Supplemental Text

Refined sediment source characterization via bivariate non-negative matrix factorization of detrital provenance indicators

Joel E. Saylor

Kurt E. Sundell

Three Sources: Sources were derived from a global database of U–Pb and εHfT data from detrital zircons (Puetz and Condie, 2019). Sundell and Macdonald (in review) grouped the samples in the database by landmass following the suture boundaries identified by Domeier and Torsvik (2014). We use data from Africa (n=12,235), Antarctica (n=1,373), and South America (n=7,872) (**Figure S1A and Supplemental Table S1**). These 3 sources were mixed into 24 sink samples where all sink samples have 1,373 age/εHfT pairs (**Supplemental Table S2**).

Breakpoint analysis of the 24 sink samples mixed from three sources indicates that the optimum rank for factorization is three (**Figure S2A**). In contrast, bi-cross validation yields a range of optimum ranks between 3 and 13 for quadrants A–D (**Figure S2B**). The total residual from bi-cross-validation indicates that the optimum rank is 5 (**Figure S2C**). Factorization of the 24 sink samples to 3 endmembers yields distributions and weightings that closely match the known distributions and weightings (**Figure S1B and S3**).

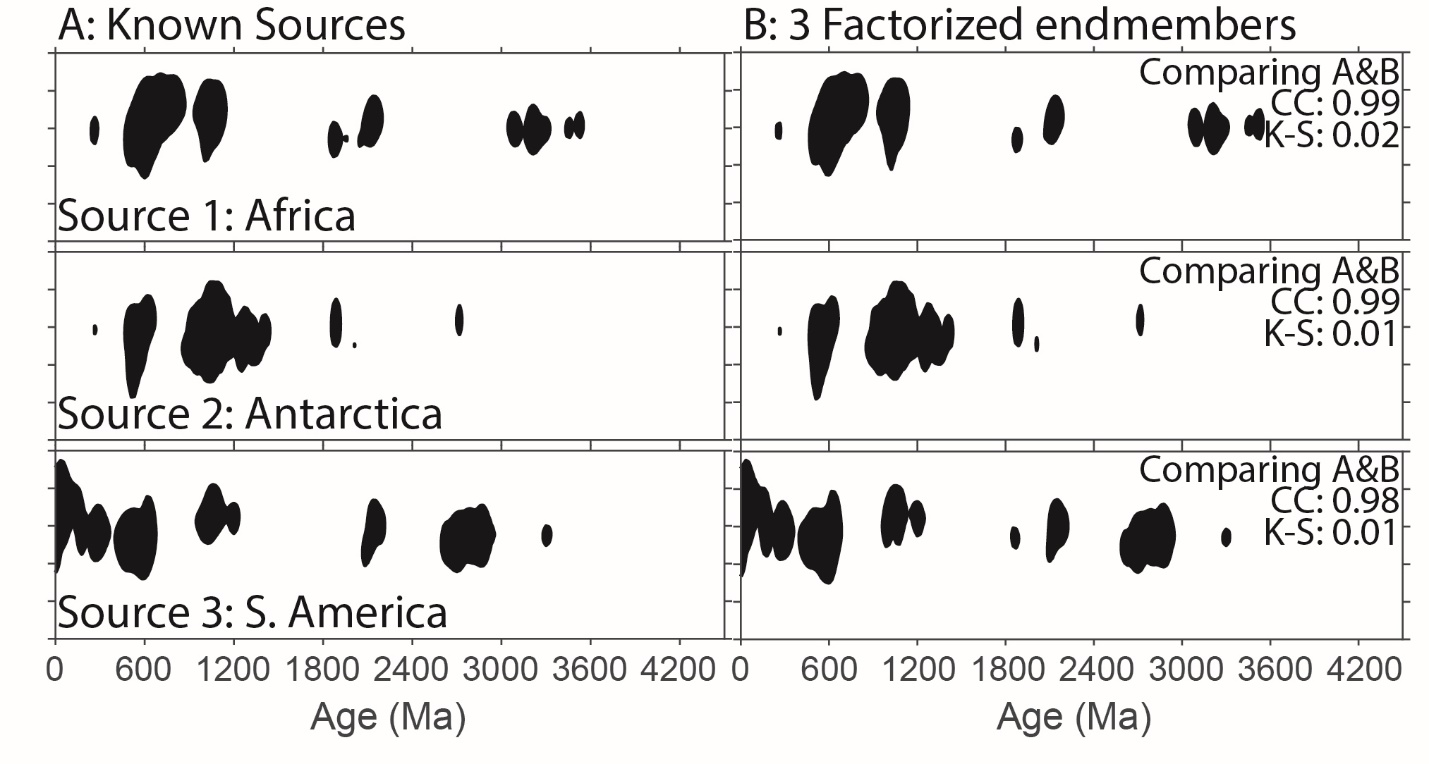
Five sources: Sources are based on a compilation of Paleozoic–Proterozoic samples from North America (**Figure S4A and Supplemental Table S4 and S5**). Source 1 includes 845 age/εHfT pairs from Proterozoic–Cambrian basement and cover sequences from the Mojave desert in California and Nevada (Wooden et al., 2013). Source 2 comprises 371 age/εHfT pairs from Mississippian–Permian samples from the Appalachian foreland (Thomas et al., 2017) and detritus recycled from the southern Appalachian Orogen into Pleistocene beach ridge sands in Florida (Mueller et al., 2008). Source 3 is an amalgamation of 95 age/εHfT pairs from the Neoproterozoic Caddy and Mutual quartzites exposed in Utah (Gehrels and Pecha, 2014). Source 4 is composed of 91 age/εHfT pairs from Mesoproterozoic strata of the upper Belt Supergroup of Idaho and western Montana (Stewart et al., 2010). Finally, Source 5 is a combination of 334 age/εHfT pairs from upper Neoproterozoic and Paleozoic strata from Newfoundland (Pollock et al., 2015). We mixed these five sources into 40 sinks where all sink samples have 91 age/εHfT pairs.

Breakpoint analysis of the 40 sink samples mixed from five sources indicates that the optimum rank for factorization is 4, although rank 5 is only slightly worse (**Figure S5A, Supplemental Table S8**). Bi-cross validation of this data set yields a range of optimum ranks between 5 and 10 (**Figure S5B**) for the individual quadrants. The total residual indicates that the optimum rank is 10 (**Figure S5C**).

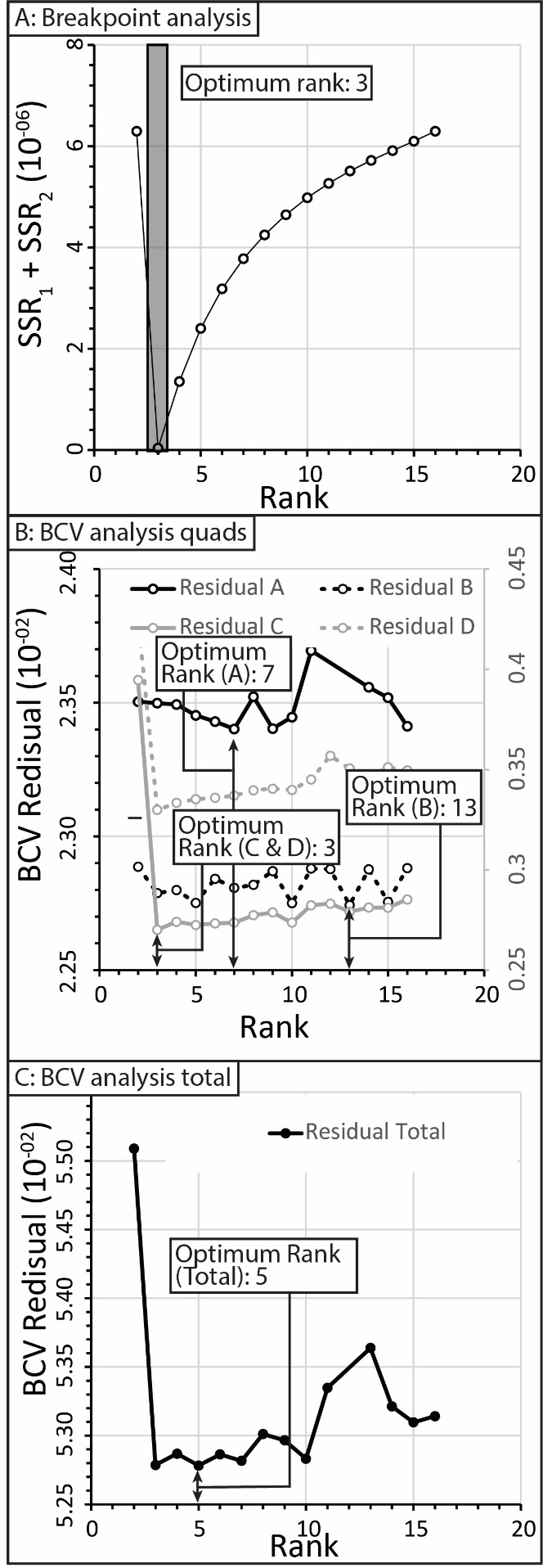
Since breakpoint analysis indicates that a rank of either 4 or 5 yields optimum factors (**Figure S5A**), we factorized the 40 sink samples mixed from 5 sources to both 4 and 5 endmembers (**Figure S4B and C**). The Cross-correlation coefficient and KS D value indicate close similarity between the known sources to both four or five endmembers. However, there is a slightly better match between the sources factorized for rank five than for rank four (compare **Figure S4B and C**). The mean Cross-correlation coefficient between factorized endmembers for rank five and the known sources is 0.98. In comparison the mean Cross-correlation coefficient between factorized endmembers for rank four and the known sources is 0.96, or 0.83 if the best match to Input 3 is included.

Factorization of the 5-source data set to 4 or 5 endmembers yielded a mean Cross-correlation coefficient between the factorized sources and their closest known matches of 0.96 and 0.98, respectively. When comparing mixing weights for the 5-source data set and its factorized equivalents, the R2 between known and factorized mixing weights is slightly higher for the 5-source factorization (0.93) than for the 4-source factorization (0.87) (**Figure S4**).

There is close correspondence between the mixture weighting used to mix the five sources into 40 sink samples and the weightings based on factorization of the 40 sinks (**Figure S6**). Cross-plotting mixture weights from the factorization to rank five and known mixture weights yields a linear fit with a slope of 0.99 and a coefficient of determination of 0.93 (**Figure S6A**). Cross-plotting mixture weights from the factorization to rank four and known mixture weights yields slightly more poor fit with a linear fit with a slope (1.04) which deviates from one by more than the slope for rank five (**Figure S6B**). It also has a non-zero intercept (0.03) and a lower coefficient of determination (0.87).



**Figure S1**. Comparison of known sources and factorized sources for the 3-source data set. (A) Eight source kernel density estimates that were used as sources and mixed to form 24 random sink samples. (B) Factorization of the 24 samples to 3 endmembers yields distributions that closely match the original five sources as indicated by Cross-correlation coefficients approaching 1 and D values approaching 0. KDEs are based on a 20 Myr bandwidth for age, and a 4 ε-unit bandwidth for εHfT and are plotted it as the contour of the median (50th percentile).

Figure S2. (A) Breakpoint analysis based on factorization of 24 sink samples from three sources to ranks between 2–16 yields a minimum sum of squared residuals at rank three. (B) Bi-cross validation of based on factorization of four quadrants (i.e., A, B, C, D in equation 5) from the 24 sink samples for ranks 2–16 yields minima in residual at ranks ranging from three to 13. (C) The sum of residuals of each of the four quadrants yields a minimum at a rank of five. Data for this figure are presented in Supplemental Table S3.

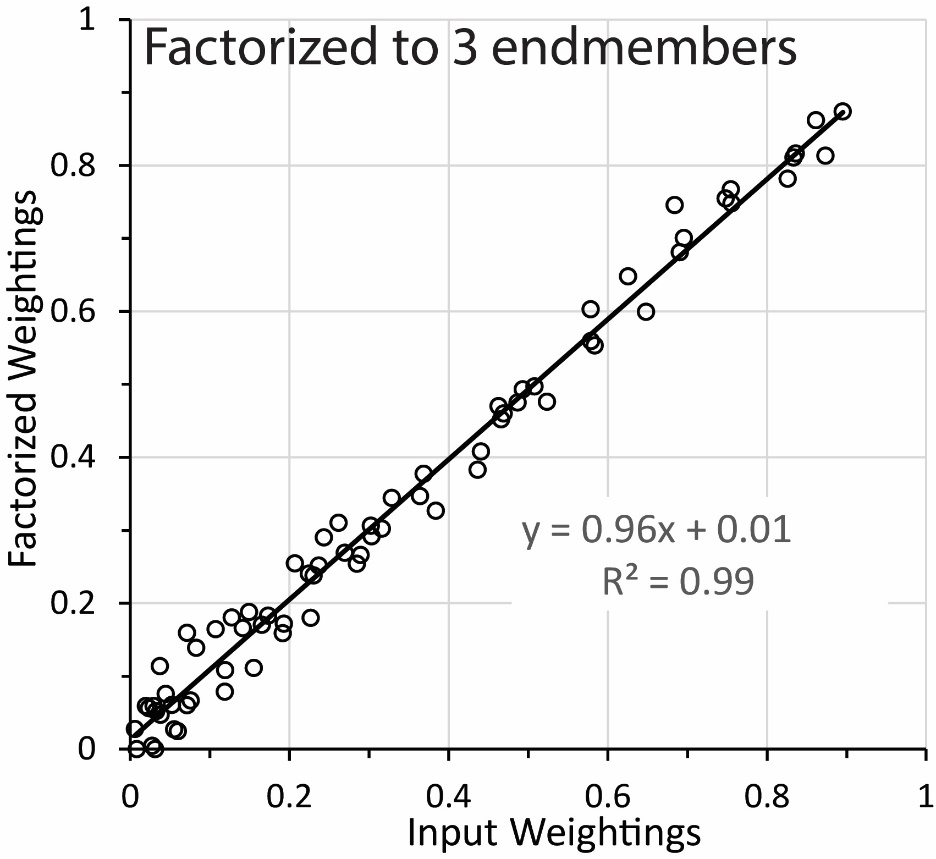
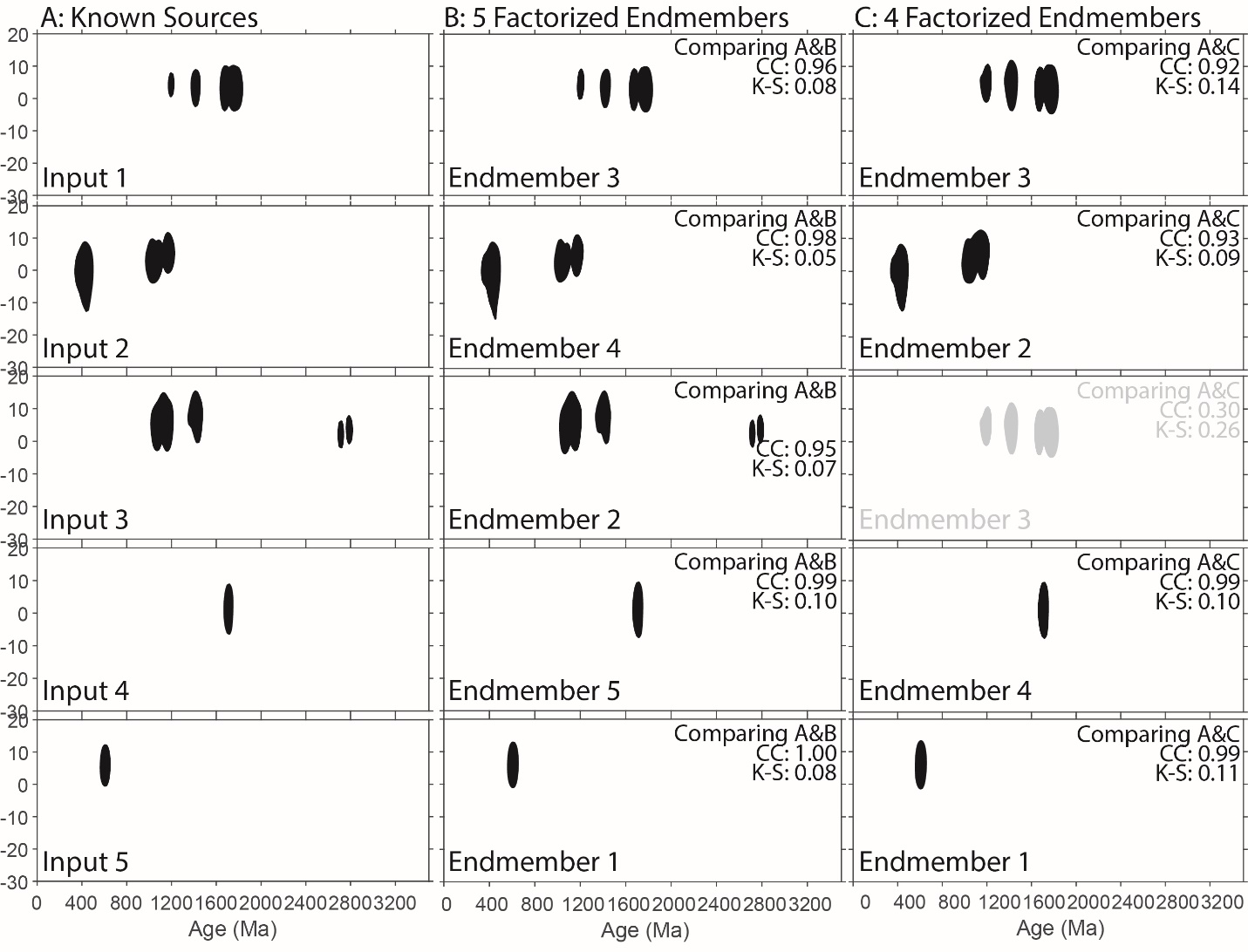
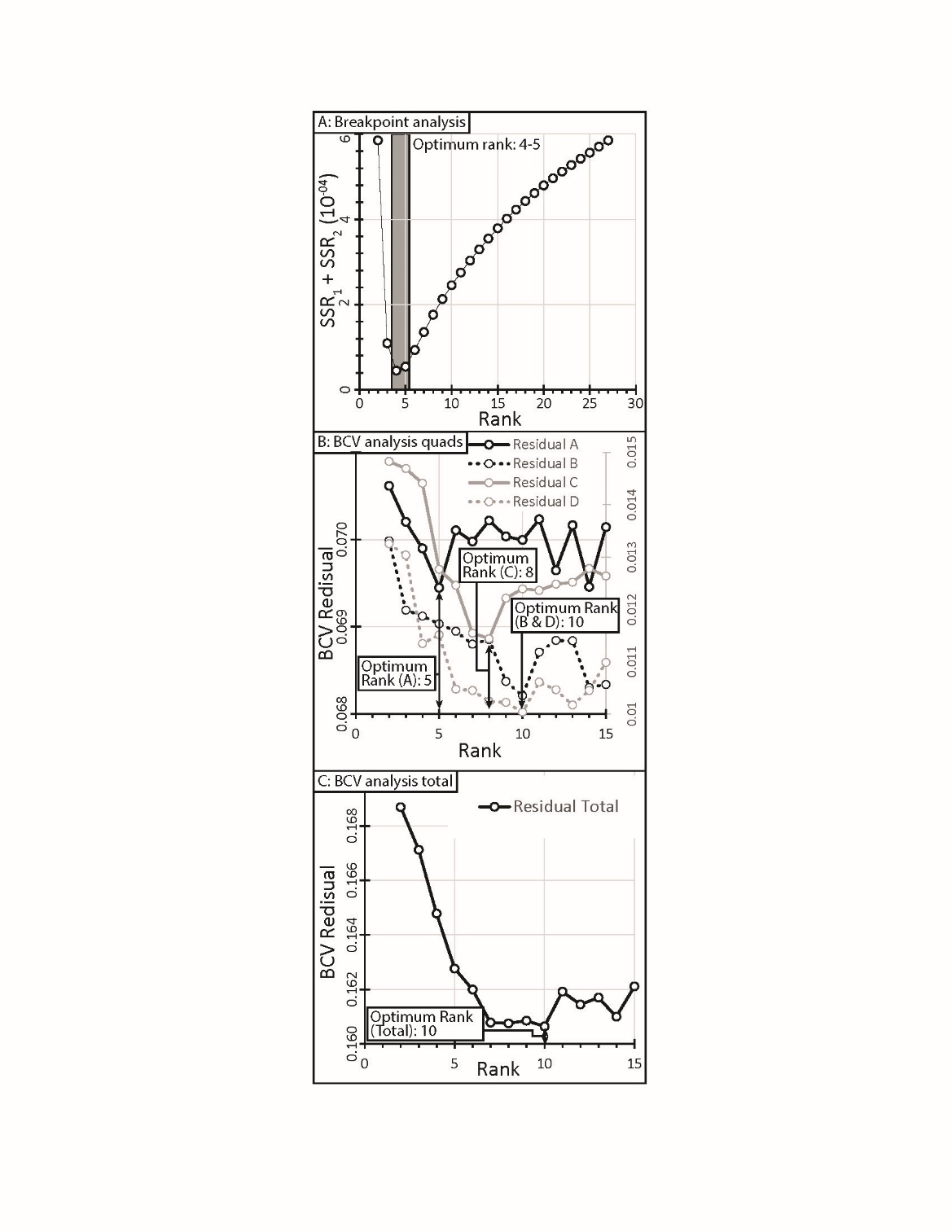
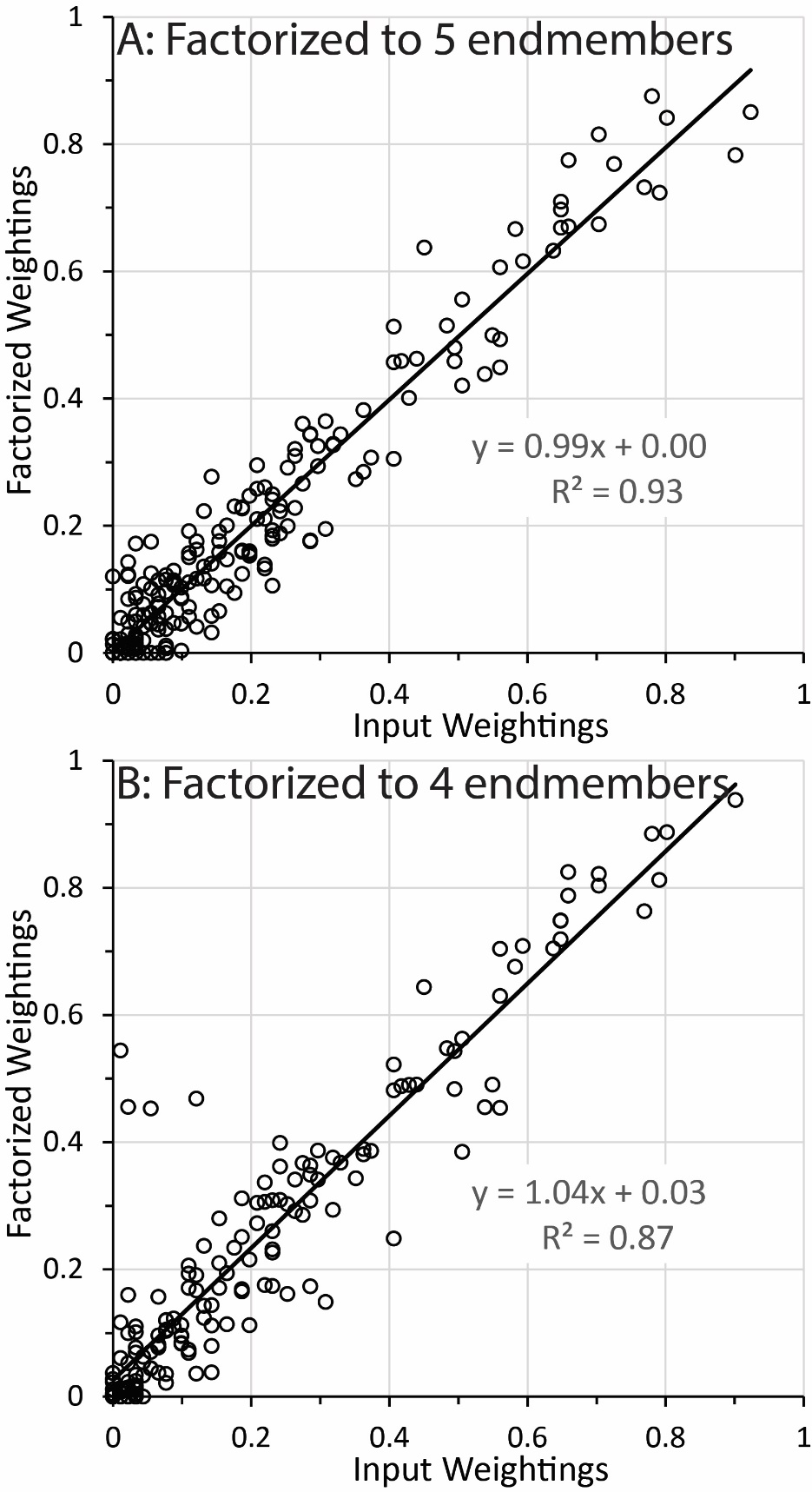


Figure S3. Comparison of the known weights for the 24 sink samples mixed from three known sources to the mixing weights based on factorization of the sink samples to rank three.

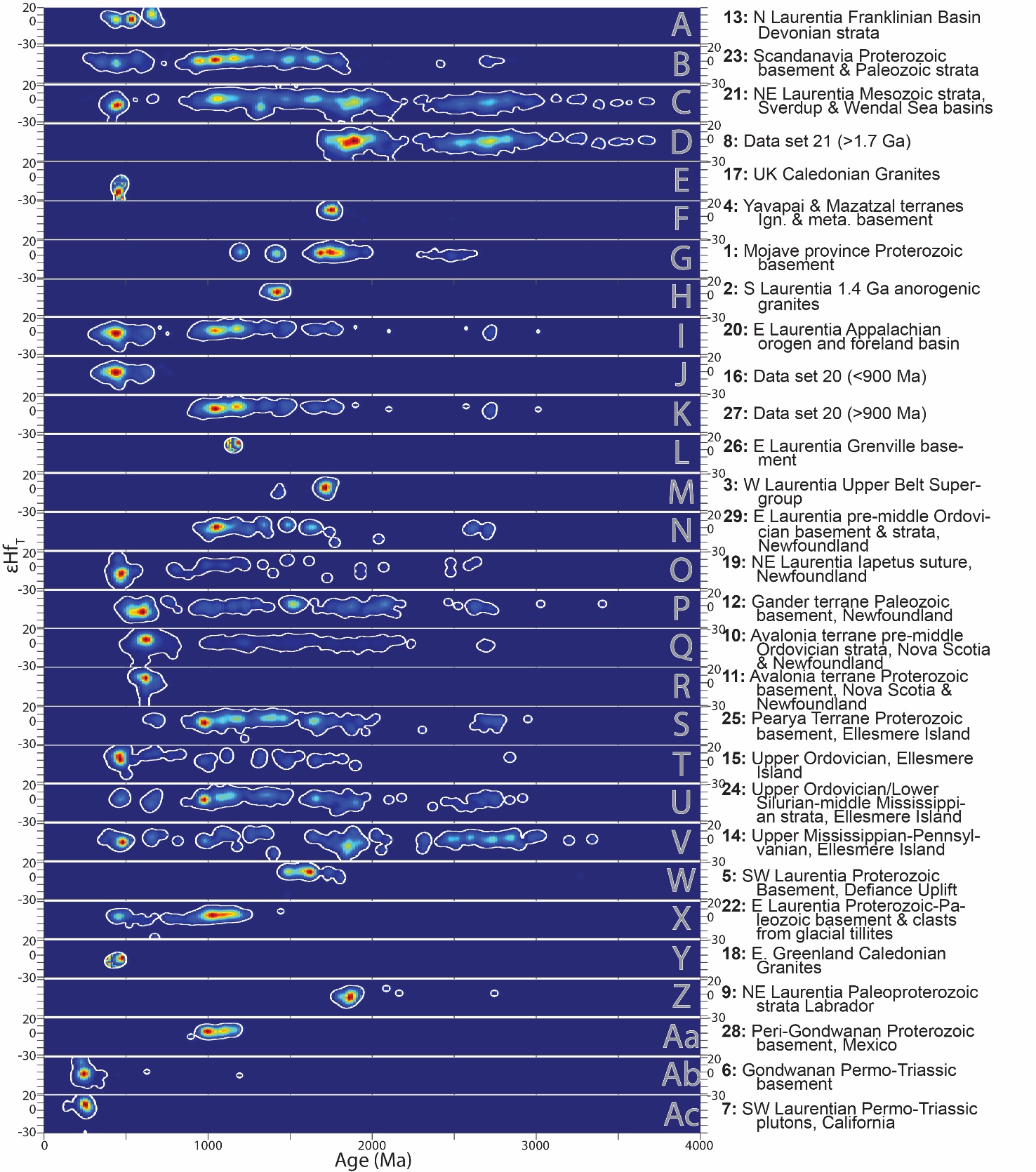


**Figure S4.** Comparison of known sources and factorized endmembers for the 5-source data set. (A) Five source kernel density estimates that were used as sources and mixed to form 40 random sink samples. (B) Factorization of the 40 samples to 5 endmembers yields distributions that closely match the original five sources as indicated by Cross-correlation values (CC) approaching 1 and Kolmogorov-Smirnov D values (K-S) approaching 0. (C) Factorization of the 40 sink samples to 4 endmembers yields distribution that match the original five more poorly. KDEs are based on a 20 Myr bandwidth for age, and a 4 ε-unit bandwidth for εHfT and are plotted it as the contour of the median (50th percentile).

**Figure S5.** (A) Breakpoint analysis based on factorization of 40 sink samples from five sources to ranks between 2–27 yields a minimum sum of squared residuals at rank four, closely followed by rank five. (B) Bi-cross validation of based on factorization of four quadrants (i.e., A, B, C, D in equation 5) from the 40 sink samples for ranks 2–27 yields minima in residual at ranks ranging from five to 10. (C) The sum of residuals of each of the four quadrants yields a minimum at a rank of 10. Data for this figure are presented in Supplemental Table S8



**Figure S6.** Comparison of the known weights for the 40 sink samples mixed from five known sources to the mixing weights based on factorization of the sink samples. A) Comparison based on factorization to rank five. B) Comparison based on factorization to rank four. For factorization to rank four the comparison is based on the weightings for the four factorized sources that most closely match the known counterparts (see **Figure 2C**). Data for this figure are presented in Supplemental Table S8.



**Figure S7**. Bivariate (U-Pb age and εHft) distributions of the 29 empirical source samples to which factorized sources were compared. Densities are shown as heat maps with greater probability shown in red and lower probability shown in blue. White contour is at the 95th percentile.

Works Cited

Domeier, M., Torsvik, T.H., 2014. Plate tectonics in the late Paleozoic. Geoscience Frontiers 5, 303-350.

Gehrels, G., Pecha, M., 2014. Detrital zircon U-Pb geochronology and Hf isotope geochemistry of Paleozoic and Triassic passive margin strata of western North America. Geosphere 10, 49–65.

Mueller, P.A., Kamenov, G.D., Heatherington, A.L., Richards, J., 2008. Crustal evolution in the southern Appalachian orogen: Evidence from Hf isotopes in detrital zircons. The Journal of Geology 116, 414-422.

Pollock, J.C., Sylvester, P.J., Barr, S.M., 2015. Lu–Hf zircon and Sm–Nd whole-rock isotope constraints on the extent of juvenile arc crust in Avalonia: examples from Newfoundland and Nova Scotia, Canada. Can. J. Earth Sci. 52, 161-181.

Puetz, S.J., Condie, K.C., 2019. Time series analysis of mantle cycles Part I: Periodicities and correlations among seven global isotopic databases. Geoscience Frontiers 10, 1305-1326.

Stewart, E.D., Link, P.K., Fanning, C.M., Frost, C.D., McCurry, M., 2010. Paleogeographic implications of non–North American sediment in the Mesoproterozoic upper Belt Supergroup and Lemhi Group, Idaho and Montana, USA. Geology 38, 927–930.

Sundell, K.E., Macdonald, F.A., in review. Crustal recycling though supercontinent cycles. Science Advances.

Thomas, W.A., Gehrels, G.E., Greb, S.F., Nadon, G.C., Satkoski, A.M., Romero, M.C., 2017. Detrital zircons and sediment dispersal in the Appalachian foreland. Geosphere 13, 2206-2230.

Wooden, J.L., Barth, A.P., Mueller, P.A., 2013. Crustal growth and tectonic evolution of the Mojave crustal province: Insights from hafnium isotope systematics in zircons. Lithosphere 5, 17-28.