

A Bayesian nonparametric approach to unmixing detrital geochronologic data

John Tipton · Glenn Sharman · Samuel
Johnstone

Received: date / Accepted: date

1 **Abstract** Sedimentary deposits constitute the primary record of changing
2 environmental conditions that have acted on Earth's surface over geologic
3 time. Clastic material is eroded from source locations (parents) in sediment
4 routing systems and deposited at sink locations (children). Both parents and
5 children have characteristics that vary across many different dimensions, in-
6 cluding grain size, chemical composition, and the geochronologic age of con-
7 stituent detrital minerals. During transport, sediment from different parents
8 is mixed together to form a child, which in turn may serve as the parent for
9 other sediment further down system or later in time when buried sediment
10 is exhumed. The distribution of detrital mineral ages observed in parent and
11 child sediments allows for investigation of the proportion of each parent in the
12 child sediment which reflects the properties of the sediment routing system. To
13 model the proportion of dates in a child sample that comes from each of the
14 parent distributions, we use a Bayesian mixture of Dirichlet processes. This
15 model allows for estimation of the mixing proportions with associated uncer-
16 tainty while making minimal assumptions. We also present an extension to the
17 model whereby we reconstruct unobserved parent distributions from multiple
18 observed child distributions using mixtures of Dirichlet processes. The model
19 accounts for uncertainty in both the number of mineral formation events that
20 compose each parent distribution and the mixing proportions of each parent
21 distribution that composes a child distribution. To demonstrate the model,
22 we perform analyses using simulated data where the true age distribution is

John R. Tipton
Department of Mathematical Sciences, University of Arkansas, Fayetteville, AR, USA
E-mail: jrtipton@uark.edu

Glenn R. Sharman
Department of Geosciences, University of Arkansas, Fayetteville, AR, USA

Samuel A. Johnstone
U.S. Geological Survey, Geosciences and Environmental Change Science Center, Denver,
USA

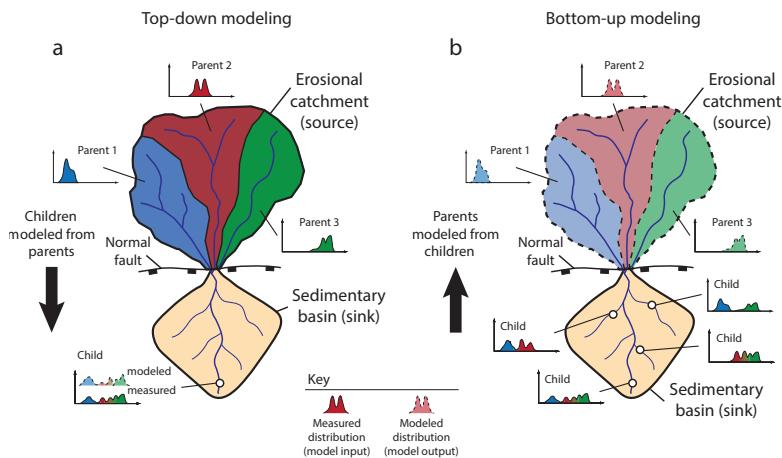


Fig. 1: Schematic depiction of a hypothetical sediment routing system with an erosional source region characterized by three parents (black, red, and green) and an associated sedimentary basin (yellow). The histograms represent the age distributions of detrital minerals from the respective parents and children with lines mapping these distributions to their respective locations in the landscape. Subfigure (a) shows the top-down modeling framework (*sensu* Sharman and Johnstone (2017)) where one or more children are modeled as a mixture of two or more parents. Subfigure (b) shows the bottom-up modeling framework where multiple children are used to reconstruct unobserved end-member sources (parents).

23 known as well as using a real world case study from the central California,
24 USA coast.

25 1 Introduction

26 To understand the origins of modern and ancient landscapes one must un-
27 derstand how erosional processes and associated sedimentary basins evolve
28 through time (Romans et al., 2016). As clastic material is generated by weath-
29 ering and erosion, it is subsequently transported downstream, mixed, and ult-
30 imately deposited into a depositional sink. Statistical modeling of sediment
31 mixing allows for inference about processes that generated the modern land-
32 scape. The ability to decipher the relative proportions of sources that eroded to
33 produce sediment informs understanding of the underlying geologic processes
34 controlling the evolution of the Earth's surface (Stock et al., 2006; Sharman
35 et al., 2019; Mason et al., 2017; Kimbrough et al., 2015).

36 One of the most common ways to characterize the provenance of sediment
37 is detrital geochronology – dating the time at which the individual miner-

als that make up a sedimentary rock formed or cooled. These mineralization events typically reflect the timing of igneous rock forming events or metamorphic alteration of source rocks (Gehrels, 2014). In other cases mineral ages reflect the history of mineral cooling (e.g., ‘thermochronology’, (Reiners and Brandon, 2006)). Detrital geochronologic ages are most commonly determined from measurements of radiogenic isotopes contained within individual mineral crystals. The decay of uranium (U) to lead (Pb) within zircon, a relatively robust mineral, makes this approach ideally suited for tracking sedimentary mixing (Amidon et al., 2005b; Sundell and Saylor, 2017; Sharman and Johnstone, 2017). While the date of a geochronometer from a particular source may be insensitive to the post-formation history of the mineral because of the high closure-temperature of geochronometers (e.g., on the order of 1000 °C for zircon U-Pb (Cherniak and Watson, 2001)), the same can not be expected of low-temperature thermochronometers (e.g., at a typical cooling rate apatite U-Th/He has a closure temperature of ~60 °C (Flowers et al., 2009)). Detrital minerals with low-closure temperatures derived from a single source may have ages that were impacted by the burial of sediments (Fosdick et al., 2015) or progressive exhumation of sources (Garver et al., 1999). In addition, there is increasing recognition that using multiple isotopic approaches to characterize individual minerals, such as measuring zircon chemistry in addition to age (Gehrels, 2014) or both the U-Pb and U-Th/He age of a zircon (Reiners et al., 2005), are critical for discriminating between sources with indistinct U-Pb ages. While new methods can provide increasing specificity to provenance studies, many past and ongoing studies rely on measurements of age alone and thus we focus on developing an approach to better characterize the uncertainty in inferences made from those data, recognizing that the method we present here can be extended to multiple dimensions as the need arises.

We will follow the convention that sediment sources are called *parents* and sink locations are called *children*. Because the statistical model is not defined mechanistically, the definition of the parent source is flexible enough to accommodate different scales. For example, a river could be considered the parent source sediment to a marine basin (as in the natural case study presented later in the manuscript). If rivers are considered the children, then the parent sources may be the upstream tributaries or distinctive bedrock domains within the upstream catchment. Using this language, the manuscript aims to address two questions. First, can we estimate the proportion of each parent age distribution in a child age distribution with associated uncertainty? Second, can we estimate the age distributions for unobserved parents given a set of child age distributions? These questions are answered using “top-down” and “bottom-up” approaches to sediment unmixing, respectively (see Fig. 1; (Sharman and Johnstone, 2017)). The top-down approach (Fig. 1a) models one or more child samples as a mixture of specified parent samples. The bottom-up approach (Fig. 1b) uses multiple child samples to model likely parents which are more generally referred to as end-members in mixture modeling efforts (Weltje, 1997; Paterson and Heslop, 2015).

Bayesian mixture modeling of geologic data, including detrital data, has the advantage of explicitly quantifying uncertainty (Ward et al., 2010; Cooper and Krueger, 2017; Blake et al., 2018). For geochronologic data, Bayesian methods have been used to estimate the age distribution when addressing single samples with probabilistic estimates (Jasra et al., 2006). We extend the work of Jasra et al. (2006) to multiple sample locations by modeling the geologic mixing of sediments derived from source areas containing minerals recording different crystallization events. A key feature of these data are that the age distributions are multimodal due to a range of mineral formation events for a given mineral phase. Thus, an implicit assumption is that each mode in the age distribution is the result of a distinct mineral formation event.

We demonstrate the utility of the Bayesian approach using both a synthetic dataset and a well-constrained, natural case study in central California, USA (Sickmann et al., 2016). In the simulated data, the model is shown empirically to recover the simulated age distributions. After better understanding model performance with simulated data, the model is applied to the case study dataset from central California. The top-down mixing approach is able to successfully reconstruct parent contributions in both the synthetic and natural datasets. Although the bottom-up unmixing model is able to successfully reconstruct parents in the synthetic dataset, there is evidence of non-identifiability – where parents cannot be uniquely characterized from the children – when applied to the natural dataset. Although the focus of this work is on sediment age distributions, the framework presented here can also give guidance about other scientific questions that relate to mixing of non-parametric sum-to-one data in Earth sciences and other disciplines (e.g., unmixing sediment grain size distribution; Weltje (1997) and references within)

2 Model Overview

To define the statistical model, we follow the convention that letters represent data and Greek symbols represent parameters. A plaintext symbol (y/θ) represents a scalar, a bold lowercase symbol ($\mathbf{y}/\boldsymbol{\theta}$) represents a vector, and a bold uppercase symbol ($\mathbf{Z}/\boldsymbol{\Theta}$) is a matrix whose columns are vectors written as $\mathbf{Z} = (\mathbf{z}_1, \dots, \mathbf{z}_p)$ or $\boldsymbol{\Theta} = (\boldsymbol{\theta}_1, \dots, \boldsymbol{\theta}_p)'$ where the appropriate dimensions are implied. We use the notation $[y]$ to represent the probability distribution/mass function (pdf/pmf) and let $[y|\theta]$ represent the conditional pdf/pmf of the random variable y given θ .

Following Berliner (2003), the statistical model described below is divided into three components: the data model, the process model, and the prior model. In general, the data model defines probability distributions that describe the variability in the data due to the measurement process. The data model can be modified to account for non Gaussian measurement processes like counts, outliers, spatial/temporal correlations, etc (Tipton et al., 2017; Hefley et al., 2017). Process models describe the best scientific understanding of the process of interest. For example, process models have been used to

126 describe the monthly response of trees to climate (Tipton et al., 2016), the
 127 relationship between climate and pollen in sediments (Tipton et al., 2019),
 128 and the movement of ice sheets in Antarctica (Chang et al., 2016; Guan et al.,
 129 2018). For the sediment mixing model, the two processes we attempt to cap-
 130 ture with the statistical model are 1) the mineral formation events and 2) the
 131 erosional/weathering, transport, and filtering of grain ages from parent sources
 132 to children sinks. The prior model describes the range of parameter values that
 133 are plausible and completes the formal mathematical definition of the model
 134 by guaranteeing that that posterior distribution is proper (i.e., integrates to
 135 1).

136 **3 Top-down mixing model**

137 The model framework presented below, which is appropriate for situations
 138 where the parent and children sediment have been independently character-
 139 ized, will answer the first research question: can one estimate the proportion
 140 of each parent that comprises the child sediment with appropriate estimates
 141 of uncertainty?

142 **3.1 Top-down mixing data model**

143 Let \mathbf{y} be a n_y -vector of observed age measurements of a single child of inter-
 144 est and let \mathbf{z}_b be a n_b -vector of observed date measurements for each of the
 145 $b = 1, \dots, B$ parents. Because the observed ages include measurement uncer-
 146 tainty reported as a standard deviation, we explicitly account for this source
 147 of uncertainty in the data model. In the case of U-Pb dating of detrital zircon
 148 grains, dates are most commonly determined using laser ablation-inductively
 149 coupled plasma-mass spectrometry (Gehrels, 2012). Such date measurements
 150 typically have relative 2σ analytical precision of 1-4%, with relative uncer-
 151 tainty increasing for younger analyses (Puetz et al., 2018). For each detrital
 152 mineral, the analytical uncertainty (in standard deviations) for the child is
 153 reported as the n_y -vector $\boldsymbol{\sigma}_y$ and for each of the $b = 1, \dots, B$ parents as a n_b -
 154 vector $\boldsymbol{\sigma}_{z_b}$. We assume the date measurement uncertainty follows a Gaussian
 155 distribution where the observed sediment grain date is

$$\begin{aligned} \mathbf{y} | \tilde{\mathbf{y}}, \boldsymbol{\sigma}_y^2 &\sim N(\tilde{\mathbf{y}}, \text{diag}(\boldsymbol{\sigma}_y^2)), \\ \mathbf{z}_b | \tilde{\mathbf{z}}_b, \boldsymbol{\sigma}_{z_b}^2 &\sim N(\tilde{\mathbf{z}}_b, \text{diag}(\boldsymbol{\sigma}_{z_b}^2)), \end{aligned} \quad (1)$$

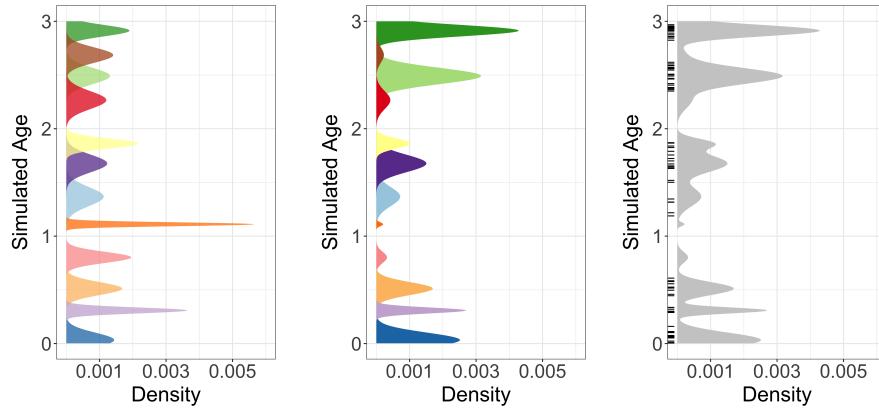
156 where $N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ is a multivariate normal distribution with mean vector $\boldsymbol{\mu}$
 157 and covariance matrix $\boldsymbol{\Sigma}$. The notation $\text{diag}(\boldsymbol{\sigma}^2)$ represents a diagonal covari-
 158 ance matrix with i, i th element σ_i^2 and off diagonal elements all equal to 0. We
 159 break the variable naming convention and let $\tilde{\mathbf{y}}$ and $\tilde{\mathbf{z}}_b$ be latent parameters
 160 that represent the true, unobserved age of the sediments where \mathbf{y} (\mathbf{z}_b) will be
 161 close to $\tilde{\mathbf{y}}$ ($\tilde{\mathbf{z}}_b$) because the measurement uncertainty is small relative to the

variability in the data (i.e., the average coefficient of variation of measured dates, defined as the dating standard deviation divided by the estimated date, is about 2-3%). The reason we account for age dating uncertainty is twofold. First, in a perfect world we could measure the mineral dates exactly; however, in practice our measurements introduce some uncertainty such that our data are approximations of a true, unknown age. Second, because of this, the model can give more weight to dates with less uncertainty and vice versa. In this way, it is possible within the model framework to combine data from different datasets that have different dating uncertainties in a principled manner. In addition, it is possible to account for more uncertainty in the data or to account for asymmetric measurement uncertainties using a Student's-t, log-normal, or other distribution instead of the normal distribution (Jasra et al., 2006). However, these extensions were not explored in the current work to minimize the number of parameters to estimate in this proof of concept model framework.

3.2 Top-down mixing process model

The process model addresses two scientific questions. First, what are the estimates of the true, unobserved detrital mineral age distributions at the parent and child locations? Second, what proportions of those detrital minerals did each parent source contribute to the child? There are many different methods available to model geochronological age distributions from sample data, including kernel density estimation (Vermeesch, 2012), non-negative matrix factorization (Saylor et al., 2019), and Bayesian nonparametric models of mineral formation event mixing (Jasra et al., 2006; Tye et al., 2019). The following section details how our understanding of the geologic processes that generated the observed data inform the development of the statistical model.

Over geologic time, individual minerals may be repeatedly recycled into sedimentary rocks by erosion, transport, deposition, and exhumation. However, in many cases the dates recorded by individual minerals contained in these deposits are distinctive and unaffected by these recycling processes (e.g., excluding burial reheating of low temperature thermochronometers (Fosdick et al., 2015)). We assume that minerals created by the same geologic event share an age distribution that is relatively homogeneous with only small variability. Furthermore, we characterize episodes of rock and mineral formation as punctuated events (typically lasting 10^5 to 10^7 years (Chen and Moore, 1982; Irwin and Wooden, 1999; Wotzlaw et al., 2013)), which are nearly discrete events relative to the age of Earth (approximately 4.5×10^9 years). While minerals often show overgrowths of different ages, this provides a useful approximation. Under the conceptual model (Fig. 1), sediment at each parent is formed by the decomposition of rocks containing minerals created at different times where the potential mineral creation events are shared across parents. Although all parents share the same potential mineral formation events, this does not imply that each parent will actually contain minerals from these po-



(a) Distribution of simulated mineral formation events. Each color represents a different formation event. Notice that some formation events have wider standard deviations (i.e., resulting from longer durations of mineral formation), while other formation events are shorter.

(b) The mineral formation events from Fig. 2a are re-weighted to account for relative abundance of potentially replaced by gray because observable mineral ages for the formation events are unknown. The observed data are shown as a rug plot along the y-axis.

Fig. 2: A cartoon of the mixing model over hypothetical mineral formation events for a single parent distribution. The y-axis of each plot is the age of formation and the x-axis is the probability density of the hypothetical parent distribution.

tential events. The parent sediments are then mixed producing sediment that has the potential to be present at every child location. The model assumes the system is closed so that the sediment at each child comes entirely from the parent sources. Therefore, each child sediment is composed of minerals from the parent sources that are created by an unknown number of mineral formation events at source locations.

A common choice for modeling a mixture of unknown distributions is the Dirichlet process. Figure 2 demonstrates visually an example simulation from the Dirichlet process model. Mathematically, a Dirichlet Process is constructed over an infinite-dimensional mixture of base measures $G(\theta)$ for a family of probability measures $G(\cdot)$ and parameters θ . However, despite the Dirichlet Process being infinite-dimensional, the expected number of clusters scales logarithmically with respect to sample size (Ghosal, 2010) which induces sparsity and makes it easy to numerically approximate the infinite mixture with a finite number of clusters K much, much less than the sample size n . Thus, even though the true number of mineral formation events K recorded by a detrital sample is unknown, the *a priori* expected number of clusters can be

222 determined using the sample size with a fixed K chosen much larger than the
 223 expected number of clusters to ensure a good finite approximation.

224 Figure 2a shows an example simulation from the Dirichlet process prior us-
 225 ing a finite approximation with a mixture of K Gaussian distributions. Each of
 226 the shaded colors is a Gaussian distribution where the each distinct color shade
 227 depicts a discrete mineral formation event. The means and variances differ for
 228 each of the distributions reflecting that formation events can arise from various
 229 geologic processes (Fig. 2a). The transition from Fig. 2a to Fig. 2b represents
 230 the statistical process model which accounts for geological characteristics and
 231 processes including aerial extent, differential erosion, the abundance of min-
 232 erals of different ages within different rocks, and other factors that influence
 233 the proportion of minerals of a given age in sediment at a site (Amidon et al.,
 234 2005a,b). Figure 2b shows the age distributions in Fig. 2a that have been re-
 235 weighted to account for all of the geological factors determining the mixing
 236 proportions of the possible formation events for the parent.

237 Figure 2c depicts a realization from the process model for a given parent.
 238 The labels ($k = 1, \dots, K$) for each mineral formation event are not observed.
 239 Thus, the simulated observation distribution only records the age of the de-
 240 trital minerals in sediment and the color shading is dropped representing this
 241 lack of knowledge about the labels. The data are not observations of the mix-
 242 ture density in Fig. 2c but are actually a finite sample taken from the
 243 mixture which is shown as a rug plot where each tick on the y axis represents
 244 an observed detrital mineral grain date. Thus, the number of mineral forma-
 245 tion events is potentially challenging to extract from the data as neither the
 246 true mineral age density nor the labels that identify the underlying formation
 247 events are known.

248 3.2.1 Modeling parent sediment ages

249 Let the $i = 1, \dots, n_b$ latent detrital sediment grain ages from parent b be rep-
 250 resented by \tilde{z}_{ib} . The sediment grain from which we estimate the latent age \tilde{z}_{ib}
 251 from the observed age z_{ib} is assumed to come from a single mineral formation
 252 event implying the mixture distribution over mineral formation events

$$\tilde{z}_{ib} | \boldsymbol{\mu}_b, \boldsymbol{\sigma}_b^2, \gamma_{ib} \sim \begin{cases} N(\tilde{z}_{ib} | \mu_{b1}, \sigma_{b1}^2) & \text{if } \gamma_{ib} = 1 \\ \vdots & \vdots \\ N(\tilde{z}_{ib} | \mu_{bK}, \sigma_{bK}^2) & \text{if } \gamma_{ib} = K, \end{cases} \quad (2)$$

253 where γ_{ib} is a categorical random variable whose value indicates which
 254 formation event k the detrital mineral comes from. We assume the probability
 255 of a detrital mineral coming from formation event k is $p_{bk} \equiv P(\gamma_{ib} = k)$. Then,
 256 we write the joint distribution over all mineral grains from parent b as

$$\tilde{\mathbf{z}}_b | \boldsymbol{\mu}_b, \boldsymbol{\sigma}_b^2, \boldsymbol{\gamma}_b \sim \prod_{i=1}^{n_b} N(\tilde{z}_{ib} | \mu_{b1}, \sigma_{b1}^2)^{I\{\gamma_{ib}=1\}} N(\tilde{z}_{ib} | \mu_{b2}, \sigma_{b2}^2)^{I\{\gamma_{ib}=2\}} \cdots N(\tilde{z}_{ib} | \mu_{bK}, \sigma_{bK}^2)^{I\{\gamma_{ib}=K\}} \quad (3)$$

where $I\{\gamma_{ib} = k\}$ is an indicator function that takes the value 1 if $\gamma_{ib} = k$ and 0 otherwise. Because there are a large number of indicator functions, we integrate them out of the process model to improve mixing and model fit. The integrated age distribution model is

$$\tilde{\mathbf{z}}_b | \boldsymbol{\mu}_b, \boldsymbol{\sigma}_b^2, \mathbf{p}_b \sim \prod_{i=1}^{n_b} \sum_{k=1}^K p_{bk} N(\tilde{z}_{ib} | \mu_{bk}, \sigma_{bk}^2) \quad (4)$$

where $\mathbf{p}_b = (p_{b1}, \dots, p_{bK})'$ is a vector of positive mixing probabilities with $\sum_{k=1}^K p_{bk} = 1$. To account for uncertainty in the number of formation events K , the probabilities \mathbf{p}_b can be modeled using the Dirichlet process as described in detail in Sect. 3.3. In the model as written, the assumption is that there could be a different set of mineral formation events for each parent. However, as these different events could be jointed to form a superset of all events common across locations, we will assume that all formation events are potentially shared among parents to simplify the fitting algorithm.

3.2.2 Modeling children sediment ages

The process model specifies the proportion of each parent distribution in the child distribution. We represent the mixing proportions of the B parent distributions for the child of interest with the parameter $\phi = (\phi_1, \dots, \phi_B)'$, with $\sum_{b=1}^B \phi_b = 1$ and $\phi_b > 0$. The parameter ϕ_b is the proportion of the child distribution that comes from parent b and accounts for differential mixing of parents. As the statistical model is not mechanistic, the specific, geologic interpretation of ϕ changes based on the context of the sediment transport system. For example, when the parents that are comprised of bedrock, ϕ is a function of each parent's relative aerial extent in the drainage catchment, average erosion rate, and average concentration of the detrital mineral of interest (Amidon et al., 2005a). If parents are sediment inputs (e.g., rivers), then ϕ is a function of each parent's relative sediment supply and the average concentration of the detrital mineral of interest within the sediment.

For a single latent child sediment mineral date \tilde{y}_i , the sediment grain for that mineral comes from only one parent. Using a categorical variable, the distribution of the child sediment grain can be written as a weighted mixture of components (similar to (2)) where each component is a weighted sum of the parent distributions (i.e., the sum is over the K weighted densities in (4) but the mixture density is evaluated with child observations rather than the parent observations). Then, the latent indicator variables can be integrated

290 out of the mixture, similar to (4), giving the integrated child age distribution
 291 model

$$\tilde{y}_i | \{\boldsymbol{\mu}_b, \boldsymbol{\sigma}_b^2, \mathbf{p}_b\}_{b=1}^B, \boldsymbol{\phi} \sim \prod_{i=1}^{n_y} \sum_{b=1}^B \phi_b \sum_{k=1}^K p_{bk} N(\tilde{y}_i | \mu_{bk}, \sigma_{bk}^2),$$

292 where the notation $\{\boldsymbol{\theta}_b\}_{b=1}^B$ denotes the set of parameters $\{\boldsymbol{\theta}_1, \dots, \boldsymbol{\theta}_B\}$.

293 3.3 Top-down mixing prior model

294 The conceptual process model assumes the number of mineral formation events
 295 K is known. In practice, the number of formation events is unknown and is a
 296 parameter to be estimated. In fact, it is likely that the different parent sites
 297 will have different numbers of mineral formation events based on site-specific
 298 history. The prior model addresses the fundamental question of estimating the
 299 number of mineral formation events.

300 There are a variety of potential approaches to model the unknown number
 301 of formation events. First, one can treat the number of formation events as a
 302 fixed parameter, perform a grid search over the different number of formation
 303 events, and choose the model that best fits the data based on some information
 304 theoretic criteria (Miller and Harrison, 2018). A second approach is to sample
 305 over the latent unknown number of formation events using a reversible jump
 306 algorithm (Green, 1995). The third approach is to assign a Dirichlet process
 307 prior over the number of formation events. The Dirichlet process estimates an
 308 unknown number of components without *a priori* specifying the number.

309 The Dirichlet process is an infinite dimensional stochastic process which is
 310 a distribution over distributions (Ferguson, 1973). We assign the range of min-
 311 eral ages for the k th formation event the base probability distribution $G(\boldsymbol{\theta}_{bk})$
 312 which depends on parameters $\boldsymbol{\theta}_{bk}$. There are many possible choices for the
 313 base distribution $G(\boldsymbol{\theta}_{bk})$; we assume a normal distribution $N(\mu_{bk}, \sigma_{bk}^2)$ with
 314 mean μ_{bk} and variance σ_{bk}^2 , therefore $\boldsymbol{\theta}_{bk} = (\mu_{bk}, \sigma_{bk}^2)'$. Other possible choices
 315 include a log-normal or gamma distribution that enforces a positive support
 316 on the observed age dates or a skew-t distribution that allows for asymmetry
 317 in the duration of formation events (Jasra et al., 2006). While other distribu-
 318 tions are possible, and may better capture the effects of natural dispersion in
 319 geochronometers, we rely on normal distributions here to minimize the num-
 320 ber of model parameters and emphasize inclusion of the multiple parent and
 321 children sediments withing the bottom-up unmixing framework. Because we
 322 assume the age distribution of a single mineral formation event is relatively
 323 short with respect to the overall time of interest, the variance parameters that
 324 model the duration of the mineral formation events σ_{bk}^2 will be small relative
 325 to the scale of the observed age distribution. Note that the variance σ_{bk}^2 rep-
 326 resents the process variance due to mineral formation events and is different
 327 than the measurement process variances, σ_y^2 and $\{\sigma_{z_b}^2\}_{b=1}^B$, which are fixed
 328 and known.

329 We use the stick-breaking representation of a Dirichlet process

$$\sum_{k=1}^{\infty} p_{bk} G(\boldsymbol{\theta}_{bk}), \quad (5)$$

330 where p_{bk} are the positive mixing weights with $\sum_{k=1}^{\infty} p_{bk} = 1$. In practice,
331 $p_{bk} \approx 0$ for large k , therefore, the infinite sum is well approximated by the
332 finite sum $\sum_{k=1}^K p_{bk}$ for a large enough K (for most problems $K=10$ or $K=20$
333 is sufficiently large). The stick-breaking representation for \mathbf{p}_b is constructed
334 by transforming auxiliary variables $\tilde{\mathbf{p}}_b = (\tilde{p}_{b1}, \dots, \tilde{p}_{bK-1})'$ using the stick-
335 breaking representation

$$p_{bk} = \begin{cases} \tilde{p}_{b1} & \text{for } k = 1, \\ \tilde{p}_{bk} \prod_{k'=1}^{k-1} (1 - \tilde{p}_{bk'}) & \text{for } k = 2, \dots, K-1, \\ \prod_{k'=1}^{k-1} (1 - \tilde{p}_{bk'}) & \text{for } k = K. \end{cases}$$

336 Priors on the \tilde{p}_{bk} are assigned exchangeable Beta($1, \alpha_b$) priors giving rise
337 to the stick-breaking Dirichlet process. The hyperparameters α_b are given ex-
338 changeable Gamma($1, 1$) priors that control the Dirichlet process concentra-
339 tion (i.e., smaller α_b give fewer formation events, larger α_b give more formation
340 events). Because our study site is constrained geographically, the parent and
341 child sites contain mineral grains derived from common formation events. As
342 such, we follow Lock and Dunson (2015) and used shared kernels by letting
343 $\boldsymbol{\theta}_{bk} \equiv \boldsymbol{\theta}_k = (\mu_k, \sigma_k^2)'$ for all $b = 1, \dots, B$. In theory, there could be no over-
344 lap at all among the formation events, although in this situation it would be
345 possible to choose a K that is large enough such that the number of shared
346 kernels is larger than the total number of formation events among the parents
347 and would result in equivalent inference.

348 The mixing kernel means μ_k are assigned vague, independent $N(\mu_\mu, \sigma_\mu^2)$
349 priors with $\mu_\mu = 150$ Million years (Myr) and $\sigma_\mu^2 = 150^2$ Myr². The standard
350 deviations for the ages of formation are assigned truncated half-Cauchy priors
351 $\sigma_k \sim \text{Cauchy}^+(0, s) I\{0 < \sigma_k < \omega\}$, where we choose s to be small relative
352 to the range of dates observed and ω provides an upper limit to the duration
353 of formation events. For the case study where the majority of dates span the
354 range of 0 to about 300 Myr, we set s to be 25 Myr and set ω to be 50 Myr.
355 The truncation is important to prevent the Dirichlet process mixture from
356 generating unrealistically long formation events which does not match our *a*
357 *priori* geologic knowledge.

358 The mixing parameter $\boldsymbol{\phi}$ is assigned a Dirichlet($\alpha_\phi \mathbf{1}$) prior where $\mathbf{1}$ is a
359 vector of ones and the hyperparameter α_ϕ is assigned a Gamma($1, 1$) prior.
360 When α_ϕ is small the mixing proportions concentrate with a large probability
361 on a single parent component. When α_ϕ is one $\boldsymbol{\phi}$ will be uniformly distributed
362 over all possible mixing proportions. When α_ϕ is large the mixing proportion
363 will be concentrated at equal mixing proportions $(\frac{1}{B}, \dots, \frac{1}{B})$.

³⁶⁴ 3.4 Top-down mixing posterior distribution

³⁶⁵ The top-down mixing model posterior distribution is

$$\begin{aligned} & [\tilde{\mathbf{y}}, \{\tilde{\mathbf{z}}_b\}_{b=1}^B, \boldsymbol{\mu}, \boldsymbol{\sigma}^2, \{\mathbf{p}_b\}_{b=1}^B, \phi, \alpha_\phi, \boldsymbol{\alpha} | \mathbf{y}, \boldsymbol{\sigma}_y, \{\mathbf{z}_b, \boldsymbol{\sigma}_b^2\}_{b=1}^B] \propto \\ & [\mathbf{y} | \tilde{\mathbf{y}}, \boldsymbol{\sigma}_y] \prod_{b=1}^B [\mathbf{z}_b | \tilde{\mathbf{z}}_b, \boldsymbol{\sigma}_b] \times \\ & [\tilde{\mathbf{y}} | \boldsymbol{\mu}, \boldsymbol{\sigma}^2, \{\mathbf{p}_b\}_{b=1}^B, \phi] \prod_{b=1}^B [\tilde{\mathbf{z}}_b | \boldsymbol{\mu}, \boldsymbol{\sigma}^2, \mathbf{p}_b] \times \\ & [\boldsymbol{\mu}] [\boldsymbol{\sigma}^2] [\phi | \alpha_\phi] [\alpha_\phi] \left(\prod_{b=1}^B [\mathbf{p}_b | \alpha_b] [\alpha_b] \right), \end{aligned} \quad (6)$$

³⁶⁶ where each line on the right hand side of the proportional symbol is the
³⁶⁷ data, process, and prior model, respectively. We estimate the posterior using
³⁶⁸ Markov Chain Monte Carlo (MCMC) with the *R* package *NIMBLE* (de
³⁶⁹ Valpine et al., 2017) using an adaptive block Metropolis-Hastings algorithm
³⁷⁰ (Haario et al., 2001). The constrained auxiliary variables $\tilde{\mathbf{p}}_b$ and standard
³⁷¹ deviations $\boldsymbol{\sigma}^2$ are transformed to unconstrained support (logit- and log-scale
³⁷² transformations) for tuning the Metropolis-Hastings block proposals, with cor-
³⁷³ responding Jacobian adjustments to the acceptance probabilities. The sam-
³⁷⁴ pling of the sum-to-one mixing proportion ϕ is performed by introducing aux-
³⁷⁵ illary variables $\tilde{\phi}$, assigning a stick-breaking prior on $\tilde{\phi}$, then sampling on a
³⁷⁶ logit-scale after correcting for the transformation using the Jacobian to induce
³⁷⁷ a Dirichlet($\alpha_\phi \mathbf{1}$) prior on ϕ .

³⁷⁸ 4 Bottom-up unmixing model

³⁷⁹ The second research question is: can we reconstruct unobserved parent age
³⁸⁰ distributions from multiple child observations? In previous work, this analysis
³⁸¹ has been variably termed “end-member mixing analysis”, “end-member mod-
³⁸²eling”, or “end-member analysis” as applied to unmixing grain size or detrital
³⁸³ age distributions (Sharman and Johnstone, 2017; Saylor et al., 2019). The end-
³⁸⁴ member unmixing model has two components. First, the number of parents B
³⁸⁵ is unknown and needs to be estimated. Second, given the number of parents
³⁸⁶ B , what are the unobserved mineral formation age distributions for the B
³⁸⁷ parents? For this paper, we assume the number of parents B is known. There
³⁸⁸ are a number of criteria for selecting the number of parents including using
³⁸⁹ Bayesian information criteria, reversible jump MCMC (Jasra et al., 2006), as-
³⁹⁰ suming a Dirichlet process over the number of parents, or fitting a mixture of
³⁹¹ finite mixtures (Miller and Harrison, 2018). Rather than explore these ideas,
³⁹² we devote our effort on developing the unmixing model for a fixed number of
³⁹³ parents. The end-member model uses the same general framework presented
³⁹⁴ in the mixture of Gaussians model (6) with some modifications.

395 4.1 Bottom-up unmixing data model

396 Let $d = 1, \dots, D$ index the D child sediments that are each composed of $i =$
 397 $1, \dots, n_d$ samples. As before, we assume a Gaussian measurement distribution
 398 for child d given by

$$\mathbf{y}_d | \tilde{\mathbf{y}}_d, \boldsymbol{\sigma}_d^2 \sim N(\tilde{\mathbf{y}}_d, \text{diag}(\boldsymbol{\sigma}_d^2)),$$

399 where $\tilde{\mathbf{y}}_d$ is the true, unobserved n_d -vector of sediment dates and $\boldsymbol{\sigma}_d$ is a
 400 fixed and known n_d -vector of reported dating standard deviations.

401 Unlike in the top-down mixing model above, none of the parent **zs** are
 402 observed. Hence, the parent distributions are estimated entirely using child
 403 sediment observations. Assuming a fixed and known number of parents B , the
 404 bottom-up process model for a mineral grain age arising from the d th child is

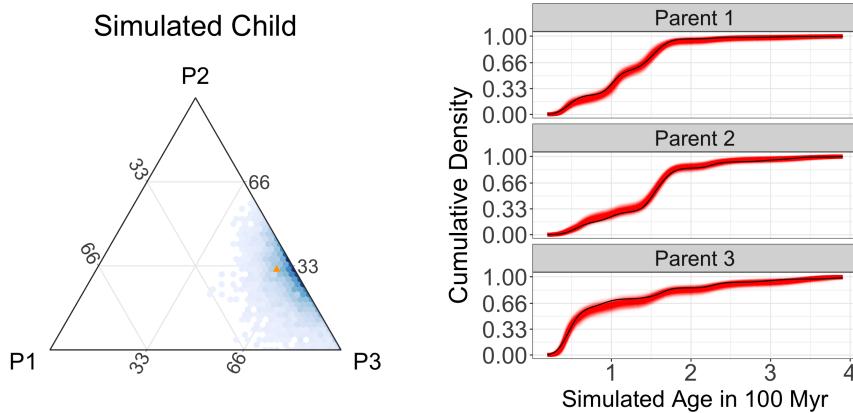
$$\tilde{y}_{id} | \boldsymbol{\mu}, \boldsymbol{\sigma}^2, \{\mathbf{p}_b\}_{b=1}^B, \boldsymbol{\phi}_d \sim \sum_{b=1}^B \phi_{db} \sum_{k=1}^K p_{bk} N(\mu_k, \sigma_k^2), \quad (7)$$

405 where, like before, we assume a Gaussian mixing distribution using shared
 406 kernels across the B parents. The B -dimensional mixture proportion $\boldsymbol{\phi}_d =$
 407 $(\phi_{d1}, \dots, \phi_{dB})'$ models the proportion of the d th child sediment that can be
 408 attributed to each of the B parents. Like the top-down mixing model, these
 409 equations can be derived by introducing categorical random variables then
 410 marginalizing out the latent indicator variables from the model. The b th un-
 411 known parent age distribution is given by $\sum_{k=1}^K p_{bk} N(\mu_k, \sigma_k^2)$. The prior model
 412 for the bottom-up unmixing model is the same as for the top-down mixing
 413 model, except for the variables have different dimensions.

414 4.2 Bottom-up unmixing posterior distribution

415 The posterior distribution that we estimate with the end member unmixing
 416 model is

$$\begin{aligned} & [\{\tilde{\mathbf{y}}_d\}_{d=1}^D, \boldsymbol{\mu}, \boldsymbol{\sigma}^2, \{\mathbf{p}_b\}_{b=1}^B, \{\boldsymbol{\phi}_d\}_{d=1}^D, \boldsymbol{\alpha}_\phi, \boldsymbol{\alpha} | \{\mathbf{y}_d, \boldsymbol{\sigma}_d^2\}_{d=1}^D] \propto \\ & \prod_{d=1}^D [\mathbf{y}_d | \tilde{\mathbf{y}}_d, \boldsymbol{\sigma}_d] \times \\ & \prod_{d=1}^D [\tilde{\mathbf{y}}_d | \boldsymbol{\phi}_d, \{\mathbf{p}_b\}_{b=1}^B, \boldsymbol{\mu}, \boldsymbol{\sigma}] \times \\ & [\boldsymbol{\mu}] [\boldsymbol{\sigma}^2] \left(\prod_{b=1}^B [\mathbf{p}_b | \alpha_b] [\alpha_b] \right) \left(\prod_{d=1}^D [\boldsymbol{\phi}_d | \alpha_{\phi d}] [\alpha_{\phi d}] \right), \end{aligned} \quad (8)$$



(a) Ternary plot showing posterior mixing proportion estimates shaded relative to posterior density in blue and the simulated true mixing proportions as an orange triangle.

(b) Plot of simulated parents with fitted posterior CDF estimates in red and the simulated true CDF in black. Each red line represents a posterior sample of the cumulative parent age density.

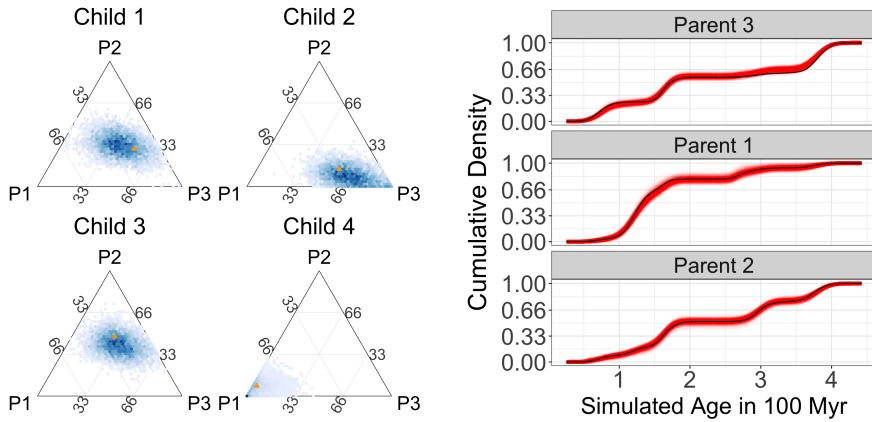
Fig. 3: Results for the top-down mixture modeling approach based on simulated data. As can be seen in the figures, the top-down mixture model is performing well in estimating the simulated mixing distributions.

417 where the priors are the same as those in (6) except that there are now D
 418 children which implies there are now $D \alpha_{\phi ds}$. Likewise, the MCMC algorithm
 419 is the same as presented in Sect. 3 except for a change in dimension for some
 420 parameters.

421 5 Simulation of synthetic detrital age distributions

422 We explore the performance of the model using a synthetic detrital age dis-
 423 tribution dataset. The aim of the simulation study is to understand how the
 424 model performs using realistic data and verify the model is capable of recov-
 425 ering the simulated parameters. The simulation study framework can also be
 426 used to understand how uncertainty in estimation varies with respect to sample
 427 size, variability in the data, and consequences of prior assumptions, although
 428 these details are not explored in this work (Vehtari et al., 2017). For example,
 429 by simulating data with smaller sample sizes than observed, the impact on the
 430 uncertainty estimates can be explored and perhaps used to guide sample size
 431 recommendations for the collection of new data. Similarly, simulations with
 432 different observation errors can be used to quantify the amount of analytic
 433 precision needed for an inferential question analogous to how sample size and
 434 power calculations can be used in experimental design.

435 For the simulation study, a synthetic dataset is created using the top-down
 436 mixing model in (6) for $B = 3$ parents and a single child (Fig. 1a). The parent



(a) Posterior estimates of mixing proportions for 4 of the 20 children from the unmixing model shown. The blue shading is relative to posterior density and the simulated true mixing proportions are shown as orange triangles.

(b) Posterior estimates of the unobserved parent CDFs in red. The simulated parent CDF is shown in black.

Fig. 4: Results for the bottom-up, end-member unmixing model using simulated data. The bottom-up unmixing model does a good job of estimating the true, unobserved parent age CDFs despite the model not using any simulated parent data.

distributions were composed of 200, 250, and 150 simulated sediment grains, respectively, and the child distribution was composed of 150 sediment grains. In simulation, we used dating uncertainties (σ_y, σ_z) that were about 1-3% of the total range of the simulated age distribution of 0-400 Myr. These are similar to measurement uncertainties in the case study and demonstrate the model is capable of accounting for measurement uncertainty.

The posterior samples for the mixing proportion ϕ are shown in Fig. 3a in hexagonal bins with blue shading proportional to the posterior density (Hamilton, 2018), and the simulated mixing proportion is represented by the orange triangle. The simulated mixing proportion (orange triangle) lies within the region of high posterior density demonstrating that the model is accurately estimating the mixing proportions. Figure 3b shows the estimated cumulative distribution functions (CDFs) with posterior samples in red and the simulated CDF in black. The results in Fig. 3 demonstrate that the model is accurately estimating the simulated mixing proportions ϕ as well as the parent age distributions, validating the effectiveness of the top-down mixing model to recover simulated parameters of interest.

The second simulation generated data from the bottom-up, end-member unmixing model (Fig. 1b) to test how well the proposed framework can reconstruct unobserved parent distributions from a set of child observations. For the simulation, we used $B = 3$ parents and $D = 20$ children where each

458 child consisted of 250 measured sediment grains following the model in (8).
459 The range of simulated mixing proportions was simulated uniformly over the
460 three-dimensional simplex resulting in some end members being close to pure
461 end members (e.g., close to 90% of sediment grains coming from a single par-
462 ent). The dating uncertainties (σ_y) were set at about 1-3% of the total range
463 of the simulated age distribution of 0-400 Myr.

464 Figure 4a shows the posterior samples for the mixing proportion of four
465 representative child samples in hexagonal bins with shading in blue propor-
466 tional to posterior density and the orange triangle at the simulated mixing
467 proportion. Based on the simulation study, the model can recover the mixing
468 proportions in this simulation example with high accuracy as the orange tri-
469 angle is within regions of high posterior mass. Even though the model uses
470 none of the data from the parents, the bottom-up unmixing model produces
471 reasonable end-member parent age distribution estimates. Figure 4b shows the
472 estimated CDF produced by the bottom-up unmixing model which shows the
473 model is estimating the unobserved parent distributions. However, the preci-
474 sion for the bottom-up unmixing model is lower in the bottom-up unmixing
475 simulation relative to the top-down mixture simulation despite the bottom-
476 up umixing simulated data containing many more sediment grain samples.

477 6 Application to a Natural Case Study

478 We apply the top-down mixing and bottom-up unmixing models to a well-
479 constrained modern dataset from the central California coast (Sickmann et al.,
480 2016) shown in Fig. 5. Following the same mixing framework presented in Shar-
481 man and Johnstone (2017), there are five samples (river and beach sediment)
482 used to characterize three distinct sediment inputs (parents) to the region,
483 each with a distinct detrital age distribution. Parents 1 and 2 (P1 and P2) are
484 comprised of river samples (CAR and SAL) that represent sediment sources
485 along the Big Sur coastline and Salinas River drainage, respectively. Parent 3
486 (P3) is comprised of two river samples (SNR and PAR) and one beach sample
487 (ANB) that represent northern sediment sources in the Santa Cruz Mountains
488 and western Diablo Range (Sickmann et al., 2016; Sharman and Johnstone,
489 2017). Twelve child samples (beach and submarine canyon sediment) are used
490 to characterize how these parents are mixed in littoral and marine environ-
491 ments. In total, this dataset (Fig. 6) consists of 4,026 individual detrital zircon
492 U-Pb analyses, with individual samples having 82 to 316 analyses each (me-
493 dian of 290 analyses per sample) (Sickmann et al., 2016). The age range of the
494 case study data covered approximately 80-3000 Myr; however, there were very
495 few observation older than 300 Myr (approximately 6% of the data) so even
496 though the model was fit to the entire age range, the presentation of results
497 focuses on the period 0-300 Myr to better interpret the results.

498 We first examine the top-down mixture model (Fig. 1a). Figure 7a shows
499 the reconstruction of the mixing proportions for a sample from a submarine
500 canyon (T693) modeled as a mixture of the three specified parent distribu-

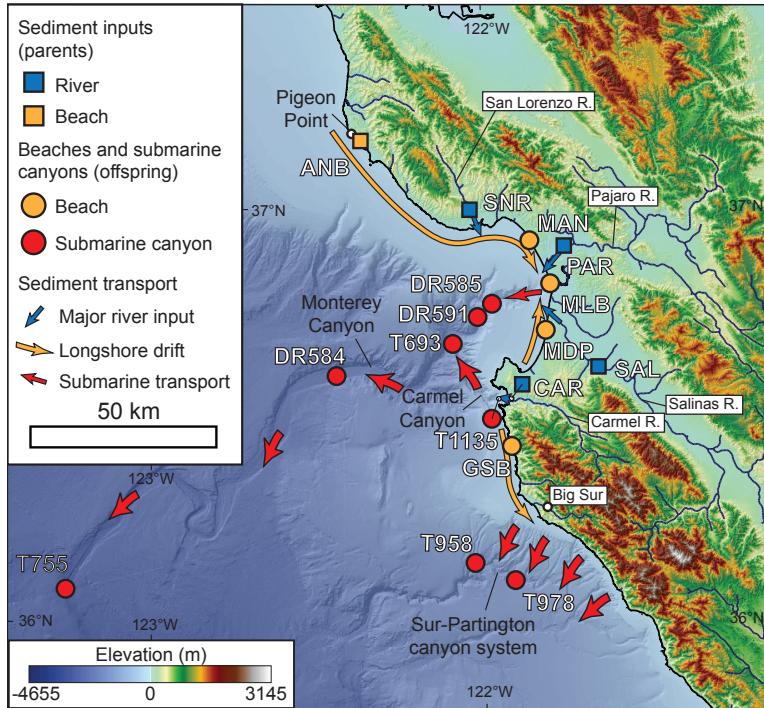


Fig. 5: Locations of the parents and children data for the study region in California, USA (Sharman and Johnstone, 2017).

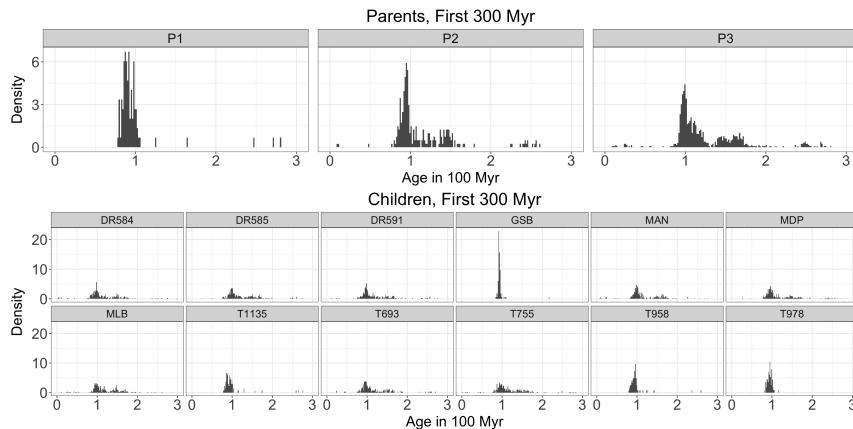
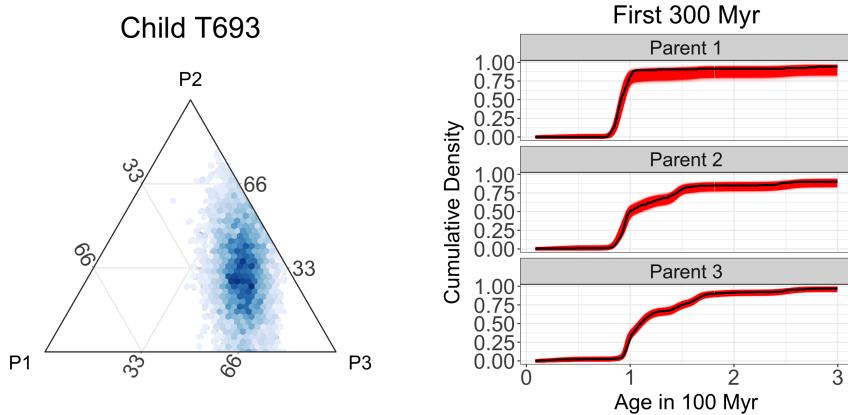


Fig. 6: The sediment age data for the study region in California, USA used for the mixing and unmixing models (Sickmann et al., 2016). The three parent age distributions are shown in the top plot and the 12 child age distributions are shown in the bottom plot. The x -axes represent the measured age in 100 Myr and the y -axes shows the empirical density.

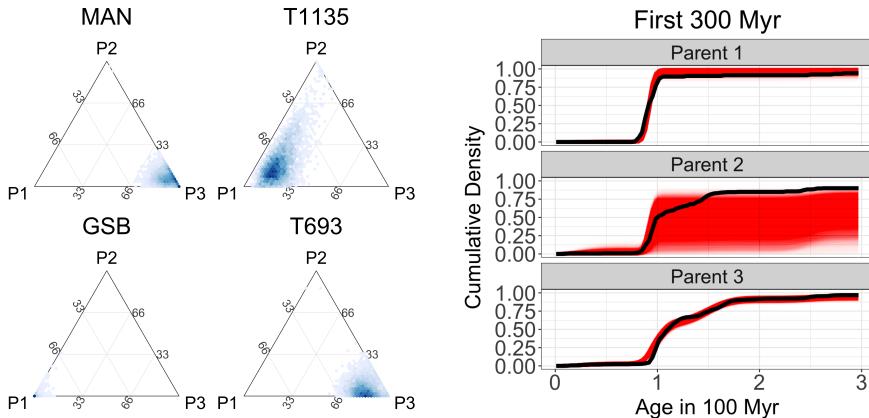


(a) Ternary plot showing posterior density estimates of mixing proportions. The relative posterior density is shown in blue shading. (b) Posterior estimates of the parent and child CDFs shown in red. The empirical CDFs calculated from the raw data are shown in black.

Fig. 7: Results from the top-down mixing model using data from the study region in California, USA. The model results, when applied to sample T693, demonstrate that the top-down mixing model is able to accurately reconstruct the parent and child distributions and produce estimates of the mixing proportions with associated uncertainty.

tions (P1-P3). Visual inspection of the histograms of the data (Fig. 6) would suggest that this child sample is a mixture composed mostly of P3 with some contribution from P2. The posterior estimates of the mixing proportion of each parent for child T693 confirms that the primary component of the mixture is from the P3 with parent P2 as the secondary component (Fig. 7a). Figure 7b shows the model is capturing the basic patterns in the parent CDFs as the estimated CDFs are very close to the empirical CDFs for the parents.

The bottom-up, end-member unmixing model results for the case study data are shown in Fig. 8. The posterior density for the mixing proportions are shown for four children in Fig. 8a as hexagonal bins with shading in blue proportional to posterior density. The posterior distribution of CDFs for the reconstructed parents in Fig. 8b show that the model is very uncertain about the age distribution for parent P2. This is not a totally unsurprising result as the age distributions for parent P2 and parent P3 are very similar and thus are difficult to tease apart in an unsupervised learning situation like that in the bottom-up unmixing model. In Fig. 8a the mixture probability for child T693 is more concentrated at P3 when compared to the top-down mixing proportion estimates in Fig. 7a. Because the model expresses uncertainty about the distribution of parent P2, the difference in estimation between the top-down and bottom-up mixing proportions for child T693 can be attributed to this lack of identifiability among parents P2 and P3. Similar to how the



(a) Posterior estimates for the mixing proportions of each parent for four child sediments. Notice that without observing the parents, the posterior distribution of mixing proportions for child T693 is generally similar to the top-down mixing model in Fig. 7a but has a slightly different shape.

(b) Posterior estimates for the unobserved parent cumulative distribution functions shown in red over 0–300 Myr. The black lines show the empirical cumulative distribution functions.

Fig. 8: Results of the end-member unmixing model fit to data from the study region in California, USA. These figures show that the end-member unmixing model is estimating the parameters of interest, but with some inaccuracies due to a lack of identifiability. However, these issues are easily identified by the end user due to the large amount of posterior uncertainty which provides a check on overly strong inferential claims.

522 important variables in a linear regression can change with small perturbations
 523 of the data when the covariates are highly co-linear (i.e., in models with a
 524 high variance inflation factor), the attribution of the child sediment to a given
 525 parent would change which parent distribution (P2 or P3) is estimated with
 526 accuracy and which parent shows evidence of non-identifiability. Thus, the non-
 527 identifiability manifests as large posterior uncertainty in the CDF for parent
 528 P2 and provides a caution about being overly confident in the inference about
 529 the bottom-up unmixing model.

530 A large overlap in the distribution of parent ages is a feature that often
 531 occurs in detrital zircon geochronology studies. The preservation of zircons
 532 through multiple cycles of erosion and re-sedimentation means that overlapping
 533 zircon ages will be present in many rocks. For parent age distributions
 534 that are quite similar to one another, the reconstruction of the unknown parent
 535 distributions suffers from weak identifiability. In these situations, the es-
 536 timated parents jointly contain all of the formation events, but the model is
 537 unable to attribute the formation events to the correct parents. In other words,
 538 while the model identifies the correct age components, the model sometimes
 539 struggles to correctly group these components into the correct parent dis-

tributions. This aliasing effect is not an unexpected result because Bayesian nonparametric models are well understood to suffer from non-identifiability issues (Ferguson, 1983; Diebolt and Robert, 1994; Richardson and Green, 1997; Frühwirth-Schnatter, 2006).

Non-identifiability is inherent in all end-member unmixing models (Weltje and Prins, 2007). To overcome the non-identifiability in other modeling frameworks, a potential solution is to impose constraints on the end-members and/or provide informative initial conditions for maximum likelihood optimization algorithms (Donoho and Stodden, 2004; Miao and Qi, 2007; Chen and Guillame, 2012). Therefore, any end-member unmixing model that uses only child age distributions will have issues in accurately reconstructing the parent distributions if the assumption of the constraints is not met (i.e., the parent age distributions are structurally similar). Bottom-up unmixing models provide a useful way to explore large detrital datasets with unknown sedimentary sources. Our proposed model framework provides a way to identify those datasets that either are or are not susceptible to non-identifiability by producing uncertainty estimates that are larger when the model is weakly identifiable (Fig. 8b, parent 2). Thus, the uncertainty intervals are a useful diagnostic check for identifiability.

Direct, probabilistic estimates of uncertainty and the ability to calculate derived quantities with uncertainty is a benefit of the proposed method and of Bayesian methods in general. Thus, we can answer questions like what is the probability that at least 50% of child sediment T693 comes from parent P3 using the top-down mixing model applied to the case study data? The answer is calculated directly from the posterior samples using the Monte Carlo approximation $\frac{1}{L} \sum_{\ell=1}^L I\{\phi_3^{(\ell)} \geq 0.5\} = 0.672$, where $\ell = 1, \dots, L$ are the indices of the MCMC samples and $\phi_3^{(\ell)}$ is the estimated mixing proportion for the ℓ th MCMC iteration. The probability that at least 50% of the child sediment comes from parent P3 and at least 25% comes from parent P2 is $\frac{1}{L} \sum_{\ell=1}^L I\{\phi_3^{(\ell)} \geq 0.5\} \times I\{\phi_2^{(\ell)} \geq 0.25\} = 0.283$. Another question of interest that can be expressed in term of the model is “How many mineral formation events occurred for each parent?” We assumed that there was an upper bound of K possible mineral formation events for each parent distribution modeled by the K -dimensional probability vector \mathbf{p}_b for each of the B parents. For a given parent b and a fixed threshold τ where we conclude that a potential formation event is realized (say $\tau = 0.01$ or $\tau = 0.05$), the estimated number of formation events is $\frac{1}{L} \sum_{\ell=1}^L \sum_{k=1}^K I\{p_{bk}^{(\ell)} \geq \tau\}$. For example, the posterior mean estimate for the number of formation events for parent P1 under the top-down mixing model is 4.37 (95% CI 4-6) when the threshold $\tau = 0.05$ and 8.34 (95% CI 7-10) when $\tau = 0.01$. Because the model produces a posterior probability, any other such probabilistic questions like those above can be calculated as derived quantities. For example, we can ask questions like: what proportion of a given sample contains grains older than a given age? or what is the probability that an unobserved parent contains grains within a particular age range. Once the

584 posterior samples have been calculated, any questions that can be evaluated
585 using derived quantities can be answered probabilistically.

586 In addition, the ability to include prior information in the Bayesian frame-
587 work is a useful tool that can be used to improve estimation and test geologic
588 hypotheses. For example, certain geologic events, such as the Grenville orogeny,
589 produced large amounts of zircon that have since been broadly dispersed and
590 recycled in sedimentary rocks (Moecher and Samson, 2006). Priors that ac-
591 count for the likelihood of observing zircons of Grenville-age (or other known
592 zircon-producing events) can be introduced into this model framework to im-
593 prove performance. In addition, our framework can accommodate a variety of
594 detrital data with different magnitudes of uncertainty. As analytic techniques
595 for dating minerals improve, it is important to account for dating uncertainties
596 that might vary widely, making our method more robust to future improve-
597 ments in analytic laboratory techniques.

598 7 Conclusion

599 Starting from a conceptual model of how sediments mix over a landscape,
600 we developed a generative Bayesian nonparametric statistical model for detri-
601 tal mineral age data. This model allows us to characterize the uncertainty in
602 the age distributions of parents and children and the mixing proportions for
603 sediments while explicitly accounting for the uncertainties in measured dates
604 (Jasra et al., 2006; Tye et al., 2019). Because the model can generate sedi-
605 ment age distributions, we can directly explore the assumptions of the model
606 by simulating synthetic data. Running a simulation experiment demonstrated
607 the model is capable of recovering simulated distributions which supports the
608 usefulness of the framework when applied to observed data.

609 We proposed two frameworks to model the sediment mixing mechanisms:
610 the top-down mixing model where mineral dates are measured for both par-
611 ent and child sediments and a bottom-up unmixing framework where mineral
612 dates are only measured for the children. The top-down model estimated the
613 parent and child distributions and the mixing proportions with high precision
614 and accuracy. The bottom-up model occasionally showed evidence of non-
615 identifiability in the simulation experiments and did show non-identifiability
616 in the case study, suggesting the inference for the bottom-up model is less
617 precise than for the top-down mixing model. Because the variances of these
618 estimates are larger in our bottom-up unmixing model, the user is provided
619 with feedback about the potential pitfalls in being overly confident about the
620 reconstructed parent distributions.

621 Obtaining correct inference is vitally important for any statistical model.
622 However, many models make identifying when inference is suspect challenging.
623 The explicit modeling of uncertainty presented in this manuscript provides
624 a check on overly confident inference. As such, the inference in the model
625 presented herein provides a useful diagnostic on the quality of model fit. In
626 the case study, the posterior distribution for one of the parent distributions

627 was estimated with a large amount of uncertainty. Therefore, the inference
628 based on the model fit is less reliable than the model that includes the parent
629 data. Thus, explicit modeling of uncertainty is critical in providing information
630 about the quality of the inference.

631 While this manuscript focuses on estimating the mixture of geochronological
632 measurements from sediments, the methods discussed can be applied to
633 mixtures of any univariate variable of interest. For example, the top-down mixing
634 and bottom-up unmixing models can be applied to mixtures of sediment
635 grain size (Weltje and Prins, 2007). In addition to applying the model frame-
636 works to other variables, the extension of the mixing and unmixing models to
637 multivariate data would allow for the inclusion of more nuanced and detailed
638 data. In the cases where the distributions are only weakly identified based on
639 one variable, the mixing distributions might be identifiable using other vari-
640 ables. Thus, the results presented in this manuscript can provide a roadmap for
641 future development and extension to better characterize geologic landscapes.

642 8 Declarations

643 The authors declare that they have no known competing financial interests or
644 personal relationships that could have appeared to influence the work reported
645 in this paper. Support for SAJ came from the FEDMAP component of the
646 US Geological Survey National Cooperative Geologic Mapping Program. This
647 draft manuscript is distributed solely for purposes of scientific peer review.
648 Its content is deliberative and predecisional, so it must not be disclosed or
649 released by reviewers. Because the manuscript has not yet been approved for
650 publication by the U.S. Geological Survey (USGS), it does not represent any
651 official USGS finding or policy. Any use of trade, firm, or product names is
652 for descriptive purposes only and does not imply endorsement by the U.S.
653 Government.

654 Code and data for replication of results presented in this manuscript can be
655 found freely available under the permissive MIT license on GitHub at <https://github.com/jtipton25/mixing-manuscript>.

656 **Keywords** Detrital sediment age distributions · Sediment unmixing ·
657 Bayesian nonparametrics · Uncertainty quantification

659 References

- 660 Amidon WH, Burbank DW, Gehrels GE (2005a) Construction of detrital min-
661 eral populations: insights from mixing of U-Pb zircon ages in Himalayan
662 rivers. *Basin Research* 17(4):463–485
663 Amidon WH, Burbank DW, Gehrels GE (2005b) U-Pb zircon ages as a sed-
664 iment mixing tracer in the Nepal Himalaya. *Earth and Planetary Science
665 Letters* 235(1-2):244–260

- 666 Berliner LM (2003) Physical-statistical modeling in geophysics. *Journal of*
667 *Geophysical Research: Atmospheres* 108(D24)
- 668 Blake WH, Boeckx P, Stock BC, Smith HG, Bodé S, Upadhyay HR, Gas-
669 par L, Goddard R, Lennard AT, Lizaga I, et al. (2018) A deconvolutional
670 Bayesian mixing model approach for river basin sediment source apportion-
671 ment. *Scientific reports* 8(1):1–12
- 672 Chang W, Haran M, Applegate P, Pollard D (2016) Calibrating an ice sheet
673 model using high-dimensional binary spatial data. *Journal of the American*
674 *Statistical Association* 111(513):57–72
- 675 Chen JH, Moore JG (1982) Uranium-lead isotopic ages from the Sierra
676 Nevada batholith, California. *Journal of Geophysical Research: Solid Earth*
677 87(B6):4761–4784
- 678 Chen W, Guillaume M (2012) HALS-based NMF with flexible constraints for
679 hyperspectral unmixing. *EURASIP Journal on Advances in Signal Process-
680 ing* 2012(1):54
- 681 Cherniak D, Watson E (2001) Pb diffusion in zircon. *Chemical Geology* 172(1-
682 2):5–24
- 683 Cooper RJ, Krueger T (2017) An extended Bayesian sediment fingerprint-
684 ing mixing model for the full bayes treatment of geochemical uncertainties.
685 *Hydrological Processes* 31(10):1900–1912
- 686 de Valpine P, Turek D, Paciorek C, Anderson-Bergman C, Temple Lang D,
687 Bodik R (2017) Programming with models: writing statistical algorithms
688 for general model structures with NIMBLE. *Journal of Computational and*
689 *Graphical Statistics* 26:403–413, DOI 10.1080/10618600.2016.1172487
- 690 Diebolt J, Robert CP (1994) Estimation of finite mixture distributions
691 through Bayesian sampling. *Journal of the Royal Statistical Society Series*
692 *B (Methodological)* pp 363–375
- 693 Donoho D, Stodden V (2004) When does non-negative matrix factorization
694 give a correct decomposition into parts? In: *Advances in Neural Information*
695 *Processing Systems*, pp 1141–1148
- 696 Ferguson TS (1973) A Bayesian analysis of some nonparametric problems. *The*
697 *Annals of Statistics* pp 209–230
- 698 Ferguson TS (1983) Bayesian density estimation by mixtures of normal distri-
699 *butions*. In: *Recent Advances in Statistics*, Elsevier, pp 287–302
- 700 Flowers RM, Ketcham RA, Shuster DL, Farley KA (2009) Apatite (U–Th)/He
701 thermochronometry using a radiation damage accumulation and annealing
702 model. *Geochimica et Cosmochimica acta* 73(8):2347–2365
- 703 Fosdick JC, Grove M, Graham SA, Hourigan JK, Lovera O, Romans BW
704 (2015) Detrital thermochronologic record of burial heating and sediment re-
705 cycling in the Magallanes foreland basin, Patagonian Andes. *Basin Research*
706 27(4):546–572
- 707 Frühwirth-Schnatter S (2006) Finite mixture and Markov switching models.
708 Springer Science & Business Media
- 709 Garver JI, Brandon MT, Roden-Tice M, Kamp PJ (1999) Exhumation history
710 of orogenic highlands determined by detrital fission-track thermochronology.
711 Geological Society, London, *Special Publications* 154(1):283–304

- 712 Gehrels G (2012) Detrital zircon U-Pb geochronology: Current methods and
713 new opportunities. *Tectonics of Sedimentary Basins: Recent Advances* pp
714 45–62
- 715 Gehrels G (2014) Detrital zircon U-Pb geochronology applied to tectonics.
716 *Annual Review of Earth and Planetary Sciences* 42:127–149
- 717 Ghosal S (2010) The Dirichlet process, related priors and posterior asymptotics.
718 *Bayesian nonparametrics* 28:35
- 719 Green PJ (1995) Reversible jump Markov chain Monte Carlo computation and
720 Bayesian model determination. *Biometrika* 82(4):711–732
- 721 Guan Y, Haran M, Pollard D (2018) Inferring ice thickness from a glacier dy-
722 namics model and multiple surface data sets. *Environmetrics* 29(5-6):e2460
- 723 Haario H, Saksman E, Tamminen J (2001) An adaptive Metropolis algorithm.
724 *Bernoulli* 7(2):223–242
- 725 Hamilton N (2018) ggtern: An Extension to ggplot2, for the creation of ternary
726 diagrams. URL <https://CRAN.R-project.org/package=ggtern>, R pack-
727 age version 2.2.2
- 728 Hefley TJ, Brost BM, Hooten MB (2017) Bias correction of bounded location
729 errors in presence-only data. *Methods in Ecology and Evolution* 8(11):1566–
730 1573
- 731 Irwin WP, Wooden JL (1999) Plutons and accretionary episodes of the Kla-
732 math Mountains, California and Oregon. Tech. rep., US Geological Survey
- 733 Jasra A, Stephens DA, Gallagher K, Holmes CC (2006) Bayesian mixture
734 modelling in geochronology via Markov chain Monte Carlo. *Mathematical
735 Geology* 38(3):269–300
- 736 Kimbrough DL, Grove M, Gehrels GE, Dorsey RJ, Howard KA, Lovera O,
737 Aslan A, House PK, Pearthree PA (2015) Detrital zircon U-Pb provenance
738 of the Colorado River: A 5 my record of incision into cover strata overlying
739 the Colorado Plateau and adjacent regions. *Geosphere* 11(6):1719–1748
- 740 Lock EF, Dunson DB (2015) Shared kernel Bayesian screening. *Biometrika*
741 102(4):829–842
- 742 Mason CC, Fildani A, Gerber T, Blum MD, Clark JD, Dykstra M (2017)
743 Climatic and anthropogenic influences on sediment mixing in the Missis-
744 sippi source-to-sink system using detrital zircons: Late Pleistocene to recent.
745 *Earth and Planetary Science Letters* 466:70–79
- 746 Miao L, Qi H (2007) Endmember extraction from highly mixed data using min-
747 imum volume constrained nonnegative matrix factorization. *IEEE Transac-
748 tions on Geoscience and Remote Sensing* 45(3):765–777
- 749 Miller JW, Harrison MT (2018) Mixture models with a prior on the number of
750 components. *Journal of the American Statistical Association* 113(521):340–
751 356
- 752 Moehler DP, Samson SD (2006) Differential zircon fertility of source terranes
753 and natural bias in the detrital zircon record: Implications for sedimentary
754 provenance analysis. *Earth and Planetary Science Letters* 247(3-4):252–266
- 755 Paterson GA, Heslop D (2015) New methods for unmixing sediment grain size
756 data. *Geochemistry, Geophysics, Geosystems* 16(12):4494–4506

- 757 Puetz SJ, Ganade CE, Zimmermann U, Borchardt G (2018) Statistical anal-
758 yses of global U-Pb database 2017. *Geoscience Frontiers* 9(1):121–145
- 759 Reiners PW, Brandon MT (2006) Using thermochronology to understand oro-
760 genic erosion. *Annual Review Earth Planetary Sciences* 34:419–466
- 761 Reiners PW, Campbell I, Nicolescu S, Allen CM, Hourigan J, Garver J, Mat-
762 tinson J, Cowan D (2005) (U-Th)/(He-Pb) double dating of detrital zircons.
763 *American Journal of Science* 305(4):259–311
- 764 Richardson S, Green PJ (1997) On Bayesian analysis of mixtures with an
765 unknown number of components (with discussion). *Journal of the Royal
766 Statistical Society: series B (Statistical Methodology)* 59(4):731–792
- 767 Romans BW, Castelltort S, Covault JA, Fildani A, Walsh J (2016) Environ-
768 mental signal propagation in sedimentary systems across timescales. *Earth-
769 Science Reviews* 153:7–29
- 770 Saylor JE, Sundell K, Sharman G (2019) Characterizing sediment sources by
771 non-negative matrix factorization of detrital geochronological data. *Earth
772 and Planetary Science Letters* 512:46–58
- 773 Sharman GR, Johnstone SA (2017) Sediment unmixing using detrital
774 geochronology. *Earth and Planetary Science Letters* 477:183–194
- 775 Sharman GR, Sylvester Z, Covault JA (2019) Conversion of tectonic and cli-
776 matic forcings into records of sediment supply and provenance. *Scientific
777 reports* 9(1):4115
- 778 Sickmann ZT, Paull CK, Graham SA (2016) Detrital-zircon mixing and parti-
779 tioning in fluvial to deep marine systems, Central California, USA. *Journal
780 of Sedimentary Research* 86(11):1298–1307
- 781 Stock GM, Ehlers TA, Farley KA (2006) Where does sediment come from?
782 Quantifying catchment erosion with detrital apatite (U-Th)/He ther-
783 mochronometry. *Geology* 34(9):725–728
- 784 Sundell K, Saylor JE (2017) Unmixing detrital geochronology age distribu-
785 tions. *Geochemistry, Geophysics, Geosystems*
- 786 Tipton J, Hooten M, Goring S (2017) Reconstruction of spatio-temporal tem-
787 perature from sparse historical records using robust probabilistic principal
788 component regression. *Advances in Statistical Climatology, Meteorology and
789 Oceanography* 3(1):1–16
- 790 Tipton JR, Hooten MB, Pederson N, Tingley MP, Bishop D (2016) Recon-
791 struction of late Holocene climate based on tree growth and mechanistic
792 hierarchical models. *Environmetrics* 27(1):42–54
- 793 Tipton JR, Hooten MB, Nolan C, Booth RK, McLachlan J (2019) Predicting
794 paleoclimate from compositional data using multivariate Gaussian process
795 inverse prediction. *Annals of Applied Statistics* 13(4):2363–2388, DOI 10.
796 1214/19-AOAS1281
- 797 Tye A, Wolf A, Niemi N (2019) Bayesian population correlation: A probabilis-
798 tic approach to inferring and comparing population distributions for detrital
799 zircon ages. *Chemical Geology* 518:67–78
- 800 Vehtari A, Gelman A, Gabry J (2017) Practical Bayesian model evaluation
801 using leave-one-out cross-validation and WAIC. *Statistics and Computing*
802 27:1413–1432, DOI 10.1007/s11222-016-9696-4

- 803 Vermeesch P (2012) On the visualisation of detrital age distributions. *Chemical
804 Geology* 312:190–194
- 805 Ward EJ, Semmens BX, Schindler DE (2010) Including source uncertainty
806 and prior information in the analysis of stable isotope mixing models. *En-
807 vironmental Science & Technology* 44(12):4645–4650
- 808 Weltje GJ (1997) End-member modeling of compositional data: Numerical-
809 statistical algorithms for solving the explicit mixing problem. *Mathematical
810 Geology* 29(4):503–549
- 811 Weltje GJ, Prins MA (2007) Genetically meaningful decomposition of grain-
812 size distributions. *Sedimentary Geology* 202(3):409–424
- 813 Wotzlaw JF, Schaltegger U, Frick DA, Dungan MA, Gerdes A, Günther D
814 (2013) Tracking the evolution of large-volume silicic magma reservoirs from
815 assembly to supereruption. *Geology* 41(8):867–870