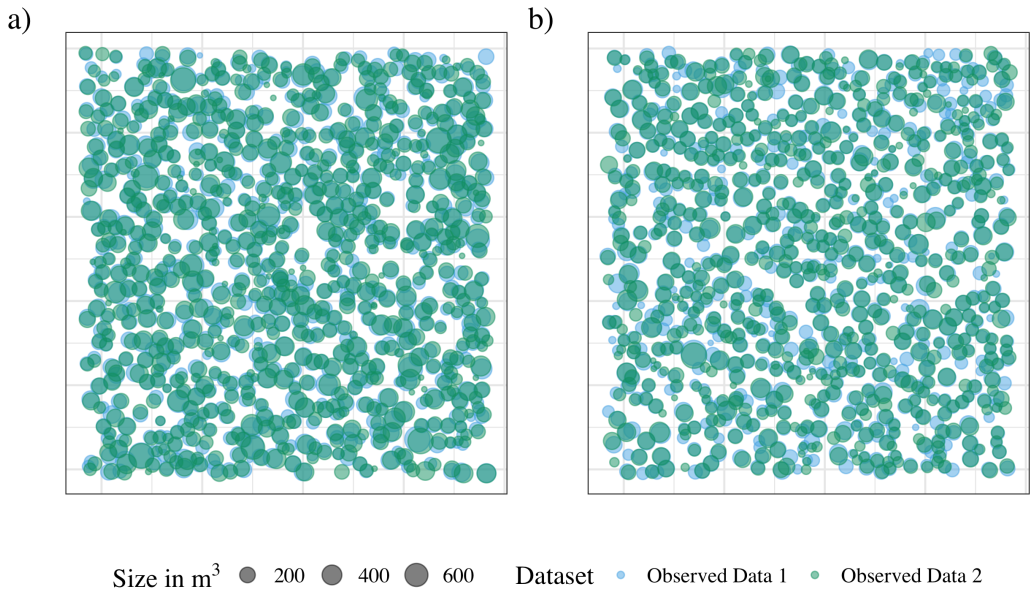

FIG 7. *Comparison of the data from a medium density* $100\,\mathrm{m}^2$ *subset where plot a) is the simulated data and plot b) is the empirical data subset used to build the predictive model for the medium density validation data.*

in the observed locations according to the data process of the spatial record linkage model in Section 3.1 such that each observed location is generated from a bivariate normal distribution centered at the latent parent point $\boldsymbol{s}_{j'}$ with measurement error σ^2. Finally, we translate and rotate the points to achieve the final spatial configurations for each file. Given the observed marks for the points from the first file, for each point we predict the growth from the first observed time to the second according to the Michaelis-Menten style mean function μ with measurement error τ^2. We note that growth is also predicted for the generated recruits, but without the measurement error component as these points are technically unobserved in the first file due to their sizes being below the detection threshold. The final set of observed data from two files is obtained by truncating the generated patterns to the center $100\,\mathrm{m}^2$ area. We apply the outlined simulation framework to produce a collection of datasets with known generating parameters and linkage structure as a baseline for assessing model performance, in the absence of a field inventory validation dataset, with varying levels of measurement error deemed to be plausible. A comparison of the simulated and empirical data for a $100\,\mathrm{m}^2$ medium density subset may be seen in Figure 7. A detailed discussion of the data generation algorithm and underlying assumptions may be found in Appendix D of the Supplementary Materials (Drew, Kaplan and Breckheimer (2025)).

6.2. *Simulation Performance.* In this section, we assess the performance of the LA two-stage modeling approach on data generated using the simulation scheme discussed in Section 6.1. We are able to gauge the efficacy of both the record linkage model and the downstream growth model given that the true linkage structure and parameter values used to generate the simulated data are known. We consider the performance of the model on 100 simulated datasets from low, medium, and high densities, which correspond to point intensities of approximately .04, .06, and .08 respectively over $100\,\mathrm{m}^2$ areas. We first consider the performance of the spatial record linkage model and then explore credible interval coverage rates for the growth model parameters across 100 simulated data sets for each density and with varying levels of noise in the observed locations.

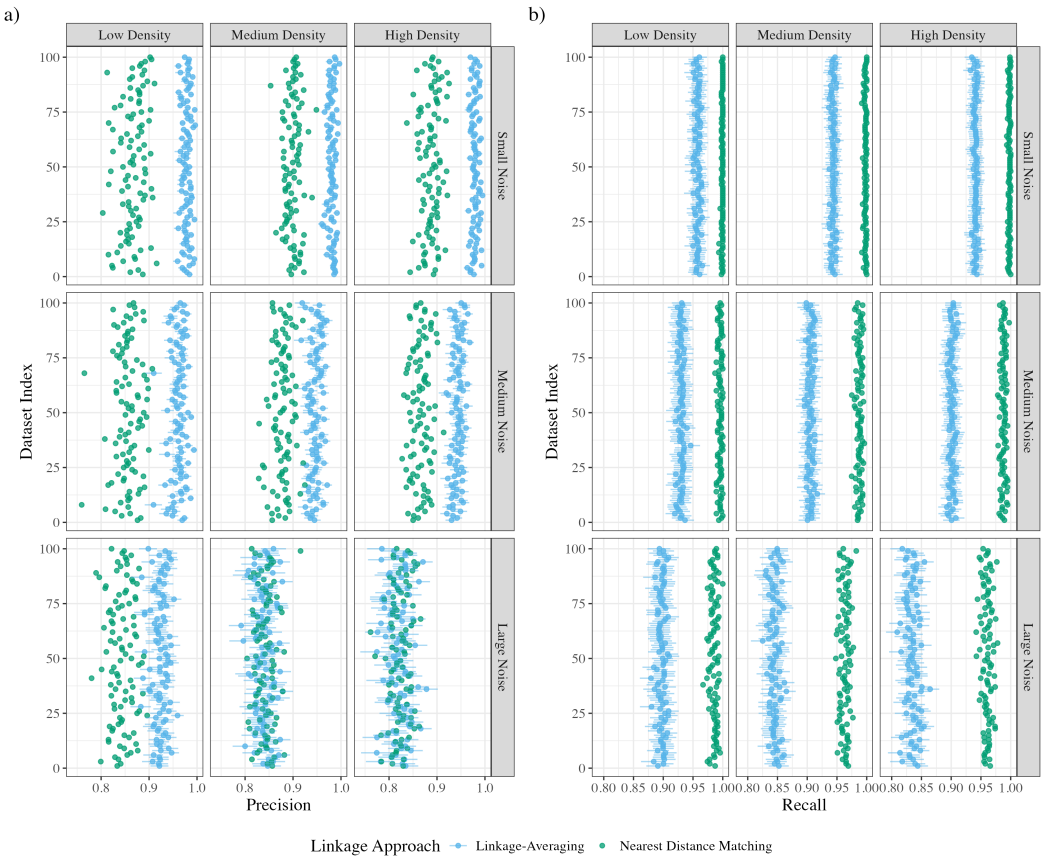

FIG 8. *Plots comparing the precision a) and recall b) performance for the LA and NDM linkage approaches on 100 simulated datasets for each density and noise combination with known true linkage and with* $\alpha = 1$.

For the linkage model, we consider the metrics precision and recall, which are standard evaluation criteria for classification tasks defined as

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \quad \& \quad \text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}},$$

where TP, FP, and FN are the number of true positives, false positives, and false negatives respectively. Precision measures the proportion of correctly identified matches out of all identified matches while recall measures the proportion of correctly identified matches out of all possible matches (Christen (2012)). We summarize the linkage performance for $\alpha = 1$ over 100 datasets with known true linkage in Figure 8. We considered three noise levels for the generating process for the observed spatial locations corresponding to $\sigma = .25$, $\sigma = .35$, and $\sigma = .45$ which we term small, medium, and large respectively. The model was run with $q = 1.25$ when determining N to match the value selected for the empirical data analysis. The results were similar for two alternative α values, which may be found in Appendix D of the Supplementary Materials (Drew, Kaplan and Breckheimer (2025)).

We note that the NDM algorithm only produces a single linkage estimate for each dataset and consequently a single estimate of precision and recall, while we obtain a full posterior distribution of the linkage for each dataset from the spatial record linkage model. We see from Figure 8 that the average precision performance for the record linkage model tends to be a notable improvement over the NDM approach across the varying levels of density and noise. We note that the increase in precision for the record linkage model comes at a cost in the

form of reduced recall compared to the NDM approach. The NDM approach matches each point from the first file with a point in the second file, and resultingly captures more matches overall but the accuracy of those matches is reduced relative to the LA approach. We note that incorrect links function much like random noise and may result in a form of attenuation bias, effectively driving the estimates of the downstream model covariate coefficients to zero. By prioritizing precision, we are able to draw more accurate conclusions about the impacts of the covariates on growth for the trees that are linked. However, this may result in additional bias when attempting to generalize the findings beyond the observed data.

We next consider the performance of the downstream growth model in terms of nominal coverage rates for 90% credible intervals for the growth model parameters of interest. We examine the same density, noise, and α settings as for the linkage model with growth parameter values $\gamma = 12$, $\boldsymbol{\beta} = \begin{bmatrix} 3 & .5 & -.5 & .5 & -.5 \end{bmatrix}^{\mathsf{T}}$, and $\tau^2 = .5$. Our prior specifications for overlapping parameters match those we use in the real data application to provide as direct an analogue between the two modeling scenarios as possible. The coverage results for the growth model parameters over the 100 datasets for each setting may be found in Table 3.

We note that the coverage rates for the true linkage model are consistently around the 90% nominal coverage rate, which serves as the gold standard for the growth model performance. Comparing the LA and NDM approaches, we see that the LA approach tends to outperform the NDM approach and often by a substantial margin. In the instances where the NDM coverage is closer to the nominal level, the coverage for the LA approach is generally more conservative due to the uncertainty propagation from the linkage stage of the modeling pipeline. These results are in line with our expectations regarding the performance of the different linkage approaches, and provide evidence that the two-stage LA framework can reliably recover the parameters of interest from a downstream model. In particular, the LA approach consistently has better coverage for the growth asymptote parameter β_0, which may explain the differences that we observed in the estimates from the empirical analysis shown in Figure 5. We note that coverage rates for τ^2 are generally lower than the nominal level, most notably in the high noise scenarios for both the LA and NDM approaches. This lower than nominal empirical coverage can be attributed to the incorrect links that are introduced by both the LA and NDM approaches as compared to the true linkage structure, which results in an overestimation of the value of τ^2. However, even at higher noise levels the empirical coverage for the covariate coefficients remain at or above the nominal level for the LA approach. This suggests that our interpretation of the impact of the covariates on growth is robust to the error in the linkage. We also note that in the presence of large amounts of noise, we may encounter identifiability issues with the growth model parameters, as evidenced by the reduced coverage rates for β_0, γ, and τ^2 in the LA and NDM models. Coverage results for the growth model parameters for datasets with $\alpha = 2$ and $\alpha = 3$ were similar and may be found in Appendix D of the Supplementary Materials (Drew, Kaplan and Breckheimer (2025)).

7. Discussion and Future Work. In this paper we have established a two-stage modeling framework built around a record linkage model for spatial location data which serves as the first step in a modeling pipeline constructed for multi-temporal location data. We demonstrated the efficiency and scalability of this approach for analyzing LiDAR derived individual tree characteristic data and provided a general schematic for using the LA approach for two-stage modeling to obtain equivalent inference for the downstream model parameters to the marginal inference obtained from a joint model for the linkage, latent spatial process, and downstream model. This framework enables researchers to investigate growth trends as a function of topographic information at a spatial scale that was previously difficult to achieve and provides flexibility in examining a variety of downstream models for different modeling

TABLE 3

Empirical coverage rates for 90% credible intervals for covariate parameters of the downstream growth model across the true linkage, LA, and NDM linkage approaches. We consider the coverage over 100 datasets for each setting with $\alpha = 1$ where the bolded coverages are closest to the nominal level (excluding the true linkage).

			Empirical Coverage by Parameter							
Density	Noise	Linkage Approach	α	β_0	β_1	β_2	β_3	β_4	γ	τ^2
Low	Small	TL	0.93	0.90	0.86	0.92	0.87	0.85	0.90	0.91
		LA	**0.91**	**0.89**	**0.90**	**0.95**	**0.94**	**0.88**	**0.90**	**0.89**
		NDM	0.70	0.38	0.75	**0.85**	0.80	0.72	0.45	0.17
	Medium	TL	0.85	0.83	0.92	0.89	0.92	0.88	0.91	0.89
		LA	**0.89**	**0.80**	**0.92**	**0.92**	**0.92**	**0.91**	**0.87**	**0.77**
		NDM	0.56	0.20	0.75	0.84	0.86	0.83	0.25	0.06
	Large	TL	0.90	0.90	0.92	0.93	0.94	0.89	0.90	0.92
		LA	**0.93**	**0.76**	**0.98**	1.00	0.95	**0.97**	**0.76**	**0.42**
		NDM	0.70	0.08	0.79	**0.85**	**0.87**	0.82	0.12	0.02
Medium	Small	TL	0.92	0.91	0.88	0.94	0.92	0.92	0.91	0.92
		LA	**0.91**	**0.89**	**0.88**	**0.95**	0.94	**0.93**	**0.92**	**0.75**
		NDM	0.43	0.36	0.64	0.83	**0.92**	0.79	0.36	0.23
	Medium	TL	0.90	0.92	0.89	0.94	0.97	0.93	0.91	0.90
		LA	**0.85**	**0.84**	**0.90**	0.98	0.98	**0.95**	**0.83**	**0.21**
		NDM	0.24	0.10	0.65	**0.91**	**0.86**	0.80	0.16	0.02
	Large	TL	0.86	0.84	0.85	0.90	0.88	0.90	0.89	0.87
		LA	**0.79**	**0.38**	**0.96**	1.00	0.99	0.98	**0.39**	0.00
		NDM	0.32	0.01	0.74	**0.94**	**0.87**	**0.85**	0.01	0.00
High	Small	TL	0.85	0.83	0.87	0.93	0.94	0.85	0.85	0.87
		LA	**0.84**	**0.79**	**0.88**	**0.94**	**0.95**	**0.88**	**0.89**	**0.67**
		NDM	0.36	0.30	0.74	0.95	0.82	0.66	0.29	0.17
	Medium	TL	0.87	0.85	0.90	0.84	0.91	0.83	0.88	0.88
		LA	**0.84**	**0.81**	**0.90**	0.97	1.00	**0.92**	**0.78**	0.06
		NDM	0.28	0.16	0.64	**0.96**	**0.89**	0.67	0.14	**0.08**
	Large	TL	0.86	0.81	0.83	0.92	0.97	0.88	0.86	0.92
		LA	**0.89**	**0.37**	**0.97**	1.00	1.00	1.00	**0.40**	0.00
		NDM	0.28	0.03	0.82	**0.99**	**0.96**	**0.83**	0.02	0.00

objectives. It would also be straight forward to extend for use in a Bayesian model averaging construction, as introduced by Raftery, Madigan and Hoeting (1997), when considering a variety of candidate models for the same downstream modeling objective in lieu of a model selection procedure, as we employed in this application. Another natural extension of our record linkage model would be to applications with more than two files, though some care would need to be taken in specifying the prior for the linkage structure to ensure the correct identification of clusters containing records from multiple files. This extension could also be

applied in the context of streaming data where the linkage structure is updated as new data is collected as discussed by Taylor, Kaplan and Betancourt (2024).

We applied this two-stage framework to investigate individual specific growth-size curves of conifer species on Snodgrass Mountain in the Southern Rocky Mountains of Colorado. We were able to quantify the impact of several key topographic covariates that serve as proxies for energy and water availability, which have been hypothesized to be limiting factors on growth for conifer species in this region. We demonstrated the effectiveness of our modeling approach in a series of numerical experiments on simulated data, in the absence of a ground truth dataset, and implemented a simulation framework for generating data arising from a multi-temporal process as a function of the data model from the linkage model and a general downstream growth model. This approach provides researchers with an alternative tool for model testing and validation in the absence of data with known linkage structure and degrees of measurement error from the various processes involved.

As an alternative to the image alignment framework of Green and Mardia (2006), one could consider an additional pre-processing step utilizing an image registration approach to transform the observed point clouds. The image registration literature is extensive, as discussed by Zitová and Flusser (2003), and includes a variety of approaches that may be utilized in the context of forestry data. For example, Ferraz et al. (2018) introduced an approach for generating a fused high density vegetation point cloud using a time series of low density LiDAR scans taken at different times from the NASA-JPL Airborne Snow Observatory. Their method similarly estimates a transformation matrix to align the point clouds, although it uses an iterative procedure and relies on the use of a collection of "tie objects" to estimate the transformation. However, in the context of our application, in addition to misalignment, we observe variability in the point clouds and derived digital surface models such that the estimated tree crowns have different shapes and sizes across scans due to the growth of the trees as seen in the inset of Figure 1 panel c). As a result of the noise in the LiDAR data and the biological processes of tree growth, we have limited "tie objects" available in our study domain with fixed shapes and locations that would be usable in an image registration procedure. Consequently, we believe that the image alignment framework of Green and Mardia (2006), which enables a fully Bayesian implementation of the record linkage model that incorporates the uncertainty inherent in the LiDAR scanning process when identifying which records correspond to the same unobserved latent locations, is a more appropriate choice.

While our modeling approach is flexible and scalable, it can be sensitive to the specification of hyperparameters, like the maximum number of unique individuals across datasets, in facilitating the linkage. There is also a clear relationship with the amount of noise in the observed spatial locations and the performance of the linkage model, so care must be taken when considering the efficacy of a record linkage approach with extremely noisy data. Our modeling approach attempts to decompose the observed distortions in the data to more accurately address the possible sources of error, but these mechanisms are somewhat dependent on the spatial scale of the data (i.e. systematic rotation in a scan). We also note that while the LA approach gives researchers a high degree of freedom, a joint modeling approach where the downstream modeling objective influences the linkage will likely lead to improved performance across the modeling pipeline if the downstream task can be well specified. In our future work, we plan to explore the viability of a joint modeling approach for similar problems which depend on multi-temporal spatial location data.

Acknowledgments. I.B. is also affiliated with the Clark Family School of Environment and Sustainability at Western Colorado University.

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

Funding. A.K. was partially supported by NSF CAREER SES-2338428.

I.B. was partially supported by the Environmental System Science program, U.S. Department of Energy, Office of Biological and Environmental Research - DOE-DE-SC0023029.

SUPPLEMENTARY MATERIAL

Web Supplement

The web supplement contains an appendix with additional model specification and implementation details, a proof of Theorem 4.1, additional details for the empirical analysis, and additional simulation study details and results.

Code Supplement

All code used in the analysis along with step-by-step instructions for recreating the analysis, simulation study performed, and all figures and tables from the main paper and appendix. The code may also be found online in the following repository. (https://github.com/lanedrew/SpRL)

Data Supplement

The empirical data used in the analysis can be obtained from the following repository. (https://data.ess-dive.lbl.gov/datasets/ess-dive-32482a38131d613-20240503T212244619)

REFERENCES

AUBRY-KIENTZ, M., DUTRIEUX, R., FERRAZ, A., SAATCHI, S., HAMRAZ, H., WILLIAMS, J., COOMES, D., PIBOULE, A. and VINCENT, G. (2019). A Comparative Assessment of the Performance of Individual Tree Crowns Delineation Algorithms from ALS Data in Tropical Forests. *Remote Sensing* **11** 1086. https://doi.org/10.3390/rs11091086

BABCOCK, C., FINLEY, A. O., COOK, B. D., WEISKITTEL, A. and WOODALL, C. W. (2016). Modeling Forest Biomass and Growth: Coupling Long-Term Inventory and LiDAR Data. *Remote Sensing of Environment* **182** 1–12. https://doi.org/10.1016/j.rse.2016.04.014

BARBOSA, R., RAMÍREZ-NARVÁEZ, P., FEARNSIDE, P., VILLACORTA, C. and CARVALHO, L. (2019). Allometric Models to Estimate Tree Height in Northern Amazonian Ecotone Forests. *Acta Amazonica* **49** 81–90. https://doi.org/10.1590/1809-4392201801642

BERKELHAMMER, M., STILL, C. J., RITTER, F., WINNICK, M., ANDERSON, L., CARROLL, R., CARBONE, M. and WILLIAMS, K. H. (2020). Persistence and Plasticity in Conifer Water-Use Strategies. *Journal of Geophysical Research: Biogeosciences* **125** e2018JG004845. https://doi.org/10.1029/2018JG004845

BETANCOURT, B., ZANELLA, G. and STEORTS, R. C. (2022). Random Partition Models for Microclustering Tasks. *Journal of the American Statistical Association* **117** 1215–1227. https://doi.org/10.1080/01621459.2020.1841647

BOLIN, D. and WALLIN, J. (2023). Local Scale Invariance and Robustness of Proper Scoring Rules. *Statistical Science* **38** 140–159. https://doi.org/10.1214/22-STS864

BOLKER, B. M. (2008). *Ecological Models and Data in R*. Princeton University Press. https://doi.org/10.2307/j.ctvcm4g37

BRAHMA, B., SILESHI, G. W., NATH, A. J. and DAS, A. K. (2017). Development and Evaluation of Robust Tree Biomass Equations for Rubber Tree (Hevea Brasiliensis) Plantations in India. *Forest Ecosystems* **4** 14. https://doi.org/10.1186/s40663-017-0101-3

BRECKHEIMER, I. (2023). Integrated Snow and Air Temperature Metrics for Ecological Applications. https://doi.org/pending

BUECHLING, A., MARTIN, P. H. and CANHAM, C. D. (2017). Climate and Competition Effects on Tree Growth in Rocky Mountain Forests. *Journal of Ecology* **105** 1636–1647. https://doi.org/10.1111/1365-2745.12782

CARROLL, R. W. H., GOCHIS, D. and WILLIAMS, K. H. (2020). Efficiency of the Summer Monsoon in Generating Streamflow Within a Snow-Dominated Headwater Basin of the Colorado River. *Geophysical Research Letters* **47** e2020GL090856. https://doi.org/10.1029/2020GL090856

CHEN, T., HE, T., BENESTY, M., KHOTILOVICH, V., TANG, Y., CHO, H., CHEN, K., MITCHELL, R., CANO, I., ZHOU, T., LI, M., XIE, J., LIN, M., GENG, Y., LI, Y. and YUAN, J. (2023). Xgboost: Extreme Gradient Boosting.

CHRISTEN, PETER. (2012). Data Matching : Concepts and Techniques for Record Linkage, Entity Resolution, and Duplicate Detection. In *Data Matching : Concepts and Techniques for Record Linkage, Entity Resolution, and Duplicate Detection. Data-Centric Systems and Applications.* Springer, Berlin.

CONTRERAS, M. A., AFFLECK, D. and CHUNG, W. (2011). Evaluating Tree Competition Indices as Predictors of Basal Area Increment in Western Montana Forests. *Forest Ecology and Management* **262** 1939–1949. https://doi.org/10.1016/j.foreco.2011.08.031

DALPONTE, M. and COOMES, D. A. (2016). Tree-Centric Mapping of Forest Carbon Density from Airborne Laser Scanning and Hyperspectral Data. *Methods in Ecology and Evolution* **7** 1236–1245. https://doi.org/10.1111/2041-210X.12575

DE LA CRUZ, R. and BRANCO, M. D. (2009). Bayesian Analysis for Nonlinear Regression Model under Skewed Errors, with Application in Growth Curves. *Biometrical Journal* **51** 588–609. https://doi.org/10.1002/bimj.200800154

DREW, L., KAPLAN, A. and BRECKHEIMER, I. (2025). Supplement to "A Bayesian Record Linkage Approach to Applications in Tree Demography Using Overlapping LiDAR Scans". https://doi.org/tobefilledin

EDDELBUETTEL, D., FRANCOIS, R., ALLAIRE, JJ., USHEY, K., KOU, Q., RUSSELL, N., UCAR, I., BATES, D. and CHAMBERS, J. (2023a). Rcpp: Seamless R and C++ Integration.

EDDELBUETTEL, D., FRANCOIS, R., BATES, D., NI, B. and SANDERSON, C. (2023b). RcppArmadillo: 'rcpp' Integration for the 'Armadillo' Templated Linear Algebra Library.

FAGERBERG, N., OLSSON, J.-O., LOHMANDER, P., ANDERSSON, M. and BERGH, J. (2022). Individual-Tree Distance-Dependent Growth Models for Uneven-Sized Norway Spruce. *Forestry: An International Journal of Forest Research* **95** 634–646. https://doi.org/10.1093/forestry/cpac017

FELLEGI, I. P. and SUNTER, A. B. (1969). A Theory for Record Linkage. *Journal of the American Statistical Association* **64** 1183–1210. https://doi.org/10.1080/01621459.1969.10501049

FERRAZ, A., SAATCHI, S., BORMANN, K. J. and PAINTER, T. H. (2018). Fusion of NASA Airborne Snow Observatory (ASO) Lidar Time Series over Mountain Forest Landscapes. *Remote Sensing* **10** 164. https://doi.org/10.3390/rs10020164

FORD, K. R., BRECKHEIMER, I. K., FRANKLIN, J. F., FREUND, J. A., KROISS, S. J., LARSON, A. J., THEOBALD, E. J. and HILLERISLAMBERS, J. (2017). Competition Alters Tree Growth Responses to Climate at Individual and Stand Scales. *Canadian Journal of Forest Research* **47** 53–62. https://doi.org/10.1139/cjfr-2016-0188

GELMAN, A. and RUBIN, D. B. (1992). Inference from Iterative Simulation Using Multiple Sequences. *Statistical Science* **7** 457–472. https://doi.org/10.1214/ss/1177011136

GNEITING, T. and RAFTERY, A. E. (2007). Strictly Proper Scoring Rules, Prediction, and Estimation. *Journal of the American Statistical Association* **102** 359–378. https://doi.org/10.1198/016214506000001437

GOULDEN, T., HASS, B., BRODIE, E., CHADWICK, K. D., FALCO, N., MAHER, K., WAINWRIGHT, H. and WILLIAMS, K. (2020). NEON AOP Survey of Upper East River CO Watersheds: LAZ Files, LiDAR Surface Elevation, Terrain Elevation, and Canopy Height Rasters.

GREEN, P. J. and MARDIA, K. V. (2006). Bayesian Alignment Using Hierarchical Models, with Applications in Protein Bioinformatics. *Biometrika* **93** 235–254. https://doi.org/10.1093/biomet/93.2.235

GUTMAN, R., AFENDULIS, C. C. and ZASLAVSKY, A. M. (2013). A Bayesian Procedure for File Linking to Analyze End-of-Life Medical Costs. *Journal of the American Statistical Association* **108** 34–47. https://doi.org/10.1080/01621459.2012.726889

HANSEN, B. E. (1994). Autoregressive Conditional Density Estimation. *International Economic Review* **35** 705. https://doi.org/10.2307/2527081

HEILMAN, K. A., DIETZE, M. C., ARIZPE, A. A., ARAGON, J., GRAY, A., SHAW, J. D., FINLEY, A. O., KLESSE, S., DEROSE, R. J. and EVANS, M. E. K. (2022). Ecological Forecasting of Tree Growth: Regional Fusion of Tree-Ring and Forest Inventory Data to Quantify Drivers and Characterize Uncertainty. *Global Change Biology* **28** 2442–2460. https://doi.org/10.1111/gcb.16038

HUO, L. and LINDBERG, E. (2020). Individual Tree Detection Using Template Matching of Multiple Rasters Derived from Multispectral Airborne Laser Scanning Data. *International Journal of Remote Sensing* **41** 9525–9544. https://doi.org/10.1080/01431161.2020.1800127

HYYPPÄ, J., HYYPPÄ, H., LECKIE, D., GOUGEON, F., YU, X. and MALTAMO, M. (2008). Review of Methods of Small-Footprint Airborne Laser Scanning for Extracting Forest Inventory Data in Boreal Forests. *International Journal of Remote Sensing* **29** 1339–1366. https://doi.org/10.1080/01431160701736489

JOHNSON, D. J., MAGEE, L., PANDIT, K., BOURDON, J., BROADBENT, E. N., GLENN, K., KADDOURA, Y., MACHADO, S., NIEVES, J., WILKINSON, B. E., ALMEYDA ZAMBRANO, A. M. and BOHLMAN, S. A. (2021). Canopy Tree Density and Species Influence Tree Regeneration Patterns and Woody Species Diversity in a Longleaf Pine Forest. *Forest Ecology and Management* **490** 119082. https://doi.org/10.1016/j.foreco.2021.119082

KAPLAN, A., BETANCOURT, B. and STEORTS, R. C. (2022). A Practical Approach to Proper Inference with Linked Data. *The American Statistician* **76** 384–393. https://doi.org/10.1080/00031305.2022.2041482

LEININGER, T. J. (2014). Bayesian Analysis of Spatial Point Patterns, PhD thesis, Duke University.

LISEO, B. and TANCREDI, A. (2011a). Some Advances on Bayesian Record Linkage and Inference for Linked Data.

LISEO, B. and TANCREDI, A. (2011b). Bayesian Estimation of Population Size via Linkage of Multivariate Normal Data Sets. *Journal of Official Statistics* **27** 491–505.

LÓPEZ, S., FRANCE, J., GERRITS, W. J. J., DHANOA, M. S., HUMPHRIES, D. J. and DIJKSTRA, J. (2000). A Generalized Michaelis-Menten Equation for the Analysis of Growth. *Journal of Animal Science* **78** 1816–1828. https://doi.org/10.2527/2000.7871816x

LU, X., HOOTEN, M. B., KAPLAN, A., WOMBLE, J. N. and BOWER, M. R. (2022). Improving Wildlife Population Inference Using Aerial Imagery and Entity Resolution. *Journal of Agricultural, Biological and Environmental Statistics*. https://doi.org/10.1007/s13253-021-00484-w

MA, Q., SU, Y., TAO, S. and GUO, Q. (2018). Quantifying Individual Tree Growth and Tree Competition Using Bi-Temporal Airborne Laser Scanning Data: A Case Study in the Sierra Nevada Mountains, California. *International Journal of Digital Earth* **11** 485–503. https://doi.org/10.1080/17538947.2017.1336578

MAES, S. L., PERRING, M. P., VANHELLEMONT, M., DEPAUW, L., VAN DEN BULCKE, J., BRŪMELIS, G., BRUNET, J., DECOCQ, G., DEN OUDEN, J., HÄRDTLE, W., HÉDL, R., HEINKEN, T., HEINRICHS, S., JAROSZEWICZ, B., KOPECKÝ, M., MÁLIŠ, F., WULF, M. and VERHEYEN, K. (2019). Environmental Drivers Interactively Affect Individual Tree Growth across Temperate European Forests. *Global Change Biology* **25** 201–217. https://doi.org/10.1111/gcb.14493

MARCHANT, N. G., KAPLAN, A., ELAZAR, D. N., RUBINSTEIN, B. I. P. and STEORTS, R. C. (2021). D-Blink: Distributed End-to-End Bayesian Entity Resolution. *Journal of Computational and Graphical Statistics* **30** 406–421. https://doi.org/10.1080/10618600.2020.1825451

MARKS, C. O., YELLEN, B. C., WOOD, S. A., MARTIN, E. H. and NISLOW, K. H. (2020). Variation in Tree Growth along Soil Formation and Microtopographic Gradients in Riparian Forests. *Wetlands* **40** 1909–1922. https://doi.org/10.1007/s13157-020-01363-9

MØLLER, J., GHORBANI, M. and RUBAK, E. (2016). Mechanistic Spatio-Temporal Point Process Models for Marked Point Processes, with a View to Forest Stand Data. *Biometrics* **72** 687–696. https://doi.org/10.1111/biom.12466

MURRAY, J. S. (2015). Probabilistic Record Linkage and Deduplication after Indexing, Blocking, and Filtering. *Journal of Privacy and Confidentiality* **7**. https://doi.org/10.29012/jpc.v7i1.643

NOBRE, A. D., CUARTAS, L. A., HODNETT, M., RENNÓ, C. D., RODRIGUES, G., SILVEIRA, A., WATERLOO, M. and SALESKA, S. (2011). Height Above the Nearest Drainage – a Hydrologically Relevant New Terrain Model. *Journal of Hydrology* **404** 13–29. https://doi.org/10.1016/j.jhydrol.2011.03.051

PADMANABHAN, S., CARTY, L., CAMERON, E., GHOSH, R. E., WILLIAMS, R. and STRONGMAN, H. (2019). Approach to Record Linkage of Primary Care Data from Clinical Practice Research Datalink to Other Health-Related Patient Data: Overview and Implications. *European Journal of Epidemiology* **34** 91–99. https://doi.org/10.1007/s10654-018-0442-4

POMMERENING, A. and SÁNCHEZ MEADOR, A. J. (2018). Tamm Review: Tree Interactions between Myth and Reality. *Forest Ecology and Management* **424** 164–176. https://doi.org/10.1016/j.foreco.2018.04.051

POORAZIMY, M., RONOUD, G., YU, X., LUOMA, V., HYYPPÄ, J., SAARINEN, N., KANKARE, V. and VASTARANTA, M. (2022). Feasibility of Bi-Temporal Airborne Laser Scanning Data in Detecting Species-Specific Individual Tree Crown Growth of Boreal Forests. *Remote Sensing* **14** 4845. https://doi.org/10.3390/rs14194845

RAFTERY, A. E., MADIGAN, D. and HOETING, J. A. (1997). Bayesian Model Averaging for Linear Regression Models. *Journal of the American Statistical Association* **92** 179–191. https://doi.org/10.1080/01621459.1997.10473615

ROUSSEL, J.-R., AUTY, D., COOPS, N. C., TOMPALSKI, P., GOODBODY, T. R. H., MEADOR, A. S., BOURDON, J.-F., DE BOISSIEU, F. and ACHIM, A. (2020). lidR: An R Package for Analysis of Airborne Laser Scanning (ALS) Data. *Remote Sensing of Environment* **251** 112061. https://doi.org/10.1016/j.rse.2020.112061

SAATCHI, S. S., HARRIS, N. L., BROWN, S., LEFSKY, M., MITCHARD, E. T. A., SALAS, W., ZUTTA, B. R., BUERMANN, W., LEWIS, S. L., HAGEN, S., PETROVA, S., WHITE, L., SILMAN, M. and MOREL, A. (2011). Benchmark Map of Forest Carbon Stocks in Tropical Regions across Three Continents. *Proceedings of the National Academy of Sciences of the United States of America* **108** 9899–9904.

SADINLE, M. (2017). Bayesian Estimation of Bipartite Matchings for Record Linkage. *Journal of the American Statistical Association* **112** 600–612. https://doi.org/10.1080/01621459.2016.1148612

SADINLE, M. (2018). Bayesian Propagation of Record Linkage Uncertainty into Population Size Estimation of Human Rights Violations. *The Annals of Applied Statistics* **12** 1013–1038. https://doi.org/10.1214/18-AOAS1178

STEORTS, R. C. (2015). Entity Resolution with Empirically Motivated Priors. *Bayesian Analysis* **10** 849–875. https://doi.org/10.1214/15-BA965SI

STEORTS, R. C., HALL, R. and FIENBERG, S. E. (2016). A Bayesian Approach to Graphical Record Linkage and Deduplication. *Journal of the American Statistical Association* **111** 1660–1672. https://doi.org/10.1080/01621459.2015.1105807

STEORTS, R. C., VENTURA, S. L., SADINLE, M. and FIENBERG, S. E. (2014). A Comparison of Blocking Methods for Record Linkage. In *Privacy in Statistical Databases* (J. DOMINGO-FERRER, ed.). *Lecture Notes in Computer Science* 253–268. Springer International Publishing, Cham. https://doi.org/10.1007/978-3-319-11257-2_20

TAYLOR, I., KAPLAN, A. and BETANCOURT, B. (2024). Fast Bayesian Record Linkage for Streaming Data Contexts. *Journal of Computational and Graphical Statistics* **33** 833–844. https://doi.org/10.1080/10618600.2023.2283571

STAN DEVELOPMENT TEAM (2023). RStan: The R Interface to Stan.

WENSEL, L., MEERSCHAERT, W. and BIGING, G. (1987). Tree Height and Diameter Growth Models for Northern California Conifers. *Hilgardia* **55** 1–20. https://doi.org/10.3733/hilg.v55n08p020

YONGTAO, G. and AFSHARTOUS, D. R. (2007). Test for Independence between Marks and Points of Marked Point Processes: A Subsampling Approach. *Environmental and Ecological Statistics* **14** 101–111. https://doi.org/10.1007/s10651-007-0010-7

ZITOVÁ, B. and FLUSSER, J. (2003). Image Registration Methods: A Survey. *Image and Vision Computing* **21** 977–1000. https://doi.org/10.1016/S0262-8856(03)00137-9