Received: 13 July 2021 | Revised: 26 November 2022 | Accepted: 9 December 2022

DOI: 10.1002/env.2785

RESEARCH ARTICLE

WILEY

Multivariate receptor modeling with widely dispersed Lichens as bioindicators of air quality

Matthew Heiner[1] | Taylor Grimm[1] | Hayden Smith[1] | Steven D. Leavitt[2,3] | William F. Christensen[1] | Gregory T. Carling[4] | Larry L. St. Clair[2,3]

[1] Department of Statistics, Brigham Young University, Provo, Utah, USA

[2] Department of Biology, Brigham Young University, Provo, Utah, USA

[3] M. L. Bean Life Science Museum, Brigham Young University, Provo, Utah, USA

[4] Department of Geological Sciences, Brigham Young University, Provo, Utah, USA

Correspondence
Matthew Heiner, Department of Statistics, Brigham Young University, Provo, UT, USA.
Email: heiner@stat.byu.edu

Funding information
U.S. Forest Service

Abstract

Biomonitoring studies evaluating air quality via airborne element accumulation patterns in lichens typically control variability by focusing on narrow geographic regions and short time windows. Using samples of the widespread "rock-posy" lichen sampled across the Intermountain Region of the United States, we investigate whether accumulation patterns of generic pollution sources are detectable on broad geographic and temporal scales. We develop a novel Bayesian multivariate receptor modeling (BMRM) approach that sharpens detection and discrimination of candidate pollution sources through (i) regularization of source contributions to each sample and (ii) incorporating estimated lichen secondary chemistry as a factor. Through a simulation study, we demonstrate a distinct advantage in shrinking contributions when they are truly sparse, as would be expected with heterogeneous samples from dispersed collection sites. We contrast analyses employing both standard and sparse BMRMs, and positive matrix factorization (PMF). The sparse model better maintains source identity, as specified though informative prior distributions on elemental profiles. We advocate quantitative profile matching, which reveals that PMF primarily captures variations of the baseline profile for lichen secondary chemistry. Both PMF and BMRM results suggest that the most detectable signatures relate to aeolian dust deposition, while spatial patterns hint at sporadic anthropogenic influence.

KEYWORDS

Aeolian dust, Bayesian methods, elemental analysis, pollution source apportionment, positive matrix factorization, regularization

1 | INTRODUCTION

Multivariate receptor models (MRMs) are well-established tools for identifying pollutant signatures in air-quality monitoring data, particularly for quantifying specific contributions from point sources (Hopke, 2016; Krall & Chang, 2019). Encoding external information about source signatures through prior distributions provides an effective methodology for constraining MRM solutions to better identify sources (Hackstadt & Peng, 2014; Lingwall et al., 2008). In this article, we address the related but distinct problem of flexibly estimating contributions of modeled pollution sources to a collection of heterogeneous samples observed across sites or time. We propose a regularization method that decouples these objectives in a novel way, resulting in sharper estimation of both signatures and contributions when source influence is truly sparse. Such is often the case with air-quality monitoring across wide, or dynamic, environments.

Apportionment studies employing MRMs typically rely on concentrations of multiple chemical species as measured from filter instrumentation at fixed locations. Costs associated with establishing and maintaining these monitors have partially motivated efforts to find and exploit naturally occurring bioindicators that can provide similar quantitative information. One promising alternative is in situ sampling of lichens, which are among a suite of sensitive organisms/communities that serve as direct surrogates for assessing disturbances to biological communities (Hodkinson & Jackson, 2005; Nimis et al., 2002).

The utility of lichens as bioindicators of air-quality derives, in part, from the fact that they obtain mineral nutrients as well as co-occurring contaminants through atmospheric deposition rather than from their substrates (Agnan et al., 2014). Differential accumulation of atmospheric pollutants in thalli (tissues) of sensitive indicator lichens, through dry or wet deposition (Sloof, 1995; Szczepaniak & Biziuk, 2003), can provide means to quantitatively assess ecosystem health (Geiser & Neitlich, 2007; Henderson-Sellers & Seaward, 1979). Unlike with filter instruments, however, correlations between element air concentrations and air pollutant accumulations by lichens may be confounded by complex processes and factors including growth form, substrate chemistry and pH, and physiological differences (Lawrey & Hale Jr, 1981; Riddell et al., 2011; Rola, 2020; Saeki et al., 1977; St. Clair et al., 2002; Will-Wolf et al., 2006). Gradients in local climate, lichen age and growth rates (on decadal or century scales), and collection timing can further contribute to variation in elemental concentrations (Root et al., 2021).

Despite the challenges associated with using natural instruments, lichen-biomonitoring studies evaluating airborne element accumulation patterns relative to a specific pollution source are becoming more common (Aznar et al., 2008; Seaward, 1993; St. Clair et al., 1994). For example, the epiphytic lichen *Hypogymnia physodes* has been used to monitor the impact of industrial operations in the Athabasca Oil Sands Region in Alberta, Canada over the past three decades. Using MRM approaches, Landis et al. (2012); Landis, Studabaker, et al. (2019) identified signatures of production-related airborne pollutants.

These and other recent studies in pollution-source apportionment with lichens manage or reduce variability by focusing on narrow geographic regions, using samples collected within short time frames, collecting replicate samples, or imposing experimental controls (Boamponsem et al., 2010; Contardo et al., 2020; Loppi et al., 2019). In contrast, we fit MRMs to data from lichens collected across the Intermountain Region of the United States and investigate the strength of pollution signals when the spatial and temporal scope are increased to subcontinental and multi-annual scales. In our application, replication exists only spatially, with nearest-neighboring samples typically collected kilometers apart. Such a general, and cost effective, sampling approach introduces substantial statistical challenges.

In this article, we examine whether state-of-the-art MRMs can sufficiently detect and discriminate signals of generic (rather than specific point) pollution sources with noisy elemental data from widely dispersed lichen samples. Perhaps the most standard tool for apportionment is positive matrix factorization (PMF), which decomposes elemental concentration profiles from samples into contributions from estimated pollution source profiles. We contrast PMF results with those of two Bayesian multivariate receptor models (BMRMs). Through informative priors, these models incorporate information about potential pollution source profiles while accommodating for uncertainties arising with use of a natural bioindicator. Because our elemental measurements represent lichen physiology in addition to accumulated pollution, we customize the models to include a pre-estimated factor (source) representing background lichen secondary chemistry.

The second Bayesian model extends the standard BMRM to accommodate sparsity of sources within a site by regularizing source contribution estimates through a novel, flexible prior that draws from the literature on variable selection via shrinkage. We demonstrate that this extension assists in maintaining source identity without fixing profiles, thereby more effectively balancing desirable features of factor models and chemical-mass-balance regression (Christensen et al., 2006).

This article is organized as follows. We first describe the elemental analysis data set and note features and limitations of the data in Section 2. Section 3 describes our modeling approaches and details methodologies for regularized source contributions and source profile engineering, including background lichen profiles, as well as quantitative profile matching. We empirically verify model characteristics with synthetic data designed to resemble the lichen analysis in Section 4. We present results in Section 5 and conclude with discussion in Section 6.

2 | LICHEN BIOMONITORING DATA

Effective elemental bioindicators should be widespread, pollution tolerant, and amenable to collection (Will-Wolf et al., 2017). For this study, we focused on the common umbilicate "rock-posy" lichen, characterized by its fungal symbiont, *Rhizoplaca melanophthalma* (Leavitt, Fernández-Mendoza, Pérez-Ortega, Sohrabi, Divakar, Lumbsch, & St. Clair, L. L., 2013). Members of this species complex are commonly found throughout the Intermountain Region of the United States in well-developed populations (Leavitt, Fernández-Mendoza, Pérez-Ortega, Sohrabi, Divakar, Vondrák, et al., 2013), and are assumed to indiscriminately accumulate atmospheric outwash (Dillman, 1996). While rock-posy lichens are commonly used for inferring the impact of potential pollutants, elemental accumulation patterns across broad geographic and temporal scales have not yet been characterized.

As part of the larger biomonitoring program at Brigham Young University (https://lichenairquality.com; Leavitt & St. Clair, 2015), samples of *R. melanophthalma* aggregate were collected from over 96 sites primarily distributed throughout the Great Basin and adjacent areas between 2009 and 2015 (six samples were collected prior to 2009) for elemental analysis using inductively coupled plasma optical emission spectrometry (ICP-OES) analysis. Section S1.2 further describe sampling and ICP protocols. Sampling locations were generally selected within the National Wilderness Preservation System and adjacent National Forest lands to provide broad-scale representation of wilderness areas and were determined by accessibility and availability of diverse lichen communities. Measured concentrations of 25 potential air-pollutant or geogenic elements were used in all analyses. Measurements flagged as below detection limits are labeled "bdl," and "nd," which we take to mean "not detected."

An exploratory spatial analysis for individual elements did not return consistent results, underscoring the need for dimension reduction afforded by multivariate statistical methods to help identify signals and spatial patterns. Without replicate samples, we rely on work by Landis, Berryman, et al. (2019), who analyzed samples of *Hypogymnia physodes* to produce a general breakdown of variability attributable to (i) environmental differences and deposition gradients, accounting for inter-site variability (>85%); ii) individual lichen characteristics (thallus size, age, metabolic factors) that account for intra-site variability; and (iii) laboratory and ICP measurement error, resulting in replicate variability (<1%).

3 | SOURCE APPORTIONMENT MODELS

Let $y_{i,\ell}$ represent the measured concentration, in parts per million (ppm), of element $\ell \in \{1, \dots, L\}$ from sample $i \in \{1, \dots, n\}$. In this analysis, we have $L = 25$ and $n = 96$. We treat occasional missing values as data missing completely at random.

3.1 | Positive matrix factorization

Multivariate receptor models represent the measurement of element ℓ from sample i as

$$y_{i,\ell} = \sum_{k=1}^{K} f_{i,k}\lambda_{k,\ell} + \epsilon_{i,\ell} \ , \tag{1}$$

where $\{f_{i,k} : k = 1, \dots, K\}$ are positive-valued contributions (with the same units as y) from K sources, $\{\boldsymbol{\lambda}_k \equiv (\lambda_{k,1}, \dots, \lambda_{k,L}) : k = 1, \dots, K\}$ contain nonnegative source profiles such that $\sum_{\ell=1}^{L} \lambda_{k,\ell} = 1$ for all k, and $\epsilon_{i,\ell}$ captures model error. If we let $\boldsymbol{Y}$ be an n by L matrix such that $(\boldsymbol{Y})_{i,\ell} = y_{i,\ell}$, then the model can be equivalently written as

$$\underset{n\times L}{\boldsymbol{Y}} = \underset{n\times K}{\boldsymbol{F}}\,\underset{K\times L}{\boldsymbol{\Lambda}} + \underset{n\times L}{\boldsymbol{E}} \ , \tag{2}$$

where $(\boldsymbol{F})_{i,k} \equiv f_{i,k}$, row k of $\boldsymbol{\Lambda}$ contains $\boldsymbol{\lambda}_k$, and $(\boldsymbol{E})_{i,\ell} \equiv \epsilon_{i,\ell}$. One instructive interpretation of the factorization in (2) views each sample as a weighted sum of K source profile vectors, plus error, that is, $\boldsymbol{y}_i$ and $\boldsymbol{\epsilon}_i$ are L-length vectors representing row i of $\boldsymbol{Y}$ and $\boldsymbol{E}$, respectively. Then we have $\boldsymbol{y}_i = \sum_{k=1}^{K} f_{i,k}\boldsymbol{\lambda}_k + \boldsymbol{\epsilon}_i$. The model is distinct from traditional factor models in that the parameters are constrained to be nonnegative, although these constraints do not resolve scale indeterminacy and rotational ambiguity in $\boldsymbol{F\Lambda}$ (Park et al., 2002). Chemical-mass-balance models assume profiles in $\boldsymbol{\Lambda}$ are known, allowing contributions in $\boldsymbol{F}$ to be estimated by regression methods (Christensen et al., 2006).

One standard method for estimating the model in (2), called PMF (Paatero & Tapper, 1994), is implemented in free software provided by the US Environmental Protection Agency (EPA; U.S. Environmental Protection Agency, 2014). EPA PMF employs a gradient search algorithm to minimize the objective function Q, a sum of squares of weighted residuals. User-supplied uncertainty values individually down-weight observations when calculating Q, affording flexibility and robustness to the fit. The number of sources (factors, K) is a fixed model input. We fit the EPA PMF model as a standard comparison to the Bayesian models, and refer to it throughout as the PMF fit.

3.2 | Bayesian multivariate receptor models

Our Bayesian implementation is closely related to the multivariate receptor model of Lingwall et al. (2008), which flexibly incorporates prior information on source profiles, is readily adaptable to accommodate data-specific features, and yields uncertainty along with point estimates. Our strategy is to estimate source contributions in $\boldsymbol{F}$ with minimal imposition, and concentrate a priori knowledge in specifying profiles in $\boldsymbol{\Lambda}$ for candidate pollutant sources.

To naturally satisfy positivity constraints on the data, profiles, and contributions, we employ a log-normal likelihood and modify (1) to have multiplicative, independent errors, that is, $y_{i,\ell} = (\sum_k f_{i,k} \lambda_{k,\ell})\epsilon_{i,\ell}$. In matrix notation, this becomes $\boldsymbol{Y} = \boldsymbol{F\Lambda} \odot \boldsymbol{E}$, where $\odot$ denotes elementwise multiplication. It is equivalently expressed as

$$\log(\boldsymbol{Y}) = \log(\boldsymbol{F\Lambda}) + \log(\boldsymbol{E}) , \tag{3}$$

where the natural logarithm function with a vector or matrix argument signifies applying the logarithm to each element. We consider using normal and Student-t distributed log-errors, $\log(\epsilon_{i,\ell})$. The multiplicative model applied on the log scale is appropriate especially when working with elemental concentrations that differ by several orders of magnitude. The cost of using (3) is that estimating $\boldsymbol{F}$ and $\boldsymbol{\Lambda}$ is more complex relative to modeling with (2). We assume throughout that K, the number of sources, is known and fixed.

We begin with a BMRM that modifies that of Lingwall et al. (2008) in two important ways. First, we parameterize the likelihood to model the median and coefficient of variation on the original y scale, both interpretable parameters that are amenable to prior elicitation. Second, we separate a baseline source, for which we employ a distinct prior on contributions in $\boldsymbol{F}$. We will refer to the model with these modifications as the base BMRM.

Excluding the baseline source, the general specification for the base BMRM is as follows. For $i = 1, \dots, n$ indexing samples (also referred to as observations), $\ell = 1, \dots, L$ indexing elements, and $k = 1, \dots, K$ indexing sources, we have

$$\begin{aligned} \log(y_{i,\ell}) \mid \{\boldsymbol{F}, \boldsymbol{\Lambda}, \{\sigma_\ell\}\} &\overset{\text{ind.}}{\sim} \mathrm{N}\left(\log\left(\sum_{k=1}^{K} f_{i,k}\lambda_{k,\ell} \right),\ \sigma_\ell \right) , \\ \lambda_k &\overset{\text{ind.}}{\sim} \mathrm{GDir}(\boldsymbol{\alpha}_k,\ \boldsymbol{\beta}_k) , \\ v_\ell &\overset{\text{ind.}}{\sim} \mathrm{Ga}(a_\ell,\ b_\ell) , \end{aligned} \tag{4}$$

where $\mathrm{N}(\mu, \sigma)$ refers to a normal distribution with location μ and scale σ. Thus the median of $y_{i,\ell}$ is $(\boldsymbol{F\Lambda})_{i,\ell}$. Here we parameterize $\sigma_\ell = \sqrt{\log(v_\ell^2 + 1)}$ with v_ℓ being interpreted as the the coefficient of variation (i.e., $\sqrt{\mathrm{Var}(y_{i,\ell})}/E(y_{i,\ell})$) when the likelihood is log-normal. $\mathrm{Ga}(a, b)$ refers to a gamma distribution with mean ab, and $\mathrm{GDir}(\boldsymbol{\alpha}, \boldsymbol{\beta})$ refers to what Lingwall et al. (2008) call a generalized Dirichlet distribution with shape vector $\boldsymbol{\alpha} = (\alpha_1, \dots, \alpha_L)$ and scale vector $\boldsymbol{\beta} = (\beta_1, \dots, \beta_L)$. A random probability vector $\lambda \sim \mathrm{GDir}(\boldsymbol{\alpha}, \boldsymbol{\beta})$ is constructed by normalizing a latent random vector $\boldsymbol{z} = (z_1, \dots, z_L)$ as $\lambda_\ell = z_\ell / \sum_{j=1}^{L} z_j$, where each z_ℓ is independently drawn from a gamma distribution with shape α_ℓ and scale β_ℓ, thus inducing a prior on λ_k. This formulation for source profiles facilitates prior elicitation, as discussed in Section 3.3. Although informative priors on $\boldsymbol{\Lambda}$ can alleviate the need for strict enforcement of likelihood identifiability conditions, such conditions proved necessary for the lichen analyses. We follow the strategy of Hackstadt and Peng (2014); see Section S2.

Log-normal priors on $\{f_{i,k}\}$ offer some flexibility. However, they tend to encourage contribution of every source to every observation, precipitating alterations in $\boldsymbol{\Lambda}$ to accommodate a dense $\boldsymbol{F}$. We propose a novel structured prior for $\boldsymbol{F}$ intended to shrink contributions from sources whose signatures do not appear in the sample, thus avoiding undue influence on profiles in $\boldsymbol{\Lambda}$. Borrowing from the literature on variable selection in regression, we replace the log-normal priors on $\{f_{i,k} : i = 1, \dots, n;\ k = 1, \dots, K\}$ in (4) with

$$\begin{aligned} f_{i,k} \mid \gamma_i &\overset{\text{ind.}}{\sim} \mathrm{Ga}\left(1/K,\ \gamma_i\right) , \\ \gamma_i &\overset{\text{iid}}{\sim} \mathit{Exp}(2\gamma_0) , \end{aligned} \tag{5}$$

where $Exp(\mu)$ denotes to an exponential distribution with mean μ. We will refer to the model in (4) with this modification as the sparse Bayesian multivariate receptor model (S-BMRM). While sparse factor modeling is not uncommon, most applications apply shrinkage to factor loadings (profiles in MRMs). See Frühwirth-Schnatter and Lopes (2018) for a review.

The hierarchical formulation in (5) provides a global-local shrinkage prior for source contribution weights that simultaneously induces sparsity and allows for large contributions. It derives from a positive-valued equivalent of the Dirichlet–Laplace prior for regression coefficients (Bhattacharya et al., 2015), which possesses excellent shrinkage qualities. While the original Dirichlet–Laplace prior mimics joint posterior behavior of popular spike-and-slab priors (George & McCulloch, 1993; George and McCulloch, 1997), our prior in (5) corresponds with the marginal distribution arising from $f_{i,k} | \phi_{i,k} \overset{\text{ind.}}{\sim} Exp(\phi_{i,k}\ \gamma_i)$ with $\phi_{i,k} \overset{\text{iid}}{\sim} \text{Beta}(1/K,\ 1-1/K)$ integrated out. This form can also be viewed as a continuous extension of the Spike-and-Slab LASSO prior of Ročková and George (2018), in which $\phi_{i,k}$ is dichotomous. If less aggressive shrinkage is desired, $1/2$ can replace $1/K$, yielding a symmetric U-shaped distribution for the local shrinkage factor, $\phi_{i,k}$.

Within the context of our application, we add a baseline source for secondary lichen chemistry to both BMRMs, which is further detailed in Section 3.4. We treat the baseline profile similarly to an intercept term, and grant this source priority in allocating mass to individual samples by assigning a distinct, nonpenalized prior to its contributions. Specifically, we add a zero-indexed term to the median of $y_{i,\ell}$ in the likelihood of (4), which becomes $f_{i,0}\lambda_{0,\ell} + \sum_{k=1}^{K} f_{i,k}\lambda_{k,\ell}$. The priors for $\{f_{i,0} : i = 1, \ldots, n\}$ are independent uniform on the interval $(0, B_i)$, where B_i represents an upper bound for baseline mass of sample i. The prior for the baseline profile, $\lambda_0 \equiv (\lambda_{0,1}, \ldots, \lambda_{0,L})$, follows the pattern in (4), with elicited hyperparameters $\boldsymbol{\alpha}_0$ and $\boldsymbol{\beta}_0$.

Aside from source-profile priors discussed in Section 3.3, the prior specification for all models requires values for B_i, $\{a_\ell\}$, and $\{b_\ell\}$. The upper bounds for the uniform priors on baseline contributions were set to $B_i = 175{,}000$ ppm for all samples. We used separate error coefficients of variation, $\{\nu_\ell\}$, to accommodate the wide range of relative abundance across elements, and to allow for elements that prove irrelevant or that show inconsistent measurement to lose influence. However, each were assigned identical gamma priors with mean 0.4 and *SD* 0.15, allowing for considerable uncertainty in the measurements.

For the base BMRM, we used $\mu_f = \log(10^4)$ to scale the prior median contribution up to 10^4 parts per million for each source, and $\sigma_f = \sqrt{\log(10)}$, which corresponds with a prior coefficient of variation of 3. In the S-BMRM, γ_i is loosely interpretable as the portion of the median total mass in sample i not explained by baseline lichen chemistry. We observe that the total mass contributed from all measured elements mostly varies between 20,000 and 90,000 ppm. Assuming a significant portion can be attributed to baseline chemistry, we selected a value for γ_0 of 30,000, resulting in a prior mean of 60,000 ppm for γ_i.

Several observations contain missing concentrations, attributable to differences in ICP protocols across samples and not to the missing values themselves. We therefore assume these measurements missing at random. Other concentrations flagged as below detection limits ("bdl") were considered left-censored. In both cases, the likelihood (truncated for "bdl" values) supplied the sampling model. Our implementation treated missing and censored values as latent variables and sampled (imputed) from their posterior distributions within the Markov chain Monte Carlo (MCMC) algorithm.

Long tails in the empirical distributions of log-residuals across i for for several elements ℓ in the lichen analysis required that we consider replacing the normal distribution in the likelihood of (4) with a Student-t distribution. Models employing this likelihood, with location $\log((\boldsymbol{F\Lambda})_{i,\ell})$ and scale σ_ℓ in log units, follow the same specification given to this point with the following exceptions. We modeled the degrees of freedom, ν, with an informative inverse-gamma prior truncated below at 4, and using a shape parameter of 19 and scale parameter of 200 to yield a prior mode of 10. The $\{\nu_\ell\}$ parameters are no longer directly interpretable as coefficients of variation, and high values led to extreme prior behavior. Thus, with the Student-t likelihoods, we replaced the gamma priors on $\{\nu_\ell\}$ with bounded Beta(1.2, 4.8) distributions, yielding $E(\nu_\ell) = 0.2$ a priori.

Our use of the log-normal (or log-Student-t) likelihood with multiplicative error precludes convenient Gibbs sampling for posterior computation. To facilitate general implementation with other applications, we fit these BMRMs in the probabilistic programming language STAN, which utilizes a general Hamiltonian Monte Carlo algorithm (Carpenter et al., 2017; Stan Development Team, 2020).

3.3 | Candidate source profiles

Apportionment studies typically utilize elemental profiles of airborne (or terrestrial) samples from known sources collected locally (Landis et al., 2012), or rely on generic, published source profiles to inform the models (Krall & Chang, 2019).

We take the latter approach of incorporating external profile estimates for important potential sources from public databases, scientific literature, and expert opinion.

We draw primarily from the EPA's SPECIATE database (version 5.0, U.S. Environmental Protection Agency, 2019), initially obtaining a variety of pollution sources spanning agriculture, excavation, industrial processes and waste, and transportation. However, high correlation among profiles, together with high noise in the lichen data, prompted restriction to the following candidate sources in the Bayesian models: Brake Wear, Motor Vehicle Exhaust, and Unpaved Road Dust. In addition to the SPECIATE profiles, we include a profile for fine playa dust adapted from Goodman et al. (2019). Playa dust comes from dry lakebeds, which commonly occur in the Great Basin area of the western United States. Finally, we consider two generic sources and assign weaker priors with the intent of capturing otherwise missed sources. We label the two generic sources Natural and Anthropogenic.

Quantified profile signatures, with uncertainty, facilitate the prior-specification approach of Lingwall et al. (2008). We rely on the constructive definition of the generalized Dirichlet distribution to set values of $\boldsymbol{\alpha}_k$ and $\boldsymbol{\beta}_k$ for each candidate source. Details are given in Section S3.

3.4 | Baseline Rhizoplaca biochemistry

Bioindicators differ from standard pollution-measuring instruments in that they possess a complex secondary chemistry. Instead of filtering particulates from the atmosphere, lichens incorporate surface-deposited particles as they grow. Indeed, Landis et al. (2012) attribute at least one estimated "source" to lichen biochemistry. To partially account for this process, we constructed a source profile that corresponds to baseline *R. melanophthalma* agg. secondary chemistry. Ideally, this profile would be estimated with specimens grown in an environment free of pollutants. Without such samples, we relied on the spatial extent of the collection to provide a typical profile.

We estimated what we call the Baseline *Rhizoplaca* profile in a separate spatial modeling step to mitigate the influence of (i) regional anomalies in pollutant abundance and (ii) bias from nonrandom, cluster-type sampling. Each of the $L - 1$ models estimated the log-ratio of one element's concentration to a reference element, in this case Ca (the most abundant). Modeling relative concentration is important because the total proportion of mass from unmeasured elements varies by sample.

If $y_{i,1}$ refers to the calcium measurement (in ppm) for sample $i \in \{1, \dots, n\}$, then for each other element $\ell = 2, \dots, L$, the model is

$$\log\left(\frac{y_{i,\ell}}{y_{i,1}}\right) = \mu_\ell + \nu_{i,\ell} + \varepsilon_{i,\ell} \ , \tag{6}$$

where μ_ℓ represents the mean log-ratio, $\nu_{i,\ell}$ is a spatial random effect, and $\varepsilon_{i,\ell}$ provides mutually independent (Student-t distributed) noise to accommodate long tails apparent in the observed log-ratios. Spatial random effects, $\boldsymbol{\nu}_\ell \equiv (\nu_{1,\ell}, \dots, \nu_{n,\ell})$, were each modeled with Gaussian process (GP) priors utilizing isotropic, exponential covariance functions of the geodesic distances between sampling locations. Specific prior settings are reported in Section S3.

Posterior samples of the mean log-ratios, $\{\mu_\ell\}$, were used to construct samples of the baseline profile, denoted $\tilde{\boldsymbol{\pi}} \equiv (\tilde{\pi}_1, \dots, \tilde{\pi}_L)$, using the inverse multi-logit relation

$$\tilde{\pi}_\ell = \frac{\exp(\mu_\ell)}{1 + \sum_{j=2}^{L} \exp(\mu_j)} \ , \tag{7}$$

for $\ell = 1, \dots, L$, with $\mu_1 = 0$. This nonlinear transformation does not yield the mean profile, but rather an estimate constructed from marginal median ratios. Marginal posterior summaries of $\tilde{\boldsymbol{\pi}}$ were used to specify $\boldsymbol{\alpha}_0$ and $\boldsymbol{\beta}_0$ for Baseline, by element, following the procedure used with other source-profile priors.

Although the spatial model for the source Baseline could be embedded into the BMRM (4), we separate this lichen-specific elicitation step for simplicity in the proposed framework. We further note that this approach of building an informative prior to be used by the primary models involves two passes with the data; thus, one could view this as an empirical Bayes-type procedure. Finally, we recognize that pollution sources inevitably contribute to the estimate of the Baseline profile. However, centering is an important step in separating the signals.

3.5 | Profile matching

Source identification is an important step of PMF analysis, which yields profile estimates that are not constrained by the priors for $\mathbf{\Lambda}$. We employ a quantitative approach to profile matching, comparing estimated profiles for each factor to candidate source profiles. In addition to the fact that our candidate profiles were not measured with lichens, we must account for uncertainty in matching profiles as well as differing scales of variation in elemental concentration. To address profile uncertainty, including mismatching sets of reported elements, we use Monte Carlo-based comparisons by simulating profiles from modified priors.

The challenge of differing scales is common in apportionment problems, where discrimination often hinges on signatures of relatively less-abundant elements. To avoid comparisons being dominated by elements with high concentrations, we match profiles using distance metrics designed for compositional data. These metrics are often defined on ratios, which can be sensitive to extremely small values and arbitrary assignment of effective zeros (Aitchison & Egozcue, 2005). We instead base our comparisons on the Hellinger distance, which transforms the space on a polynomial rather than a logarithmic scale. Because these are computed on Monte Carlo samples, we replace the squared deviations typical of Euclidean and Hellinger metrics with absolute deviations to dampen the penalty induced by uncertainty (i.e., mean-squared distances increase with higher uncertainty variances).

To precisely define our chosen profile dissimilarity score, let $\{\hat{\lambda}_{k,\ell}\}$ represent PMF estimates in $\hat{\mathbf{\Lambda}}$, and $\{q^{(s)}_{k',\ell}\}$ denote Monte Carlo samples (indexed by $s = 1, \ldots, S$) of the proportion of element ℓ in the profile for database source k'. The dissimilarity between the estimated profile k and candidate source profile k' is then computed as

$$D^{(s)}(k,k') = \sum_{\ell=1}^{L} \left| \sqrt{\hat{\lambda}_{k,\ell}} - \sqrt{q^{(s)}_{k',\ell}} \right| . \tag{8}$$

This score is then averaged across samples for each (k, k') pair.

4 | SIMULATION STUDY

This section presents a simulation study conducted to assess the performance of the base and sparse BMRMs under different scenarios, including various types of model misspecification and prior settings. Both BMRMs were fit to replicate data sets generated under all combinations of the following two factors: (1) number of sources present and (2) degree of sparsity in source contributions. Additionally, combinations of model-fitting settings were considered: (1) whether one of the sources present in the simulated data was included as a candidate source; (2) the number of generic, template sources with relatively weak profile priors included; and (3) strength of source profile prior distributions. Each of these factors and their levels are summarized in Table 1.

The case where src = 1 and sp = 0% could represent studies conducted over a small geographic area dominated by one pollution source. The case where src = 1 and sp = 50% could apply when the signal from the lone source varies and may not be detectable at certain times or specific locations. It can also be appropriate for geographically wide studies conducted in areas that are free of the pollution sources under investigation, which may be true of several locations in our lichen study. The case where src = 3 and sp = 0% could represent studies conducted over a small geographic area impacted by a number of pollution sources. This scenario is commonly encountered in the source apportionment literature, including those using lichens as bioindicators. The case where src = 3 and sp = 50% could represent studies conducted over small

TABLE 1 Factors and their levels used in the simulation study

Scope	Factor	Short name	Levels		
Data	Number of non-Baseline sources present (K)	src	1	3	
	Sparsity in non-Baseline contributions ($\boldsymbol{F}$)	sp	0%	50%	
Fit	Number of sources in fit (including Baseline)	K_0	3	4	
	Number of extra template sources with weak priors	K_e	0	1	2
	Strength of profile priors (inflation factor on st. dev)	inf	0.2	1	3

or large geographic areas that are potentially impacted by a number of pollution sources, but signals from all sources are not expected at each site/time. We consider this scenario the most similar to the lichen study in this article.

In addition to the Baseline *Rhizoplaca* source, the SPECIATE source Unpvaved Road Dust was always present in the simulations. The Playa and Brake Wear sources were also included when src = 4. Varying the number of underlying sources enables us to assess each model's ability to discriminate among active and extraneous sources. To assess the consequences of under-fitting, we varied the number of sources included in the model fits. When $K_0 = 3$, the fitted model used Baseline, Playa, and Brake Wear, omitting Unpaved Road Dust. Runs with $K_0 = 4$ include Unpaved Road Dust as a candidate source. In addition to database-derived candidate sources, we consider adding $K_\mathrm{e} = 0$, 1, or 2 generic template sources from Section 3.3 to the fits with relatively weak profile priors. When $K_\mathrm{e} = 1$, this is the Natural source, and the second alternate (when $K_\mathrm{e} = 2$) is Anthropogenic. We were particularly interested in learning whether an under-specified BMRM could identify and estimate a source that presents a weak signal in the data, but is not included explicitly as a candidate source.

The last two experimental factors are sparsity, which refers to the fraction of contributions in $\boldsymbol{F}$ that are fixed at zero, and prior strength. Prior strength is quantified by an inflation factor on the standard deviation of the gamma variates used to inform the priors on elements in $\boldsymbol{\Lambda}$. In all, the five factors yield a total of 72 treatment combinations.

The simulations were designed to mimic the conditions for the lichen analysis, with each set consisting of $n = 100$ observations of $L = 25$ elemental concentrations. Simulated data sets used a generative model similar to (4); see Section S4 for a detailed specification. Similar to the lichen analysis, elemental contributions from the Baseline source dominate the median concentrations in $\boldsymbol{F\Lambda}$. To induce the prescribed sparsity when sp = 50, half of the nK nonbaseline contributions were replaced with zeros prior to simulating observations in $\boldsymbol{Y}$. Additionally, 150 values in $\boldsymbol{Y}$ were randomly selected to be missing, and elemental detection limits were defined as the fifth percentile of observed concentrations for each element. Baseline and other source profiles in $\boldsymbol{\Lambda}$ were not simulated, but fixed at values derived from the gamma priors.

The following criteria were used to assess the relative merits of the two BMRMs: error in estimating overall concentration medians ($\boldsymbol{F\Lambda}$); ability to recover a focal source profile; and ability to recover source contributions and discriminate between presence and absence of sources. We used mean-squared error (MSE) across all $\log\left(\sum_{k=0}^{K} f_{i,k}\ \lambda_{k,\ell}\right) - \log\left((\widehat{\boldsymbol{F\Lambda}})_{i,\ell}\right)$ to assess overall estimation, where $(\widehat{\boldsymbol{F\Lambda}})_{i,\ell}$ is a posterior median estimate of $\sum_{k=0}^{K} f_{i,k}\ \lambda_{k,\ell}$. The modified Hellinger divergence, defined in (8), applied to the true λ_k and posterior mean $\hat{\lambda}_k$ (instead of across Monte Carlo samples), was used to evaluate source profile estimation. We calculated mean-squared errors from $f_{i,k}^{0.2} - \hat{f}_{i,k}^{0.2}$ to evaluate estimation of contributions (posterior medians). The power transformation de-emphasizes errors in contributions with large magnitudes.

4.1 | Simulation results

We generated 10 replicate data sets at each combination of the two data factors, src and sp. Each combination of modeling factors was used in fitting the two BMRMs to the resulting 40 data sets. We employed similar model settings to those used for the lichen analysis. Section S4.1 gives further details, including MCMC convergence diagnostics specific to each model-setting group.

Tables 2,3, and 4 report aggregated simulation results on overall MSE, Unpaved Road Dust profile recovery, and Unpaved Road Dust contribution recovery, respectively. We briefly summarize the relative strengths of the two models from these perspectives below. Section S4.1 presents analyses specific to each of the four data-generating scenarios and include plots indicating replicate-data variation in each of the reported criteria (see Figures S1–S8).

Preference between the base and sparse BMRM depends not only on the data scenario, but on the aspect of modeling that motivates the analysis. As expected, the S-BMRM performs well in the presence of sparsity in source contributions. It is also more successful at recovering sources omitted from the candidate pool. However, it is not uniformly preferred when the model is over-specified with respect to candidate sources, as one would expect when using regularization for variable selection in a regression setting.

In terms of matrix decomposition, the S-BMRM generally outperforms the base model (in overall MSE, Table 2) in the presence of *heterogeneous* sparsity patterns in $\boldsymbol{F}$ (sp = 50), while the base BMRM provides superior estimation otherwise. This is true even if the model is over-specified with respect to candidate sources, as well as with the addition of generic sources to compensate for model under-specification. As proposed, the S-BMRM's strength lies in accommodating sparsity across sources and within observations (i.e., sites or times), but not necessarily when a source is missing across

TABLE 2 Overall MSE in $\log(\widehat{\boldsymbol{F}\boldsymbol{\Lambda}})$, averaged over 10 replicate data sets generated at each combination of data and model settings

				sp	0			50		
src	K_0	K_e	**inf**	**model**	**base**	**Sparse**	**(Ratio)**	**base**	**Sparse**	**(Ratio)**
1	3 (misspecified)	0	0.2		0.143	0.132	(0.923)	1.324	0.547	(0.413)
			1		0.050	0.051	(1.029)	0.198	0.066	(0.339)
			3		0.043	0.042	(0.985)	0.187	0.054	(0.295)
		1	0.2		0.068	0.066	(1.005)	0.506	0.112	(0.223)
			1		0.047	0.053	(1.128)	0.243	0.063	(0.280)
			3		0.041	0.047	(1.154)	0.175	0.054	(0.309)
		2	0.2		0.053	0.048	(0.918)	0.190	0.053	(0.277)
			1		0.037	0.044	(1.187)	0.164	0.044	(0.270)
			3		0.037	0.043	(1.142)	0.228	0.046	(0.242)
	4 (over-specified)	0	0.2		0.228	0.355	(1.560)	0.704	0.875	(1.243)
			1		0.049	0.052	(1.080)	1.059	0.813	(0.767)
			3		0.047	0.053	(1.142)	0.941	0.055	(0.079)
		1	0.2		0.055	0.067	(1.221)	0.256	0.082	(0.321)
			1		0.039	0.063	(1.615)	0.163	0.081	(0.504)
			3		0.041	0.064	(1.599)	0.167	0.081	(0.501)
		2	0.2		0.057	0.055	(0.974)	0.354	0.065	(0.211)
			1		0.040	0.047	(1.185)	0.240	0.051	(0.264)
			3		0.041	0.048	(1.175)	0.218	0.048	(0.250)
3	3 (under-specified)	0	0.2		0.152	0.147	(0.971)	8.217	7.511	(0.915)
			1		0.078	0.077	(0.992)	7.662	6.792	(0.887)
			3		0.076	0.076	(0.992)	7.996	7.090	(0.887)
		1	0.2		0.057	0.065	(1.140)	6.937	1.750	(0.251)
			1		0.060	0.072	(1.181)	6.291	1.460	(0.230)
			3		0.060	0.074	(1.254)	7.091	2.875	(0.395)
		2	0.2		0.048	0.058	(1.234)	7.000	1.571	(0.223)
			1		0.048	0.056	(1.164)	6.394	0.775	(0.119)
			3		0.048	0.057	(1.181)	6.426	0.821	(0.128)
	4 (correct sources)	0	0.2		1.624	1.614	(0.994)	10.355	8.462	(0.820)
			1		0.076	0.077	(1.015)	6.753	4.385	(0.652)
			3		0.066	0.069	(1.032)	6.638	4.056	(0.611)
		1	0.2		0.048	0.079	(1.655)	6.782	2.944	(0.426)
			1		0.049	0.079	(1.612)	6.077	1.661	(0.274)
			3		0.050	0.080	(1.596)	6.058	1.623	(0.268)
		2	0.2		0.047	0.061	(1.290)	6.786	1.532	(0.218)
			1		0.047	0.059	(1.248)	6.302	0.244	(0.039)
			3		0.048	0.060	(1.255)	6.332	0.239	(0.038)

Notes: Short names for settings are given in Table 1. The ratio column refers to the average pairwise ratio of MSE from the S-BMRM to MSE from the base BMRM fit to the same data.

TABLE 3 Modified Hellinger divergence between the true Unpaved Road Dust profile and the estimated profile using the Natural generic source when the model is under-specified ($K_0 = 3$), and using Unpaved Road Dust as a candidate source when $K_0 = 4$

				sp	0			50		
src	K_0	K_e	**inf**	**model**	**base**	**Sparse**	**(Ratio)**	**base**	**Sparse**	**(Ratio)**
1	3 (misspecified)	1	0.2		6.30	4.93	(0.78)	17.04	7.66	(0.45)
			1		12.83	12.94	(0.99)	20.71	11.56	(0.56)
			3		15.55	14.90	(0.95)	20.24	13.41	(0.66)
		2	0.2		21.04	17.45	(0.82)	23.09	19.97	(0.86)
			1		18.34	16.68	(0.88)	21.97	19.60	(0.90)
			3		18.59	17.71	(0.93)	22.74	20.21	(0.90)
	4 (over-specified)	0	0.2		1.72	1.06	(0.62)	1.86	1.05	(0.57)
			1		3.05	3.16	(1.04)	7.79	2.38	(0.33)
			3		4.09	3.82	(0.96)	11.15	10.13	(0.98)
		1	0.2		2.32	1.68	(0.71)	3.29	1.90	(0.60)
			1		3.52	3.68	(1.10)	15.00	2.96	(0.20)
			3		4.12	4.29	(1.05)	17.75	4.58	(0.26)
		2	0.2		1.97	1.61	(0.91)	3.66	1.46	(0.43)
			1		3.70	2.98	(0.83)	14.22	2.54	(0.21)
			3		4.55	3.61	(0.83)	17.00	4.31	(0.26)
4	3 (under-specified)	1	0.2		10.73	12.99	(1.21)	12.99	9.74	(0.75)
			1		13.49	14.96	(1.13)	13.25	16.37	(1.27)
			3		17.40	17.58	(1.03)	22.24	17.55	(0.80)
		2	0.2		23.23	19.21	(0.83)	27.85	8.19	(0.29)
			1		24.26	20.87	(0.86)	27.25	11.57	(0.43)
			3		24.44	22.34	(0.91)	27.64	15.36	(0.56)
	4 (correct sources)	0	0.2		1.64	2.27	(1.41)	1.49	0.71	(0.49)
			1		2.95	3.36	(1.13)	4.56	1.67	(0.38)
			3		3.25	3.73	(1.14)	6.18	1.82	(0.35)
		1	0.2		1.11	1.32	(1.19)	1.54	0.76	(0.50)
			1		2.46	2.80	(1.14)	4.61	1.91	(0.46)
			3		2.84	3.19	(1.12)	5.04	2.83	(0.65)
		2	0.2		1.65	2.28	(1.41)	1.36	0.77	(0.59)
			1		2.81	2.94	(1.07)	4.88	1.51	(0.34)
			3		3.28	3.38	(1.04)	6.59	1.64	(0.31)

Notes: Values are averaged over 10 replicate data sets generated at each combination of data and model settings, and scaled up by a factor of 10. Short names for settings are given in Table 1. The ratio column refers to the average pairwise ratio of modified Hellinger divergence from the S-BMRM to the divergence from the base BMRM fit to the same data.

all observations (i.e., when src = 1). To model the latter, global shrinkage parameters (γ_i) should correspond to sources (columns of $\boldsymbol{F}$) rather than to observations (rows).

From the perspective of estimating a present, specific source profile, the S-BMRM tends to outperform the base model (Table 3). A notable exception is when $\boldsymbol{F}$ is dense (src = 3 and sp = 0). Even in this case, however, the sparse model is capable of better utilizing extra generic sources to detect a source missed in modeling. We attribute sharper profile estimation in the S-BMRM to its not needing to allocate every source at every site/time, since forced contribution at every site would tend to encourage profile estimates toward global averages.

TABLE 4 Mean squared error between the true Unpaved Road Dust contributions and the estimated contributions using the Natural generic source when the model is under-specified ($K_0 = 3$), and using Unpaved Road Dust as a candidate source when $K_0 = 4$

				sp	0			50		
src	K_0	K_e	**inf**	**model**	**base**	**Sparse**	**(Ratio)**	**base**	**Sparse**	**(Ratio)**
1	3 (misspecified)	1	0.2		0.14	0.19	(1.36)	6.94	.72	(0.10)
			1		0.58	2.29	(4.36)	7.44	.83	(0.12)
			3		0.76	2.64	(3.74)	7.02	.87	(0.13)
		2	0.2		1.72	6.74	(3.91)	10.47	5.54	(0.56)
			1		1.16	6.22	(5.35)	8.30	4.89	(0.62)
			3		1.13	6.21	(5.75)	9.82	5.28	(0.62)
	4 (over-specified)	0	0.2		0.28	0.71	(2.54)	5.46	2.36	(0.43)
			1		0.61	2.39	(3.98)	6.18	2.48	(0.40)
			3		0.79	2.83	(3.74)	7.87	4.48	(0.59)
		1	0.2		0.52	3.11	(5.89)	6.90	2.20	(0.32)
			1		0.71	3.87	(5.49)	10.51	2.37	(0.22)
			3		0.81	3.94	(4.99)	11.66	2.67	(0.23)
		2	0.2		1.72	6.09	(3.50)	7.62	2.18	(0.30)
			1		1.40	6.57	(4.69)	10.01	2.89	(0.29)
			3		1.38	6.64	(4.86)	10.94	2.95	(0.27)
3	3 (under-specified)	1	0.2		0.42	0.73	(1.74)	12.34	10.28	(0.83)
			1		0.62	0.86	(1.51)	12.24	21.03	(1.73)
			3		0.94	1.26	(2.01)	16.98	16.08	(0.94)
		2	0.2		2.15	2.97	(1.52)	20.11	5.96	(0.30)
			1		2.34	3.91	(2.00)	19.54	6.39	(0.33)
			3		2.33	4.64	(2.28)	19.50	8.80	(0.45)
	4 (correct sources)	0	0.2		0.38	0.89	(2.43)	8.40	3.75	(0.44)
			1		0.59	1.91	(3.31)	8.88	3.65	(0.41)
			3		0.57	1.87	(3.39)	9.36	3.79	(0.40)
		1	0.2		0.43	2.46	(6.26)	8.40	2.97	(0.35)
			1		0.51	2.90	(5.99)	8.89	2.96	(0.33)
			3		0.53	3.01	(5.88)	9.14	2.99	(0.32)
		2	0.2		0.82	3.80	(4.73)	8.24	2.42	(0.29)
			1		0.78	4.53	(5.88)	8.82	2.35	(0.26)
			3		0.79	4.67	(6.00)	9.29	2.37	(0.25)

Notes: Values are averaged over 10 replicate data sets generated at each combination of data and model settings. Contribution values are power transformed (by 0.2). Short names for settings are given in Table 1. The ratio column refers to the average pairwise ratio of MSE from the S-BMRM to MSE from the base BMRM fit to the same data.

The proposed methodology targets the model for $\boldsymbol{F}$, directly impacting estimation of source contributions. As expected, the performance of S-BMRM relative to the base alternative depends on the sparsity level in $\boldsymbol{F}$ (Table 4). Similar to its performance with overall MSE, the S-BMRM does not improve estimation of contributions from a lone source in a over-specified but otherwise non-sparse model; it often successfully detects that single source, but contributions may be biased low. The S-BMRM appears to assist in detection of global source presence only if (1) sparsity exists in contributions across observations or (2) a modeler employs strong profile priors.

5 | LICHEN ANALYSES

Apportionment studies typically identify sources by inspecting both estimated profiles in $\hat{\boldsymbol{\Lambda}}$ and the spatial distribution of estimated contributions in $\hat{\boldsymbol{F}}$. We compare these results and defer details on model fit, settings, and assessment to the Supplementary Materials.

5.1 | EPA PMF analysis

Multiple iterations of the EPA PMF model were run with the number of factors ranging from $K = 3$ to $K = 8$. The percent marginal decrease to an outlier-omitting version of Q is nearly constant with each additional factor, yielding no clear choice of K. The $K = 4$ model returns what are essentially four slight variants of the profile for Baseline *Rhizoplaca*. We proceed with the $K = 6$ run, which includes additional potentially interpretable signals.

Figure 1 shows estimated profiles for each of the six factors. To investigate how these estimates relate to candidate profiles, we performed quantitative profile matching, as described in Section 3.5. Results are sensitive to both the metric used and approach to incorporating uncertainty, and we stress the importance of balancing plausibility of results with the desired properties of the matching method. Table 5 reports the closest three matches for each estimated factor profile against prior and BMRM-posterior estimated profiles. Note that best matches against a restricted set of candidates are not necessarily plausible matches. Graphical comparisons of the estimated and candidate profiles, with Monte Carlo uncertainty, appear in Section S7 as Figures S14–S19. Although Baseline is the closest profile for five of the six factors, differences among more distant matches are instructive.

Estimated contributions, shown in Figure 2, appear to support Factors 2, 5, and 6 as versions of a universal or baseline lichen profile. In all cases, dissimilarity scores increase sharply after the first match. The estimated profiles for Factors 2, 5, and 6 are very similar to the baseline profile, with the exception of omitting one or two elements in each case. These factors account for three of the four highest median contributions, and contribution maps reveal that Factor 2 is ubiquitous.

Factor 1 uniquely matches away from the baseline profile, with elevated levels of Ca, Al, Fe, and Si causing it to resemble the Cement dust candidate profile. This factor concentrates extreme contributions with a few samples from eastern Utah near the South Unit of the Ashley National Forest (Duchesne), which hosts oil and gas extraction activity. Biomonitoring reference sites in the South Unit are almost exclusively located near dirt roads heavily traveled by large trucks and other vehicles. We broadly classify this factor as dust associated with industrial activity (Branquinho et al., 2008).

Factor 4 has a high urban concentration south of Salt Lake City, Utah, which could indicate cement or similar anthropogenic sources. However, it is also prominent in southwestern Wyoming and central-eastern Nevada. These locations are downwind of substantial playas of the Great Basin, suggesting that this factor at least partially contains a strong playa-type aeolian dust signature. However, low contributions of Factor 4 from samples in these areas complicate the interpretation.

Factor 3 is characterized by relatively low Ca and elevated Fe, and does not particularly correlate with any of our candidate profiles. It concentrates in the White River National Forest in Colorado, with sampling sites among the most

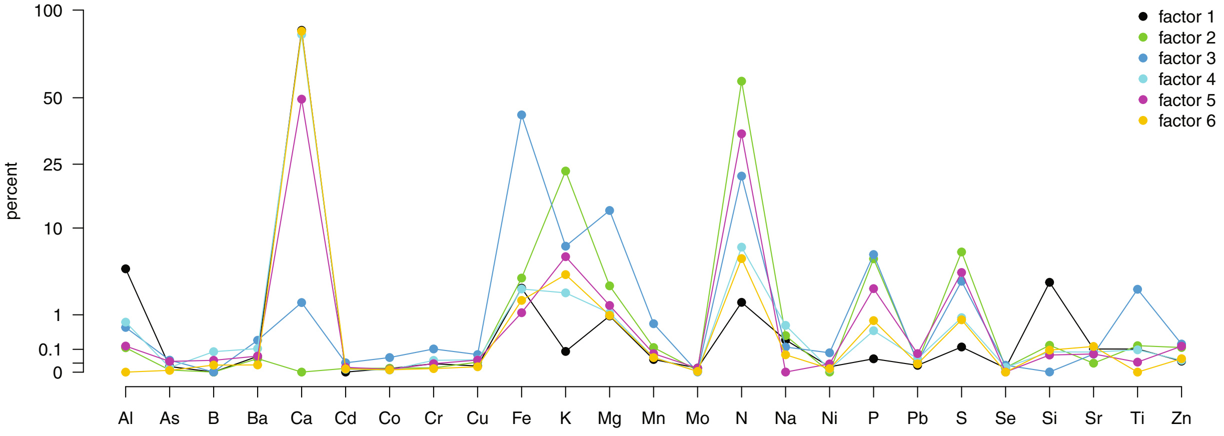

FIGURE 1 Elemental source profiles from the six-factor Environmental Protection Agency positive matrix factorization model fit

TABLE 5 Nearest potential source profiles to each of the six factor profiles estimated from the Environmental Protection Agency positive matrix factorization model, by the modified Hellinger dissimilarity score

PMF Factor	Comparison	1st	2nd	3rd
1	prior	Cement (1.10)	Baseline (1.48)	Playa Dust (1.89)
	base BMRM	Playa Dust (0.78)	Natural (1.01)	Baseline (1.37)
	S-BMRM	Baseline (1.16)	Playa Dust (1.71)	Unpaved Road Dust (2.61)
2	prior	Baseline (1.64)	Highway Dust (3.02)	Playa Dust (3.03)
	base BMRM	Baseline (2.05)	Exhaust (2.08)	Natural (2.49)
	S-BMRM	Baseline (1.98)	Playa Dust (3.00)	Natural (3.24)
3	prior	Baseline (1.96)	Brake Wear (2.43)	Coal Power (2.76)
	base BMRM	Brake Wear (1.92)	Exhaust (2.31)	Baseline (2.31)
	S-BMRM	Baseline (2.22)	Brake Wear (2.24)	Unpaved Road Dust (2.33)
4	prior	Baseline (0.90)	Cement (1.40)	Playa Dust (1.87)
	base BMRM	Natural (0.71)	Baseline (0.79)	Playa Dust (0.82)
	S-BMRM	Baseline (0.55)	Playa Dust (1.73)	Unpaved Road Dust (2.74)
5	prior	Baseline (0.61)	Cement (1.89)	Playa Dust (2.04)
	base BMRM	Baseline (0.69)	Exhaust (1.07)	Natural (1.41)
	S-BMRM	Baseline (0.69)	Playa Dust (1.89)	Natural (3.01)
6	prior	Baseline (0.90)	Cement (1.55)	Playa Dust (1.99)
	base BMRM	Baseline (0.62)	Natural (0.81)	Playa Dust (0.82)
	S-BMRM	Baseline (0.49)	Playa Dust (1.81)	Unpaved Road Dust (2.91)

Notes: Comparisons are made against prior profiles and posterior profiles estimated with the base and sparse BMRMs. Highway Dust and Exhaust refer to SPECIATE sources Paved Road Dust – Highway and Motor Vehicle Exhaust, respectively.

remote and elevated in the collection. Thus Factor 3 may exist as a unique, local adjustment to the other baseline signals, which appears more likely than the Brake Wear and Coal Power Plant alternatives.

5.2 | Bayesian models

We compare the results from the base BMRM and S-BMRM analyses, both of which incorporated profile priors for the following candidate sources: Baseline *Rhizoplaca*, Brake Wear, Motor Vehicle Exhaust, Playa Dust, and Unpaved Road Dust. By strengthening priors on source profiles, we were able to consider adding the two complementary generic sources labeled Natural and Anthropogenic used in the simulation study.

Information-criterion-based model selection (reported in Section S6.3) favors using the log-Student-t likelihood, as well as two extra generic sources ($K_e = 2$), which settings we use exclusively hereafter. Long-tailed distributions of log-transformed residuals for several elements corroborate this preference. Although final inferences from the S-BMRM suggest that the Anthropogenic source does not contribute meaningfully, we proceed with the full model using seven total sources for illustration.

Some estimates, particularly source contributions, are sensitive to the number of sources included in the model. Most notable are changes to the contribution patterns for Brake Wear and Unpaved Road Dust estimated by S-BMRM when a second generic source is added (see Figures S9–S11). We examine sensitivity to model and prior settings in Section S5.

Inferences reported hereafter are based on runs from four MCMC chains, using 15,000 iterations per chain and discarding the first 10,000 draws as burn-in, for a total of 20,000 posterior samples per model. As with the simulation study, we detected slight multimodalities in the posterior distributions, with local modes yielding similar estimates. We report inferences from groups of chains returning positive diagnostics and exploring the clearly dominant modes.

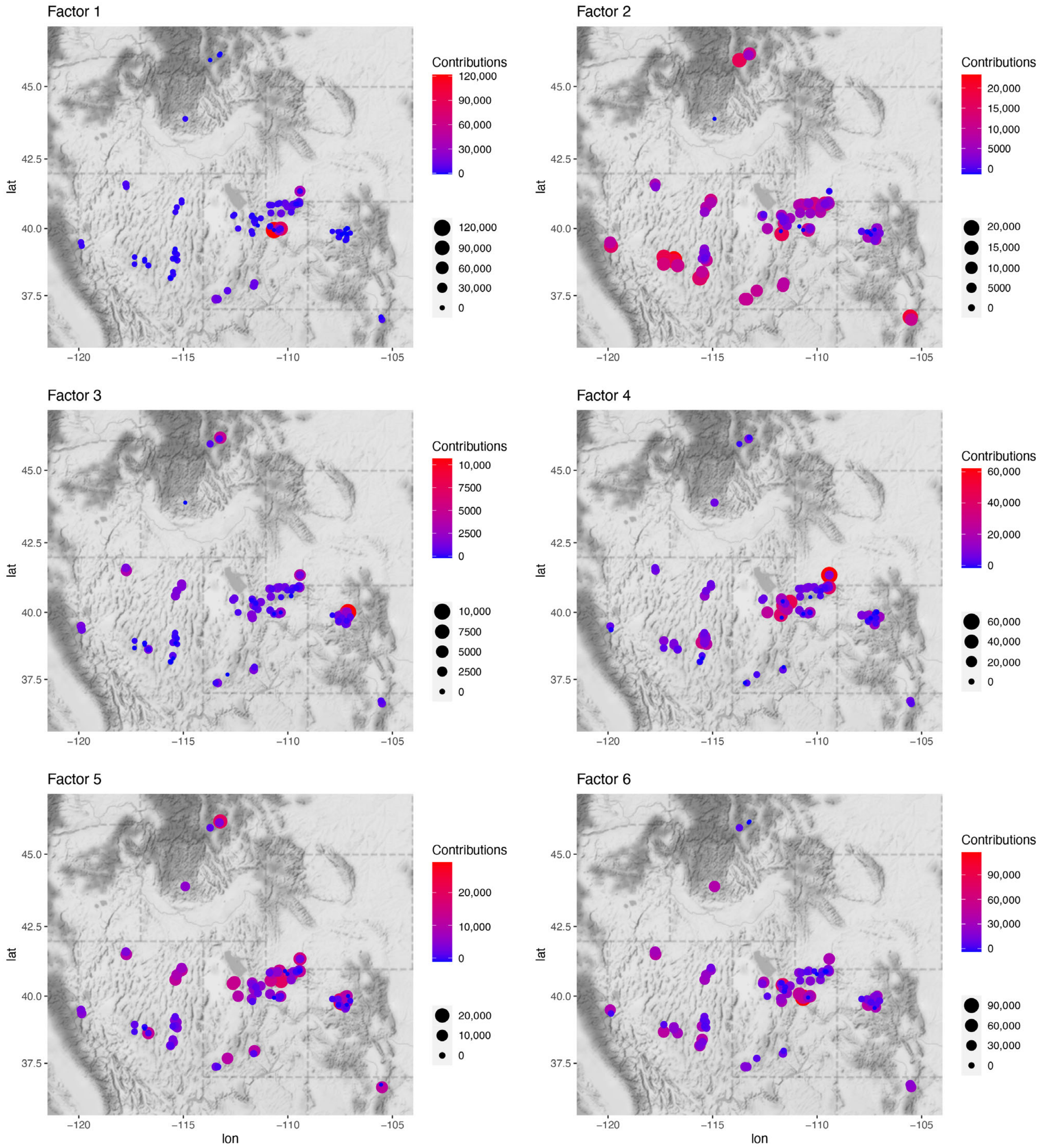

FIGURE 2 Maps of estimated factor (source) contributions from the six-factor Environmental Protection Agency positive matrix factorization model fit. Note the distinct scales for each panel

Bar plots in Figure 3 display, for all sources, prior and posterior median profile estimates with uncertainty. We continue to refer to each factor by using the source name associated with its prior. However, similarity of profiles from distinct sources, and prior-posterior differences in profiles demand broad and inclusive interpretations beyond assigned labels.

Posterior source profile estimates for the base BMRM are similar in some cases to the sparse model estimates. As expected, both models characterize Baseline as being primarily composed of Ca, N, and K, which is also similar to the prior profile. Playa Dust was estimated by both models to be high in Ca; the base model especially characterizes this source as

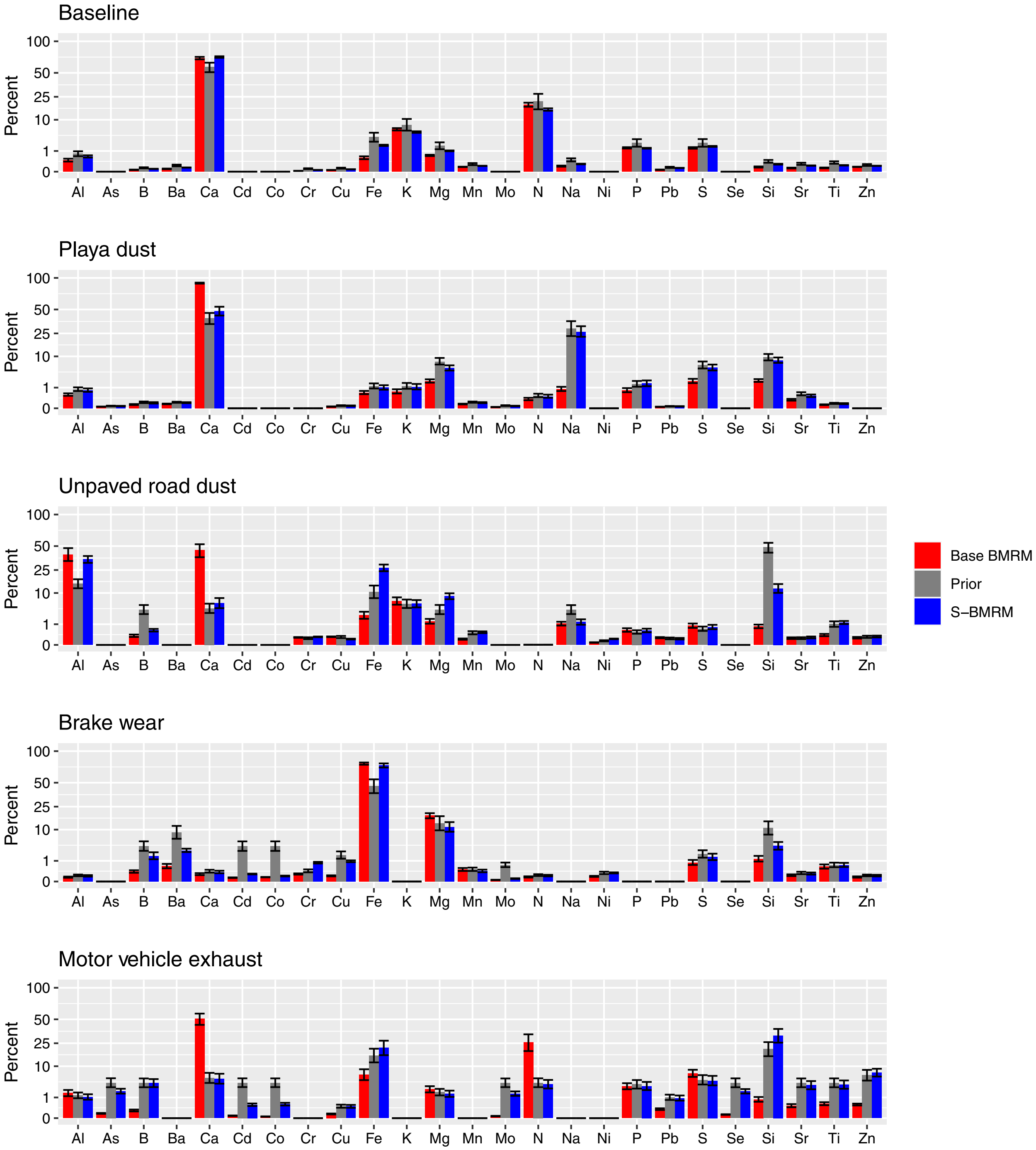

FIGURE 3 Prior (gray, center) and posterior (base Bayesian multivariate receptor model red, left; sparse Bayesian multivariate receptor model blue, right) elemental profiles estimated with the Bayesian models. Uncertainty whiskers extend to the 25th and 75th percentiles.

dominated by Ca, while the S-BMRM estimates are more consistent with the prior. While we expect high concentrations of Ca in the Baseline and Playa Dust profiles, appreciable levels of Na, Si, and N, are needed to differentiate between them. Similarity between these two profiles suggests that dust contributes significantly and universally to elemental profiles of *R. melanophthalma* agg. in the Intermountain Region, and suggests that the profile Baseline does not exclusively reflect lichen secondary chemistry.

The base-BMRM profile for Unpaved Road Dust is dominated by Al and Ca, with moderate levels of Fe, K, and Mg. The sparse model largely replicates the prior profile with respect to estimated concentrations of Ca, Fe, K, and Mg. However, the estimated Al is much higher, while Si is estimated to be much lower than the prior suggests. Overall, the posterior profiles from both models still resemble dust sources.

Brake Wear was estimated by both profiles to be primarily composed of Fe and Mg, consistent with the prior. In contrast, the Motor Vehicle Exhaust profile was estimated by the base BMRM to be composed primarily of Ca and N, a significant departure from the prior, while the S-BMRM estimates resemble the prior profile, being high in Si and Fe.

Generally, the posterior source profile estimates from the sparse model tend to more closely resemble the priors, while posterior estimates for the base BMRM tend to load on one or two elements. Two elements account for more than 75% of the profile composition in five of the seven profiles estimated with the base BMRM, while the same is true for only two of the seven profiles estimated by S-BMRM.

Figure 4 reports maps of the posterior median-estimated contributions for database sources. Estimated contributions for Baseline (not shown) have very similar spatial patterns between the two models, while the S-BMRM estimates are often higher, especially in northern Utah and northern Nevada. In general, the estimated contributions from the other sources are much higher in the base BMRM than the sparse model, likely attributable to shrinkage. Estimated contributions from the S-BMRM appear to exhibit more spatial variability, facilitating discrimination among sites.

Estimated source contributions indicate relatively high levels of aeolian dust sources across the study area, including Playa Dust and Unpaved Road Dust, and potentially the generic Natural source. The ubiquitous Baseline source likewise contains a dust signature. The Brake Wear source is prominent in Colorado and northeastern Nevada, consistent with the Fe-dominated Factor 3 in the PMF analysis. A similar correspondence exists between the BMRM Motor Vehicle Exhaust profiles and contributions and the N-dominated Factor 5 from PMF.

The spatial distribution of the Motor Vehicle Exhaust profile from both models appears to be plausible for true combustion exhaust, including at the most northern sites in Nevada, which are situated immediately east of agricultural complexes. However, low contribution estimates suggest that this signal is weak in the S-BMRM, partially because its dominating elements are generally low-concentration. Note also that samples with low estimated contributions of this source were collected in close (spatial) proximity to those with the highest contributions, including near urban areas where we would expect the highest concentrations of vehicle exhaust. Such differences may be attributable to short ranges in spatial correlation or temporal variation in sampling, including seasonal effects, which was not controlled for in sampling or analysis (see Section S6.4, including Figures S12 and S13). Challenges like this underscore the need to avoid over-interpretation of results.

Figure 5 shows estimated profiles and contribution maps for the two extra generic sources, Natural and Anthropogenic. The South Unit of the Ashley National Forest (eastern Utah) features significant contributions from the generic Natural source, whose map resembles that from Factor 1 in the PMF model. The Anthropogenic source appears to accommodate sites with elevated Al concentration, and is effectively selected out from the S-BMRM fit.

6 | DISCUSSION

Our objectives with the lichen analysis were to (i) investigate whether deposition patterns of areal pollutants are detectable with sensitive indicator lichens sampled across broad spatial and temporal scales, under extremely limited sampling replication; (ii) to compare results from a standard analysis tool, routinely applied to air-monitoring data and with demonstrated success using lichens under more favorable circumstances, and extended Bayesian models; and (iii) to evaluate their suitability in this challenging scenario. Here, we briefly summarize and assess these objectives, and comment on possible future work.

With respect to the first objective, results were limited. All analyses detected signals with weak spatial patterns, with some agreement between them. The most prominent signatures appear consistent with variations of aeolian dust. Specific athropogenic sources were more difficult to identify. Inferences proved somewhat sensitive to modeling choices and influential observations, likely as evidence of model misspecification (i.e., restrictive assumptions, unknown measurement error) and noisy data.

In the present analysis, PMF primarily estimates regional variations of the Baseline *Rhizoplaca* profile. Local anomalies possibly correspond with industrial activity, but the results are not definitive. The BMRM analyses highlight the utility of softly identifying candidate sources, including secondary chemistry, with uncertainty. Estimates from the base BMRM are characterized by malleable profiles and widespread contributions.

In contrast with the other two models, the S-BMRM behaves like a chemical-mass-balance regression model that retains flexibility in source profiles. Without adjusting the strength of profile priors, the S-BMRM more readily allows for

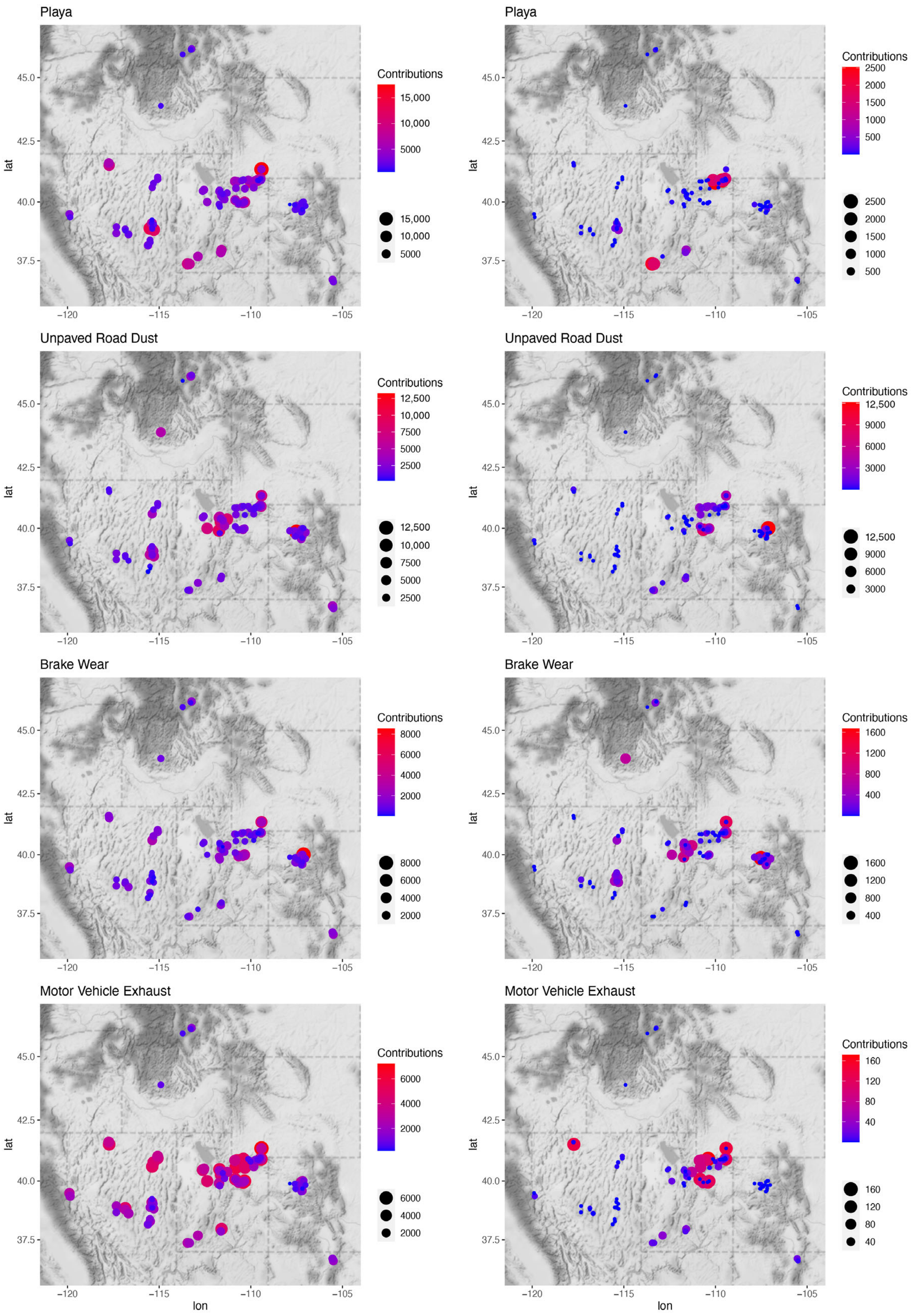

FIGURE 4 Posterior median estimates of source contributions from the base Bayesian multivariate receptor model (left) and sparse Bayesian multivariate receptor model (right). Note the different scales on each panel.

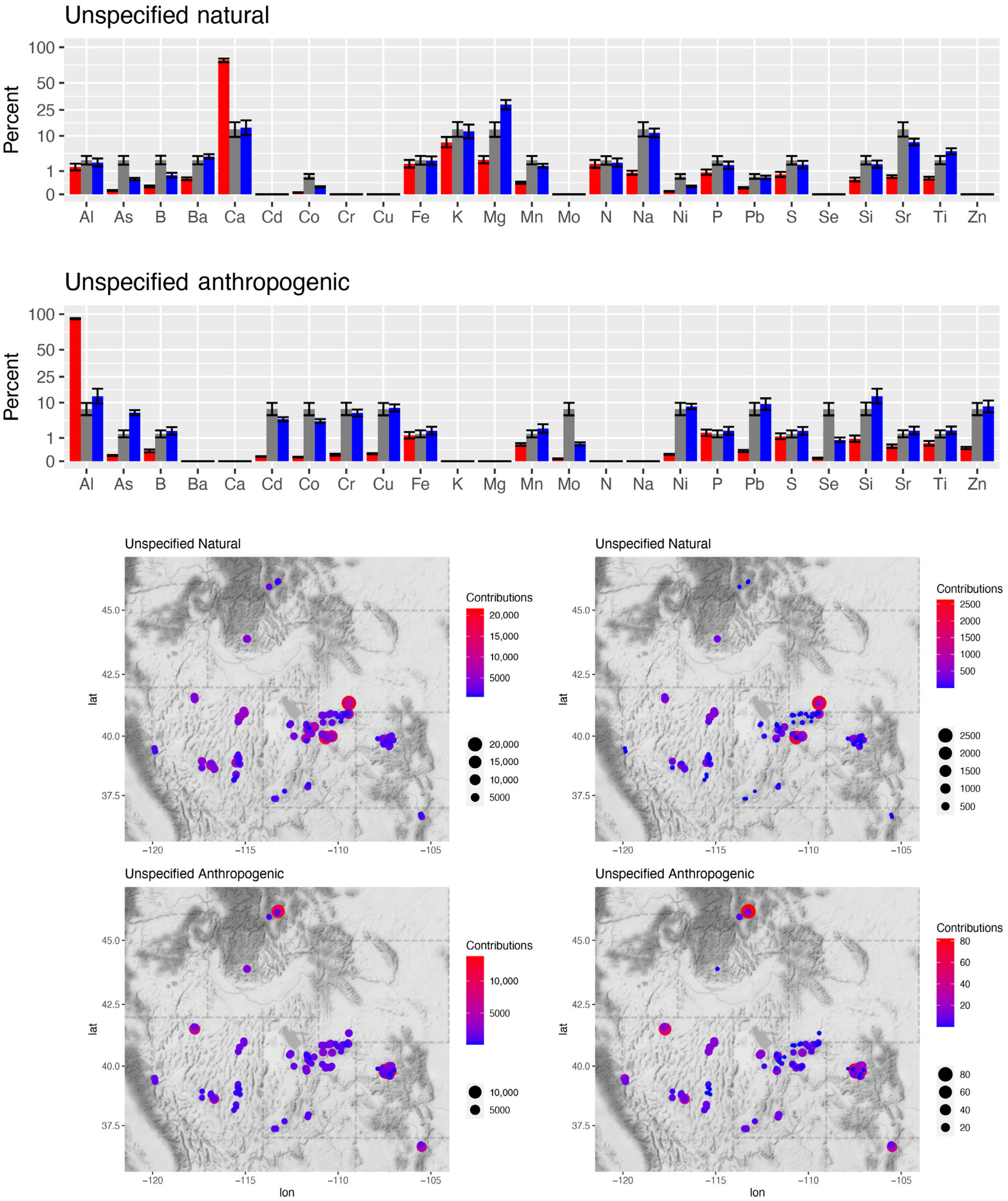

FIGURE 5 Prior (gray, center) and posterior (base Bayesian multivariate receptor model [BMRM] red, left; sparse Bayesian multivariate receptor model [S-BMRM] blue, right) elemental profiles for the two extra generic sources (bar plots), and posterior median estimates of source contributions (maps) from the base BMRM (left) and S-BMRM (right) for the two extra generic sources. Uncertainty whiskers in bar plots extend to the 25th and 75th percentiles. Note the different scales on each panel of maps.

sources to *not* contribute to the decomposition of an observation. The simulation study demonstrates clear advantages of regularizing contribution estimates when samples contain contributions from only a subset of the modeled candidate sources. Strong heterogeneity across the topography of the Intermountain Region, while likely precluding use of stationary process-based spatial modeling, provides a reasonable use case for the S-BMRM.

Although only 10 replicate data sets were generated for each combination of the data factors src and sp, the simulation study employed a full factorial experiment using each level of the factors in Table 1. While using more than 40 total generated data sets would be ideal, the computationally intensive nature of the BMRM and S-BMRM models, together with their repeated use for all combinations of model-fitting settings (18 fits for each data set), makes fitting an increased number of replications impractical. The simulation results nevertheless clearly identify salient differences among methods and settings.

Currently, it is difficult to discern whether measurements are the product of local sampling conditions or greater spatial patterns. Accurate characterization of lichen secondary chemistry is important, but challenging and costly. Within-site and within-sample replication would further sharpen our characterization of the sources of variability, enabling us to (i) estimate the reliability of results and (ii) improve the resolution on estimates of pollutant signatures.

As noted above, the motivating objectives of this research each relate to the evaluation of the feasibility of pollution source apportionment when lichen data are used in place of more conventional measures of ambient pollution. We have demonstrated potential for the proposed Bayesian approaches to yield physically meaningful insights about air pollutants. Notwithstanding, some of the perennial challenges associated with the multivariate receptor modeling of conventional air pollution data are exacerbated when using lichen-based measurements. These heightened challenges include the unique separation of distinct but correlated sources and the identification of the number of sources. The newly proposed Bayesian multivariate receptor models and a careful characterization of a source related to the lichen's physiology provide insights into how source apportionment might be used to exploit widely available lichen data as a means for characterizing historical pollution sources.

ACKNOWLEDGMENTS

We thank Bruce Webb and Rachel Buck for help with ICP analysis, and Benjamin Wright and David Wimmer for help compiling elemental analysis data from technical reports. We acknowledge the USDA Forest Service for funding and logistical support. We likewise thank an associate editor and three anonymous reviewers for helpful comments that led to significant improvements in this work. Figures with maps were generated using the *ggmap* R package (Kahle & Wickham, 2013; R Core Team, 2021).

DATA AVAILABILITY STATEMENT

The data sets analyzed in the current study are available in the `LichenReceptorModels` GitHub repository: https://github.com/mheiner/LichenReceptorModels.

ORCID

Matthew Heiner https://orcid.org/0000-0002-7944-5517
William F. Christensen https://orcid.org/0000-0002-8068-1031

REFERENCES

Agnan, Y., Séjalon-Delmas, N., & Probst, A. (2014). Origin and distribution of rare earth elements in various lichen and moss species over the last century in France. *Science of the Total Environment*, *487*, 1–12.

Aitchison, J., & Egozcue, J. J. (2005). Compositional data analysis: Where are we and where should we be heading? *Mathematical Geology*, *37*, 829–850.

Aznar, J.-C., Richer-Laflèche, M., & Cluis, D. (2008). Metal contamination in the lichen Alectoria Sarmentosa near the copper smelter of Murdochville, Québec. *Environmental Pollution*, *156*, 76–81.

Bhattacharya, A., Pati, D., Pillai, N. S., & Dunson, D. B. (2015). Dirichlet–Laplace priors for optimal shrinkage. *Journal of the American Statistical Association*, *110*, 1479–1490.

Boamponsem, L., Adam, J., Dampare, S., Nyarko, B., & Essumang, D. (2010). Assessment of atmospheric heavy metal deposition in the Tarkwa gold mining area of Ghana using epiphytic lichens. *Nuclear Instruments and Methods in Physics Research Section B: Beam Interactions with Materials and Atoms*, *268*, 1492–1501.

Branquinho, C., Gaio-Oliveira, G., Augusto, S., Pinho, P., Máguas, C., & Correia, O. (2008). Biomonitoring spatial and temporal impact of atmospheric dust from a cement industry. *Environmental Pollution*, *151*, 292–299.

Carpenter, B., Gelman, A., Hoffman, M., Lee, D., Goodrich, B., Betancourt, M., Brubaker, M., Guo, J., Li, P., & Riddell, A. (2017). Stan: A probabilistic programming language. *Journal of Statistical Software, Articles*, *76*, 1–32.

Christensen, W. F., Schauer, J. J., & Lingwall, J. W. (2006). Iterated confirmatory factor analysis for pollution source apportionment. *Environmetrics*, *17*, 663–681.

Contardo, T., Vannini, A., Sharma, K., Giordani, P., & Loppi, S. (2020). Disentangling sources of trace element air pollution in complex urban areas by lichen biomonitoring. A case study in Milan (Italy). *Chemosphere*, *256*, 127155.

Dillman, K. L. (1996). Use of the lichen Rhizoplaca melanophthalma as a biomonitor in relation to phosphate refineries near Pocatello, Idaho. *Environmental Pollution*, *92*, 91–96.

Frühwirth-Schnatter, S., & Lopes, H. F. (2018). Sparse Bayesian factor analysis when the number of factors is unknown. https://arxiv.org/abs/1804.04231

Geiser, L. H., & Neitlich, P. N. (2007). Air pollution and climate gradients in western Oregon and Washington indicated by epiphytic macrolichens. *Environmental Pollution*, *145*, 203–218.

George, E. I., & McCulloch, R. E. (1993). Variable selection via Gibbs sampling. *Journal of the American Statistical Association*, *88*, 881–889.

George, E. I., & McCulloch, R. E. (1997). Approaches for Bayesian variable selection. *Statistica Sinica*, *7*, 339–373.

Goodman, M. M., Carling, G. T., Fernandez, D. P., Rey, K. A., Hale, C. A., Bickmore, B. R., Nelson, S. T., & Munroe, J. S. (2019). Trace element chemistry of atmospheric deposition along the Wasatch front (Utah, USA) reflects regional playa dust and local urban aerosols. *Chemical Geology*, *530*, 119317.

Hackstadt, A. J., & Peng, R. D. (2014). A Bayesian multivariate receptor model for estimating source contributions to particulate matter pollution using national databases. *Environmetrics*, *25*, 513–527.

Henderson-Sellers, A., & Seaward, M. (1979). Monitoring lichen reinvasion of ameliorating environments. *Environmental Pollution (1970)*, *19*, 207–213.

Hodkinson, I. D., & Jackson, J. K. (2005). Terrestrial and aquatic invertebrates as bioindicators for environmental monitoring, with particular reference to mountain ecosystems. *Environmental management*, *35*, 649–666.

Hopke, P. K. (2016). Review of receptor modeling methods for source apportionment. *Journal of the Air & Waste Management Association*, *66*, 237–259.

Kahle, D., & Wickham, H. (2013). ggmap: Spatial visualization with ggplot2. *The R Journal*, *5*, 144–161. https://journal.r-project.org/archive/2013-1/kahle-wickham.pdf

Krall, J. R., & Chang, H. H. (2019). *Ch. 23. Statistical methods for source apportionment*. In A. E. Gelfand, M. Fuentes, J. A. Hoeting, & R. L. Smith (Eds.), *Handbook of environmental and ecological statistics* (pp. 523–546). CRC Press.

Landis, M., Pancras, J., Graney, J., Stevens, R., Percy, K., & Krupa, S. (2012). *Receptor modeling of epiphytic lichens to elucidate the sources and spatial distribution of inorganic air pollution in the Athabasca Oil Sands Region*. In *Developments in environmental science* (Vol. *11*, pp. 427–467). Elsevier.

Landis, M. S., Berryman, S. D., White, E. M., Graney, J. R., Edgerton, E. S., & Studabaker, W. B. (2019). Use of an epiphytic lichen and a novel geostatistical approach to evaluate spatial and temporal changes in atmospheric deposition in the Athabasca Oil Sands Region, Alberta, Canada. *Science of the Total Environment*, *692*, 1005–1021.

Landis, M. S., Studabaker, W. B., Pancras, J. P., Graney, J. R., Puckett, K., White, E. M., & Edgerton, E. S. (2019). Source apportionment of an epiphytic lichen biomonitor to elucidate the sources and spatial distribution of polycyclic aromatic hydrocarbons in the Athabasca Oil Sands Region, Alberta, Canada. *Science of the Total Environment*, *654*, 1241–1257.

Lawrey, J. D., & Hale, M. E., Jr. (1981). Retrospective study of lichen lead accumulation in the northeastern United States. *The Bryologist*, *84*, 449–456.

Leavitt, S., Fernández-Mendoza, F., Pérez-Ortega, S., Sohrabi, M., Divakar, P., Lumbsch, T., & St. Clair, L. L. (2013). DNA barcode identification of lichen-forming fungal species in the Rhizoplaca melanophthalma species-complex (Lecanorales, Lecanoraceae), including five new species. *MycoKeys*, *7*, 1.

Leavitt, S. D., Fernández-Mendoza, F., Pérez-Ortega, S., Sohrabi, M., Divakar, P. K., Vondrák, J., Lumbsch, H. T., & St. Clair, L. L. (2013). Local representation of global diversity in a cosmopolitan lichen-forming fungal species complex (Rhizoplaca, Ascomycota). *Journal of Biogeography*, *40*, 1792–1806.

Leavitt, S. D., & St. Clair, L. L. (2015). *Bio-monitoring in Western North America: What can lichens tell us about ecological disturbances?* In *Recent advances in lichenology* (pp. 119–138). Springer.

Lingwall, J. W., Christensen, W. F., & Reese, C. S. (2008). Dirichlet based Bayesian multivariate receptor modeling. *Environmetrics*, *19*, 618–629.

Loppi, S., Ravera, S., & Paoli, L. (2019). Coping with uncertainty in the assessment of atmospheric pollution with lichen transplants. *Environmental Forensics*, *20*, 228–233.

Nimis, P. L., Scheidegger, C., & Wolseley, P. A. (Eds.). (2002). *Monitoring with lichens—Monitoring lichens NATO Science Series* (Vol. *7*). Kluwer Academic Publishers.

Paatero, P., & Tapper, U. (1994). Positive matrix factorization: A non-negative factor model with optimal utilization of error estimates of data values. *Environmetrics*, *5*, 111–126.

Park, E. S., Spiegelman, C. H., & Henry, R. C. (2002). Bilinear estimation of pollution source profiles and amounts by using multivariate receptor models. *Environmetrics*, *13*, 775–798.

R Core Team. (2021). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing https://www.R-project.org/

Riddell, J., Jovan, S., Padgett, P., & Sweat, K. (2011). Tracking lichen community composition changes due to declining air quality over the last century: The Nash legacy in Southern California. *Bibliotheca Lichenologica*, *106*, 263–277.

Ročková, V., & George, E. I. (2018). The spike-and-slab LASSO. *Journal of the American Statistical Association*, *113*, 431–444.

Rola, K. (2020). Insight into the pattern of heavy-metal accumulation in lichen thalli. *Journal of Trace Elements in Medicine and Biology*, *61*, 126512.

Root, H. T., Jovan, S., Fenn, M., Amacher, M., & Hall, J. (2021). Lichen bioindicators of nitrogen and sulfur deposition in dry forests of Utah and New Mexico, USA. *Ecological Indicators*, *127*, 107727.

Saeki, M., Kunii, K., Seki, T., Sugiyama, K., Suzuki, T., & Shishido, S. (1977). Metal burden of urban lichens. *Environmental Research*, *13*, 256–266.

Seaward, M. (1993). Lichens and sulphur dioxide air pollution: Field studies. *Environmental Reviews*, *1*, 73–91.

Sloof, J. E. (1995). Lichens as quantitative biomonitors for atmospheric trace-element deposition, using transplants. *Atmospheric Environment*, *29*, 11–20.

St. Clair, L. L., St. Clair, S. B., & Newberry, C. C. (1994). *Establishment of a lichen air quality biomonitoring program and baseline in the anaconda pintler wilderness area and adjacent areas west of the anaconda copper smelter* (Technical report), US Department of Agriculture, Forest Service.

St. Clair, S. B., St. Clair, L. L., Mangelson, N. F., & Weber, D. J. (2002). Influence of growth form on the accumulation of airborne copper by lichens. *Atmospheric Environment*, *36*, 5637–5644.

Stan Development Team (2020). *RStan: The R interface to stan*. R package version 2.19.3. https://mc-stan.org/.

Szczepaniak, K., & Biziuk, M. (2003). Aspects of the biomonitoring studies using mosses and lichens as indicators of metal pollution. *Environmental Research*, *93*, 221–230.

U.S. Environmental Protection Agency. (2014). *EPA positive matrix factorization (PMF) 5.0 fundamentals and user guide* (Report No. EPA/600/R-14/108), Office of Research and Development, Washington, DC.

U.S. Environmental Protection Agency. (2019). *SPECIATE version 5.0 database development documentation* (Report No. EPA/600/R-19/098), Office of Research and Development, Washington, DC.

Will-Wolf, S., Geiser, L. H., Neitlich, P., & Reis, A. H. (2006). Forest lichen communities and environment—How consistent are relationships across scales? *Journal of Vegetation Science*, *17*, 171–184.

Will-Wolf, S., Jovan, S., & Amacher, M. C. (2017). Lichen elemental content bioindicators for air quality in upper Midwest, USA: A model for large-scale monitoring. *Ecological Indicators*, *78*, 253–263.

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: Heiner, M., Grimm, T., Smith, H., Leavitt, S. D., Christensen, W. F., Carling, G. T., & St. Clair, L. L. (2023). Multivariate receptor modeling with widely dispersed Lichens as bioindicators of air quality. *Environmetrics*, *34*(3), e2785. https://doi.org/10.1002/env.2785