Received: 21 December 2023 | Revised: 4 July 2024 | Accepted: 26 August 2024

DOI: 10.1002/eap.3077

ARTICLE

ECOLOGICAL APPLICATIONS
ECOLOGICAL SOCIETY OF AMERICA

A probabilistic approach to estimating timber harvest location

Jakub Truszkowski[1,2] | **Roi Maor**[3] | **Raquib Bin Yousuf**[4] | **Subhodip Biswas**[4] | **Caspar Chater**[3] | **Peter Gasson**[3] | **Scot McQueen**[5] | **Marigold Norman**[6] | **Jade Saunders**[6] | **John Simeone**[7] | **Naren Ramakrishnan**[4] | **Alexandre Antonelli**[1,2,3,8] | **Victor Deklerck**[3,6,9]

[1]Department of Biological and Environmental Sciences, University of Gothenburg, Gothenburg, Sweden

[2]Gothenburg Global Biodiversity Centre, Gothenburg, Sweden

[3]Royal Botanic Gardens, Kew, Richmond, UK

[4]Department of Computer Science, Virginia Tech, Arlington, Virginia, USA

[5]Forest Stewardship Council International, Technology and Information Unit, Bonn, Germany

[6]World Forest ID, Washington, DC, USA

[7]Simeone Consulting, LLC, Littleton, New Hampshire, USA

[8]Department of Plant Sciences, University of Oxford, Oxford, UK

[9]Meise Botanic Garden, Meise, Belgium

Correspondence
Jakub Truszkowski
Email: jakub.truszkowski@worldforestid.org

Funding information
Royal Botanical Gardens, Kew; Department for Environment, Food and Rural Affairs, UK Government, Grant/Award Number: 29084; Vetenskapsrådet, Grant/Award Number: 2019-05191; Stiftelsen för Miljöstrategisk Forskning

Handling Editor: Juan C. Corley

Abstract

Determining the harvest location of timber is crucial to enforcing international regulations designed to protect natural resources and to tackle illegal logging and associated trade in forest products. Stable isotope ratio analysis (SIRA) can be used to verify claims of timber harvest location by matching levels of naturally occurring stable isotopes within wood tissue to location-specific ratios predicted from reference data ("isoscapes"). However, overly simple models for predicting isoscapes have so far limited the confidence in derived predictions of timber provenance. In addition, most use cases have limited themselves to differentiating between a small number of predetermined location options. Here, we present a new analytic pipeline for SIRA data, designed to predict the harvest location of a wood sample in a continuous, arbitrarily large area. We use Gaussian processes to robustly estimate isoscapes from reference wood samples, and overlay with species distribution data to compute, for every pixel in the study area, the probability of it being the harvest location of the examined timber. This is the first time, to our knowledge, that this approach is applied to determining timber provenance, providing probabilistic results rather than a binary outcome. Additionally, we include an active learning tool to identify locations from which additional reference data would maximize the improvement to model performance, allowing for optimisation of subsequent field efforts. We demonstrate our approach on a set of SIRA data from seven oak species in the United States as a proof of concept. Our method can determine the harvest location up to within 520 km from the true origin of the sample and outperforms the state-of-the-art approach. Incorporating species distribution data improves accuracy by up to 36%. The future sampling locations proposed by our tool decrease the variance of resultant isoscapes by up to

Alexandre Antonelli and Victor Deklerck are co-senior authors.

86% more than sampling the same number of locations at random. Accurate prediction of harvest location has the potential to transform worldwide efforts to enforce anti-deforestation legislation and protect natural resources.

KEYWORDS
Gaussian processes, illegal logging, isoscapes, origin traceability, stable isotopes, timber provenance

INTRODUCTION

Unsustainable exploitation of natural resources is the largest driver of terrestrial biodiversity loss after land-use change (Díaz et al., 2019) and a major conservation challenge globally. To avoid a sixth mass extinction (Barnosky et al., 2011), nearly 200 nations have recently agreed on a new set of targets and goals under the Kunming-Montreal Global Biodiversity Framework. In particular, Target 5 of the agreement includes the objective to "ensure that the use, harvesting and trade of wild species is sustainable, safe and legal, preventing overexploitation" (2022 UN Biodiversity Conference, 2022). Meeting this ambitious target will require overcoming a key element of unsustainable use of natural resources: the illegal harvest of threatened tree species.

Legal frameworks have been established to combat illegal logging and trade in illegally harvested timber, such as the Convention on the International Trade in Endangered Species (CITES), the US Lacey Act (amended 2008), the UK Timber Regulation (2021), the EU Deforestation Regulation (EUDR; 2023), and the Australian Illegal Logging Prohibition Act (2012). The new policies place additional traceability and reporting requirements on companies trading in wood and agricultural products (Dormontt et al., 2015). For example, the EUDR requires operators to record and report the coordinates of production location (forest or farm), and enforcement officials will be expected to scrutinize those claims of harvest location. Despite the comprehensive legislation already in place and the international commitments under current adoption, enforcement of such regulations remains a challenge. Illegally harvested timber is shipped under false declarations of origin or mixed into legal shipments, and methods for verifying geographical location have so far only been able to determine the correct location out of a few predetermined options, mostly at country-level resolution (Horacek et al., 2009; Muñoz-Redondo et al., 2021; Watkinson, Rees, Gwenael, et al., 2022). This challenge is greatly intensified by the new EUDR legislation adopting precise geographical location (GPS point or polygon for plot of land) as a determinant of the legal status of timber.

Stable isotope ratio analysis to verify provenance

Well-established scientific techniques enable measurement of the chemical, anatomical, and genetic features of plants from a tissue sample such as wood (Deklerck, 2023), with ever increasing precision and availability. When compared against a robust physical reference collection, these attributes of the wood tissue can be used to (in-)validate declared species and origin claims, and support enforcement officials in their efforts to detect, for example, illegally harvested timber or fraud in supply chains.

Stable isotope ratio analysis (SIRA) is one of the most promising technologies in this context. Several chemical elements within biological tissues (mainly hydrogen, oxygen, carbon, sulfur, nitrogen) have multiple naturally occurring stable isotopes, whose ratios vary predictably across space, in correlation with environmental conditions (Gay et al., 2022; Pederzani & Britton, 2019; Siegwolf et al., 2022; West et al., 2010). The heavy isotopes of these elements do not undergo radioactive decay, and their proportion can be readily detected by mass spectrometry (Boner et al., 2007). The isotopic composition of elements incorporated into the tissues of a plant is determined by soil properties, climate, metabolic fractionation, and other biotic and abiotic conditions characteristic of the species and the habitat in which the individual grows (Camin et al., 2017; Gay et al., 2022; Horacek et al., 2009; Siegwolf et al., 2022; van der Sleen et al., 2017). Hence, differences in stable isotope ratios among individuals correspond to the environment they grew in and can be used to discriminate between plants from different geographic areas. SIRA has proven useful in determining the risk of illegally harvested material in a wide variety of contexts, for example, forest products (Boner et al., 2007; Watkinson et al., 2020), wildlife trafficking (Bowen et al., 2005; Koehler et al., 2019; Vander Zanden et al., 2015; Wunder & Norris, 2008), ivory trade (Van der Merwe et al., 1990; Ziegler et al., 2016), agricultural products (Camin et al., 2016; Saadat et al., 2022), fish/seafood (Cusa et al., 2022; Kroetz et al., 2020; Silva et al., 2021), precious metals (Kirk et al., 2003), and

natural and synthetic illegal drugs (Casale et al., 2005; Kurashima et al., 2004), but without the spatial precision aspect required by the new timber legislation.

Modeling stable isotope ratios

Current modeling practices for the use of SIRA to verify harvest location of both legally and illegally harvested forest products require improvement. The use of SIRA is currently limited by the simplistic models used, as well as by the limited number of reference wood samples used as input data for such models. Reference sample collection campaigns are costly and budgetary needs are often underestimated so that the choice of collection locations may be based on relative ease of sampling rather than areas that yield a gain in model prediction accuracy (Schmitz et al., 2019). There has been considerable development of isoscapes ("isotope landscapes"), which are geospatial maps that show the isotopic ratio variation of the material of interest (West et al., 2010). While the potential of isoscapes for determining forest product origins has long been recognized, few rigorous methods exist to achieve this task. The existing methods use simple prediction strategies such as linear regression (Watkinson et al., 2020; Watkinson, Rees, Hofem, et al., 2022), which do not fully leverage the information contained in isotope ratio (IR) data. For example, Watkinson et al. (2020) use linear regression to estimate isoscapes based on spatial maps of IRs in precipitation, water vapor, reflected shortwave radiation, and several other atmospheric and climate variables. They then report the set of locations, in which each SIRA value is within the 95% predicted CI, as the set of plausible locations for the wood sample in hand.

Gaussian process (GP) regression is a class of flexible regression models that use the values measured at sampled locations to predict the values in surrounding areas (Li & Heap, 2008; Williams & Rasmussen, 2006). A key advantage of GP regression is that it can quantify the uncertainty of its own predictions based on the inferred spatial covariance structure of the population. The importance of quantifying the uncertainty of predictions is increasingly recognized in safety-critical (Jankowiak et al., 2020) and forensic (Chang & Srihari, 2010; Swofford & Champod, 2022) machine learning applications. Additionally, GP regression facilitates inference of a sparsely sampled variable of interest from variables that are highly correlated with it but more densely sampled (Adhikary et al., 2017; Kanankege et al., 2018). In the context of plant harvest location prediction, this translates to inferring stable isotope ratios from atmospheric drivers (such as precipitation, temperature, and water vapor pressure) known to influence the stable isotope signal in wood (Horacek et al., 2009; Siegwolf et al., 2022). This is a powerful tool for predicting the isotopic composition in areas that have not yet been sampled. However, previous work on timber isoscapes used GP regression primarily as a spatial interpolation technique without a probabilistic interpretation (Gori et al., 2018; Watkinson, Rees, Gwenael, et al., 2022). Others used approximate GP models to derive variance estimates for origin determination in animals (Ma et al., 2020; St. John Glew et al., 2019).

Here, we develop GP-based probabilistic models to predict timber harvest location by inferring timber isoscapes directly from SIRA data, with the aid of atmospheric predictors and species distribution data. We show that probabilistic modeling greatly enhances the utility of SIRA in predicting the harvest location of timber, and, based on a reference data set, can guide future sample collection by identifying locations from which data will contribute the most to minimizing prediction uncertainty.

MATERIALS AND METHODS

Study area

We used a rectangular grid to represent the study area, the contiguous United States, delimited by latitudes 25° and 50° north, and longitudes 126° and 66° west. Grid points were placed every 0.125° latitude ($\approx$14 km) and every 0.06° longitude ($\approx$4.3–6.0 km), which allowed us to approximate possible harvest locations with high precision.

Data sets

We used data from 87 trees of the genus *Quercus* located across the contiguous United States, as described in Watkinson et al. (2020)—see Figure 1 for an overview of sampling locations. Stable isotope ratio measurements were done following the protocol described in Boner et al. (2007). Each entry contained stable isotope ratio measurements of oxygen $\delta^{18}O$ (ratio between ^{18}O and ^{16}O), hydrogen $\delta^{2}H$ (ratio between ^{2}H and ^{1}H), carbon $\delta^{13}C$ (ratio between ^{13}C and ^{12}C), and sulfur $\delta^{34}S$ (ratio between ^{34}S and ^{32}S) as well as the GPS coordinates of the sampled tree. As stable isotope ratios are largely driven by environmental conditions such as precipitation, temperature, humidity, and so on, publicly available data sets for these factors can be used to improve the inference of isoscapes. We used the following atmospheric data: $\delta^{2}H$ and $\delta^{18}O$ isotopic composition of precipitation (Bowen & Revenaugh, 2003), water vapor (Borbas

State	Training samples	Test samples
Kentucky	24	23
Maryland	9	9
Wisconsin	9	9
Pennsylvania	8	8
Maine	4	4
Tennessee	4	4
Vermont	3	3
Washington	3	2
Michigan	3	2
Mississippi	3	2
New Hampshire	2	2
Virginia	2	2
West Virginia	2	2
New York	3	1
Ohio	2	1
California	6	0
Total	87	74

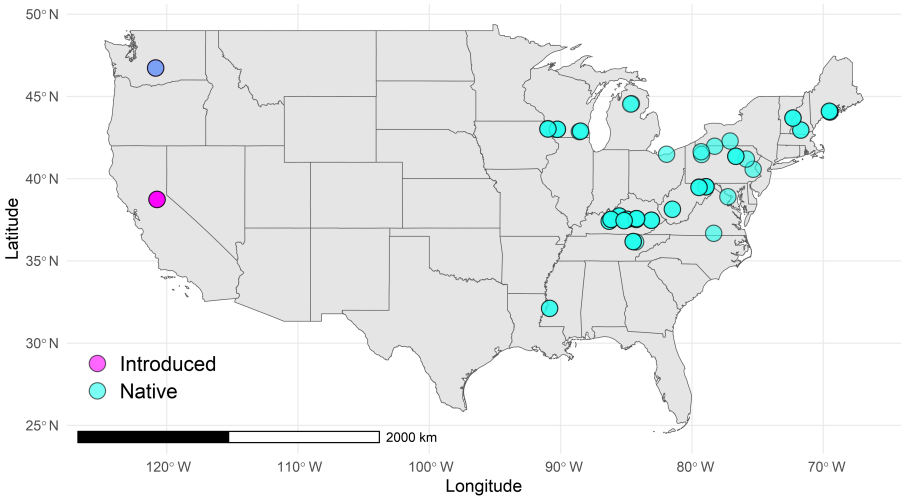

FIGURE 1 Distribution of sampling locations of oak trees included in this study. The table shows how many oak samples from each state were used as test data. The map shows the harvest location of each sample. The blue dot in the northwest corner of the map indicates a location where both introduced and native samples have been collected.

et al., 2015) (found to be associated with $\delta^{13}C$ by Watkinson et al., 2020), reflected shortwave radiation (NEO, 2023) and precipitation (multi-satellite) (Huffman et al., 2020), both of which were found to be associated with $\delta^{34}S$ (Watkinson et al., 2020). For each of those data types, we used monthly means averaged over 12–40 years to minimize the impact of weather patterns in specific years (see Watkinson et al., 2020 for precise year ranges).

To inform the models about the range of possible tree harvest locations, we used species inventory data across the natural range of each species within the United States (Wilson et al., 2013). The data are available as species-specific raster layers of tree abundance at 250 m resolution. We then applied bilinear aggregation, implemented in the function project() of the R package terra (Hijmans, 2022) to bring the abundance data to the same resolution as other spatial data in the pipeline.

Model architecture

Our method consists of four stages. First, we use the training data to fit a GP regression model for every IR. Second, we use the GP models to compute the mean and variance of each IR at every point in the study area. Third, for each sample in the test set, we compute the probability density of observing the IRs at every point in the study area based on the means and variances computed in the previous step. Finally, we use Bayes' theorem to compute the posterior probability distribution of possible harvest locations. See Figure 2 for an overview of the data sets and components comprising our model and output. We provide more details of each step in the following paragraphs.

Gaussian processes

We fit a GP regression model for each IR in a set of training data, to obtain the mean and variance of that IR at every point of the grid (see Appendix S1 for the full detail on implementation). This model is conceptually equivalent to a mixed-effects model with a covariance structure to account for spatially correlated random effects (Bose et al., 2018; Littell et al., 2006). The GP regression model assumes that the responses—that is, IRs—at different locations are jointly normally distributed (Gaussian). This model is defined by three elements: (1) the mean, for which we use a constant; (2) the covariance function, which consists of a Matern term (Williams & Rasmussen, 2006) and a linear term to model spatial and atmospheric effects respectively (see *Covariance function*); and (3) the noise parameter. The choice of mean and covariance functions reflects domain knowledge and modeling assumptions about the regression problem. The covariance function expresses the amount of information that observed values reveal about nearby locations. The function parameters as well as the noise parameter were estimated by maximizing the so-called *marginal likelihood* of the model (Williams & Rasmussen, 2006), in contrast to standard kriging approaches in geostatistics literature, which use least squares estimation (Kitanidis, 1997) or approximate techniques based on summary statistics

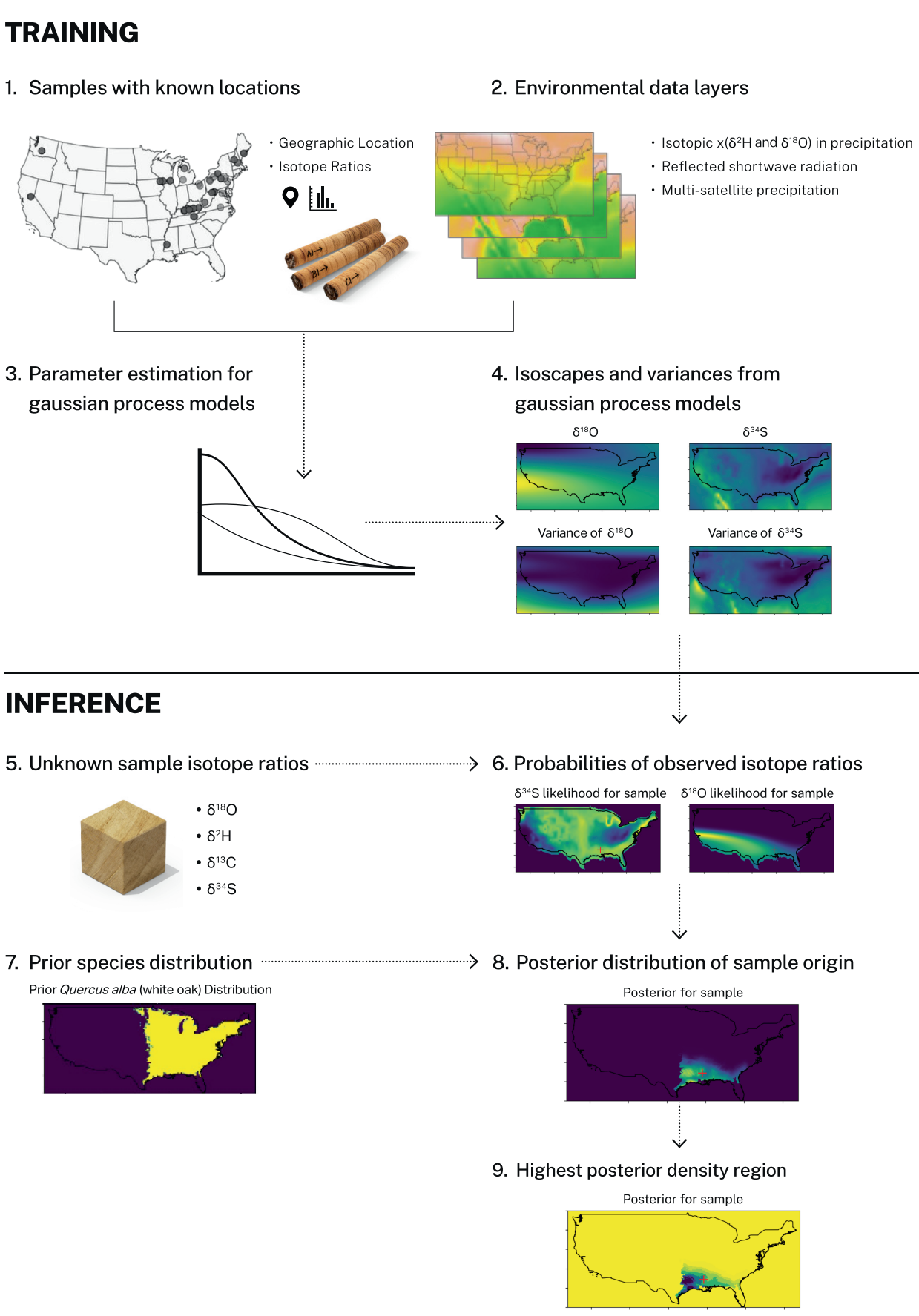

FIGURE 2 Legend on next page.

(Oliver & Webster, 2015). We used GPyTorch (Gardner et al., 2018), a flexible python package for GPs, to efficiently find the maximum likelihood parameter estimates.

Isoscape computation

After parameter estimation, the GP regression model can be used to predict the value of response variables at unsampled locations. Since the responses at training and test points are assumed to be jointly Gaussian, the conditional distribution of the jth IR y_j at test point $\mathbf{x}^*$ given the training data $(\mathbf{X},\mathbf{Y})$ is Gaussian (Bilodeau & Brenner, 1999; Valliant et al., 2000) with mean

$$\mathbb{E}\left[y_j|\mathbf{x}^*,\mathbf{X},\mathbf{Y}\right] = \mu_j + \mathbf{k}^{(j)}\left(\mathbf{K}^{(j)} + \sigma_j^2\mathbf{I}\right)^{-1}\left(\mathbf{y}_{.j} - \mu_j\right). \quad (1)$$

Here, $\mathbf{X}$ is the $n\times 2$ matrix containing the coordinates for the n training points and $\mathbf{Y}$ is the $n\times m$ matrix containing the IR values for each of the m isotopes at each training point. $\mathbf{y}_{.j}$ is the column vector containing IR values of the jth IR. $\mathbf{k}^{(j)} = \left[k^{(j)}(\mathbf{x}^*,\mathbf{x}_{1,.}),k^{(j)}(\mathbf{x}^*,\mathbf{x}_{2,.}),...,k^{(j)}(\mathbf{x}^*,\mathbf{x}_{n,.})\right]$ is the vector of covariances between values of isotope j at $\mathbf{x}^*$ and training data locations and $\mathbf{K}^{(j)}$ is the matrix of covariances between the training data points with $\mathbf{K}^{(j)}_{a,b} = k^{(j)}(\mathbf{x}_{a,.},\mathbf{x}_{b,.})$. The mean response μ_j, the intrinsic noise σ_j^2, and the parameters of the covariance function (see *Covariance function*) are estimated separately for each IR. The variance of y_j at $\mathbf{x}^*$ is given by

$$\mathbb{V}\left(y_j|\mathbf{x}^*,\mathbf{X}\right) = k^{(j)}(\mathbf{x}^*,\mathbf{x}^*) + \sigma_j^2 - \mathbf{k}^{(j)}\left(\mathbf{K}^{(j)} + \sigma_j^2\mathbf{I}\right)^{-1}\mathbf{k}^{(j)\mathrm{T}}. \quad (2)$$

See Williams and Rasmussen (2006) for a derivation.

Calculating the probability of observed IRs

A GP regression model predicts the stable isotope ratio based on the coordinates and/or the atmospheric variable values at each grid point. For a specific response value y, its probability of being observed at $\mathbf{x}^*$ is just the Gaussian probability density with mean and variance found by applying Equations (1 and 2)

$$p\left(y_j|\mathbf{x}^*,\mathbf{X},\mathbf{Y}\right) = \frac{1}{\sqrt{2\pi\mathbb{V}\left(y_j|\mathbf{x}^*,\mathbf{X}\right)}} \times \exp\left(\frac{-\left(y_j - \mathbb{E}\left[y_j|\mathbf{x}^*,\mathbf{X},\mathbf{Y}\right]\right)^2}{2\mathbb{V}\left(y_j|\mathbf{x}^*,\mathbf{X}\right)}\right). \quad (3)$$

Bayesian inference of harvest location

For a vector $\mathbf{y}^* = [y_1,...,y_m]$ of observed IR values (meaning $\delta^{18}O$, δ^2H, $\delta^{13}C$, $\delta^{34}S$), the Bayes' theorem gives the posterior distribution of possible harvest locations:

$$p(\mathbf{x}^*|\mathbf{y}^*,\mathbf{X},\mathbf{Y}) = \frac{p(\mathbf{x}^*)\prod_{j=1}^{m} p\left(y_j|\mathbf{x}^*,\mathbf{X},\mathbf{Y}\right)}{\int_{\mathbf{x}\in\mathcal{A}} p(\mathbf{x})\prod_{j=1}^{m} p\left(y_j|\mathbf{x},\mathbf{X},\mathbf{Y}\right)d\mathbf{x}}, \quad (4)$$

where the probability densities $p\left(y_j|\mathbf{x}^*,\mathbf{X},\mathbf{Y}\right)$ are computed from the GP models for the respective isotopes using Equation (3), and $\mathcal{A}$ is the study area. The integral in the denominator is accurately approximated by summing the probabilities over the spatial grid. For ease of interpretation, the output is a map indicating regions of highest posterior density (HPD) for several probability levels (15%, 30%, 50%, 75%, 90%, 95%).

Covariance function

The covariance function is a sum of the spatial autocorrelation term, which acts on the GPS coordinates, and the atmospheric term, which acts on the atmospheric variable values at those coordinates:

$$k(\mathbf{x_1},\mathbf{x_2}) = k_{\text{spatial}}(\mathbf{x_1},\mathbf{x_2}) + k_{\text{atm}}(\mathbf{x_1},\mathbf{x_2}).$$

For the spatial term, we used the Matern function with shape parameter $\nu = 1.5$ and separate scaling

FIGURE 2 Model workflow. We use a training set of isotope ratios from trees collected at known locations and atmospheric data layers ("Samples with known locations"). We fit a Gaussian process regression model to infer isoscapes and associated variance estimates. To predict the source of material with uncertain provenance ("Samples from unknown origin"), we compute the probability of observing the isotope ratio values for each element across the study area. These probabilities are then weighted with prior information on the geographical distribution of the species, to yield a posterior probability distribution of harvest location for the sample. We visualize predicted probability maps by plotting highest posterior density regions for a range of probability levels (15%, 30%, 50%, 75%, 90%, and 95%, dark blue to light green). Image credit (all panels): World Forest ID.

parameters for latitude and longitude, which takes the form (Abramowitz & Stegun, 1972):

$$k_{\text{spatial}}(\mathbf{x_1},\mathbf{x_2}) = A\left(1+\sqrt{3}d\right)\exp\left(-\sqrt{3}d\right),$$

where $d = \sqrt{(\mathbf{x_1}-\mathbf{x_2})\mathbf{D}^{-1}(\mathbf{x_1}-\mathbf{x_2})^{\top}}$ is the Euclidean distance between $\mathbf{x_1}$ and $\mathbf{x_2}$ with dimensions scaled by diagonal matrix $\mathbf{D}^{-1}$. A and $\mathbf{D}$ are the parameters of the spatial covariance function.

For the atmospheric term, we used monthly averages of the atmospheric variables listed in *Data sets*. We used a linear covariance function to model the covariance component corresponding to the variation in each atmospheric variable q:

$$\theta_q\left[\mathbf{u}_q(\mathbf{x_1})\right]^{\top}\mathbf{u}_q(\mathbf{x_2}), \tag{5}$$

where $\mathbf{u}_q(\mathbf{x})$ is the 12-entry vector of monthly values of atmospheric variable q at location $\mathbf{x}$ and θ_q is a parameter to be estimated during training. The linear covariance function models a linear relationship between the atmospheric variable and the response and is mathematically equivalent to Bayesian linear regression with a Gaussian prior on the regression coefficients (Williams & Rasmussen, 2006). The atmospheric covariance term is the sum of the linear terms corresponding to each atmospheric variable:

$$k_{\text{atm}}(\mathbf{x_1},\mathbf{x_2}) = \sum_{q\in V}\theta_q\left[\mathbf{u}_q(\mathbf{x_1})\right]^{\top}\mathbf{u}_q(\mathbf{x_2}), \tag{6}$$

where V is the set of atmospheric variables impacting the considered IR.

Prior distributions of possible tree locations

The accuracy of a harvest location prediction method depends on the choice of prior distribution of possible locations of sampled trees. The simplest choice is the *flat* prior, which assigns equal probability to all grid cells. We used the spatial density maps developed by Wilson et al. (2013) to design two alternative priors that account for the spatial distribution of oak species. The first, which we call the *density prior*, holds that the probability of a sample originating at a grid cell is proportional to the basal area (average area of tree stems per unit of space) at the grid cell. The second, which we call the *range prior*, assigns equal probability to every grid cell where basal area is above zero. In addition, both priors allow a small probability for a sampled tree to occur outside of its spatial range—we set that probability to 0.01 and diffused it uniformly over all terrestrial grid cells where the species does not occur according to Wilson et al. (2013).

Performance evaluation

We performed 5-fold cross-validation on the data set and report the average values of all performance metrics over all test points. Samples with incomplete or ambiguous species information and samples collected from trees growing in botanical gardens outside of their species' native range were excluded from the test sets as they are not representative of realistic harvest location testing scenarios. This resulted in a total of 74 test points across the 5 folds. We chose not to exclude any samples from the training sets to maximize the amount of information available to the model. We reported performance of our models as well as our implementation of the approach by Watkinson et al. (2020) averaged across the five cross-validation folds.

Rigorously evaluating the performance of our models is a nontrivial task as each model produces a probability distribution over all possible locations, rather than a single predicted location. For this reason, we have defined several metrics to investigate different aspects of probabilistic harvest location prediction:

1. Predictive log-probability and log-posterior-probability: We report the probability density of observing the IR values at the true location of the test sample, as well as the posterior probability assigned by the model to the grid cell corresponding to the true harvest location. These measure how well the model fits the test data, and are reported on a logarithmic scale.
2. Mode distance: We report the great circle distance between the true location and the *mode* of the posterior distribution, that is, the most probable location according to the model. This metric measures the accuracy of the highest probability locations, but it does not account for the amount of uncertainty in model predictions.
3. Mean absolute error (MAE): To account for model uncertainty, we report the distance between the true location $\mathbf{x}_t$ and a location chosen randomly from the posterior distribution returned by our model:

$$\text{MAE} = \int_{\mathbf{x}\in\mathcal{A}} d(\mathbf{x}_t,\mathbf{x})p(\mathbf{x}|\mathbf{y}^*,\mathbf{X},\mathbf{Y})d\mathbf{x},$$

where $d()$ is the great circle distance between the two points and $p(\mathbf{x}|\mathbf{y}^*,\mathbf{X},\mathbf{Y})$ is the posterior probability density of $\mathbf{x}$ being the harvest location. A perfect prediction would have the distance of 0. This metric favors predictions concentrated around the true location over equally

dispersed predictions concentrated elsewhere. It also favors less dispersed predictions generally. For the method of Watkinson et al., which only outputs a region of plausible locations, we assumed a uniform distribution within the region highlighted by the model. In practice, isoscapes often predict similar IR values at distant locations, so even a statistically efficient method might yield a high MAE value.

4. Area scored higher than the true location (ASH): The behavior of MAE is influenced by the shape of the posterior distribution, which favors unimodal over multimodal shapes. To further examine the specificity of our models, we report the total surface area of grid cells that the model considers more plausible than the true harvest location of the sample.

$$\mathrm{ASH} = \int_{\mathbf{x} \in \mathcal{A}} I[\mathrm{score}(\mathbf{x}|\mathbf{y}^*, \mathbf{X}, \mathbf{Y}) > \mathrm{score}(\mathbf{x}_t|\mathbf{y}^*, \mathbf{X}, \mathbf{Y})]d\mathbf{x},$$

where $I(.)$ is the indicator function that yields 1 when the statement is true and 0 otherwise. For all GP models, the score is the posterior probability of harvest location, whereas for the method of Watkinson et al. we take the score to be the negative of the minimum value of the threshold that results in the location being included in the highlighted region. In contrast to MAE, this metric is likely to give a low value to a posterior distribution that is concentrated in several small areas as long as one of those areas contains the true location. For example, if the true location could be a county in New York or a county in West Virginia, this would give a low ASH but high MAE as the two counties are far apart.

Choosing sampling locations with active learning

Field sample collections are time-consuming and expensive. We can optimize future field collections by searching for locations from which additional samples are most beneficial for increasing prediction accuracy. The isoscape variance estimates provided by GPs can be used to guide future sampling efforts, which in turn will maximize the performance of the model. This paradigm is known as *active learning* in the machine learning literature. Here, we propose a strategy to minimize the error of our isoscape estimates by carefully choosing future sampling locations.

Early attempts at efficient active learning in GPs involved collecting samples at points with highest response variance or, equivalently, picking a set of points that maximizes the entropy of responses (Cressie, 2015). Unfortunately, this approach tends to recommend collecting samples on the boundaries of the study area, which is inefficient as the newly collected samples improve isoscapes in a smaller fraction of the study area than if they were placed away from the boundary. This motivated researchers to propose several criteria for optimizing sampling (Guestrin et al., 2005; Ramakrishnan et al., 2005). Here, we adopt an approach similar to that of Guestrin et al. (2005) with a few modifications designed to address the large size of our spatial grid, which renders their original method computationally intractable for our data set.

We seek to maximize the *average* reduction in predictive variance across our study area that can be achieved by adding a sample to the training set. With S the set of sampled tree locations and G the set of grid points, we define the information gain (IG) associated with adding a new point ($\mathbf{x}^*$) to the training data set as follows:

$$\mathrm{IG}(\mathbf{x}^*) = \sum_{j=1}^{m} \sum_{\mathbf{x} \in G} \left[\log\left(\mathbb{V}\left(y_j|\mathbf{x}, \mathbf{X}\right)\right) - \log\left(\mathbb{V}\left(y_j|\mathbf{x}, \mathbf{X}, \mathbf{x}^*\right)\right)\right], \quad (7)$$

where the predictive variances are computed using Equation (2) and the outer sum is over all IRs. The algorithm then picks the point in the grid that yields the highest IG. Importantly, the predictive variances depend only on the sampled locations, not on any chemical values measured from the sample, so it is possible to propose multiple locations for future sampling before the data are collected. Our method sequentially proposes additional sampling locations until a user-specified number is reached. We assume that samples can only be collected in locations that fall within the range of at least one species. Thus, grid points that lie outside of every species range are excluded from the procedure. To reduce computation time, we randomly downsampled our grid to 15,000 points before running the analysis. In addition, we assumed that the reduction in variance is negligible for grid points situated more than 15° away from each newly proposed sampling location ($\mathbf{x}^*$) in longitude or more than 7.5° in latitude.

RESULTS

Model accuracy and comparison

The plausible location areas identified by our models consisted of points within an average distance of 520–870 km from the true location of the oak tree samples, depending on model settings. Even with a relatively small training data set of 69–70 samples (depending on the cross-validation fold), our model was able to exclude the vast majority of the study area from consideration as a possible harvest location of each sample. All our models

outperformed the state-of-the-art method for determining timber harvest location (Watkinson et al., 2020) in most or all metrics. Table 1 shows performance metrics for all the models on the test data set. Our isoscapes explained 46%–76% of the variance in $\delta^{18}O$, $\delta^{2}H$, and $\delta^{34}S$ seen in test data, but only 11%–17% of the variance in $\delta^{13}C$—see Appendix S1: Table S1. The weak link between harvest location and $\delta^{13}C$ was consistent with previous studies (Horacek et al., 2009).

Incorporating species distribution information improved prediction performance for every model and every metric examined except the predictive log-probability, which is computed independently of the prior. Informative priors improved MAE by 16%–35% and ASH by 15%–57%, with most improvement for the pure spatial model and least for the spatial + atmospheric model. The more informative *density prior* gave better accuracy than the *range prior* according to all metrics. Posterior probability maps for a few test points are shown in Figure 3 (range prior) and Figure 4 (density prior).

The spatial-only GP model gave the closest location predictions to the true location of the tree samples, except when a flat prior was used. In general, the spatial-only model and the combined spatial + atmospheric model gave similar results on all metrics and outperformed the atmospheric-only model in almost all settings. Somewhat surprisingly, the combined model did not outperform the spatial-only model. This might be due to the relatively small data set used here or the choice of atmospheric predictors, and remains to be tested as we continue to expand our reference database. The predictions of atmospheric GP models appeared qualitatively different from those from the purely spatial GP, perhaps because atmospheric model predictions emphasize geographical areas with distinct climate patterns, such as Appalachia or the Gulf Coast. Unsurprisingly, the purely spatial GP identified areas that were more spatially cohesive but did not share any obvious physical features.

Active learning reduces isoscape uncertainty

We investigated the performance of our active learning strategy on the US oak data set. For the spatial-only model, we let our method propose 10 new sampling locations to add to the training data set in the first cross-validation fold and computed the predictive variances before and after including the proposed locations.

The resulting isoscape SD maps are shown in Figure 5. Our active learning strategy proposed sampling locations in currently undersampled regions with high predictive variance. Adding samples in those areas results in a visible improvement. The highest decrease in predictive variance was observed for $\delta^{2}H$ while the lowest decrease was observed for $\delta^{18}C$. Most of the chosen locations were close to, but not at the boundary of, the allowed sampling area.

To investigate the efficiency of our active learning procedure, we compared the variances of isoscapes resulting from adding samples proposed by the active learning procedure with those resulting from adding the same number of samples from randomly selected

TABLE 1 Mean test set performance for all the models used in the study.

Model	Prior	Log prob.	Mode distance (km)	MAE (km)	Log-posterior	ASH (km^2)
Spatial-only	Flat	**−6.964**	**433**	809	−9.582	470,000
Spatial-only	Range	**−6.964**	**435**	600	−9.537	327,000
Spatial-only	Density	**−6.964**	**400**	**520**	**−9.059**	**203,000**
Atmospheric-only	Flat	−7.362	531	870	−9.972	576,000
Atmospheric-only	Range	−7.362	505	606	−9.797	450,000
Atmospheric-only	Density	−7.362	534	567	−9.428	311,000
Atmospheric + spatial	Flat	**−7.149**	**408**	794	−9.518	382,000
Atmospheric + spatial	Range	**−7.149**	**399**	627	−9.431	315,000
Atmospheric + spatial	Density	**−7.149**	**463**	536	**−8.978**	**213,000**
Watkinson et al.	NA	NA	886	859	NA	691,000

Note: Best values across all models are shown in bold and underlined whereas values that are not significantly different from the best values (Wilcoxon signed-rank test, $p = 0.05$) are shown in bold. The spatial-only GP combined with the density prior gives the highest predictive log-probability and log-posterior-probability and the lowest MAE and ASH values for all priors used. The spatial-only model outperforms the other models when range or density priors are used, while the atmospheric + spatial model performs best in terms of MAE and ASH when flat priors are used. The inclusion of species distribution information decreases MAE and ASH values for all models used. All of our models outperform the earlier method of Watkinson et al. (2020) on most or all metrics. Bold indicates values that are not significantly different from the best values (Wilcoxon signed-rank test, $p = 0.05$).
Abbreviations: ASH, area scored higher; MAE, mean absolute error.

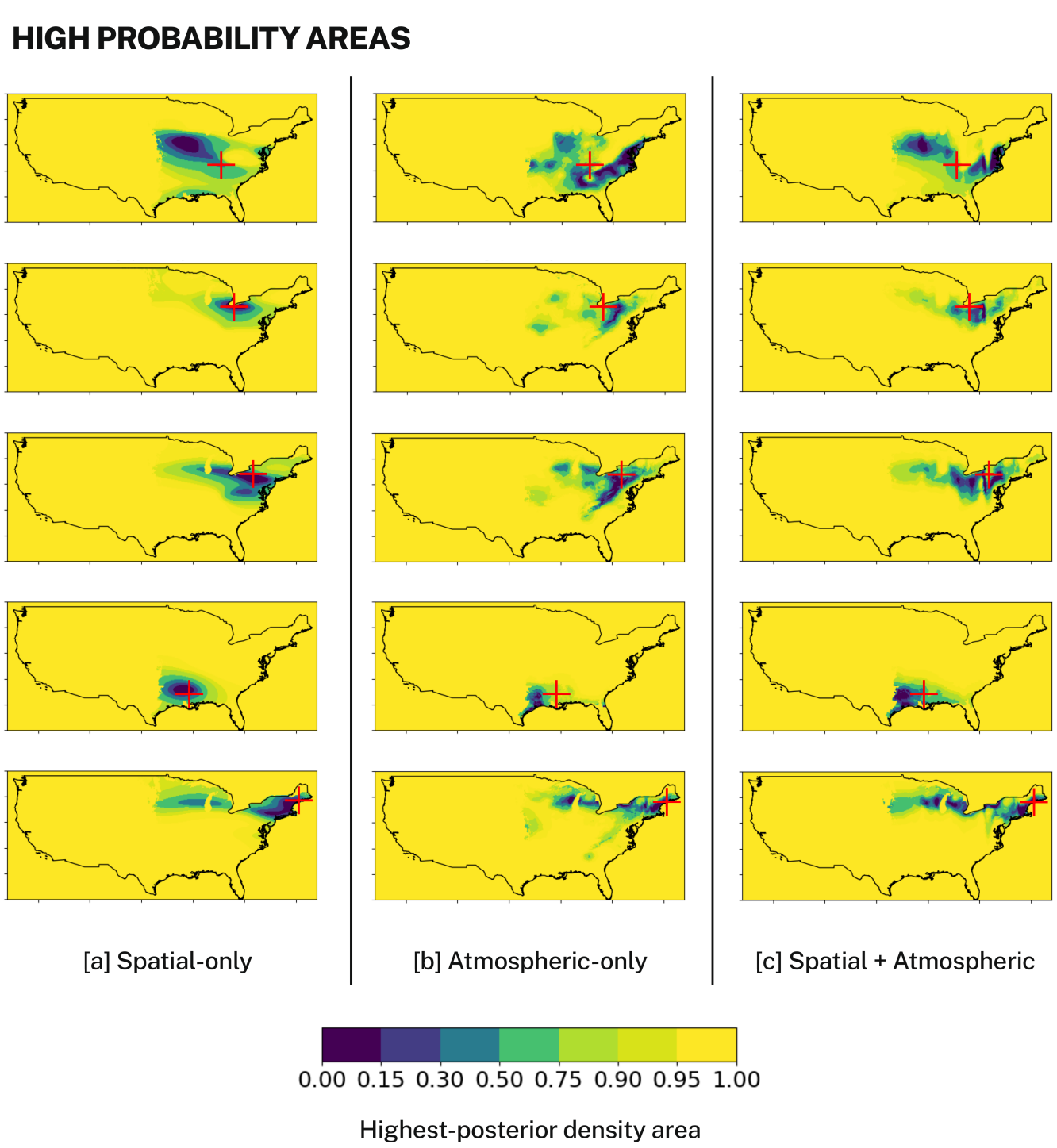

FIGURE 3 Harvest location predictions from the three models for 5 points from the test set using the range prior. Darker shades denote areas of higher posterior probability density, with thresholds set so that the total probability of a colored area is equal to a specified value (see color chart). The red cross indicates the actual location of the sampled tree.

locations within the allowed sampling area. We generated 100 such variance maps and compared the average variance (across the allowed sampling area) of those maps with the maps in Figure 5. Appendix S1: Figure S1 shows the average predictive variances as a function of the number of points added for both random and active learning sampling strategies. Our active learning strategy resulted in a substantially faster decrease in predictive variances. After adding 10 samples, the reduction in variance achieved by our active learning method was 64% ($\delta^{13}C$) to 86% ($\delta^{18}O$) greater than the average reduction achieved by randomly selected sampling locations.

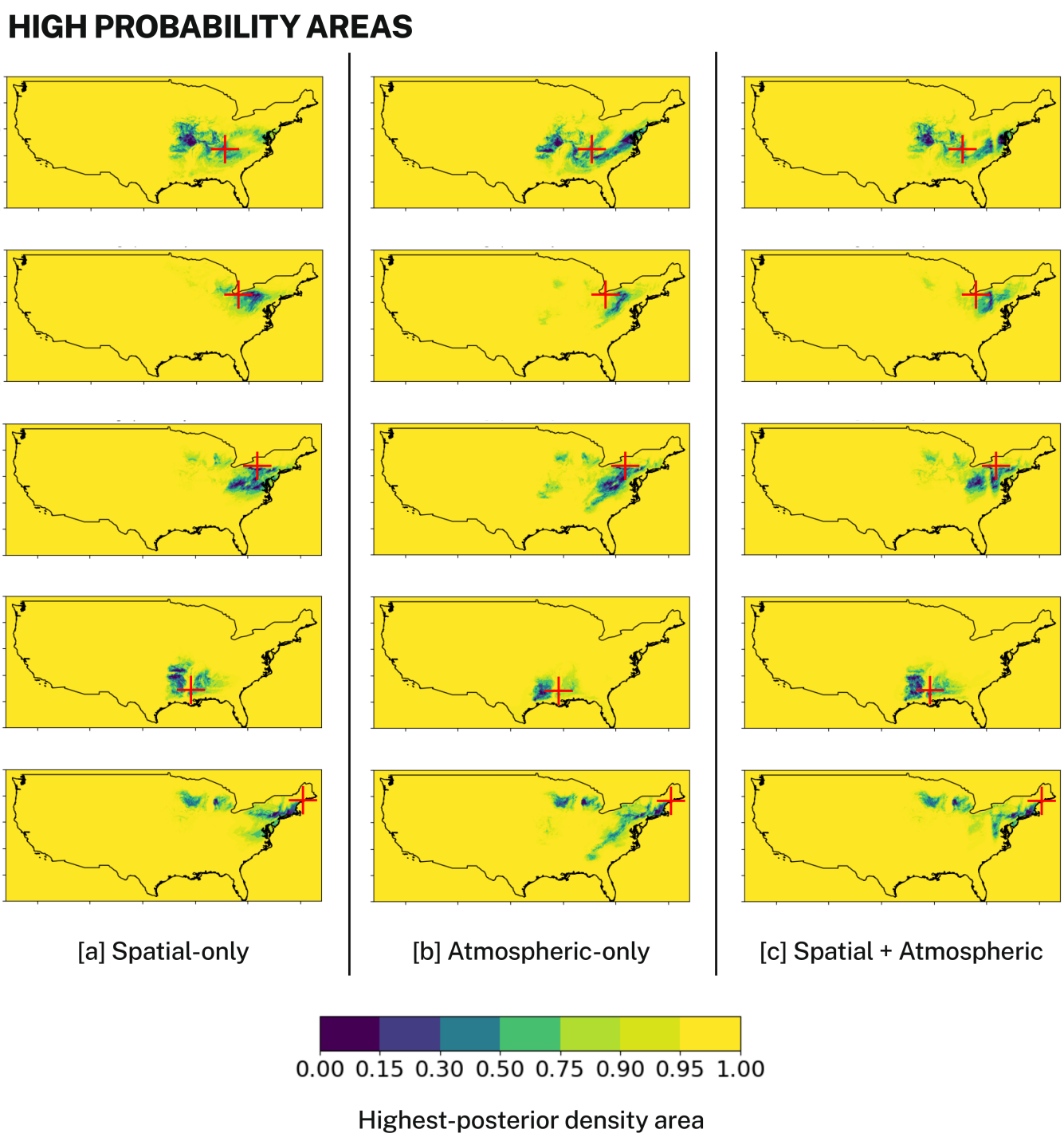

FIGURE 4 Harvest location predictions from the three models for 5 points from the test set using the density prior. Darker shades denote areas of higher posterior probability density, with thresholds set so that the total probability of a colored area is equal to a specified value (see color chart). The red cross indicates the actual location of the sampled tree.

DISCUSSION

Harvest location prediction

To halt illegal logging, to enforce timber regulations and to protect biodiversity in forested landscapes, we need to be able to accurately predict timber harvest location. There are several examples of applying SIRA to timber harvest location questions (Gori et al., 2018; Kagawa & Leavitt, 2010; Watkinson et al., 2020). However, these approaches do not take full advantage of (1) atmospheric and species distribution data sets available or (2) state-of-the-art probabilistic models enabling active learning. In addition, many SIRA use-cases limit themselves to a

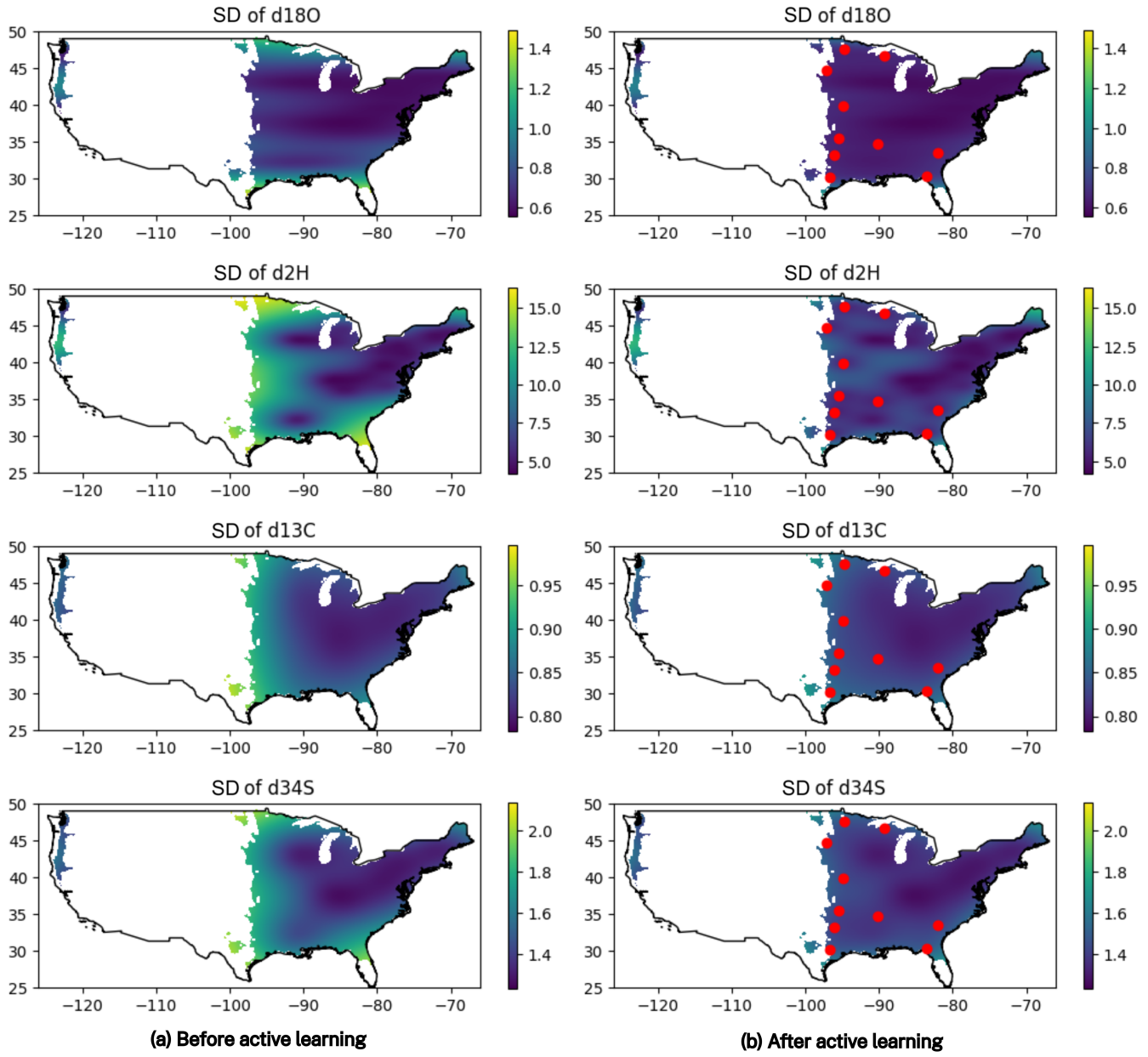

FIGURE 5 Maps showing predictive SDs for the four isotopes before and after adding 10 sample locations proposed by our active learning method for the spatial-only model. SDs are only shown within the allowed sampling area, which is the union of ranges for the species in our data set. The red dots show the proposed locations. Our method proposes locations in areas with high predictive variance, particularly for δ^2H and $\delta^{34}S$. Adding the proposed locations leads to a marked reduction of variance in the neighboring areas.

classification problem (country X vs. country Y) compared to a continuous assignment problem (true harvest location). In response to growing evidence of fraud in supply chains, legislation increasingly requires operators to trace back to plot (e.g., the EUDR). Consequently, determining the true harvest location will likely become increasingly important. In this work, we present a new computational pipeline which aims at taking advantage of both (1) and (2) while predicting the harvest location as a continuous variable.

The accuracy of our models depends on the specific modeling approach and the data used, but incorporating prior information about species distribution results in a considerable increase in accuracy, for all models and accuracy metrics we considered. The impact of adding species distribution data appears to be greater for the spatial-only model than models that use atmospheric information. This could be due to climate patterns influencing both species distributions (habitat suitability) and the values of the atmospheric variables that we incorporated

in our models, which renders species distribution information more redundant once atmospheric variables have been included in the model.

The prior distributions we investigated in this work represent just one of the many ways to derive a prior distribution from spatial forest data. Depending on data availability, other types of data could be used to inform the model about the likely logging locations, such as forest age or proximity to roads or rivers. The optimal choice of prior distribution likely depends on logging patterns, which vary between species and geographic regions. Investigating the impact of different prior distributions is a promising direction of future research.

Within timber tracing literature, our method bears the most resemblance to the work of Watkinson et al. (2020), which uses linear regression to predict isoscapes based on atmospheric data. Their approach assumes a constant variance across the study area. In contrast, our method estimates the predictive variances based on the spatial covariance structure estimated from the reference data, which enables us to translate differences in sampling density across regions into varying levels of confidence in isoscapes across space. Our method also assumes a linear relationship between atmospheric predictors and isoscapes, but it accounts for the uncertainty about regression parameter values, which should lead to more robust predictions compared to standard linear regression (Barber, 2012). In addition, our approach makes use of species distribution data, which yields substantially improved predictions compared to uninformative priors. Finally, our approach enables proposing locations for optimizing further sample collection.

The estimation of spatial covariance structure has recently attracted attention in animal stable isotope studies. Ma et al. (2020) recently proposed a method that uses probabilistic precipitation isoscapes derived from a GP (Courtiol et al., 2019), which are then calibrated to produce isoscapes for the species of interest. St. John Glew et al. (2019) introduced a model combining spatial and environmental effects using a novel marginal likelihood approximation for isoscape estimation, though the main focus of their work is isoscape modeling, not harvest location prediction. These models, like ours, use a grid to compute the distribution of posterior probability across possible harvest locations. However, they differ from our approach in that (1) they rely on marginal likelihood approximations for isoscape estimation rather than exact likelihood maximization; (2) they use ordinary least-squares regression to account for atmospheric predictors, whereas our method uses a more robust approach via a linear covariance term; and (3) they do not aim to actively improve isoscape inference through guiding additional sample collection.

Our current best performing model can predict the harvest location for *Quercus* species to 520 km across the east of the United States. Future field expeditions will lead to an improvement, especially if the identified priority locations are targeted (see *Guiding future collection efforts*). The presented model will be adapted to other use cases, with a focus largely on endangered tropical species which are under high logging pressure.

We expect that our models will be more accurate once more timber samples become available. The number of wood samples available to this study (87 samples) is quite small relative to the study area, which inevitably results in large predictive variance in many regions. In addition to reducing uncertainty about undersampled areas, larger data sets (in the range of hundreds to thousands of samples collected from across the United States) should also enable researchers to use more complex GP models, including models with heterogeneous noise (Binois et al., 2018), or deep kernel learning models where the coordinates are transformed by a neural network before being fed into the covariance function (Wilson et al., 2016). Model performance might also be improved by accounting for the uncertainty in model parameter estimates and relaxing the assumption of independence between different isotopes. We plan to investigate these questions in future work.

Guiding future collection efforts

Under the World Forest ID Programme (Gasson et al., 2021), tens of thousands of tree samples are being collected globally, and analyzed by different techniques, including SIRA, to build a georeferenced database which can be used to identify timber harvest location. Our active learning approach can be used to inform future sample collection efforts and increase the model accuracy that can be achieved within a fixed sample collection budget. This will be especially important in tropical regions, where reaching sampling sites can be difficult, time intensive, and expensive. A good spatial sampling design can substantially improve model performance (Contina et al., 2022), and our method can be used to adapt sample collection efforts as more data are analyzed. Our current approach focuses on minimizing predictive variances without considering the impact of newly sampled points on model parameters. Extending our approach to *non-myopic* sampling (Krause & Guestrin, 2007), which considers the impact on model parameters, would constitute an interesting future research direction. Another avenue for improving our approach would be to augment our IG criterion to reflect the varying investment in collecting samples as a

function of the time, logistics, and financial cost of reaching the desired sampling location.

CONCLUSION

The accurate prediction of harvest location for globally traded wood products is a critical step in combating illegal logging and associated trade, by supporting authorities' ability to verify claims made by traders at any supply chain node. In this work, we presented a novel analytical pipeline that brings together and incorporates multiple data types and algorithms. This methodology is able to accurately predict timber harvest location and can be used to optimize future sample collection in the field to further increase prediction accuracy and precision. We hope that this work will inspire more efforts to expand reference collections of wood samples and that governments and companies will more routinely use the technological tools at their disposal to have more oversight over their supply chains and promote a more sustainable use of natural resources.

ACKNOWLEDGMENTS

Jakub Truszkowski and Alexandre Antonelli are funded by the Swedish Research Council (grant number 2019-05191). Victor Deklerck is funded under the World Forest ID Timber at Kew Grant provided by the Department of Environment, Food & Rural Affairs (DEFRA), International Climate Finance (ICF) R&D Programme, UK (project 29084). Caspar Chater and Roi Maor are funded under the World Forest ID FRC at Kew grant provided by DEFRA, ICF R&D Programme, UK. Alexandre Antonelli also acknowledges financial support from the Swedish Foundation for Strategic Environmental Research MISTRA (Project BioPath) and the Royal Botanic Gardens, Kew. The authors want to thank the US Forest Service—International Programs and FSC-US for the initial collection of the US data set. The findings and conclusions in the article are those of the authors.

CONFLICT OF INTEREST STATEMENT

The authors declare no conflicts of interest.

DATA AVAILABILITY STATEMENT

Data and code (Truszkowski, 2024) are available in Figshare at https://doi.org/10.6084/m9.figshare.23068370.

ORCID

Jakub Truszkowski https://orcid.org/0000-0002-0312-2981
Roi Maor https://orcid.org/0000-0002-8876-2290
Victor Deklerck https://orcid.org/0000-0003-4880-5943

REFERENCES

2022 UN Biodiversity Conference. 2022. "The Kunming-Montreal Global Biodiversity Framework." https://www.cbd.int/gbf/targets/.

Abramowitz, M., and I. A. Stegun. 1972. *Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables*. National Bureau of Standards Applied Mathematics Series 55. Washington, DC: Tenth Printing.

Adhikary, S. K., N. Muttil, and A. G. Yilmaz. 2017. "Cokriging for Enhanced Spatial Interpolation of Rainfall in Two Australian Catchments." *Hydrological Processes* 31(12): 2143–61.

Barber, D. 2012. *Bayesian Reasoning and Machine Learning*. Cambridge, UK: Cambridge University Press.

Barnosky, A. D., N. Matzke, S. Tomiya, G. O. U. Wogan, B. Swartz, T. B. Quental, C. Marshall, et al. 2011. "Has the Earth's Sixth Mass Extinction Already Arrived?" *Nature* 471(7336): 51–57.

Bilodeau, M., and D. Brenner. 1999. *Theory of Multivariate Statistics*. New York: Springer Science & Business Media.

Binois, M., R. B. Gramacy, and M. Ludkovski. 2018. "Practical Heteroscedastic Gaussian Process Modeling for Large Simulation Experiments." *Journal of Computational and Graphical Statistics* 27(4): 808–821.

Boner, M., T. Sommer, C. Erven, and H. Förstel. 2007. "Stable Isotopes as a Tool to Trace Back the Origin of Wood." In *Proceedings of the International Workshop, Fingerprinting Methods for the Identification of Timber Origins*, 8–9.

Borbas, E., et al. 2015. "Terra/modis Temperature and Water Vapor Profiles 5-min l2 Swath 5 km." https://doi.org/10.5067/MODIS/MOD07_L2.061.

Bose, M., J. S. Hodges, and S. Banerjee. 2018. "Toward a Diagnostic Toolkit for Linear Models with Gaussian-Process Distributed Random Effects." *Biometrics* 74(3): 863–873. https://doi.org/10.1111/biom.12848.

Bowen, G. J., and J. Revenaugh. 2003. "Interpolating the Isotopic Composition of Modern Meteoric Precipitation." *Water Resources Research* 39(10): 1299. https://doi.org/10.1029/2003WR002086

Bowen, G. J., L. I. Wassenaar, and K. A. Hobson. 2005. "Global Application of Stable Hydrogen and Oxygen Isotopes to Wildlife Forensics." *Oecologia* 143(3): 337–348.

Camin, F., M. Boner, L. Bontempo, C. Fauhl-Hassek, S. D. Kelly, J. Riedl, and A. Rossmann. 2017. "Stable Isotope Techniques for Verifying the Declared Geographical Origin of Food in Legal Cases." *Trends in Food Science & Technology* 61: 176–187.

Camin, F., L. Bontempo, M. Perini, and E. Piasentier. 2016. "Stable Isotope Ratio Analysis for Assessing the Authenticity of Food of Animal Origin." *Comprehensive Reviews in Food Science and Food Safety* 15(5): 868–877.

Casale, J. F., J. R. Ehleringer, D. R. Morello, and M. J. Lott. 2005. "Isotopic Fractionation of Carbon and Nitrogen during the Illicit Processing of Cocaine and Heroin in South America." *Journal of Forensic Science* 50(6):1315-1321.

Chang, S., and S. Srihari. 2010. "Evaluation of Rarity of Fingerprints in Forensics." *Advances in Neural Information Processing Systems* 23:1207-1215.

Contina, A., S. Magozzi, H. B. Vander Zanden, G. J. Bowen, and M. B. Wunder. 2022. "Optimizing Stable Isotope Sampling Design in Terrestrial Movement Ecology Research." *Methods in Ecology and Evolution* 13(6): 1237–49.

Courtiol, A., F. Rousset, M.-S. Rohwäder, D. X. Soto, L. S. Lehnert, C. C. Voigt, K. A. Hobson, L. I. Wassenaar, and S. Kramer-Schadt. 2019. "Isoscape Computation and Inference of Spatial Origins with Mixed Models Using the r Package Isorix." In *Tracking Animal Migration with Stable Isotopes*, edited by K. A. Hobson and L. I. Wassenaar, 207–236. Cambridge, MA: Elsevier.

Cressie, N. 2015. *Statistics for Spatial Data*. New York: John Wiley & Sons.

Cusa, M., K. St. John Glew, C. Trueman, S. Mariani, L. Buckley, F. Neat, and C. Longo. 2022. "A Future for Seafood Point-of-Origin Testing Using DNA and Stable Isotope Signatures." *Reviews in Fish Biology and Fisheries* 32(2): 597–621. https://doi.org/10.1007/s11160-021-09680-w.

Deklerck, V. 2023. "Timber Origin Verification Using Mass Spectrometry: Challenges, Opportunities, and Way Forward." *Forensic Science International: Animals and Environments* 3: 100057.

Díaz, S., J. Settele, E. S. Brondízio, H. T. Ngo, J. Agard, A. Arneth, P. Balvanera, et al. 2019. "Pervasive Human-Driven Decline of Life on Earth Points to the Need for Transformative Change." *Science* 366(6471): eaax3100.

Dormontt, E. E., M. Boner, B. Braun, G. Breulmann, B. Degen, E. Espinoza, S. Gardner, et al. 2015. "Forensic Timber Identification: It's Time to Integrate Disciplines to Combat Illegal Logging." *Biological Conservation* 191: 790–98.

Gardner, J. R., G. Pleiss, D. Bindel, K. Q. Weinberger, and A. G. Wilson. 2018. "Gpytorch: Blackbox Matrix-Matrix Gaussian Process Inference with GPU Acceleration." Advances in Neural Information Processing Systems. https://proceedings.neurips.cc/paper/2018/hash/27e8e17134dd7083b050476733207ea1-Abstract.html

Gasson, P. E., C. A. Lancaster, R. Young, S. Redstone, I. A. Miles-Bunch, R. Gareth Rees, P. Guillery, M. Parker-Forney, and E. T. Lebow. 2021. "Worldforestid: Addressing the Need for Standardized Wood Reference Collections to Support Authentication Analysis Technologies; a Way Forward for Checking the Origin and Identity of Traded Timber." *Plants, People, Planet* 3(2): 130–141.

Gay, J. D., B. Currey, and E. N. J. Brookshire. 2022. "Global Distribution and Climate Sensitivity of the Tropical Montane Forest Nitrogen Cycle." *Nature Communications* 13: 12. https://doi.org/10.1038/s41467-022-35170-z.

Gori, Y., A. Stradiotti, and F. Camin. 2018. "Timber Isoscapes. A Case Study in a Mountain Area in the Italian Alps." *PLoS One* 13(2): e0192970.

Guestrin, C., A. Krause, and A. P. Singh. 2005. "Near-Optimal Sensor Placements in Gaussian Processes." In *Proceedings of the 22nd International Conference on Machine Learning*, 265–72.

Hijmans, R. J. 2022. "terra: Spatial Data Analysis." R Package Version 1.6-17. https://CRAN.R-project.org/package=terra.

Horacek, M., M. Jakusch, and H. Krehan. 2009. "Control of Origin of Larch Wood: Discrimination between European (Austrian) and Siberian Origin by Stable Isotope Analysis." *Rapid Communications in Mass Spectrometry* 23: 3688–92.

Huffman, G. J., A. Behrangi, D. T. Bolvin, and E. J. Nelkin. 2020. "Gpcp Version 3.1 Satellite-Gauge (sg) Combined Precipitation Data Set." https://disc.gsfc.nasa.gov/datasets/GPCPMON_3.1/summary.

Jankowiak, M., G. Pleiss, and J. Gardner. 2020. "Parametric Gaussian Process Regressors." In *International Conference on Machine Learning*, 4702–12. PMLR.

Kagawa, A., and S. W. Leavitt. 2010. "Stable Carbon Isotopes of Tree Rings as a Tool to Pinpoint the Geographic Origin of Timber." *Journal of Wood Science* 56(3): 175–183.

Kanankege, K. S. T., M. A. Alkhamis, N. B. D. Phelps, and A. M. Perez. 2018. "A Probability Co-Kriging Model to Account for Reporting Bias and Recognize Areas at High Risk for Zebra Mussels and Eurasian Watermilfoil Invasions in Minnesota." *Frontiers in Veterinary Science* 4: 231.

Kirk, J., J. Ruiz, J. Chesley, S. Titley, and S. Titley. 2003. "The Origin of Gold in South Africa: Ancient Rivers Filled with Gold, a Spectacular Upwelling of Magma and a Colossal Meteor Impact Combined to Make the Witwatersrand Basin a Very Special Place." *American Scientist* 91(6): 534–541.

Kitanidis, P. K. 1997. *Introduction to Geostatistics: Applications in Hydrogeology*. Cambridge, UK: Cambridge University Press.

Koehler, G., K. J. Kardynal, and K. A. Hobson. 2019. "Geographical Assignment of Polar Bears Using Multi-Element Isoscapes." *Scientific Reports* 9(1): 1–9.

Krause, A., and C. Guestrin. 2007. "Nonmyopic Active Learning of Gaussian Processes: An Exploration-Exploitation Approach." In *Proceedings of the 24th International Conference on Machine Learning*, 449–56.

Kroetz, K., G. M. Luque, J. A. Gephart, S. L. Jardine, P. Lee, K. C. Moore, C. Cole, A. Steinkruger, and C. J. Donlan. 2020. "Consequences of Seafood Mislabeling for Marine Populations and Fisheries Management." *Proceedings of the National Academy of Sciences of the United States of America* 117(48): 30318–23.

Kurashima, N., Y. Makino, S. Sekita, Y. Urano, and T. Nagano. 2004. "Determination of Origin of Ephedrine Used as Precursor for Illicit Methamphetamine by Carbon and Nitrogen Stable Isotope Ratio Analysis." *Analytical Chemistry* 76(14): 4233–36.

Li, J., and A. D. Heap. 2008. "A Review of Spatial Interpolation Methods for Environmental Scientists." https://ecat.ga.gov.au/geonetwork/dashboard/api/records/a05f7892-db6a-7506-e044-00144fdd4fa6

Littell, R. C., G. A. Milliken, W. W. Stroup, R. D. Wolfinger, and O. Schabenberger. 2006. *SAS for Mixed Models*, 2nd ed. Cary, NC: SAS Institute Inc.

Ma, C., H. B. Vander Zanden, M. B. Wunder, and G. J. Bowen. 2020. "Assignr: An r Package for Isotope-Based Geographic Assignment." *Methods in Ecology and Evolution* 11(8): 996–1001.

Muñoz-Redondo, J. M., D. Bertoldi, A. Tonon, L. Ziller, F. Camin, and J. M. Moreno-Rojas. 2021. "Tracing the Geographical Origin of Spanish Mango (*Mangifera indica* L.) Using Stable Isotopes Ratios and Multi-Element Profiles." *Food Control* 125: 107961. https://doi.org/10.1016/j.foodcont.2021.107961.

NASA Earth Observations/NEO. 2023. "Reflected Shortwave Radiation Data." https://neo.gsfc.nasa.gov/view.php?datasetId=CERES_SWFLUX_M.

Oliver, M. A., and R. Webster. 2015. *Basic Steps in Geostatistics: The Variogram and Kriging*. Technical Report. New York: Springer.

Pederzani, S., and K. Britton. 2019. "Oxygen Isotopes in Bioarchaeology: Principles and Applications, Challenges and

Opportunities." *Earth-Science Reviews* 188: 77–107. https://doi.org/10.1016/j.earscirev.2018.11.005.

Ramakrishnan, N., C. Bailey-Kellogg, S. Tadepalli, and V. N Pandey. 2005 "Gaussian Processes for Active Data Mining of Spatial Aggregates." In *Proceedings of the* 2005 *SIAM International Conference on Data Mining*, 427–38. SIAM.

Saadat, S., H. Pandya, A. Dey, and D. Rawtani. 2022. "Food Forensics: Techniques for Authenticity Determination of Food Products." *Forensic Science International* 333: 111243.

Schmitz, N., V. Haag, C. Blanc-Jolivet, M. Boner, M. T. Cervera, M. Chavesta, R. Cronn, et al. 2019. General Sampling Guide for Timber Tracking. Global Timber Tracking Network. Global Timber Tracking Network, GTTN Secretariat, European Forest Institute and Thuenen Institute. Joensuu, Finland. https://doi.org/10.13140/RG.2.2.26883.96806

Siegwolf, R. T. W., J. Renée Brooks, J. Roden, and M. Saurer. 2022. *Stable Isotopes in Tree Rings: Inferring Physiological, Climatic and Environmental Responses*. Cham, Switzerland: Springer Nature.

Silva, A. J., R. S. Hellberg, and R. H. Hanner. 2021. "Chapter 7 – Seafood Fraud." In *Food Fraud*, edited by R. S. Hellberg, K. Everstine, and S. A. Sklare, 109–137. London, UK: Elsevier.

St. John Glew, K., L. J. Graham, M. G. RA, and C. N. Trueman. 2019. "Spatial Models of Carbon, Nitrogen and Sulphur Stable Isotope Distributions (Isoscapes) across a Shelf Sea: An Inla Approach." *Methods in Ecology and Evolution* 10(4): 518–531.

Swofford, H., and C. Champod. 2022. "Probabilistic Reporting and Algorithms in Forensic Science: Stakeholder Perspectives within the American Criminal Justice System." *Forensic Science International: Synergy* 4: 100220.

Truszkowski, J. 2024. "A Probabilistic Approach to Estimating Timber Harvest Location." Figshare. Dataset. https://doi.org/10.6084/m9.figshare.23068370.v1.

Valliant, R., A. H. Dorfman, and R. M. Royall. 2000. *Finite Population Sampling and Inference: A Prediction Approach*. Number 04; QA276. 6, V3. New York: John Wiley New York.

Van der Merwe, N. J., J. A. Lee-Thorp, J. F. Thackeray, A. Hall-Martin, F. J. Kruger, H. Coetzee, R. H. V. Bell, and M. Lindeque. 1990. "Source-Area Determination of Elephant Ivory by Isotopic Analysis." *Nature* 346(6286): 744–46.

van der Sleen, P., P. A. Zuidema, and T. L. Pons. 2017. "Stable Isotopes in Tropical Tree Rings: Theory, Methods and Applications." *Functional Ecology* 31(9): 1674–89.

Vander Zanden, H. B., A. D. Tucker, K. M. Hart, M. M. Lamont, I. Fujisaki, D. S. Addison, K. L. Mansfield, et al. 2015. "Determining Origin in a Migratory Marine Vertebrate: A Novel Method to Integrate Stable Isotopes and Satellite Tracking." *Ecological Applications* 25(2): 320–335.

Watkinson, C. J., P. Gasson, G. O. Rees, and M. Boner. 2020. "The Development and Use of Isoscapes to Determine the Geographical Origin of *Quercus* spp. in the United States." *Forests* 11(8): 862.

Watkinson, C. J., G. O. Rees, M. C. Gwenael, P. Gasson, S. Hofem, L. Michely, and M. Boner. 2022. "Stable Isotope Ratio Analysis for the Comparison of Timber from Two Forest Concessions in Gabon." *Frontiers in Forests and Global Change* 4: 155.

Watkinson, C. J., G. O. Rees, S. Hofem, L. Michely, P. Gasson, and M. Boner. 2022. "A Case Study to Establish a Basis for Evaluating Geographic Origin Claims of Timber from the Solomon Islands Using Stable Isotope Ratio Analysis." *Frontiers in Forests and Global Change* 4: 645222.

West, J. B., G. J. Bowen, T. E. Dawson, and K. P. Tu. 2010. *Isoscapes: Understanding Movement, Pattern, and Process on Earth through Isotope Mapping*. Dordrecht: Springer.

Williams, C. K. I., and C. E. Rasmussen. 2006. *Gaussian Processes for Machine Learning*, Vol. 2. Cambridge, MA: MIT Press.

Wilson, A. G., H. Zhiting, R. Salakhutdinov, and E. P. Xing. 2016. "Deep Kernel Learning." In *Artificial Intelligence and Statistics*, 370–78. PMLR.

Wilson, B. T., A. J. Lister, R. I. Riemann, and D. M. Griffith. 2013. "Live Tree Species Basal Area of the Contiguous United States (2000-2009)." https://www.fs.usda.gov/rds/archive/Catalog/RDS-2013-0013.

Wunder, M. B., and R. D. Norris. 2008. "Improved Estimates of Certainty in Stable-Isotope-Based Methods for Tracking Migratory Animals." *Ecological Applications* 18(2): 549–559.

Ziegler, S., S. Merker, B. Streit, M. Boner, and D. E. Jacob. 2016. "Towards Understanding Isotope Variability in Elephant Ivory to Establish Isotopic Profiling and Source-Area Determination." *Biological Conservation* 197: 154–163.

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

How to cite this article: Truszkowski, Jakub, Roi Maor, Raquib Bin Yousuf, Subhodip Biswas, Caspar Chater, Peter Gasson, Scot McQueen, et al. 2025. "A Probabilistic Approach to Estimating Timber Harvest Location." *Ecological Applications* 35(1): e3077. https://doi.org/10.1002/eap.3077