Unmixing detrital geochronology age distributions

Kurt E. Sundell[1] (ID) **and Joel E. Saylor[1]** (ID)

[1]Department of Earth and Atmospheric Sciences, University of Houston, Houston, Texas, USA

Abstract Despite recent advances in quantitative methods of detrital provenance analysis, there is currently no widely accepted method of unmixing detrital geochronology age distributions. We developed a model that determines mixing proportions for source samples through inverse Monte Carlo modeling, wherein mixed samples are compared to randomly generated combinations of source distributions, and a range of best mixing proportions are retained. Results may then be used to constrain a forward optimization routine to find a single best-fit mixture. Quantitative comparison is based on the Kolmogorov-Smirnov (KS) test *D* statistic and Kuiper test *V* statistic for cumulative distribution functions, and the Cross-correlation coefficient for finite mixture distributions (probability density plots or kernel density estimates). We demonstrate the capacity of this model through a series of tests on synthetic data, and published empirical data from North America mixed in known proportions; this proof-of-concept testing shows the model is capable of accurately unmixing highly complex distributions. We apply the model to two published empirical data sets mixed in unknown proportions from Colombia and central China. Neither example yields perfect model fits, which provides a cautionary note of potentially inadequate characterization of source and/or mixed samples, and highlights the importance of such characterization for accurate interpretation of sediment provenance. Sample size appears to be a major control on mixture model results; small ($n < 100$) samples may lead to misinterpretation. The model is available as a MATLAB-based stand-alone executable (.exe file) graphical user interface.
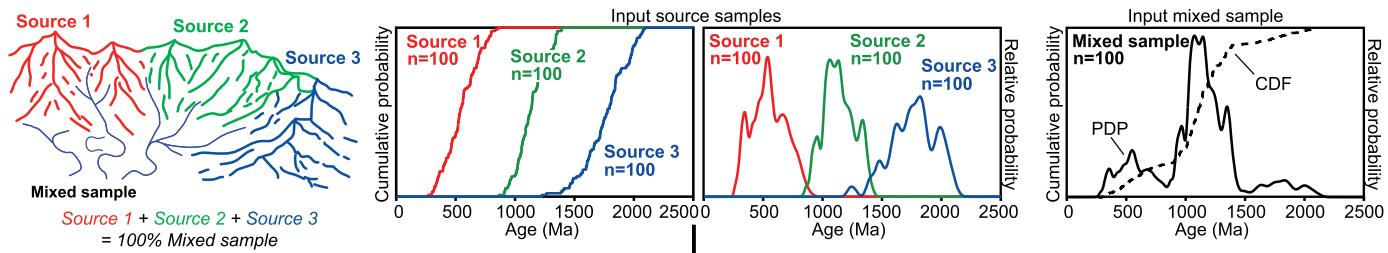
1. Introduction

Since the advent of laser ablation inductively coupled plasma mass spectrometry there has been a dramatic increase in both number of samples (N) and sample size (n) of detrital zircon U-Pb geochronological data sets [e.g., *Pullen et al.*, 2014; *Vermeesch and Garzanti*, 2015]. This technique has become a method of choice for geoscientists interested in source-to-sink questions requiring detailed sediment provenance analysis including sediment budgeting, sediment routing, paleogeography, and determining maximum depositional age [e.g., *Gehrels et al.*, 2011; *Laskowski et al.*, 2013; *Perez and Horton*, 2014]. Early interpretations of detrital age distributions relied on *qualitative* comparison based on the presence or absence of characteristic source populations for a given geologic setting, often highlighted by vertical bars on vertically stacked finite mixture distributions (probability density plots (PDPs) or kernel density estimates (KDEs)). This approach suffers from many problems, two of which are interpreter bias, and potential broad-brush provenance interpretations exacerbated by oversmoothing old and young ages in PDPs and KDEs, respectively. Furthermore, as N continues to grow larger, data management becomes a more pressing issue, and this type of visual comparison tends to break down completely.

More recently, *quantitative* techniques have been adapted to aid comparison of detrital data sets [e.g., *Gehrels*, 2000; *Amidon et al.*, 2005; *Saylor et al.*, 2013; *Vermeesch*, 2013; *Satkoski et al.*, 2013; *Kimbrough et al.*, 2015; *Licht et al.*, 2016; *Saylor and Sundell*, 2016; *Vermeesch et al.*, 2016; *Wissink and Hoke*, 2016]. Application of these methods has typically focused on the forward mixing problem of comparing source signatures mixed based on observations (e.g., relative outcrop area) to a mixed sample age distribution; few researchers have considered the inverse problem of determining the mixing proportions of potential source samples through model-based comparison to a mixed sample [e.g., *Saylor et al.*, 2013; *Kimbrough et al.*, 2015; *Licht et al.*, 2016]. Extracting mixing proportions of source samples is particularly important in geologic settings with a significant amount of sediment recycling where detrital samples are largely sourced from other clastic sedimentary rocks [e.g., *Campbell et al.*, 2005; *Perez and Horton*, 2014].
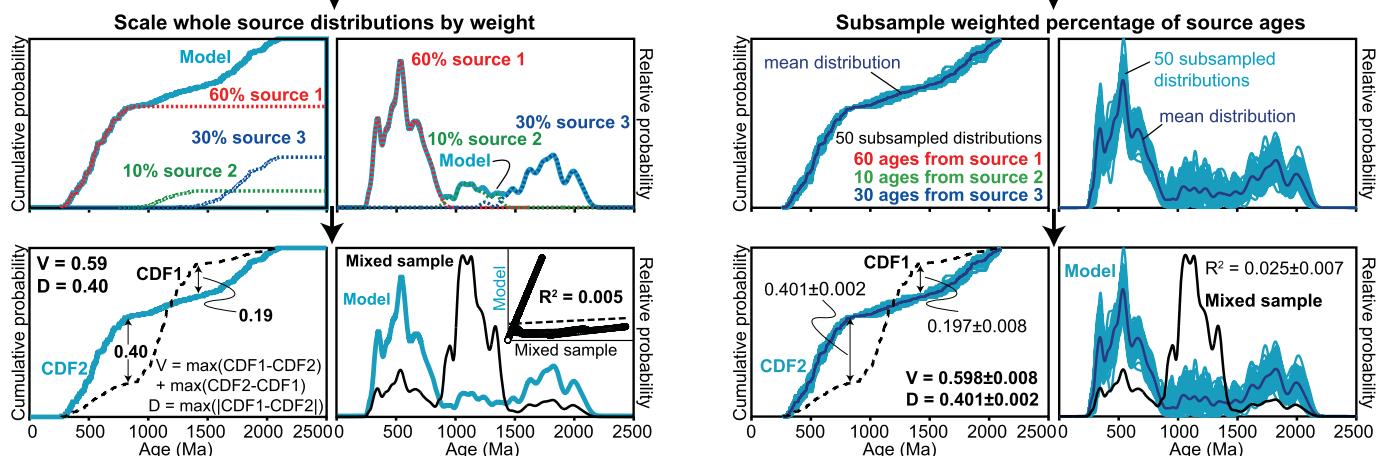
Figure 1. Inverse Monte Carlo model concept showing the steps taken to run one trial. (a) Synthetic detrital age distribution representing three potential source signatures shown as cumulative distribution plots (CDFs, left) and probability density plots (PDPs, center). Synthetic data comprised three source samples each with 100 ages and 2–5% uncertainty at the 1σ level. A mixed sample with actual distribution of 20% of source 1, 70% of source 2, and 10% of source 3 is shown on the right. (b) Example random weighting to be applied to the three source samples. (c) Randomly generated weights from B applied to source CDFs and PDPs (left) or converted to percent of ages randomly subsampled from source age distributions ($s = 100$ and $S = 50$, see section 2.1 for details). Quantitative comparison of model trial (in blue) to mixed sample (in black) using the KS test D statistic, Kuiper test V statistic, and Cross-correlation coefficient R^2. Results yield a single value when random weighting is applied to CDFs or PDPs (left), but give a range (shown as mean and standard deviation) when based on the subsampled source ages (right). (d) Steps B and C are repeated a user-specified number of times (number of trials) and a user-specified percent of best model fits are retained.

We present a new method of quantifying source mixing proportions through a combination of inverse Monte Carlo modeling and optimized forward modeling. We first demonstrate the model's ability to reproduce known source contributions using simple and complex synthetic data sets, and empirical detrital zircon U-Pb data from North America [*Laskowski et al.*, 2013]. Testing of these data mixed in known proportions shows the model is capable of perfectly determining mixing proportions of source contributions from mixed samples without any a priori information of source sample contribution. We then test the model on two published detrital zircon U-Pb empirical data sets. The first example consists of relatively low-*n* (mean *n* = 95) age distributions characterizing modern river sand mixed samples that are sourced from sedimentary units in their respective catchments [*Saylor et al.*, 2013]. The second example consists of relatively high-*n* (*n* ≥ 800), mixed loess and paleosol samples from central China [*Licht et al.*, 2016].

The model described here has been implemented as a MATLAB-based graphical user interface (GUI) stand-alone executable (.exe file), DZmix. As an executable, DZmix does not require installation of MATLAB. Both the GUI and source code are provided in supporting information, along with a step-by-step user manual and all data discussed below.

2. Model Concept

2.1. Inverse Monte Carlo Mixture Modeling
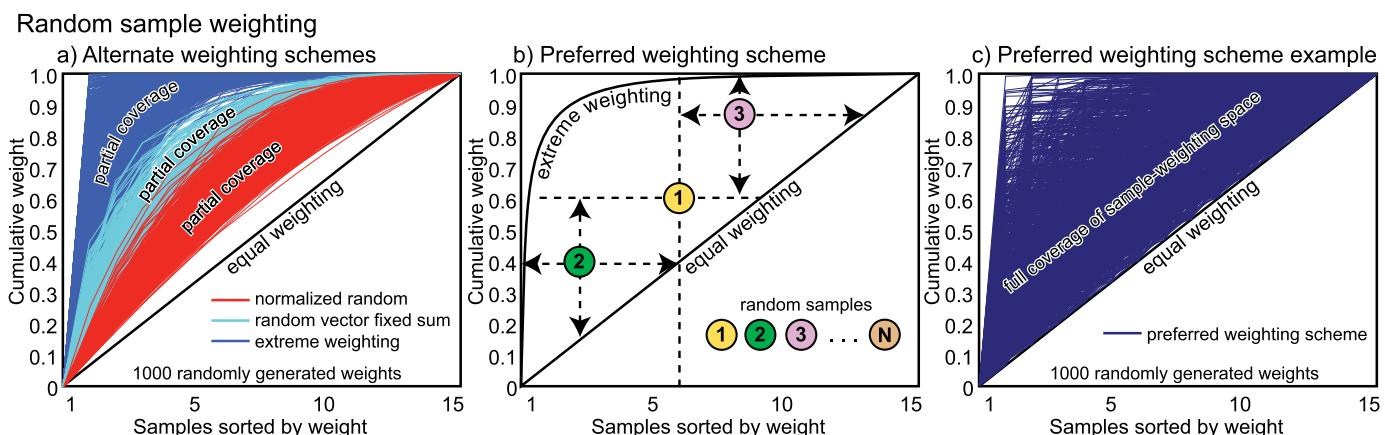
The inverse Monte Carlo model implements two different methods of constructing model source age distributions for comparison to mixed sample distributions (Figure 1). In both methods, model source distributions based on randomly generated weights are quantitatively compared to a mixed sample a user-specified number of times (number of trials) using the Kolmogorov-Smirnov (KS) test D statistic and the Kuiper test V statistic calculated from CDFs, and Cross-correlation of finite mixture distributions (PDPs or KDEs). The first method of constructing model source distributions simply scales each source's CDF and finite mixture distributions by the randomly generated weights and sums them together to produce a single CDF and PDP/KDE model source distribution [e.g., *Amidon et al.*, 2005; *Saylor et al.*, 2013]. The second method converts each randomly generated weight into an integer number of ages to be subsampled from each source, totaling a user-specified sample size (s), and repeated a user-specified number of times for each model trial (S). This latter approach was developed by *Licht et al.* [2016] as a modification of methods outlined in *Amidon et al.* [2005]. In the *Licht et al.* [2016] study, s and S are set at 800 and 200, respectively (Licht et al. use n and N). We adopt the convention of s and S to avoid confusion with number of input source samples, N, and individual sample size, n.

Both methods of constructing model source distributions are demonstrated on a synthetic data set in Figure 1. Here one example model trial with randomly generated weights of 60, 10, and 30% is applied using both methods. The first method applies these random weights to scale whole distributions that sum to a single age distribution for comparison to the mixed sample (Figures 1b and 1c, left). The second method randomly subsamples 60 ages from source 1, 10 ages from source 2, and 30 ages from source 3 ($s = 100$ ages), which is repeated $S = 50$ times to produce 50 different CDFs and PDPs for comparison to the mixed sample (Figures 1b and 1c, right). The second method is much more computationally intensive than the first because it requires multiple rounds of subsampling source ages and constructing model CDFs and finite mixture distributions for each trial.

2.2. Monte Carlo Model Random Weighting

A critical part of the mixing model is the method of randomly weighting each source sample age distribution. The ideal random weighting scheme is one that is capable of testing the entire sample-weighting space ranging from equal weighting for all samples to extremely asymmetric weighting (i.e., very high and/or low weights for one or more samples). The latter effects the model's ability to exclude noncontributing samples from the final, best fit weighting.

To test different random weighting schemes, randomly generated sample weights are plotted as CDF curves with samples sorted by weight on the x axis and cumulative weight on the y axis (Figure 2). This


Figure 2. Random weighting schemes shown as cumulative distribution functions (CDFs). (a) One-thousand randomly generated CDFs using the three different weighting schemes tested in this study (see section 2.2 for details). All three prove to be inadequate because none of them cover the full range of potential weighting combinations (area above the equal weight line). (b) Preferred weighting scheme implemented in the Monte Carlo mixing model (see section 2.2 for details). (c) One-thousand randomly generated CDF weights following preferred random weighting method shown in Figure 2b. This method of random weighting ensures testing both equal weighting (equal contribution of all source samples) and extreme weighting (contribution from only one or a few source samples).

allows visualization of the range of weights generated from each method before they are to be randomly assigned to a group of source samples. We initially tested three different random weighting schemes, all of which failed to meet the requirements mentioned above. In the first weighting scheme, N samples are each given a random number that is normalized by the sum of those random numbers ("normalized random," Figure 2a). Although this method of generating random weights is simple and intuitive, it fails to test even a third of the possible weighting combinations, is heavily bias toward the median (0.5 before normalization), and tends toward equal weighting with increased N (red lines in Figure 2a). The second weighting scheme uses a conditional distribution of uniform random variables distributed over 0 and 1 that guarantees uniform random weighting ("random vector fixed sum," Figure 2a). Although it yields a higher distribution of weights than the previous method, and resolves the issue of bias toward the median, it still cannot cover the random weighting space (light blue lines in Figure 2a). The third weighting scheme first randomly selects a weight between 0 and 1, which is then subtracted from the sum of the total sample remaining (beginning with a total of 1). Subsequent random weights may only be selected within the remainder (1 – Σ weights). The final weight assignment is simply the sum total weights subtracted from 1. This scheme tends to give very high weights to one or a few samples, and low weights (usually close to 0) to the rest, resulting in "extreme" weighting of samples (blue lines in Figure 2a). Even if all three of these weighting schemes are implemented, there is still an obvious gap in the generated weights (Figure 2a) that is exacerbated with increasing numbers of potential source samples (supporting information Figure S1), which would result in significant bias in the Monte Carlo mixing model.

An alternative, preferred weighting scheme is schematically shown in Figure 2b. Here sample weights are determined by first choosing a random sample number between 0 and N (x axis) and assigning it a random number between 0 and 1 (y axis) above the equal weight line (Figure 2b). All subsequent random samples in both dimensions are restricted in where they can be placed as to result in a monotonically increasing function. Following construction of the CDF, the weights are randomized by sample number and applied to the source samples. This weighting scheme ensures full coverage of possible weighting assignments when plotted as a CDF (Figure 2c), regardless of how many source samples are weighted (supporting information Figure S1).

2.3. Quantitative Comparison of Age Distributions
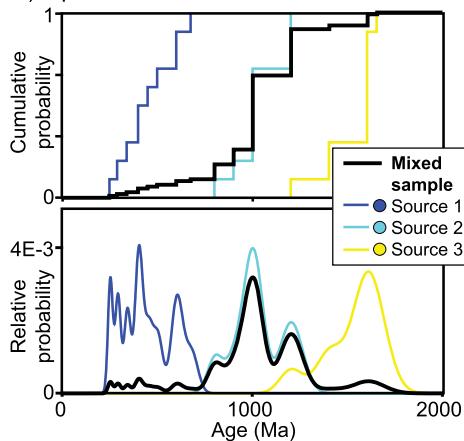
Three comparison methods are used to compare source sample age distributions to mixed samples: the two-sample KS test D statistic, the two-sample Kuiper test V statistic, and the Cross-correlation coefficient (coefficient of determination, R^2) (Figure 1c). The KS D statistic is the maximum absolute distance between two CDFs: $D = \max(|CDF_1 - CDF_2|)$ [*Stephens*, 1970]. The Kuiper test is a common variant of the KS test, and is the sum of the maximum distance between two CDFs subtracted from one another: $V = \max(CDF_1 - CDF_2) + \max(CDF_2 - CDF_1)$ [*Kuiper*, 1960; *Press et al.*, 2007]. The Cross-correlation coefficient is calculated as the coefficient of determination (R^2) of cross plots of finite mixture distribution quantiles (PDPs or KDEs) [*Saylor et al.*, 2012, 2013; *Saylor and Sundell*, 2016].

Each of these methods of comparing age distributions has its strengths and weaknesses. The KS D and Kuiper test V statistics are the basis for well-established statistical methods [*Kuiper*, 1960; *Stephens*, 1970; *Press et al.*, 2007]; using the difference between two CDFs, as opposed to p values generated from those differences, has recently been established as a useful tool for comparison of detrital age distributions [e.g., *Satkoski et al.*, 2013; *Vermeesch*, 2013; *Saylor and Sundell*, 2016]. The Cross-correlation coefficient has also recently been adapted for use in detrital geochronology quantitative comparison [*Saylor et al.*, 2012, 2013], and has been shown to have more discriminatory power than the KS and Kuiper test D and V statistics, with more sensitivity to the number and proportion of age modes in finite mixture distributions [*Saylor and Sundell*, 2016]. The Cross-correlation coefficient of PDPs also takes into account sample uncertainty by using the analytical uncertainty of each age as the kernel for individual Gaussian curves that are summed and normalized to construct the finite mixture distribution. Neither the KS test D statistic nor the Kuiper test V statistic takes into account sample uncertainty, as the empirical CDFs are constructed solely based on mean ages of an age distribution. In some cases, this may be considered a strength, for example if comparing detrital age distributions with wildly varying uncertainties which could potentially bias PDPs toward individual ages of high and low kurtosis for young and old ages, respectively. This could be resolved by using KDEs, as they too do not incorporate age uncertainty, but only if a proper bandwidth can be determined [e.g., *Botev et al.*, 2010; *Andersen et al.*, 2016]. Finally, the KS test is more sensitive about the median of the age distribution, whereas the Kuiper test guarantees equal sensitivity across the entire range of ages in the samples [*Kuiper*, 1960; *Press et al.*, 2007].
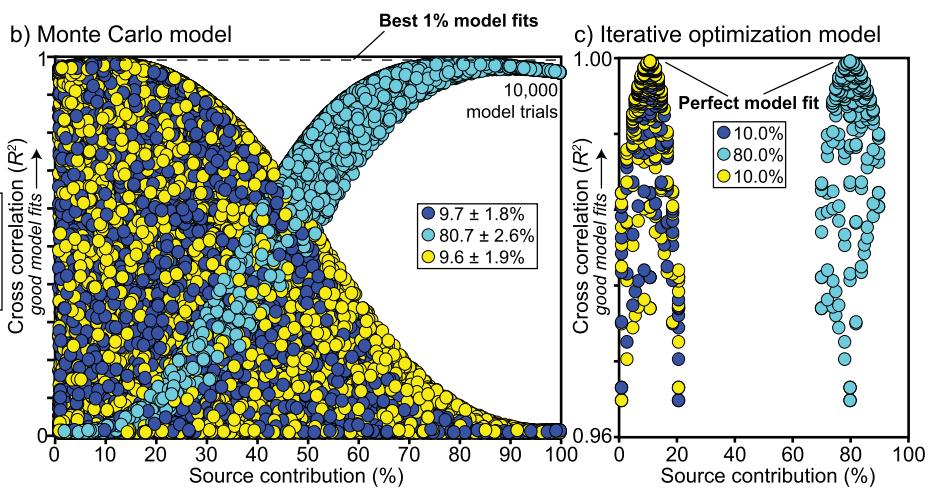
None of these comparison methods are statistical hypothesis tests. The KS D statistic and Kuiper V statistic are required to generate p values, but are themselves only measures of similarity when interpreted in the
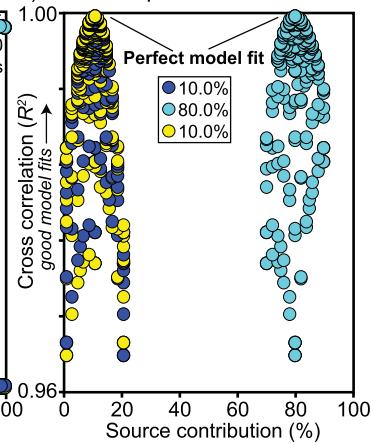
Figure 3. Simple synthetic data and model results. (a) Input synthetic source and mixed samples shown as cumulative distribution functions (top) and probability density plots (bottom). The mixed sample comprises 10% of source 1, 80% of source 2, and 10% of source 3. (b) Results of all 10,000 model trials, showing the percent contribution of each combination of sources and the Cross-correlation coefficient model result for each combination. (c) Results of the iterative optimization model, showing source combinations tested, and Cross-correlation coefficient for each combination. (d–f) Model results using the KS test *D* statistic (green) and Kuiper test *V* statistic (red) plotted as cumulative distribution functions (CDFs), and Cross-correlation coefficient R^2 (blue) as probability density plots. Black lines represent the mixed age distribution.

absence of associated *p* values. Although *p* values can give statistical measures of confidence, they are not well-suited for comparison of detrital geochronology data because they reject the null hypothesis at a higher rate than predicted for the selected *p* value [*Vermeesch*, 2013; *Saylor and Sundell*, 2016]. The Cross-correlation coefficient is a statistical measure of two-dimensional scatterplots, but in its implementation here is merely a relative measure of similarity between finite mixture distributions. Thus, results for all three tests are always relative, and no hard cutoff or absolute confidence levels can be used when interpreting model results.
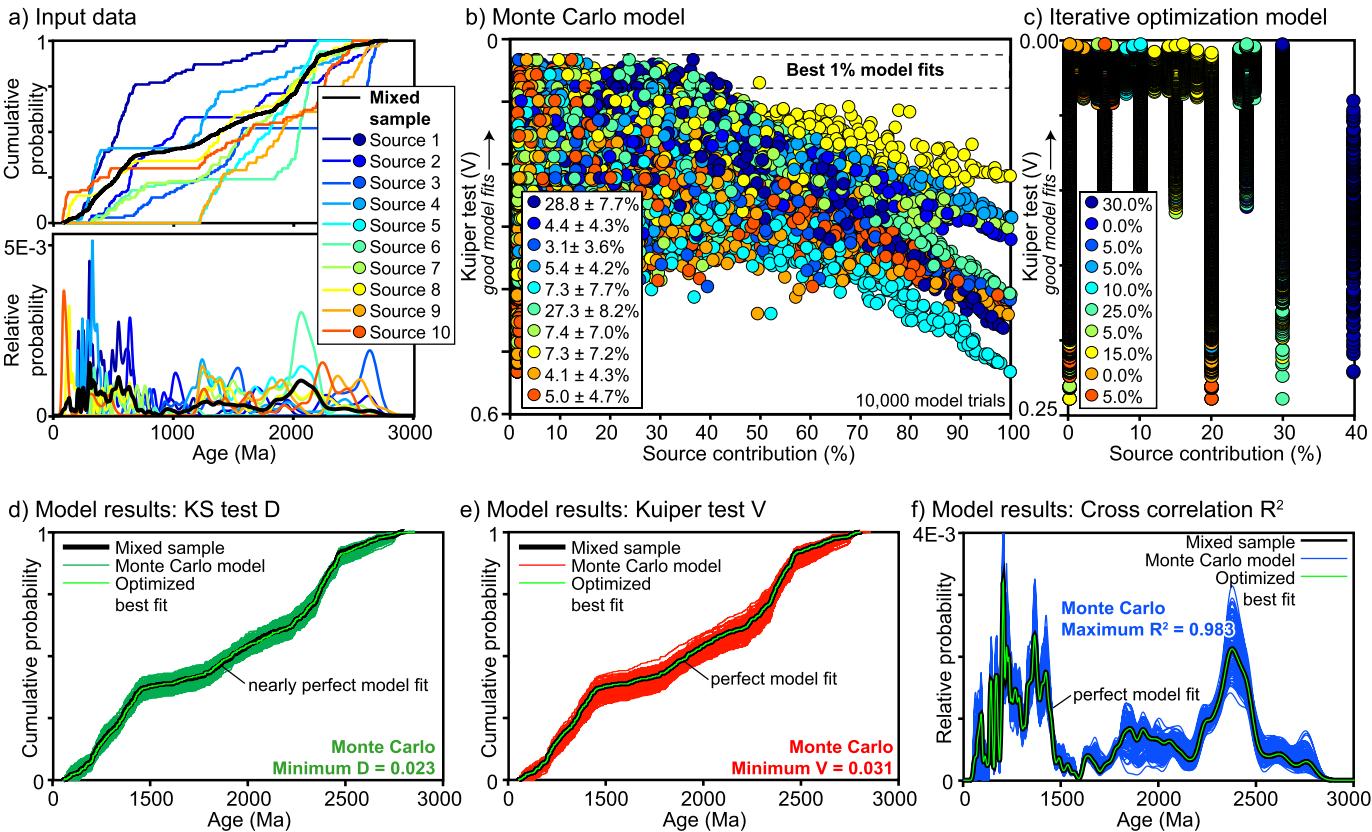
2.4. Forward Modeling

Although unmixing detrital age distributions can be approached using a forward model, it is computationally intensive to do so. If implemented as a deterministic forward model, all possible combinations of sources summing to 100% are calculated, and the best fit model result, or range of top model results, are interpreted as the most probable source contributions for a given mixed sample. Unfortunately, this method is highly computationally intensive, even for a relatively small number of source samples, because the number of possible permutations that sum to 100% for source contributions at the integer percent level is given by

$$\frac{(N+d-2)!}{(N-1)!} \times \frac{1}{(d-1)!},$$

where *N* is the number of source samples and *d* is the number of possible source weights for an individual source sample (*d* = 101 for 0–100% in increments of 1%). For low-*N* (*N* ≤ 4) data sets it is reasonable to take

Model testing: Complex synthetic data

Figure 4. Complex synthetic data and model results. (a) Input synthetic source and mixed samples shown as cumulative distribution functions (top) and probability density plots (bottom). The mixed sample comprises 30, 0, 5, 5, 10, 25, 5, 15, 0, and 5% of sources 1 through 10, respectively. (b) Results of all 10,000 model trials, showing the percent contribution of each combination of sources and the Kuiper test V statistic model result for each combination. (c) Results of the iterative optimization model, showing source combinations tested, and Kuiper test V statistic for each combination. (d–f) Model results using the KS test D statistic (green) and Kuiper test V statistic (red) plotted as cumulative distribution functions (CDFs), and Cross-correlation coefficient (blue) as probability density plots. Black lines represent the mixed age distribution.

a forward modeling approach [e.g., *Licht et al.*, 2016], as three and four samples only have 5151 and 176,851 possible combinations that sum to 100%. However, as N becomes larger, there are simply too many combinations of sources to be computationally efficient: 5 sources results in 4,598,126 potential source combinations; 10 sources results in 4,263,421,511,271. In contrast, the inverse Monte Carlo approach is capable of handling any number of input source samples.

A forward modeling approach is feasible if it is first constrained by the inverse Monte Carlo modeling results, and highly efficient if implemented as a forward optimization routine. We implement two different optimization methods, both initially constrained by results from the inverse Monte Carlo model. The first method is an iterative forward model that takes the range of source contributions based on the mean and standard deviation of Monte Carlo model results, and expands this range to be a multiple of 10. For example, a Monte Carlo modeled individual source contribution of $25 \pm 7\%$, would have a range of 18–32%, which would be used to constrain the forward model at 10–40%. The forward model then tests all possible combination of sources at a 10% "grid spacing." A user-specified number of best fits is then used to constrain subsequent iterations at smaller increments (grid spacing) of 5, 2, and finally 1%. The best model fit is reported at the 1% level. The second optimization routine utilizes the interior-point constrained nonlinear optimization algorithm (fmincon) from the MATLAB Optimization Toolbox™. In this approach, the function minimization algorithm attempts to minimize the KS test D and Kuiper test V values calculated from CDFs, and $1 - R^2$ from Cross-correlation of finite mixture distributions. A user-specified number of best fits (mean values of source contributions) from the Monte Carlo model results are used as initial guesses that the interior-point algorithm then iterates on to find a minimum function value. The lowest of these values is reported as the best model fit.

3. Model Testing

3.1. Mixtures of Synthetic Data Sets in Known Proportions

The first test of the Monte Carlo mixing model is a proof-of-concept style test on a simple synthetic data set consisting of three simple source samples with little overlap in age populations (Figure 3a). Each of the three source data sets has 10 ages with 5% uncertainty at the 1σ level. A mixed sample was generated with a known contribution from each source sample of 10% from source 1, 80% from source 2, and 10% from source 3.

A second test of the Monte Carlo mixing model is on a more complex synthetic data set consisting of 10 source samples with multimodal, overlapping age populations (Figure 4a). Each of the 10 samples has 100 ages with uncertainty between 2 and 12% at the 1σ level. For comparison, a mixed sample was generated with known contributions from each source sample of 30, 0, 5, 5, 10, 25, 5, 15, 0, and 5% from sources 1 through 10, respectively.
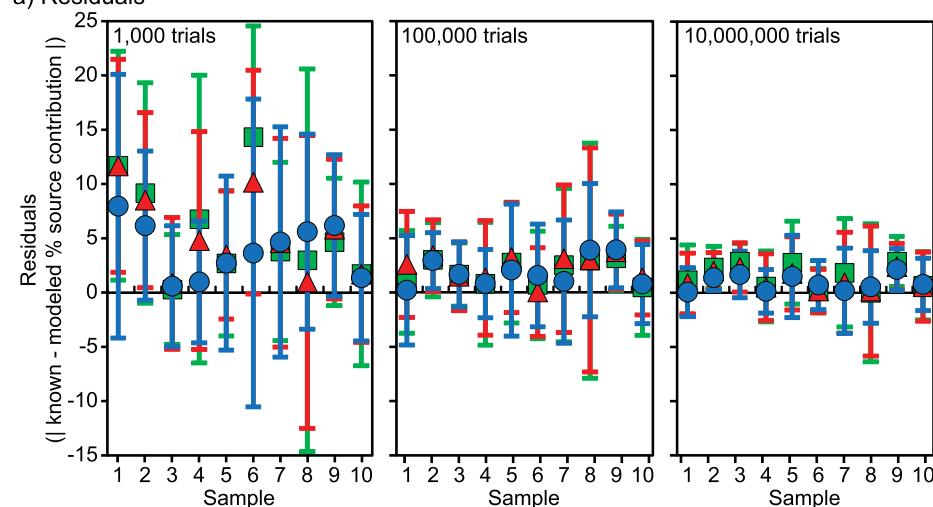
3.2. Monte Carlo Model Sensitivity Testing

We conducted two sensitivity tests of the Monte Carlo mixing model. The first is a test of the accuracy and efficiency of the model when scaling whole source distributions that uses the complex synthetic data set from section 3.1. The motivation for this test is to see how many trials are required to perfectly match complex data mixed in known proportions; if the known mixtures cannot be matched by the model then perhaps the complex data are too complicated for it. In this test the number of trials is varied, and the top percentage is reported as the best 100 model fits. Hence, percentages vary by changing the number of model trials ranging from 10^3 (10% best fits) to 10^7 (0.001% best fits) (Figure 5).
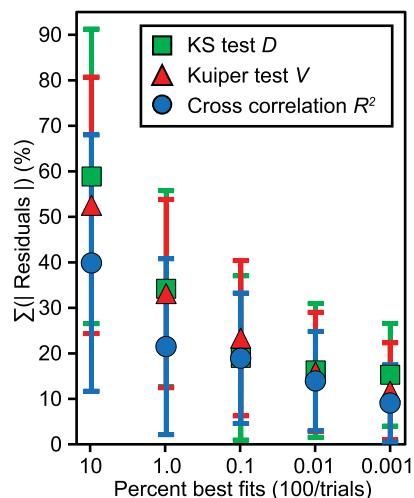
The second sensitivity test of the model uses a compilation of detrital zircon U-Pb data from North America [*Laskowski et al.*, 2013]. We constructed a data set of five source samples consisting of 500 randomly subsampled ages from five of the characteristic North American source groups: Mesozoic Eolianites, U.S. Passive Margin, Canada Passive Margin, Mogollon Highlands, and Cordilleran Arc [*Laskowski et al.*, 2013]. We then mixed these source samples in known proportions to construct a synthetically mixed sample of empirical data in proportions of 40, 30, 20, 10, and 0%, respectively. The Monte Carlo model then randomly subsamples a total of 100–500 ages from these sources with tests in increments of 100 ($s = 100$, 200, 300, 400, and 500) 50 times for each model trial ($S = 50$). The purpose of this is to test the effects of the two different source modeling methods (Figures 1b and 1c), to compare sample size effects when subsampling ages to construct source distributions [*Licht et al.*, 2016]

Monte Carlo model sensitivity
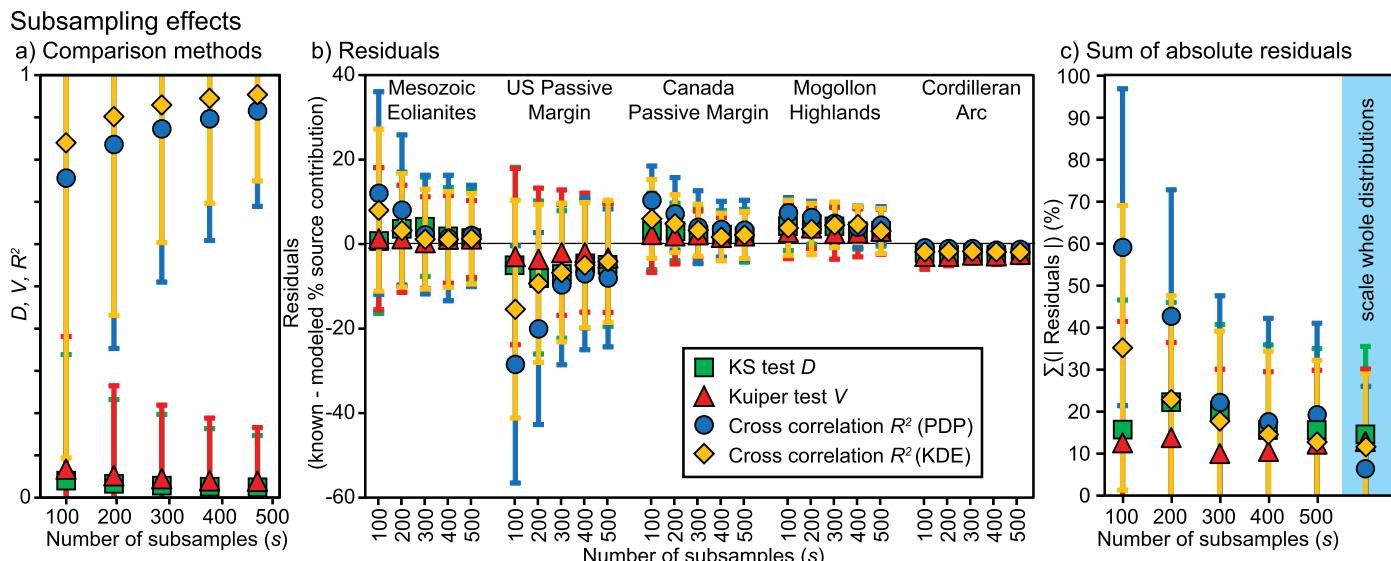
Figure 5. Monte Carlo sensitivity testing results based on the number of model trials of the complex synthetic data set shown in Figure 4 shows (a) an increasingly good fit between models and known source contributions with increasing numbers of trials run (corresponding to a lower percentage retained). (b) Summed absolute residuals show that the Cross-correlation coefficient achieves approximately the same accuracy as the Kuiper V or KS D statistic with an order of magnitude fewer model trials.

(Figures 1b and 1c, right), and to compare results from Cross-correlation of PDPs versus KDEs (using a 20 Myr kernel bandwidth) (Figure 6).

3.3. Mixtures of Empirical Data Sets in Unknown Proportions

The first empirical data set is published detrital zircon U-Pb geochronology data from north-central Colombia [*Saylor et al.*, 2013]. This data set consists of mixed sample age distributions from two modern river sand samples collected from the trunk streams of Rio Cravo Sur and Rio Cusiana, with source age distributions from sedimentary rocks in their respective drainage basins [*Saylor et al.*, 2013] (Figures 7 and 8a). Sedimentary rocks in these catchments include Cretaceous shallow and transitional marine mudstones and sandstones, Paleocene coastal-deltaic shales, coals, sandstones, and fluvial mudstones and sandstones, early Eocene alluvial fan and fluvial sandstones with subordinate conglomerates, late Eocene coastal and marine mudstones and sandstones, and Miocene nonmarine alluvial deposits [*Cooper et al.*, 1995; *Horton et al.*, 2010; *Saylor et al.*, 2013, and references therein]. Detrital zircon U-Pb age distributions for source samples were characterized either directly within the catchments (e.g., samples AM6B), or by proxy from equivalent strata outside of the catchment area (e.g., sample MA2). We chose this study area for a number of reasons. First, it is a modern system with a simple geologic setting of two drainage basins comprised of simple drainage networks, so we know what units are currently being sourced. Second, the catchments are small, and hence should be easily characterized. Third, the units have well-defined detrital zircon U-Pb age distributions from sedimentary rocks [*Horton et al.*, 2010; *Saylor et al.*, 2013], each with a characteristic age distribution presumably combined into a single mixture at the trunk streams of the drainage networks where modern sands were sampled (Figure 7a). Finally, because only sedimentary rocks are sourced, zircon fertility [e.g., *Moecher and Samson*, 2006; *Dickinson*, 2008] is a nonissue; however, differential zircon contribution based on source grain size or hydrodynamic sorting [e.g., *Garzanti et al.*, 2009; *Lawrence et al.*, 2011] could be an issue, as units comprised of dominantly mudrocks likely contribute little to the river sand detrital age distribution.
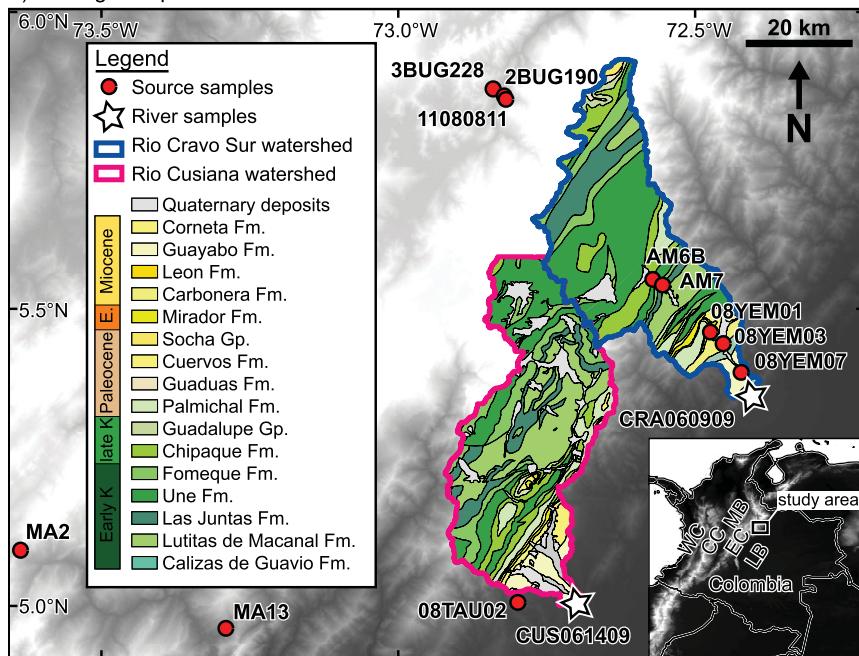
The second empirical data set is a compilation of detrital zircon U-Pb data from the Loess Plateau in central China consisting of source samples for comparison to eolian loess and paleosol mixed samples [*Licht et al.*, 2016, and references therein]. In total, 37 source samples are combined regionally into four large-*n* source samples, and eight loess samples and nine paleosol samples are combined into two mixed samples [*Licht et al.*, 2016] (Figures 9 and 10a). The four potential provenance zones are the Mu Us desert, the Central deserts, the Qaidam Basin, and the Yellow River (Figure 9a), which consist of eolian dune, yarding,
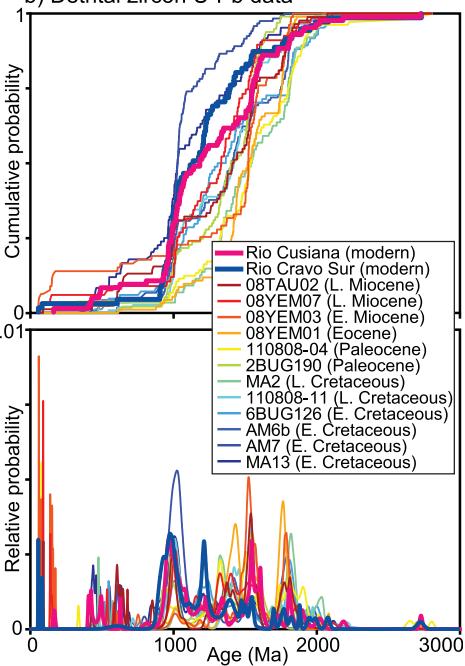

Figure 6. Inverse Monte Carlo model sensitivity based on subsample size (*s*) with the empirical North America data set from *Laskowski et al.* [2013] mixed in known proportions of 40, 30, 20, 10, and 0%. (A) While all comparison methods show a better fit between model and known distributions with increasing subsample size *s*, the Kuiper *V* and KS *D* statistics show the least improvement. (B) Residuals decrease for each of the potential source areas with increasing sample size. (C) Summed absolute residuals for the Cross-correlation coefficient decrease with increasing sample size, while those for the *D* and *V* value do not change within uncertainty. Residuals are generally higher and show more variability compared to scaling whole source distributions (highlighted in blue).

Figure 7. (a) Map of study area and sample locations in Colombia modified from *Saylor et al.* [2013]. (b) Source and mixed sample data as cumulative distribution functions (top) and probability density plots (bottom) of detrital zircon U-Pb data considered in this study.

lacustrine, and fluvial sandstone deposits [*Licht et al.*, 2016]. The first reason for testing this data set is to compare results of our mixing model directly to results in *Licht et al.* [2016], who implement construction of model source age distributions by random subsampling of source ages (Figures 1b and 1c, right). We also chose this compilation because it affords a test of large-n samples, which are predicted to provide a better characterization of age distributions [e.g., *Pullen et al.*, 2014; *Saylor and Sundell*, 2016]. Finally, we chose this compilation because of the scale of the geologic setting. Eolian transport distances would have to be >1000 km in some cases in this study area, which is in stark contrast to the short (<100 km) fluvial transport distances of the Colombia geologic setting.
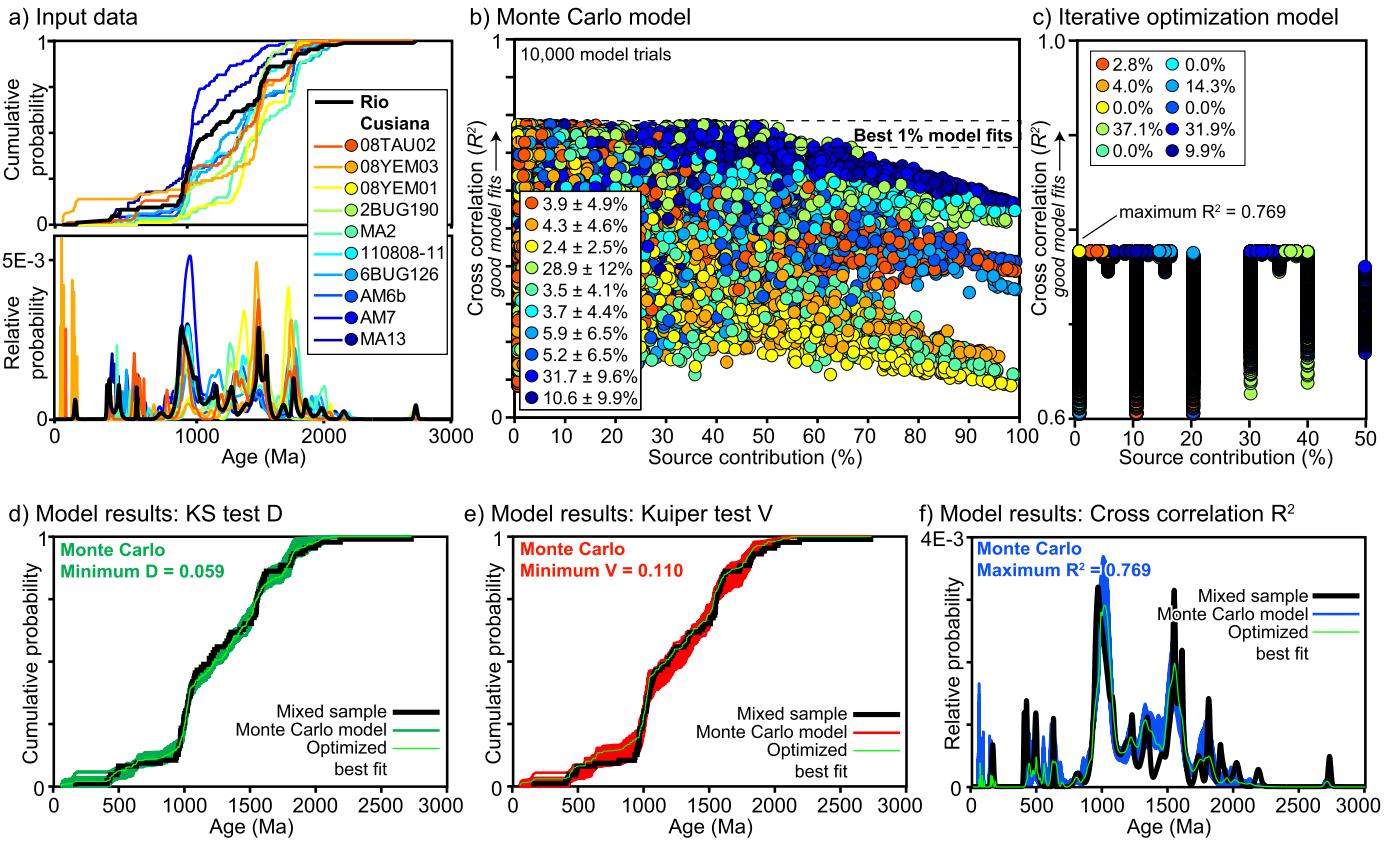
4. Results

For unmixing synthetic and empirical data (sections 3.1 and 3.3), model results generated using the KS D statistic, Kuiper V statistic, and Cross-correlation coefficient (R^2) are reported as the mean and one standard deviation of the best 100 model fits (lowest D and V, highest R^2) of 10,000 model trials, which represents the top 1% of model trials. Results are shown compared to mixed sample CDFs and finite mixture distributions. Reporting of model sensitivity testing (section 3.2) results differs slightly as described below.

4.1. Mixtures of Synthetic Sources in Known Proportions

Results from the first proof-of-concept test of simple synthetic data demonstrate that the top 1% of the Monte Carlo mixing model trials reproduce known source sample percent contributions within 1σ uncertainty (Figure 3). Figure 3b shows the range of randomly generated weight combinations of input source samples versus the resulting PDP Cross-correlation coefficient for all 10,000 trials of this model run. The top 1% of the model age distributions compared with the mixed age distributions shows that although there is a slight range about the mixed sample, each quantitative method accurately predicts the known source contribution without any a priori knowledge of it. Low minimum D and V values of 0.02 and 0.03, and a

Figure 8. Data and model results from the Rio Cusiana data set [*Saylor et al.*, 2013]. (a) Input source sample data from the river catchment (or by proxy outside the catchment) and mixed sample data from river sand shown as cumulative distribution functions (top) and probability density plots (bottom). (b) Results of all 10,000 model trials, showing the percent contribution of each combination of sources and the Cross-correlation coefficient model result for each combination. (C) Results of the iterative optimization model, showing source combinations tested, and Cross-correlation coefficient for each combination. (d–f) Model results using the KS test *D* statistic (green) and Kuiper test *V* statistic (red) plotted as cumulative distribution functions (CDFs), and Cross-correlation coefficient (blue) as probability density plots. Black lines represent the mixed age distribution.

high maximum R^2 of 0.998, also demonstrates a nearly perfect match between the known and modeled sample weights for the single best fit trial of the Monte Carlo model. Furthermore, both optimization routines yield exact matches to known source sample contributions with minimum *D* and *V* values of 0, and a maximum R^2 of 1 (Figures 3d–3f).
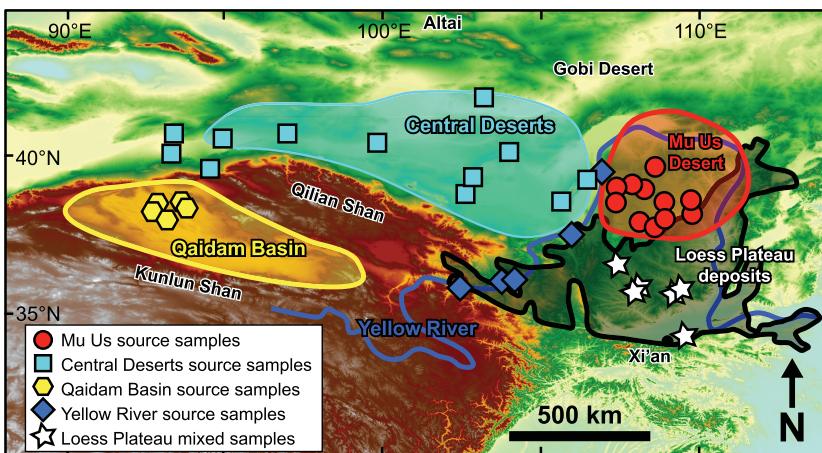
Results of the second proof-of-concept test of complex synthetic data are generally similar to those of the simple synthetic data, but show higher variability (Figure 4). There is a considerably larger range in the best 1% model fits, as shown by plotting weight combinations against resulting Kuiper test *V* value (Figure 4b), as well as a larger swath of results about the CDFs and PDP in Figures 4d–4f. Nevertheless, the top 1% of Monte Carlo model results reproduce the known source contributions within 1σ uncertainty (with one exception where a non-contributing source is 0.1% beyond the 1σ uncertainty range). Despite the increased variability, iterative optimization models constrained by inverse Monte Carlo results yield a nearly perfect model fit for the KS test *D* of 0.007, and perfect model fits for Kuiper *V* of 0, and Cross-correlation R^2 of 1; minimum search optimization yields perfect model fits for all comparison methods with $D = 0$, $V = 0$, and $R^2 = 1$, respectively.

4.2. Monte Carlo Model Sensitivity Testing
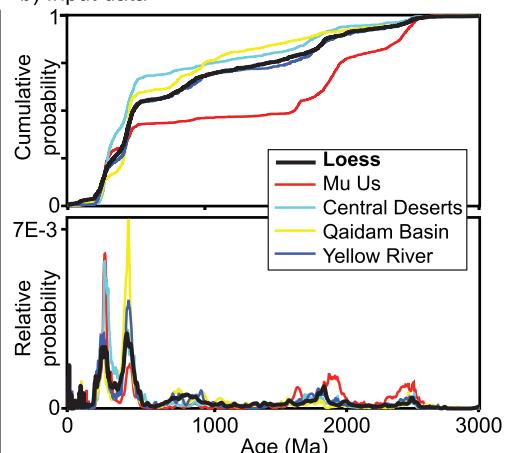
Residuals (modeled contributions subtracted from known contributions) from scaling whole age distributions of the complex synthetic data set systematically decrease with increasing trials (Figure 5). Results show relatively high residuals with large standard deviations based on the top 10% of 1000 model trials (Figure 5a). Residuals at this 10% level are on average slightly higher with larger standard deviation for the

Empirical data set: Central China

Figure 9. (a) Map of study area and sample locations in central China modified from *Licht et al.* (2016). (b) Source and mixed sample data as cumulative distribution functions (top) and probability density plots (bottom) of detrital zircon U-Pb data considered in this study.

KS D statistic at $6 \pm 10\%$, compared to the Kuiper V and Cross-correlation coefficient which give mean and standard deviations of $5 \pm 9\%$ and $4 \pm 8\%$, respectively. Model results show a dramatic improvement when attempting more model trials and retaining a lower percent of best fits for all comparison methods (Figure 5a). For this complex data set, the sum of residuals for all ten samples is as low as 0.15 ± 0.11 for the KS D statistic, 0.12 ± 0.11 for the Kuiper V statistic, and 0.09 ± 0.08 for the Cross-correlation (0.001% best model fits), requiring 10^7 model trials (Figure 5b). The Cross-correlation reaches this minimum with an order of magnitude fewer model runs, and with lower uncertainty for 10^6 trials (0.01% model fits, Figure 5b). In nearly all sensitivity tests the Cross-correlation yields both lower absolute residuals with lower associated standard deviations than the other two comparison methods (Figure 5).

Inverse Monte Carlo modeling by randomly subsampling raw ages from source distributions of empirical data from *Laskowski et al.* [2013] mixed in known proportions shows drastically different results compared to scaling whole distributions. In all cases the model gives better results when subsampling more ages (larger s), but are highly variable for all comparison methods, D, V, and R^2 (Figure 6a). Cross-correlation of finite mixture distributions shows higher variability in resulting R^2 values, and with higher residuals (modeled contributions subtracted from known contributions) than corresponding D and V model results (Figure 6a). In all subsampling cases ($s = 100$ to 500) the Kuiper test V comparison method gives the lowest residuals and Cross-correlation of PDPs gives the highest (Figures 6b and 6c); however, none give lower residuals than when scaling whole source distributions (Figure 6c).
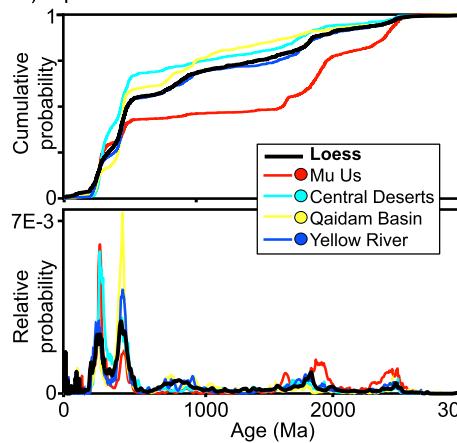
4.3. Mixtures of Empirical Sources in Unknown Proportions

A key difference between testing synthetic data and testing empirical data is we do not know the true contribution from each source sample. For brevity, and because model results are generally similar among empirical data sets, we only present and discuss results of Rio Cusiana from the Colombia data set [*Saylor et al.*, 2013], and the loess sample from the Loess Plateau data set [*Licht et al.*, 2016].
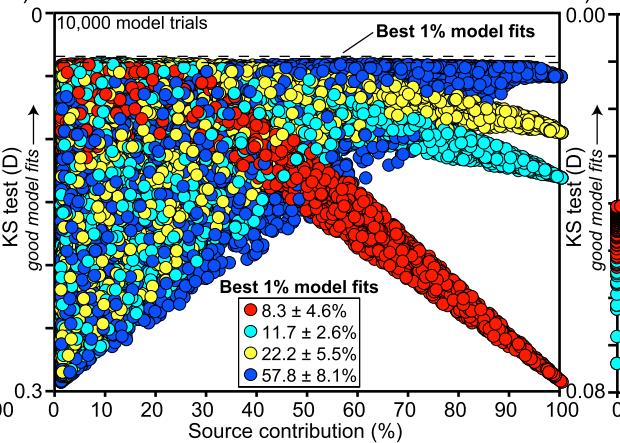
Inverse Monte Carlo model results for Rio Cusiana yield an imperfect match (Figure 8). As with the synthetic data, plotting sample weight combinations against Cross-correlation coefficient shows that the best fit scenario is indeed a mixture of the sources, but in the case of the empirical data yields an imperfect fit with a maximum $R^2 = 0.769$ (Figure 8b). Although these model results are indeed an improvement over those reported by *Saylor et al.* [2013] based on comparison of source samples weighted equally ($R^2 = 0.64$) and by percent area in the catchment ($R^2 = 0.73$), they are still far from a perfect fit, and apparently cannot match the mixed river sample age distribution based on the input source data; this is also the case for the minimum D and V values of 0.059 and 0.110, respectively. Results based on optimization routines are similar to each other, and to best fit inverse Monte Carlo
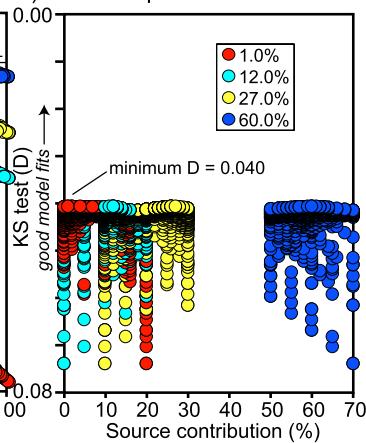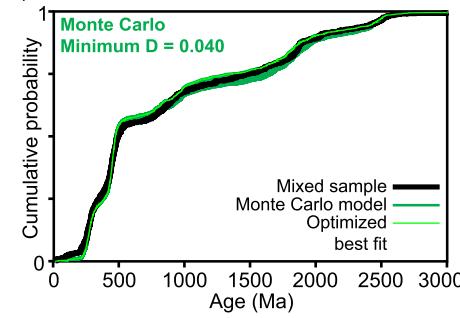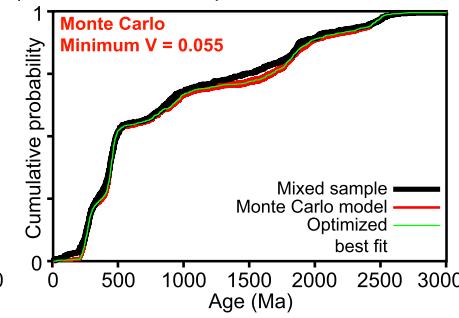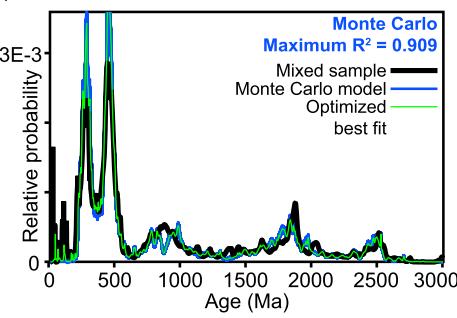
Figure 10. Data and model results from the Loess Plateau data set [*Licht et al.*, 2016]. (a) Input source sample data shown as cumulative distribution functions (top) and probability density plots (bottom). (b) Results of all 10,000 model trials, showing the percent contribution of each combination of sources and the KS test *D* value model result for each combination. (c) Results of the iterative optimization model, showing source combinations tested, and KS test *D* value for each combination. (d–f) Model results using the KS test *D* statistic (green) and Kuiper test *V* statistic (red) plotted as cumulative distribution functions (CDFs), and Cross-correlation coefficient (blue) as probability density plots. Black lines represent the mixed sample age distribution.

results, with minimum *D* of 0.053–0.055, minimum *V* of 0.098, and a maximum *R*² value of 0.778 (Figures 8d–8f).

Inverse Monte Carlo model results for the Loess plateau data are generally similar to those presented in *Licht et al.* [2016] (Figure 10), regardless the source scaling method employed (scaling whole source distributions or randomly sampling ages from within those distributions, Figure 1). Figure 10b shows Monte Carlo model source combinations plotted against KS test *D* values, with sample source contributions of 8.3 ± 4.6% for Mu Us, 11.7 ± 2.6% for the Central Deserts, 22.2 ± 5.5% for Qaidam, and 57.8 ± 8.1% for the Yellow River. These results are generally similar to those of *Licht et al.* [2016], who report source contributions of 6–8% from the Mu Us, 12–14% for the Central Deserts, 14–20% for the Qaidam basin, and 58–68% for the Yellow river. The Monte Carlo model yields minimum *D* and *V* values of 0.040 and 0.055, and a maximum *R*² value of 0.909 (Figures 10d–10f). Optimization routines yield no improvement in *D*, *V* or *R*² values, but give slightly different best fit model weights compared with those determined through inverse Monte Carlo modeling: 1% for Mu Us, 12% for the Central Deserts, 27% for Qaidam, and 60% for the Yellow River.

5. Discussion

5.1. Differences Among Comparison Methods

Quantitative comparison methods employed in the mixing model (KS *D*, Kuiper *V*, and Cross-correlation) give broadly similar results based on tests involving synthetic and empirical data. Each method is capable

of reproducing mixing proportions of source age distributions with a high degree of confidence from both simple and complex synthetic data (Figures 3 and 4) and empirical data (Figures 8 and 10). Sensitivity testing of complex synthetic data demonstrates each method's ability to accurately quantify source contribution of detrital age distributions if the source and mixed samples are adequately characterized. The mixture model is also capable of covering the entire range of combinations of a given set of source samples, as demonstrated in the visualized Monte Carlo plots in Figures 3b, 4b, 8b, and 10b. Complete coverage of each potential source sample along both the *x* and *y* axes in these plots is indicative of three things: (1) the mixture model successfully samples the full range of source combinations, including extreme weights, (2) the model disregards source samples that do not contribute to the mixed sample age distribution, and (3) a mixture of source samples, rather than any single sample, yields the closest approximation to the mixed sample for data sets tested here. This gives credibility to the random weighting method implemented in the mixing model (Figures 2b and 2c), and would likely not be possible had alternative weighting schemes been implemented (Figure 2a).

Comparison methods yield similar results when scaling whole source age distributions for comparison to mixed samples, but differ when subsampling raw source ages. Cross-correlation of PDPs has a clear advantage over the other methods when scaling whole source distributions (Figures 1b and 1c, left), as demonstrated by sensitivity test results showing lower absolute residuals and associated standard deviations with fewer model trials (Figure 5). However, when randomly subsampling ages to build source distributions (Figures 1b and 1c, right) the Cross-correlation of PDPs and KDEs appear to perform poorly compared to the KS and Kuiper *D* and *V* methods (Figure 6). We attribute these highly variable results of Cross-correlation to finite mixture distributions being inherently more complex, with the chance of higher variability in distribution shape. Further, Cross-correlation of PDPs yields highly variable results even compared to Cross-correlation of KDEs (Figure 6). This latter point is likely a result of highly differing sample uncertainties in the North America empirical data compilation [*Laskowski et al.*, 2013], and potentially exacerbated by selection of small numbers of ages from some samples (e.g., 5% of 100 ages would require that only 5 ages would be drawn from that source).

On consideration of the empirical river and catchment samples, all of the comparison methods yield imperfect model fits. Because we know the model is capable of reproducing known mixture proportions when there is indeed a viable solution (e.g., Figures 3 and 4), we interpret this poor fit as an indication the source samples cannot reproduce the mixed river sample in any mixture, and that perhaps the sediment sources in the drainage basin, the mixed sample, or both, have been inadequately characterized.

Our preferred comparison methods differ depending on type of source scaling (Figures 1b and 1c). Cross-correlation of mixture distributions is our preferred comparison method when scaling whole source distributions for several reasons which can be summarized as follows: (1) it appears to be the most sensitive with more discriminatory power than the KS test *D* statistic and the Kuiper test *V* statistic based on tests of complex synthetic data, (2) it requires fewer trials to produce model results with lower residuals than the KS *D* and Kuiper *V* statistics, and (3) it hones in more quickly to reasonable estimates for source weights when there is a viable solution from the input data (Figure 5c). Our preferred comparison method when subsampling raw source ages is the Kuiper test *V* statistic primarily based on consistently low residuals when varying the number of subsampled ages (*s*) (Figure 6c). That said, results are variable among comparison methods when subsampling source ages, and never produce model fits as good as when scaling whole source distributions (Figure 6c), which warrants further testing of this technique. However, a test of this would require very large input data sets subsampled to large model comparisons (*s* > 1000). Unfortunately, it is extremely computationally intensive to perform such a test. Users are encouraged to interpret results with caution for all comparison methods when subsampling raw source age distributions, especially given the amount of relatively low-*n* legacy data sets in the literature.

5.2. Implications for Applied Studies

Although our preferred method for quantitative comparison is the Cross-correlation coefficient of PDPs with scaling whole source distributions, we recommend the use of at least both the Cross-correlation coefficient and Kuiper test *V* statistic, as each has particular strengths and weaknesses. This is particularly important for applied studies of ancient sample data where the mixture of sources is always nonunique, and it is

more difficult to determine source contribution than, for example, the *Saylor et al*. [2013] modern river data presented above. This latter point raises an important question of what does it mean if the mixing model is not capable of reproducing a mixture of input sources with a high degree of confidence (high KS *D* and Kuiper *V*, and/or low Cross-correlation R^2 values)? For the modern river study, the mixing model can only produce a maximum Cross-correlation of PDPs of 0.77 for Rio Cusiana. For comparison, this is higher than equally weighted source samples and source samples scaled by catchment outcrop area (0.64 and 0.73, respectively) [*Saylor et al.*, 2013]. Although higher values means the Monte Carlo mixing model does a better job of matching the river sample age distributions, there is clearly something missing. It is not that the sample data are too complex to match because when there is a solution, as in the sensitivity test example (Figure 5) and forward optimization results of synthetic data (Figures 3c and 4c), the mixing model can clearly find it. We interpret such situations of poor model fits as an inadequate characterization of the source and/or mixed sample, possibly due to small sample sizes. It may also reflect failure to identify some age component(s) in the source area.

The Loess Plateau model results provide a nice comparison to the Colombia modern river results because it is a study in the ancient rather than the modern, it is on a much larger scale (1000s of km), and it incorporates much larger-*n* data sets for comparison. The Loess Plateau model results, although not perfect, give much better model results compared to the Colombia results. It appears sample size is a dominant control on model results, and low-*n* source and/or mixed samples should be treated with caution as they may lead to misinterpretation of sediment provenance.

For the *Saylor et al*. [2013] study, as well as other studies in both the modern and ancient, uncertainty can be mitigated in a number of ways: (1) by increasing the sample size (*n*) for all samples, both for source and mixed samples, (2) by eliminating proxy source samples and only sampling within the drainage basins (specific only to modern river studies), (3) by considering differences in zircon abundance within different source units to avoid bias based on zircon size due to hydrodynamic sorting [*Morton and Hallsworth*, 1994; *Gehrels*, 2000; *Garzanti et al.*, 2009; *Lawrence et al.*, 2011], or differential zircon fertility [e.g., *Moecher and Samson*, 2006; *Dickinson*, 2008], and (4) by analyzing multiple samples from the same interval or broadening the number of source intervals considered to better characterize potential sources. This aspect of the research provides important feedbacks to experiment design and execution. Having demonstrated that the model can find a perfect solution when one exists will hopefully guide future provenance research to search for a mixture of sources that provides such a perfect fit.

5.3. Mixing Model Implementation
The Monte Carlo mixing model has been developed into a MATLAB-based graphical user interface (GUI). The GUI is available as a stand-alone executable (.exe file), along with the annotated source script in supporting information. We also include a step-by-step user manual, with which the results presented in this manuscript are easily reproduced for both synthetic and empirical data sets.

All results produced by the mixing model are nonunique, as the mixing model will estimate mixing proportions regardless of the quality of the comparison fit. All results should be interpreted with a thorough understanding of geologic context, and in consideration of alternative comparison techniques such as multidimensional scaling [e.g., *Vermeesch*, 2013]. Furthermore, although there is a significant improvement in model fits with an increased number of model trials (Figure 5), there is a tradeoff between compute time and desired goodness of fit. How many models trials required for the mixing model to predict accurate and robust results is highly dependent on sample complexity; the number of model trials is only limited by the computer's memory. Sensitivity testing of the Monte Carlo method, as discussed in section 3.2, is essential in order to develop a first-order sense of source and mixed sample complexity, and to determine what is required to produce a model fits within desired limits.

6. Conclusions

We present a MATLAB-based inverse Monte Carlo method of randomly constructing source age distributions for comparison to individual mixed samples to determine mixing proportions of source sample contributions. Inverse Monte Carlo results may be used to constrain forward optimization routines to find a single best model fit. This model is applied to synthetic data in a proof-of-concept style test of

the model, and published detrital zircon U-Pb empirical data from North America [*Laskowski et al.*, 2013], Colombia [*Saylor et al.*, 2013] and central China [*Licht et al.*, 2016]. Sensitivity testing of complex synthetic data show the mixing model is capable of sampling the full range of source sample weightings, and can produce sample age distributions as CDFs and PDPs very similar to the known distribution; poor model fits are attributed to inadequate characterization of source and/or mixed samples. Sensitivity testing of subsampling source distributions for comparison to mixed samples based on empirical data from North America [*Laskowski et al.*, 2013] shows results are highly variable among comparison methods, and users are cautioned in this approach for low-*n* data sets. The mixing model presented here, source code, and user manual, are available in the supplemental material; the most up-to-date versions may be downloaded from https://www.kurtsundell.com/downloads, or https://github.com/kurtsundell/DZmix.

References

Amidon, W. H., D. W. Burbank, and G. E. Gehrels (2005), Construction of detrital mineral populations: Insights from mixing of U–Pb zircon ages in Himalayan rivers, *Basin Res.*, *17*(4), 463–485.

Andersen, T., M. Elburg, and A. Cawthorn-Blazeby (2016a), U–Pb and Lu–Hf zircon data in young sediments reflect sedimentary recycling in eastern South Africa, *J. Geol. Soc.*, *173*(2), 337–351.

Botev, Z. I., J. F. Grotowski, and D. P. Kroese (2010), Kernel density estimation via diffusion, *Ann. Stat.*, *38*(5), 2916–2957.

Campbell, I. H., P. W. Reiners, C. M. Allen, S. Nicolescu, and R. Upadhyay (2005), He–Pb double dating of detrital zircons from the Ganges and Indus Rivers: Implication for quantifying sediment recycling and provenance studies, *Earth Planet. Sci. Lett.*, *237*(3), 402–432.

Cooper, M., F. Addison, R. Alvarez, M. Coral, R. H. Graham, A. Hayward, S. Howe, J. Martinez, J. Naar, and R. Peñas (1995), Basin development and tectonic history of the Llanos Basin, Eastern Cordillera, and middle Magdalena Valley, Colombia, *AAPG Bull.*, *79*(10), 1421–1442.

Dickinson, W. R. (2008), Impact of differential zircon fertility of granitoid basement rocks in North America on age populations of detrital zircons and implications for granite petrogenesis, *Earth Planet. Sci. Lett.*, *275*(1), 80–92.

Garzanti, E., S. Andò, and G. Vezzoli (2009), Grain-size dependence of sediment composition and environmental bias in provenance studies, *Earth Planet. Sci. Lett.*, *277*(3), 422–432.

Gehrels, G. E. (2000), Introduction to detrital zircon studies of Paleozoic and Triassic strata in western Nevada and northern California, *Spec. Pap. Geol. Soc. Am.*, *347*, 1–17.

Gehrels, G. E., R. Blakey, K. E. Karlstrom, J. M. Timmons, B. Dickinson, and M. Pecha (2011), Detrital zircon U-Pb geochronology of Paleozoic strata in the Grand Canyon, Arizona, *Lithosphere*, *3*(3), 183–200.

Horton, B. K., M. Parra, J. E. Saylor, J. Nie, A. Mora, V. Torres, D. F. Stockli, and M. R. Strecker (2010), Resolving uplift of the northern Andes using detrital zircon age signatures, *GSA Today*, *20*(7), 4–10.

Kimbrough, D. L., M. Grove, G. E. Gehrels, R. J. Dorsey, K. A. Howard, O. Lovera, A. Aslan, P. K. House, and P. A. Pearthree (2015), Detrital zircon U-Pb provenance of the Colorado River: A 5 my record of incision into cover strata overlying the Colorado Plateau and adjacent regions, *Geosphere*, *11*(6), 1719–1748.

Kuiper, N. H. (1960), Tests concerning random points on a circle, in *Indagationes Mathematicae* (Proceedings), vol. 63, pp. 38–47, Elsevier.

Laskowski, A. K., P. G. DeCelles, and G. E. Gehrels (2013), Detrital zircon geochronology of Cordilleran retroarc foreland basin strata, western North America, *Tectonics*, *32*, 1027–1048, doi:10.1002/tect.20065.

Lawrence, R. L., R. Cox, R. W. Mapes, and D. S. Coleman (2011), Hydrodynamic fractionation of zircon age populations, *Geol. Soc. Am. Bull.*, *123*(1–2), 295–305.

Licht, A., A. Pullen, P. Kapp, J. Abell, and N. Giesler (2016), Eolian cannibalism: Reworked loess and fluvial sediment as the main sources of the Chinese Loess Plateau, *Geol. Soc. Am. Bull.*, *128*(5–6), 944–956.

Moecher, D. P., and S. D. Samson (2006), Differential zircon fertility of source terranes and natural bias in the detrital zircon record: Implications for sedimentary provenance analysis, *Earth Planet. Sci. Lett.*, *247*(3), 252–266.

Morton, A. C., and C. Hallsworth (1994), Identifying provenance-specific features of detrital heavy mineral assemblages in sandstones, *Sediment. Geol.*, *90*(3), 241–256.

Perez, N. D., and B. K. Horton (2014), Oligocene-Miocene deformational and depositional history of the Andean hinterland basin in the northern Altiplano plateau, southern Peru, *Tectonics*, *33*(9), 1819–1847, doi:10.1002/2014TC003647.

Press, W. H. (2007), Numerical recipes, in *The Art of Scientific Computing*, 3rd ed., Cambridge Univ. Press, Cambridge, U. K.

Pullen, A., M. Ibáñez-Mejía, G. E. Gehrels, J. C. Ibáñez-Mejía, and M. Pecha (2014), What happens when n = 1000? Creating large-n geochronological datasets with LA-ICP-MS for geologic investigations, *J. Anal. Atom. Spectrom.*, *29*(6), 971–980.

Satkoski, A. M., B. H. Wilkinson, J. Hietpas, and S. D. Samson (2013), Likeness among detrital zircon populations—An approach to the comparison of age frequency data in time and space, *Geol. Soc. Am. Bull.*, *125*(11–12), 1783–1799.

Saylor, J. E., and K. E. Sundell (2016), Quantifying comparison of large detrital geochronology data sets, *Geosphere*, *12*(1), 1–18, doi:10.1130/GES01237.1.

Saylor, J. E., D. F. Stockli, B. K. Horton, J. Nie, and A. Mora (2012), Discriminating rapid exhumation from syndepositional volcanism using detrital zircon double dating: Implications for the tectonic history of the eastern cordillera, *Geol. Soc. Am. Bull.*, *124*, 762–779.

Saylor, J. E., J. N. Knowles, B. K. Horton, J. Nie, and A. Mora (2013), Mixing of source populations recorded in detrital zircon U-Pb age spectra of modern river sands, *J. Geol.*, *121*(1), 17–33.

Stephens, M. A. (1970), Use of the Kolmogorov-Smirnov, Cramér-Von Mises and related statistics without extensive tables, *J. R. Stat. Soc., Ser. B*, *32*(1), 115–122.

Vermeesch, P. (2013), Multi-sample comparison of detrital age distributions, *Chem. Geol.*, *341*, 140–146.

Vermeesch, P., and E. Garzanti (2015), Making geological sense of "Big Data" in sedimentary provenance analysis, *Chem. Geol.*, *409*, 20–27.

Vermeesch, P., A. Resentini, and E. Garzanti (2016), An R package for statistical provenance analysis, *Sediment. Geol.*, *336*, 14–25.

Wissink, G. K., and G. D. Hoke (2016), Eastern margin of Tibet supplies most sediment to the Yangtze River, *Lithosphere*, *8*(6), 601–614.