Journal of Hydrology 409 (2011) 315–327

Contents lists available at SciVerse ScienceDirect

Journal of Hydrology

journal homepage: www.elsevier.com/locate/jhydrol

Multivariate analyses with end-member mixing to characterize groundwater flow: Wind Cave and associated aquifers

Andrew J. Long *, Joshua F. Valder

US Geological Survey, South Dakota Water Science Center, 1608 Mountain View Rd., Rapid City, SD 57702, USA

ARTICLE INFO

Article history:
Received 25 March 2011
Received in revised form 22 July 2011
Accepted 11 August 2011
Available online 22 August 2011

This manuscript was handled by Philippe Baveye, Editor-in-Chief

Keywords:
Principal component analysis
Factor analysis
Cluster analysis
End-member mixing
Karst hydrology
Multivariate analysis

SUMMARY

Principal component analysis (PCA) applied to hydrochemical data has been used with end-member mixing to characterize groundwater flow to a limited extent, but aspects of this approach are unresolved. Previous similar approaches typically have assumed that the extreme-value samples identified by PCA represent end members. The method presented herein is different from previous work in that (1) end members were not assumed to have been sampled but rather were estimated and constrained by prior knowledge; (2) end-member mixing was quantified in relation to hydrogeologic domains, which focuses model results on major hydrologic processes; (3) a method to select an appropriate number of end members using a series of cluster analyses is presented; and (4) conservative tracers were weighted preferentially in model calibration, which distributed model errors of optimized values, or residuals, more appropriately than would otherwise be the case. The latter item also provides an estimate of the relative influence of geochemical evolution along flow paths in comparison to mixing. This method was applied to groundwater in Wind Cave and the associated karst aquifer in the Black Hills of South Dakota, USA. The end-member mixing model was used to test a hypothesis that five different end-member waters are mixed in the groundwater system comprising five hydrogeologic domains. The model estimated that Wind Cave received most of its groundwater inflow from local surface recharge with an additional 33% from an upgradient aquifer. Artesian springs in the vicinity of Wind Cave primarily received water from regional groundwater flow.

Published by Elsevier B.V.

1. Introduction

Principal component analysis (PCA), cluster analysis, and end-member mixing are multivariate methods useful for characterizing groundwater on the basis of hydrochemistry to help assess groundwater mixing and flow. PCA transforms a multivariate dataset by combining the original variables into new variables, or principal components (Davis, 2002). This helps simplify interpretation of complex multivariate datasets by identifying the principal components that describe the greatest amount of the total variance of the dataset. Cluster analysis describes an array of methods used to group samples on the basis of multivariate data (Davis, 2002) and commonly is applied to results of PCA (e.g., Suk and Lee, 1999). End-member mixing is a method that can be used with multivariate data or univariate data (e.g., Fritz et al., 1976; Christophersen and Hooper, 1992; Laaksoharju et al., 1999; Carrera et al., 2004). This method assumes that the groundwater sampled is a mixture of two or more sources, or end-members, and that the mixing proportions can be estimated from the hydrochemistry.

Examples of PCA applied to karst aquifers are given in White (1977), Doctor et al. (2006), Mahler et al. (2008), and Fournier et al. (2008). Methods developed that combine PCA with end-member mixing include those of Christophersen and Hooper (1992) and Laaksoharju et al. (1999). These authors used PCA to identify extreme-value sample points, which they interpreted as representing end-member waters, assuming that the end members are well represented by the samples analyzed. However, for many studies it might not be possible to sample the actual end members. For example, a major contribution to an aquifer might be areally distributed infiltration, where one sample does not fully represent this end member. The number of end members needed might also be in question for many applications.

The method presented herein differs from previous methods that combine PCA with end-member mixing in that we focus on the assessment of major hydrogeologic domains, the number which determines the number of end members. We present an end-member mixing model that estimates end members in a way similar to that described by Carrera et al. (2004), but here initial estimates of end members are determined by PCA and

* Corresponding author.
E-mail addresses: ajlong@usgs.gov (A.J. Long), jvalder@usgs.gov (J.F. Valder).

0022-1694/$ - see front matter Published by Elsevier B.V.
doi:10.1016/j.jhydrol.2011.08.028

cluster analysis, conservative tracers (e.g., stable isotopes) are given more weight than non-conservative tracers in model calibration, and constraints can be placed on the model according to prior knowledge. We also developed and applied a method that uses a series of cluster analyses to assist in determining the number of end members and to assess the relative hydrochemical similarities between the different domains. Finally, end-member contributions were quantified in relation to the domains, which focused model results on major hydrologic processes rather than individual sample sites. This method is useful for karst groundwater applications because no assumptions need to be made regarding the presence, locations, or dimensions of conduits. Thus, we describe an application to groundwater associated with Wind Cave in the southern Black Hills of South Dakota, USA.

2. Methods

The first step in the method presented is principal component analysis (PCA) followed by a series of cluster analyses in which different numbers of clusters are tested, and the number of end members is selected on this basis. An end-member mixing model then is applied, and mixing is quantified in terms of end-member contributions to each hydrogeologic domain. Finally, PCA is applied again, except with the addition of the estimated end members, each of which is matched with a different cluster to determine which end member is associated with each of the hydrogeologic domains. This also is a verification that cluster analysis is consistent with end-member mixing.

2.1. Principal component analysis

PCA is a linear transformation of data in multi-dimensional space, where the transformed axes, or principal components, align with the greatest variances in the multivariate dataset (Davis, 2002). Each principal component is a new variable that is a linear combination of the original variables. PCA is used to elucidate data patterns that might otherwise be obscured in the original data. The term *scores* refers to the values of the new variables in the transformed space, and by plotting the sites sampled using scores as plotting positions, sample relations and groupings may become evident. PCA commonly is used to identify extreme-value points, which might be considered as possible end members (Davis, 2002). The software MATLAB (http://www.mathworks.com) was used for PCA.

2.2. Cluster analysis

The assignment of data points to a specified number of groups, or clusters, on the basis of similarity of data is referred to as cluster analysis. The method used herein iteratively assigns each data point to a cluster that minimizes the sum of Euclidian distances between data points and the nearest cluster centroid (Seber, 1984; Spath, 1985). Here, we used the scores from the PCA as the data from which clusters were generated to reduce the clustering error caused by data error or multicolinearity (Suk and Lee, 1999). MATLAB was used for the cluster analysis. Multiple aspects of the system being analyzed enter into the choice of the number of clusters to use, and the objective is to choose the number that results in the greatest amount of useful information about the system. We introduce a method that uses a series of cluster analyses to assist in this decision and provides useful information that could not be obtained from a single cluster analysis. In this method, the number of clusters also is the number of end members used in the mixing model.

2.3. End-member mixing

An assumption of the end-member mixing model is that each water sample consists of water from one or more end members in varying proportions. For example, a two end-member model adapted from Fritz et al. (1976) is described as

$$\hat{c} = f_1 E_1 + f_2 E_2, \tag{1}$$

where $\hat{c}$ is the concentration of a mixed water sample, f_1 and f_2 are the fractions, or mixing proportions, of end members 1 and 2, respectively, and E_1 and E_2 are the respective end-member concentrations. An end member is defined as water having a characteristic hydrochemical signature that best represents a source of groundwater inflow to the system. In some cases, an end member might be a point source of inflow, such as a sinking stream; in others, it might represent the integration, or characteristic hydrochemical signature, of a distributed source, such as areally distributed recharge or regional groundwater inflow.

A common approach in applying an end-member mixing model is to collect samples from assumed end-member waters and to then determine the mixing proportion, or contribution, of each end member in samples assumed to contain mixed water. In this study, we did not assume that end members had been sampled, but rather these were estimated by inverse modeling using a generalized form of Eq. (1) that allows for any number of end members and variables:

$$\hat{c}_{i,j} = \sum_{k=1}^{n} f_{i,k} E_{j,k}, \tag{2}$$

where $\hat{c}_{ij}$ is the concentration of variable j for site i, $f_{i,k}$ is the fraction, or mixing proportion, of end member k that is associated with site i, and $E_{j,k}$ is the end-member concentration for variable j and end member k. Eq. (2) was programmed in Fortran, and the mixing proportions $f_{i,k}$ and end-member hydrochemical values $E_{j,k}$ were estimated by inverse modeling using the parameter optimization software, PEST (Doherty, 2005), which uses optimization methods described by Levenberg (1944) and Marquardt (1963). This process began with user-specified initial estimates for the values of $f_{i,k}$ and $E_{j,k}$. Then, the calculated concentrations $\hat{c}_{i,j}$ were compared to observed values $c_{i,j}$, and the differences, or residuals, between calculated and observed values $(\hat{c}_{i,j} - c_{i,j})$ were minimized by optimizing the values of $f_{i,k}$ and $E_{j,k}$ iteratively. New residuals were calculated for each iteration, and $f_{i,k}$ and $E_{j,k}$ were adjusted for the next iteration until no additional reduction of residuals occurred.

The primary limitation of this model is that hydrochemical evolution of groundwater along a flow path is neglected. Here, this limitation was diminished by weighting the calibration data for conservative tracers more heavily than for other variables during inverse modeling. If contaminants are present in the system, a retardation factor as described by Zheng and Wang (1999) could be inserted into Eq. (2) for those tracers.

3. Method application

3.1. Study area

The study area is located in the Black Hills of South Dakota, USA, which is a dome-type structure with sedimentary layers of Paleozoic age dipping radially outward on the flanks (Fig. 1, Table 1). Underlying the Paleozoic sedimentary layers and exposed at the central core of the Black Hills are fractured metamorphic and igneous rocks, which compose the Precambrian aquifer (unit PC; Fig. 1, Table 1). Overlying the Precambrian aquifer are the Deadwood aquifer (unit OCd) and the Madison aquifer, the latter a karst aquifer contained within the regionally extensive Madison Limestone