Quaternary Research (2018), **89**, 365–373.
Copyright © University of Washington. Published by Cambridge University Press, 2017.
doi:10.1017/qua.2017.78

Unmixing grain-size distributions in lake sediments: a new method of endmember modeling using hierarchical clustering

Xiaonan Zhang[a]*, Aifeng Zhou[a], Xin Wang[a], Mu Song[b], Yongtao Zhao[a], Haichao Xie[a], James M. Russell[c], Fahu Chen[a,d]*

[a]Key Laboratory of Western China's Environmental Systems (Ministry of Education), College of Earth and Environmental Sciences, Lanzhou University, Lanzhou 730000, China
[b]Department of Earth Sciences, The University of Hong Kong, Hong Kong 999077, China
[c]Department of Earth, Environmental, and Planetary Sciences, Brown University, Providence, Rhode Island 02912, USA
[d]Institute of Tibetan Plateau Research, Chinese Academy of Science, Beijing 100101, China

(Received February 19, 2017; Accepted August 14, 2017)

Abstract

The grain-size distribution (GSD) of sediments provides information on sediment provenance, transport processes, and the sedimentary environment. Although a wide range of statistical parameters have been applied to summarize GSDs, most are directed at only parts of the distribution, which limits the amount of environmental information that can be retrieved. Endmember modeling provides a flexible method for unmixing GSDs; however, the calculation of the exact number of endmembers and geologically meaningful endmember spectra remain unresolved using existing modeling methods. Here we present the methodology hierarchical clustering endmember modeling analysis (CEMMA) for unmixing the GSDs of sediments. Within the CEMMA framework, the number of endmembers can be inferred from agglomeration coefficients, and the grain-size spectra of endmembers are defined on the basis of the average distance between the samples in the clusters. After objectively defining grain-size endmembers, we use a least squares algorithm to calculate the fractions of each GSD endmember that contributes to individual samples. To test the CEMMA method, we use a grain-size data set from a sediment core from Wulungu Lake in the Junggar Basin in China, and find that application of the CEMMA methodology yields geologically and mathematically meaningful results. We conclude that CEMMA is a rapid and flexible approach for analyzing the GSDs of sediments.

Keywords: Grain-size distribution; Endmember; Hierarchical clustering analysis; Unmixing; Lake sediments

INTRODUCTION

Grain-size distributions (GSDs) are one of the most widely used proxies in paleoenvironmental and paleoclimatological investigations of sedimentary deposits, as they provide information on sediment source, depositional processes, and the sedimentary environment (Visher, 1969; Ghosh and Mazumder, 1981; Qiang et al., 2007; He et al., 2015). Statistical parameters such as the mean, median, standard, deviation, kurtosis, and skewness have been used to characterize GSDs and to infer variations in hydrodynamic conditions, eolian activity, and sediment source (Inman, 1952; Blott and Pye, 2001; Fournier et al., 2014). In addition to these summary statistics, approaches to the decomposition of grain-size frequency distributions into components include graphic methods (e.g., CM patterns, where C the one percentile and M the median diameter) (Passega, 1964), analytic methods (Court, 1949; Clark, 1976), and numerical methods (Clark, 1976). A more recent analytic approach is the standard deviation method, which is used for classifying the GSD to determine the most environmentally sensitive components (Boulay et al., 2003). However, these approaches are generally based on the analysis of one percentile-median diameter part of, rather than the entire, GSD.

Several methods have been developed for mathematical unmixing of complete GSDs, such as curve fitting (Sun et al., 2002; Paterson and Heslop, 2015), eigenspace analysis (Weltje, 1997; Weltje and Prins, 2003; Dietze et al., 2012), and a recently developed Bayesian method (Yu et al., 2015). These methods have proved effective in extracting information on provenance, transport processes, and the depositional environment of sediments; however, all of these methods have deficiencies. As discussed by Weltje and Prins (2007),

*Corresponding authors at: Key Laboratory of Western China's Environmental Systems (Ministry of Education), College of Earth and Environmental Sciences, Lanzhou University, Lanzhou 730000, China. E-mail addresses: fhchen@lzu.edu.cn (F. Chen); zhangxn2012@lzu.edu.cn (X. Zhang).

the curve-fitting method is based solely on the fitting of a series of empirical curves to the GSD and ignores their geologic context. Eigenspace analysis may produce uneven spectra because of the transformation and communality of grain-size compositional data, which makes interpretation difficult and confusing (Yu et al., 2015). The Bayesian method, because of the large number of iterations needed, may require a large amount of computation time if the data set is very large; furthermore, low-probability distributions may be ignored even though they may potentially be important.

Cluster analysis is a genetic type of multivariate statistical analysis and has been widely applied in statistics, mathematics, computer science, economics, and biology. Cluster analysis applied to sedimentary grain-size data is usually used for stratigraphic subdivision or to determine the sources of sediments (Donato et al., 2009; Grimm et al., 2011; Liu et al., 2017). However, most of the studies have only employed the method to group a limited number of GSD parameters, such as the mean and standard deviation, and few studies have taken advantage of the entire distribution (Zhou et al., 1991; Nelson et al., 2014; Ordóñez et al., 2016). In addition, because of the problems of determining cluster number and cluster centers, no prior study has used cluster analysis for endmember modeling. Here we propose a new endmember model for unmixing GSDs based on hierarchical clustering endmember modeling analysis (CEMMA). Within our new model, the number of endmembers is inferred based on changes in agglomeration coefficients, and the spectra of endmembers are determined from the average distance between the samples in the clusters. Our proposed method thus provides a new way to objectively determine the number and spectrum of GSD endmembers that contribute to sediment, allowing us to unmix varying GSDs in time.

OVERVIEW OF ENDMEMBER MODELING ANALYSIS

Endmember modeling of GSDs was first proposed by Weltje (1997), who showed that the compositional variation among a series of GSDs can be regarded as the result of the physical mixing of a fixed number of endmembers. This relationship can be expressed as a linear mixing model:

$$X = M * B + E. \qquad \text{(Eq. 1)}$$

In this model, matrix $X_{(n\times p)}$ is the GSD data set in which each row represents one observed GSD with p different sizes that sum to 100%. Matrix $M_{(n\times q)}$ represents the proportional contributions of endmembers to each observation; $B_{(q\times p)}$ is a matrix containing q endmembers, each of which is a vector consisting of p elements; and $E_{(n\times p)}$ represents the errors of the model. In order to accurately apply this model, three issues must be addressed: the number of endmembers, the spectra of the endmembers, and the fractions of the endmembers in each sample (Renner, 1991). These problems can be addressed by using endmember model analysis based on hierarchical clustering (HC), described subsequently.

HIERARCHICAL CLUSTERING ENDMEMBER MODELING ALGORITHM

The sediments accumulating within lake basins have a range of provenances and are transported by various mechanisms. Sediment accumulation integrates these sources and processes, such that the GSD of a given sediment sample represents a mixture of sediments that correspond to different provenances and/or transport mechanisms. However, it can be assumed that given a sufficiently long time series of GSDs, there will be relatively brief intervals in which sediment provenance and transport are sufficiently stable to produce uniform GSDs that will represent those specific conditions. Therefore, what is required is a method of determining GSDs that represent those intervals.

The hierarchical clustering algorithm

The HC algorithm organizes data into a hierarchical structure according to a proximity matrix. HC attempts to construct a treelike nested structure that partitions original data and builds a hierarchy of clusters (Xu and Wunsch, 2005). Two major issues must be solved when using HC analysis. The first is the similarity measure that can be used as a scalar distance between different clusters, and the second is the linkage method that orders the clusters to produce a unique and meaningful solution (Johnson, 1967; Gruvaeus and Wainer, 1972; Langfelder et al., 2008). In this study, we used the average linkage as the linkage method between groups, which is defined on the basis that the similarity between two clusters is equal to the mean distance between elements of each cluster (http://home.deib.polimi.it/matteucc/Clustering/tutorial_html/hierarchical.html [accessed December 6, 2016]), and the city-block distance, which is less influenced by very large differences between just a few of the variables in high dimensional vectors (Kaufman and Rousseeuw, 2009), is used to quantify the similarity between GSDs. The city-block distance between two GSDs ($\boldsymbol{x_i}$, $\boldsymbol{x_j}$) is expressed as follows:

$$d_{(i,j)} = \sum_{k=1}^{p} \left| x_{ik} - x_{jk} \right|, \qquad \text{(Eq. 2)}$$

where $d_{(i,\,j)}$ is the distance between two GSDs $\boldsymbol{x_i}$ and $\boldsymbol{x_j}$, p is the number of size classes in each GSD, and $d_{(i,\,j)}$ expresses the degree of difference between curves $\boldsymbol{x_i}$ and $\boldsymbol{x_j}$. If observations $\boldsymbol{x_i}$ and $\boldsymbol{x_j}$ are highly similar, $d_{(i,\,j)}$ will be small, whereas d_{ij} is large if the distributions are highly dissimilar (Nelson et al., 2014).

The principal steps in calculating a hierarchical cluster, after Johnson (1967), are as follows:

Step 1: Assign each GSD to be its own cluster, and calculate the distance $d_{(i,\,j)}$ between all pairs of clusters using the city-block distance (Eq. 2).

Step 2: Find the most similar pair of clusters in the initial clustering and merge them into a new single cluster, designated pair rs, as follows:

$$d_{(rs)} = \min d_{(i,j)}, \qquad \text{(Eq. 3)}$$

where $d_{(rs)}$ is the minimum $d_{(i,\,j)}$ of all distances of clusters.

https://doi.org/10.1017/qua.2017.78 Published online by Cambridge University Press

Step 3: Compute the distance between the new cluster and the other clusters.

Step 4: Repeat steps 2 and 3 until all clusters are merged into a single cluster.

Estimating the number of clusters

A major challenge in cluster analysis is determining the optimal number of clusters. Many methods have been used, including the gap statistic (Tibshirani et al., 2001), the dynamic tree cut method (Langfelder et al., 2008), and the density peak (Rodriguez and Laio, 2014). However, all of these methods have inherent limitations, and few of them can be used efficiently in the analysis of GSDs.

In the "bottom-up" hierarchical cluster analysis used here, the agglomeration coefficient indicates the within-group variance of two clusters combined at each successive stage of the clustering and therefore provides a simple and objective means of determining the optimum stage for terminating the clustering and the total number of clusters (Everitt et al., 2011). This is because a large change in the agglomeration coefficient between two stages of clustering indicates that heterogeneous clusters are being combined, and the result has a larger total variance. Thus, this approach can be used to judge whether an optimal cluster solution has been achieved (Hill et al., 1998). A significant change in the agglomeration coefficient can be termed a "knee" (Salvador and Chan, 2004), which can be determined by calculating and graphing the change in agglomeration coefficients as a function of the stages in the cluster analysis (i.e., the total number of clusters). Large changes in the agglomeration coefficient as a function of the number of clusters will signify the merger of dissimilar clusters and therefore indicate the optimum number of clusters. As suggested by Salvador and Chan (2004) and Chiu et al. (2001), there are two efficient ways to find the "knee" of a curve. One is to look for the largest magnitude difference between two adjacent points of the change in the agglomeration coefficient; the other is to calculate the jump in ratio change between two points. Here, we use the first method to find the "knee" because it involves little computation and easy operation.

Determining the unmixed grain-size distributions

The unmixed GSDs are defined as the most typical GSDs in the data set that are representative of their clusters—in other words, they have the maximum degree of similarity to every GSD within their cluster. This is calculated to be the GSD that produces the minimum average distance in the clusters:

$$C_{(q)} = \min d_{(q,ks)}, \qquad \text{(Eq. 4)}$$

where *ks* is a cluster in which all the GSDs can be assumed to have been generated by similar depositional processes, and q denotes the most representative GSDs, which have maximum similarity to every GSD in cluster *ks*.

Fractions of each endmember within a sample

In the linear mixing model, if the number of endmembers and their compositions can be determined accurately, the fraction of each endmember contributing to each sample can be estimated using standard least squares techniques (Weltje, 1997). For compositional data (i.e., GSDs), the fraction of each endmember in each sample should be nonnegative and sum to 1, so that the fraction of each endmember can be calculated by a nonnegative least squares algorithm and scaled to a constant sum (Weltje and Prins, 2007), as follows:

$$\sum_{k=1}^{q} m_{ik} = 1, \qquad \text{(Eq. 5)}$$

where m_{ik} is the proportional contribution of the endmembers to each observation, and q is the number of the endmember. All of the m_{ik} values constitute the fractions matrix $M_{(n \times q)}$.

A CASE STUDY: THE SEDIMENTS OF WULUNGU LAKE, JUNGGAR BASIN, CHINA

A GSD data set from the sediments of Wulungu Lake in the Junggar Basin, China (Fig. 1), was used to test the CEMMA method. Wulungu Lake is a hydrologically closed terminal lake fed mainly by the Wulungu River. The sediments studied are from a 7-m-long core (WLG11E) taken from the central part of the lake (47°14.40′N, 87°13.10′E) in a water depth of 18 m. The sediments studied here accumulated during Marine Oxygen Isotope Stage 3, between ~25,300 and 51,600 cal yr BP (Zhang et al., 2016). The core is mainly composed of silty clay with occasional thin intercalated layers of silt or gypsum. Five lithologic units can be defined on the basis of sediment color and composition, as follows (see Fig. 2): unit I (901–769 cm) is dominated by gray silty clay; unit II (769–613 cm) consists of yellowish silty clay (769–711 cm), gray silty clay (711–662 cm), and brown silty clay (662–613 cm); unit III (613–400 cm) consists of yellowish silty clay; unit IV (400–294 cm) consists of gray silty clay (400–371 cm and 349–294 cm) and yellowish silty clay with gypsum (371–349 cm); and unit V (294–193 cm) consists mainly of brown silty clay.

A total of 373 samples were obtained at 2 to 3 cm intervals from the core for grain-size analysis. The samples were pretreated with 10% H_2O_2 to remove organic matter and 10% HCl to remove carbonates, rinsed with deionized water, and then dispersed with 10 mL of 0.05 mol/L $(NaPO_3)_6$ in an ultrasonic bath for 10 minutes. GSDs between 0.02 and 2000 μm were measured using a Malvern Mastersizer 2000 laser grain-size analyzer and assigned to 100 size classes. The GSDs are relatively uniform (Fig. 3a), with most samples possessing a modal grain size of around 11 μm (silt), whereas a few samples have additional peaks at around 40 μm (coarse silt) or between 300 and 800 μm (sand).

Applying the CEMMA method to the Wulungu Lake data, the agglomeration coefficients show a large change at four

https://doi.org/10.1017/qua.2017.78 Published online by Cambridge University Press

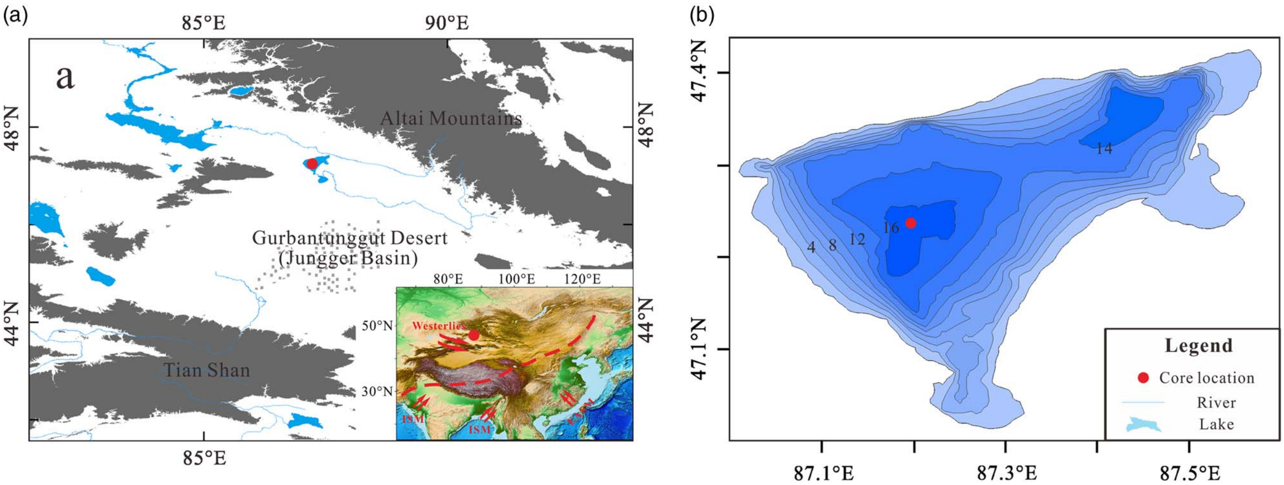

Figure 1. (a) Topography of the study area. The red dot indicates the location of Wulungu Lake. Gray-shaded areas are mountains (elevation >4374 m). Inset map shows the location of the study area within Asia, with trajectories of the major atmospheric circulation systems indicated by red arrows and the modern Asian summer monsoon limit indicated by the red dashed line (after Chen et al., 2008, 2010). EASM, East Asian summer monsoon; ISM, Indian summer monsoon. (b) Bathymetry of Wulungu Lake (contours at 2 m intervals) with the location of core WLG11E indicated by the red dot (Wu et al., 2013). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

clusters (the red dot in Fig. 4), which thus defines the “knee” and indicates that the clustering should be terminated. Therefore, four optimal endmembers can be inferred from the CEMMA model in this example. The GSDs of the clustering endmembers (CEMs) are illustrated in Figure 3c. CEM 1 exhibits a unimodal peak with a dominant mode at around 8 μm (very fine silt and coarse clay), CEM 2 exhibits a symmetrical unimodal peak in the very fine silt range (mode at 13 μm), CEM 3 is characterized by an asymmetrical unimodal peak centered at around 40 μm (coarse silt), and

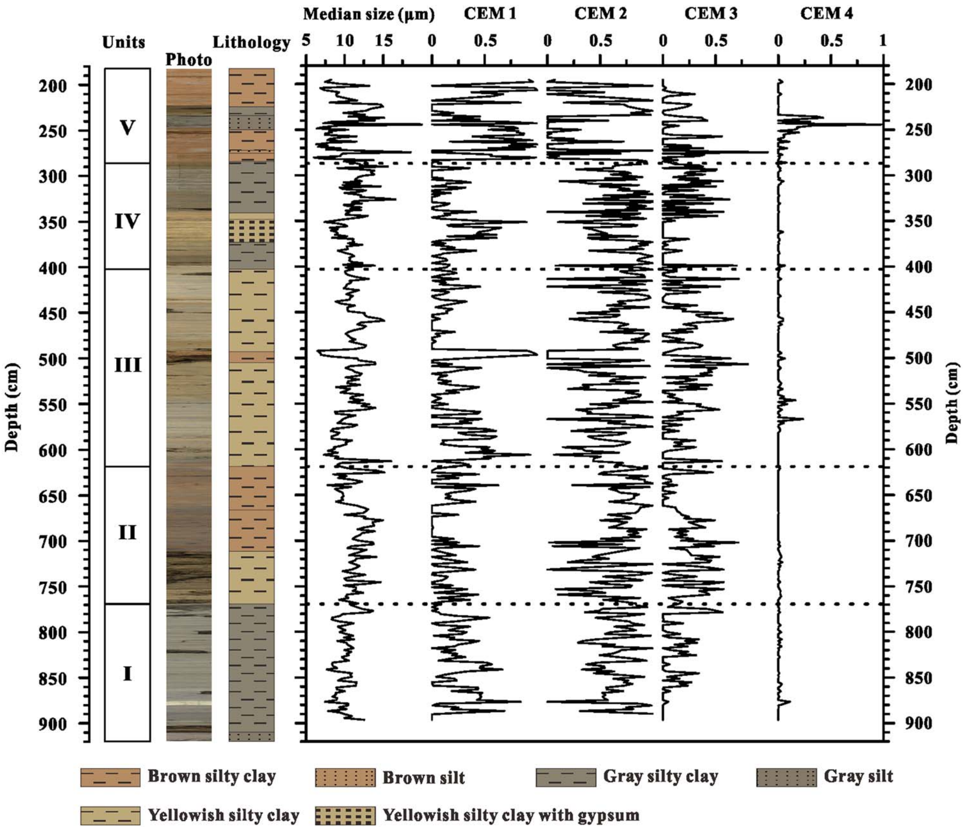

Figure 2. (color online) Lithologic units, graphic lithology, and changes in sediment median size and endmember fractions (clustering endmember [CEM] 1 to CEM 4) plotted against depth for core WLG11E.

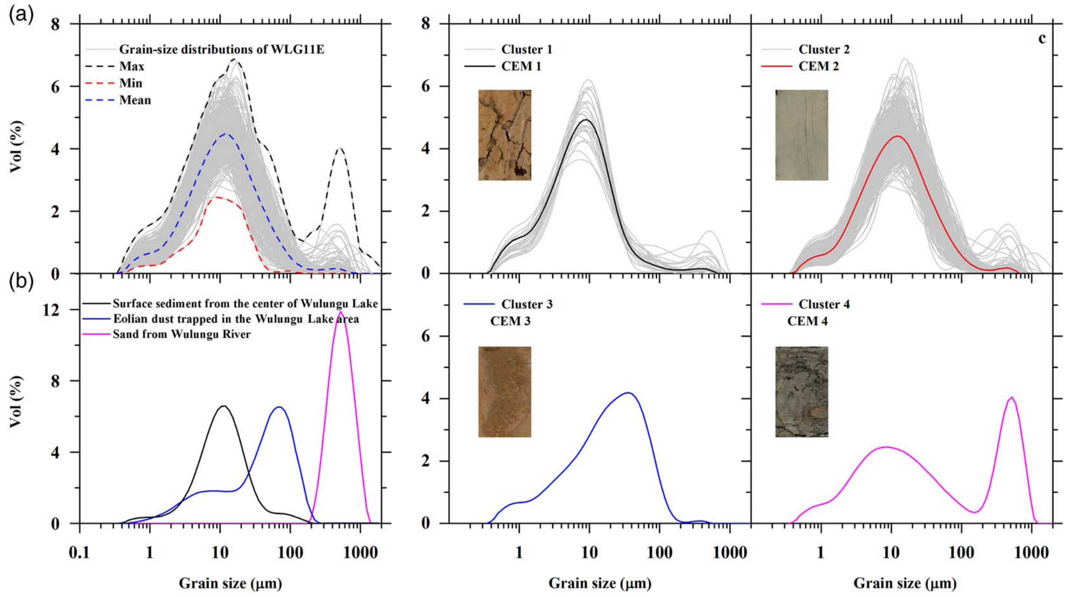

Figure 3. (a) Grain-size distributions (GSDs) of samples from sediment core WLG11E from Wulungu Lake (light-gray lines), including the maximum, minimum, and mean for each grain-size class. (b) GSD curves for surface sediment from the center of Wulungu Lake (black curve), eolian dust trapped near the ground surface in Wulungu Lake area (blue curve) (after Liu et al., 2008), and sand from the Wulungu River (magenta curve). (c) Results of clustering endmember modeling analysis for the sediments of core WLG11E from Wulungu Lake: four clustering endmember (CEM) unmixed GSDs (colored lines) and clusters (light-gray lines). Inset image shows the lithology of the relevant interval of the core. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

CEM 4 exhibits a bimodal structure with a dominant mode at around 522 μm (sand) and a minor mode at around 8 μm (very fine silt and coarse clay). The GSDs of three modern sediment samples are also plotted (Fig. 3b). Surface sediments from the center of Wulungu Lake exhibit a unimodal distribution with a peak at around 10 μm, an eolian dust sample trapped in the Wulungu Lake area exhibits a slightly bimodal GSD with a dominant mode at around 58 μm and a minor mode at around 6 μm (Liu et al., 2008), and sand from Wulungu River exhibits an asymmetrical unimodal distribution with a peak at around 522 μm.

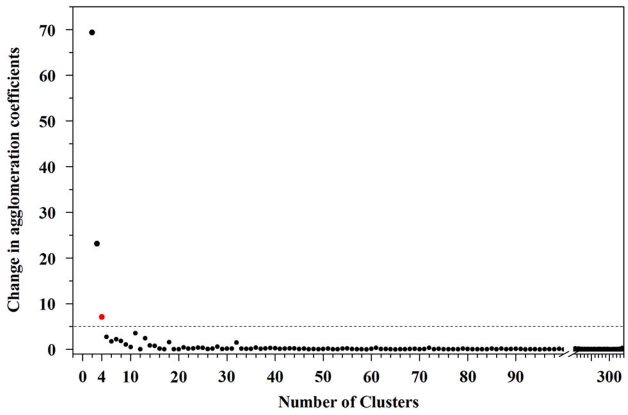

Figure 4. Change of the agglomeration coefficient versus number of clusters. The red dot is the "knee." (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Changes in the fractions of endmembers versus depth are plotted in Figure 2. The first endmember (CEM 1) is dominant in brown silty clay, corresponding to lithologic unit V, and has an average fractional abundance of 0.47. CEM 2 exhibits high values in gray silty clay (units I and IV, with average values of 0.67 and 0.69, respectively). CEM 2 is also relatively high in yellowish silty clay (units II and III, with average values of 0.57 and 0.60, respectively). The third endmember (CEM 3) exhibits relatively low values down core (average of 0.15) and comparatively high values in yellowish silty clay (average of 0.21). CEM 4 is only recorded in a few layers with gray silt; it is represented in the interval from 247–234 cm, with an average value of 0.32.

DISCUSSION

In the CEMMA results (Fig. 3), CEM 1 and CEM 2 exhibit a similar GSD to surface sediments from the center of Wulungu Lake but have slightly different modal grain-size values of 8 and 13 μm, respectively. As indicated by the lithologies corresponding to these endmembers, CEM 1 is represented in brown silty clay and CEM 2 in gray silty clay. According to the results of a previous study (Zhang et al., 2016), gray silty clay contains relatively high feldspar (17%) and low illite (35%), whereas brown silty clay has a high concentration of illite (47%) and a low concentration of feldspar (11%). Illite-rich sediments have been interpreted as representing relatively dry conditions (Singer, 1984). Together with the lithology, it can be inferred that CEM 1

https://doi.org/10.1017/qua.2017.78 Published online by Cambridge University Press

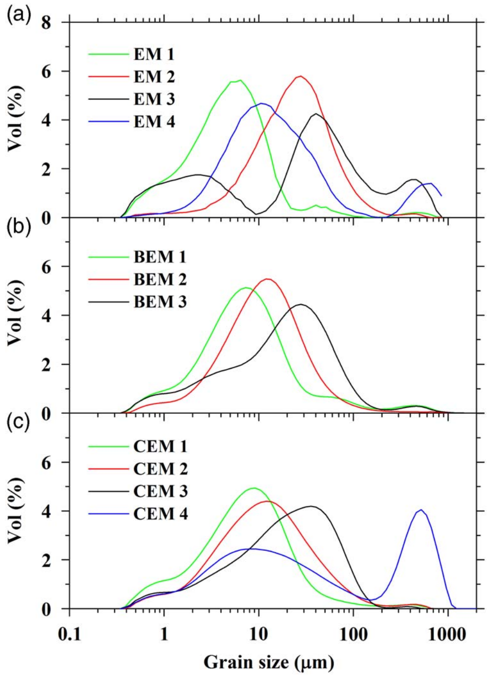

Figure 5. (color online) Comparison of endmember spectra for core WLG11E obtained using the eigenspace method (a), hierarchical Bayesian endmember modeling analysis (b), and hierarchical clustering endmember modeling analysis (this study) (c). BEM, Bayesian endmember; CEM, clustering endmember; EM, endmember.

reflects the hydrodynamic energy of a very shallow lake environment, whereas CEM 2 reflects the hydrodynamic energy of a deep lake environment. Comparison of the GSDs of the CEM endmembers with those of modern samples (Fig. 3) indicates that CEM 3 exhibits a similar GSD to the eolian dust trapped in the Wulungu Lake area. Thus, this GSD indicates eolian transport of a local dust source via short-term suspension and saltation processes, probably by local storms in winter and spring when near-surface turbulent airflow prevails (Qin et al., 2005; Qiang et al., 2007; Yu et al., 2015). CEM 4 has a similar distribution to a mixture of lacustrine suspended sediment and Wulungu River sediments and therefore can be interpreted as indicating the strength of the runoff into the lake.

To estimate the performance of our CEMMA methods, we compare our CEMMA results with the corresponding endmembers obtained from the eigenvector rotation endmember modeling analysis (EMMA) (Dietze et al., 2012) (Fig. 5a) and hierarchical Bayesian endmember modeling analysis (BEMMA) (Yu et al., 2015) (Fig. 5b). We calculate the modal size, sorting, skewness, and kurtosis of endmember spectra as suggested by Folk and Ward (1957) and also fit the endmember spectra to the WLG11E data set to calculate the error of reconstructed GSDs and the correlation coefficient between the reconstruction and original data (Table 1). The results show that the EMMA method gets the biggest error value (0.51) and poor reconstruction results (data set R^2 is 0.77). However, BEMMA and CEMMA results show similar values for the goodness-of-fit test; they both show lower error values (0.16) and higher correlation coefficient values (data set R^2 is 0.98). As to each endmember spectrum, the first endmember defined by all three methods is broadly similar, with a dominant mode at around 8 μm (Table 1). However, there are marked differences between the other three endmembers. First, as to the second endmember, the results for BEMMA (Bayesian endmember [BEM] 2) and CEMMA (CEM 2) exhibit a similar distribution with a symmetrical unimodal peak at around 13 μm. However, the EMMA result (endmember [EM] 2) is quite different, with a unimodal peak around 30 μm. With regard to the third endmember, the results for BEM 3 and CEM 3 exhibit a similar distribution with a modal peak at around 40 μm and with moderate sorting (1.94 and 1.84, respectively); however, EM 3 is poorly sorted (2.95) and exhibits a multimodal distribution with a fine peak around 3 μm and a coarse peak around 400 μm (Folk and Ward, 1957). If the second and third endmembers are combined, the results for BEMMA and CEMMA are similar, but they differ from that of EMMA. This difference results from the absence of data for the fine and coarse size fractions, because in the EMMA method problems arise as a result of the transformation and communality of grain-size compositional data if the grain-size data have many zero components (Yu et al., 2015).

BEMMA does not calculate a fourth endmember. This is because the results of the Bayesian method can be influenced by the prior distributions and maximum likelihood of samples, with the consequence that information of low

Table 1. The parameters of endmember spectra and goodness-of-fit statistics for fitting the endmember spectra to the WLG11E data set using different endmember methods. BEM, Bayesian endmember; BEMMA, Bayesian endmember modeling analysis; CEM, clustering endmember; CEMMA, clustering endmember modeling analysis; EM, endmember; EMMA, endmember modeling analysis.

	EMMA				BEMMA			CEMMA			
	EM 1	EM 2	EM 3	EM 4	BEM 1	BEM 2	BEM 3	CEM 1	CEM 2	CEM 3	CEM 4
Modal (μm)	5.71	31.53	45.47	11.87	8.23	13.42	31.53	8.23	13.42	40.24	522.30
Sorting	1.40	1.31	2.95	1.96	1.67	1.39	1.94	1.59	1.66	1.84	3.10
Skewness	0.09	0.10	0.28	−0.26	−0.06	0.06	0.26	0.10	0.03	0.24	−0.17
Kurtosis	1.14	1.08	0.84	1.48	1.34	1.10	1.02	1.12	1.05	1.00	0.63
Error		0.51				0.16			0.16		
Data set R^2		0.77				0.98			0.98		

https://doi.org/10.1017/qua.2017.78 Published online by Cambridge University Press

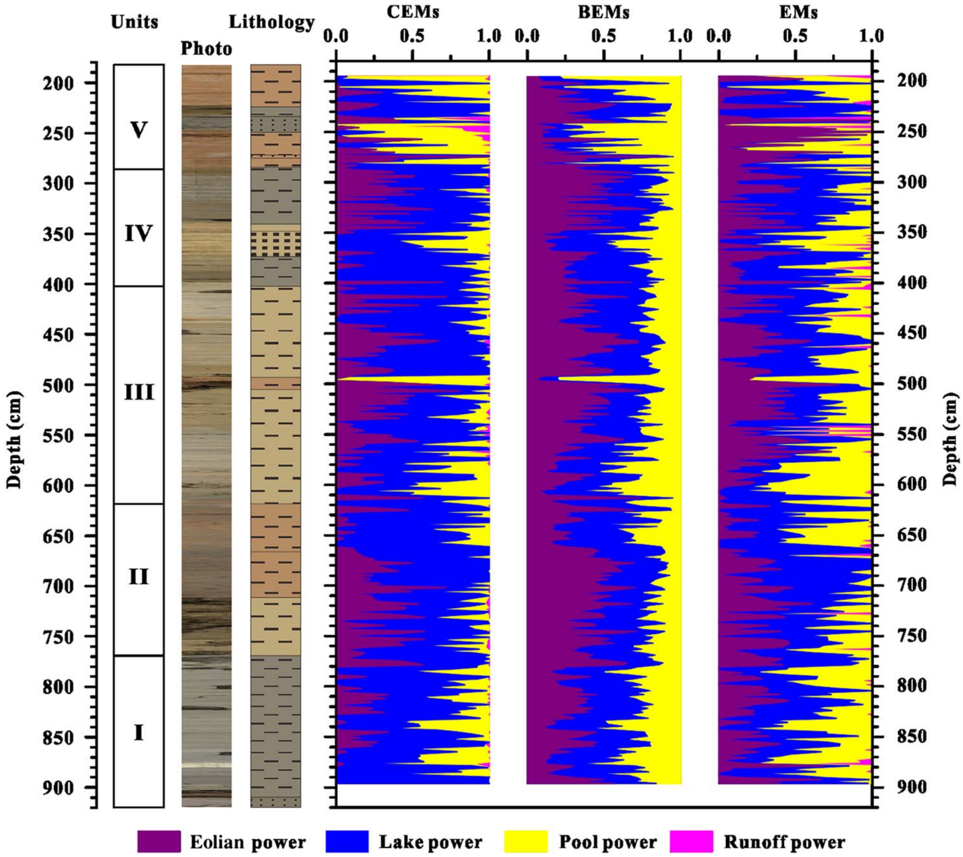

Figure 6. (color online) Lithologic units, graphic lithology, and comparison of endmember fractions from hierarchical clustering endmember modeling analysis (clustering endmembers [CEMs], this study), hierarchical Bayesian endmember modeling analysis (Bayesian endmembers [BEMs]), and the eigenspace method (endmembers [EMs]).

probability (frequency) may be ignored, even though it may potentially be of environmental significance (i.e., the fourth endmember).

With regard to the fractions of each endmember produced by the different methods (Fig. 6), the results are generally in agreement with each other, and all of them coincide with the lithologic units. However, in contrast, because of the differences between the endmember spectra, “eolian energy” (the third endmember) exhibits slight differences between the three methods, and a value for “runoff energy” (the fourth endmember) is not provided by the BEMMA results. In summary, CEMMA endmembers show much better agreement with major lithologic units in our sediment core and major sedimentary processes operating in Wulungu Lake than do the other endmember modeling methods tested here.

In addition, our CEMMA method has several advantages over the EMMA and the improved BEMMA methods: (1) the computation time for large data sets is very low because the method does not use iterations; (2) CEMMA is not influenced by zero data values in the leading and trailing sides of compositional data; and (3) it takes advantage of lithologic information, which aids understanding of the depositional environments reflected by the endmembers.

Despite the overall effectiveness of the methodology presented here, the endmembers produced by our algorithm may not correspond exactly to the hydrodynamic and eolian processes that they reflect. In addition, the effectiveness of the method is restricted by the sampling resolution and the sediment accumulation rate. Finally, because of the fact that the fractions of the endmembers are nonnegative and always sum to 1, the endmember fractions do not correspond exactly to the strength of the corresponding environmental processes.

CONCLUSIONS

Endmember modeling analysis provides an effective means of unmixing GSDs in order to determine the provenance, transport mechanisms, and depositional environment of sediments. The fractions of endmembers can potentially be used as proxies for paleoenvironmental and paleoclimatic processes through different depositional environments. In this study, the CEMMA method, based on cluster analysis combined with least squares fitting is used to define GSD endmembers. The number of endmembers can be readily determined from the agglomeration coefficients, the endmember spectra are selected based on the average distance between samples within the clusters, and the fractions of each endmember are determined by a nonnegative least squares algorithm. In the lake sediment case study presented here, the interpretation of the endmembers uses variations in lithology

https://doi.org/10.1017/qua.2017.78 Published online by Cambridge University Press

to aid the assessment of sediment provenance. We conclude that the methodology provides an efficient means of identifying the most representative samples in very large data sets.

ACKNOWLEDGMENTS

We thank Dr. Jan Bloemendal for his thorough and constructive reviews that significantly improved the manuscript. This project was supported by National Key R&D Program of China (grant number. 2017YFA0603402), the National Natural Science Foundation of China (grant number. 41130102 and 41372180), and the Fundamental Research Funds for the Central Universities (grant number. lzujbky-2016-245).

SUPPLEMENTARY MATERIAL

To view supplementary material for this article, please visit https://doi.org/10.1017/qua.2017.78

REFERENCES

Blott, S.J., Pye, K., 2001. GRADISTAT: a grain size distribution and statistics package for the analysis of unconsolidated sediments. *Earth Surface Processes and Landforms* 26, 1237–1248.

Boulay, S., Colin, C., Trentesaux, A., Pluquet, F., Bertaux, J., Blamart, D., Buehring, C., Wang, P., 2003. Mineralogy and sedimentology of Pleistocene sediments on the South China Sea (ODP Site 1144). In: Prell, W.L., Wang, P., Blum, P., Rea, D.K., Clemens, S.C. (Eds.), *Proceedings of the Ocean Drilling Program: Scientific Results*. Vol. 184. Ocean Drilling Program, Texas AM University, College Station, pp. 1–21.

Chen, F., Chen, J., Holmes, J., Boomer, I., Austin, P., Gates, J.B., Wang, N., Brooks, S.J., Zhang, J., 2010. Moisture changes over the last millennium in arid central Asia: a review, synthesis and comparison with monsoon region. *Quaternary Science Reviews* 29, 1055–1068.

Chen, F., Yu, Z., Yang, M., Ito, E., Wang, S., Madsen, D.B., Huang, X., et al., 2008. Holocene moisture evolution in arid central Asia and its out-of-phase relationship with Asian monsoon history. *Quaternary Science Reviews* 27, 351–364.

Chiu, T., Fang, D., Chen, J., Wang, Y., Jeris, C., 2001. A robust and scalable clustering algorithm for mixed type attributes in large database environment. In: Proceedings of the Seventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, pp. 263–268.

Clark, M.W., 1976. Some methods for statistical analysis of multimodal distributions and their application to grain-size data. *Journal of the International Association for Mathematical Geology* 8, 267–282.

Court, A., 1949. Separating frequency distributions into two normal components. *Science* 110, 500–501.

Dietze, E., Hartmann, K., Diekmann, B., Ijmker, J., Lehmkuhl, F., Opitz, S., Stauch, G., Wünnemann, B., Borchers, A., 2012. An end-member algorithm for deciphering modern detrital processes from lake sediments of Lake Donggi Cona, NE Tibetan Plateau, China. *Sedimentary Geology* 243–244, 169–180.

Donato, S.V., Reinhardt, E.G., Boyce, J.I., Pilarczyk, J.E., Jupp, B.P., 2009. Particle-size distribution of inferred tsunami deposits in Sur Lagoon, Sultanate of Oman. *Marine Geology* 257, 54–64.

Everitt, B.S., Landau, S., Leese, M., Stahl, D., 2011. Hierarchical clustering. In: Everitt, B.S., Landau, S., Leese, M., Stahl, D. (Eds.), *Cluster Analysis*. 5th ed. John Wiley and Sons, Chichester, UK, pp. 71–110.

Folk, R.L., Ward, W.C., 1957. Brazos River bar: a study in the significance of grain size parameters. *Journal of Sedimentary Research* 27, 3–26.

Fournier, J., Gallon, R.K., Paris, R., 2014. G2Sd: a new R package for the statistical analysis of unconsolidated sediments. *Géomorphologie: Relief, Processus, Environnement* 20, 73–78.

Ghosh, J.K., Mazumder, B.S., 1981. Size distribution of suspended particles—unimodality, symmetry and lognormality. In: Taillie, C., Patil, G.P., Baldessari, B.A. (Ed.), *Statistical Distributions in Scientific Work*. Vol. 6. Applications in Physical, Social, and Life Sciences. D. Reidal, Dordrecht, the Netherlands, pp. 21–32.

Grimm, E.C., Donovan, J.J., Brown, K.J., 2011. A high-resolution record of climate variability and landscape response from Kettle Lake, northern Great Plains, North America. *Quaternary Science Reviews* 30, 2626–2650.

Gruvaeus, G., Wainer, H., 1972. Two additions to hierarchical cluster analysis. *British Journal of Mathematical and Statistical Psychology* 25, 200–206.

He, Y., Zhao, C., Song, M., Liu, W., Chen, F., Zhang, D., Liu, Z., 2015. Onset of frequent dust storms in northern China at ~AD 1100. *Scientific Reports* 5, 17111. http://dx.doi.org/10.1038/srep17111.

Hill, E.W., Brennan, J.F., Wolman, H.L., 1998. What is a central city in the United States? Applying a statistical technique for developing taxonomies. *Urban Studies* 35, 1935–1969.

Inman, D.L., 1952. Measures for describing the size distribution of sediments. *Journal of Sedimentary Research* 22, 125–145.

Johnson, S.C., 1967. Hierarchical clustering schemes. *Psychometrika* 32, 241–254.

Kaufman, L., Rousseeuw, P.J., 2009. *Finding Groups in Data: An Introduction to Cluster Analysis*. John Wiley and Sons, Hoboken, NJ.

Langfelder, P., Zhang, B., Horvath, S., 2008. Defining clusters from a hierarchical cluster tree: the Dynamic Tree Cut package for R. *Bioinformatics* 24, 719–720.

Liu, J., Rühland, K.M., Chen, J., Xu, Y., Chen, S., Chen, Q., Huang, W., Xu, Q., Chen, F., Smol, J.P., 2017. Aerosol-weakened summer monsoons decrease lake fertilization on the Chinese Loess Plateau. *Nature Climate Change* 7, 190–194.

Liu, X., Herzschuh, U., Shen, J., Jiang, Q., Xiao, X., 2008. Holocene environmental and climatic changes inferred from Wulungu Lake in northern Xinjiang, China. *Quaternary Research* 70, 412–425.

Nelson, P.A., Bellugi, D., Dietrich, W.E., 2014. Delineation of river bed-surface patches by clustering high-resolution spatial grain size data. *Geomorphology* 205, 102–119.

Ordóñez, C., Ruiz-Barzola, O., Sierra, C., 2016. Sediment particle size distributions apportionment by means of functional cluster analysis (FCA). *Catena* 137, 31–36.

Passega, R., 1964. Grain size representation by CM patterns as a geological tool. *Journal of Sedimentary Research* 34, 830–847.

Paterson, G.A., Heslop, D., 2015. New methods for unmixing sediment grain size data. *Geochemistry, Geophysics, Geosystems* 16, 4494–4506.

Qiang, M., Chen, F., Zhou, A., Xiao, S., Zhang, J., Zhang, J., 2007. Impacts of wind velocity on sand and dust deposition during dust storm as inferred from a series of observations in the northeastern Qinghai–Tibetan Plateau, China. *Powder Technology* 175: 82–89.

https://doi.org/10.1017/qua.2017.78 Published online by Cambridge University Press

Qin, X., Cai, B., Liu, T., 2005. Loess record of the aerodynamic environment in the east Asia monsoon area since 60,000 years before present. *Journal of Geophysical Research: Solid Earth* 110, B01204. http://dx.doi.org/10.1029/2004JB003131.

Renner, R.M., 1991. An examination of the use of the logratio transformation for the testing of endmember hypotheses. *Mathematical Geology* 23, 549–563.

Rodriguez, A., Laio, A., 2014. Clustering by fast search and find of density peaks. *Science* 344, 1492–1496.

Salvador, S., Chan, P., 2004. Determining the number of clusters/ segments in hierarchical clustering/segmentation algorithms. In: *Proceedings 16th IEEE International Conference on Tools with Artificial Intelligence*, Boca Raton, FL, pp. 576–584.

Singer, A., 1984. The paleoclimatic interpretation of clay minerals in sediments—a review. *Earth-Science Reviews* 21, 251–293.

Sun, D., Bloemendal, J., Rea, D.K., Vandenberghe, J., Jiang, F., An, Z., Ruixia, S., 2002. Grain-size distribution function of polymodal sediments in hydraulic and aeolian environments, and numerical partitioning of the sedimentary components. *Sedimentary Geology* 152, 263–277.

Tibshirani, R., Walther, G., Hastie, T., 2001. Estimating the number of clusters in a data set via the gap statistic. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 63, 411–423.

Visher, G.S., 1969. Grain size distributions and depositional processes. *Journal of Sedimentary Research* 39, 1074–1106.

Weltje, G.J., 1997. End-member modeling of compositional data: numerical-statistical algorithms for solving the explicit mixing problem. *Mathematical Geology* 29, 503–549.

Weltje, G.J., Prins, M.A., 2003. Muddled or mixed? Inferring palaeoclimate from size distributions of deep-sea clastics. *Sedimentary Geology* 162, 39–62.

Weltje, G.J., Prins, M.A., 2007. Genetically meaningful decomposition of grain-size distributions. *Sedimentary Geology* 202, 409–424.

Wu, J., Ma, L., Zeng, H., 2013. Water quantity and quality change of Ulungur Lake and its environmental effects. [In Chinese with English abstract.]. *Journal of Nature Resources* 28, 844–853.

Xu, R., Wunsch, D., 2005. Survey of clustering algorithms. *IEEE Transactions on Neural Networks* 16, 645–678.

Yu, S., Colman, S.M., Li, L., 2015. BEMMA: a hierarchical Bayesian end-member modeling analysis of sediment grain-size distributions. *Mathematical Geosciences*, 1–19.

Zhang, X., Zhou, A., Zhang, C., Hao, S., Zhao, Y., An, C., 2016. High-resolution records of climate change in arid eastern central Asia during MIS 3 (51 600–25 300 cal a BP) from Wulungu Lake, north-western China. *Journal of Quaternary Science* 31, 577–586.

Zhou, D., Chen, H., Lou, Y., 1991. The logratio approach to the classification of modern sediments and sedimentary environments in northern South China Sea. *Mathematical Geology* 23, 157–165.

https://doi.org/10.1017/qua.2017.78 Published online by Cambridge University Press