Electronic Supplementary Material (ESI) for Environmental Science: Processes & Impacts.

Identifying groundwater compartmentalisation for hydraulic fracturing risk assessments

M. P. Wilson[1], F. Worrall[1], R. J. Davies[2], and A. Hart[3]

[1] Department of Earth Sciences, Durham University, Science Labs, Durham, DH1 3LE, UK

[2] School of Natural and Environmental Sciences, Newcastle University, Newcastle, NE1 7RU, UK

[3] Environment Agency, Research Assessment and Evaluation, Sapphire East, Streetsbrook Road, Solihull, B91 1QT, UK

Corresponding author: Miles Wilson (miles.wilson@durham.ac.uk)

Supplementary Material

Global Moran's I test

The Global Moran's I test from ArcMap 10.3 is a tool used to evaluate whether a spatial dataset, for example total dissolved solid concentrations in groundwater, is spatially autocorrelated (i.e. are the data clustered, dispersed or randomly located) based on the data locations and values. The tool initially calculates the mean value for a dataset. The mean is then subtracted from all location values, creating deviation values at each location. Deviation values from neighbouring locations (specified by the user) are then multiplied together creating cross-products. Comparison of cross-product values of neighbouring locations in relation to the mean value describes whether neighbours show similar or different values of that property. When summed together and normalised by the variance the observed index value is calculated and lies between -1.0 and 1.0. The tool also calculates an expected index, which is used with the observed index, the number of data points and the variance of the dataset to create z-scores and p-values. The z-scores and p-values indicate whether any difference between the observed and expected indexes is statistically significant. If the difference is not statistically significant it cannot be ruled out that the spatial distribution is the result of random spatial processes.

Principal Components Analysis

Principal components analysis (PCA) is a multivariate statistical technique that can be used for a number of reasons: to reduce the number of variables in a dataset; to identify underlying controls in a dataset; and to orthogonally transform a set of potentially correlated observations, for example major cation and anion concentrations in groundwater, into a number of linearly uncorrelated variables on a new coordinate system. These linearly uncorrelated variables are called principal components. The aim of PCA is to explain the maximum amount of variance in the observations with the fewest number of linear combinations of the existing variables – these are the principal components. There will always be the same number of principal components as variables in the original dataset, but the first principal component accounts for the largest proportion of the variance in the data and this amount decreases with each following principal component. The number of principal components to be considered can be selected by the eigenvalue of each component or the amount of the cumulative variance explained. Interpretation of principal components in terms of the original observations involves computing the loadings between the principal components and observations – these loadings can be interpreted as if they are correlation coefficients though no direct hypothesis testing of their magnitude exists. From these loadings the relative importance of particular observations in a principal component can be assessed. Principal component values for each observation (scores) can also be plotted graphically, for example PC2 versus PC1, which may help identifying groups and trends that are not obvious in the original data. Furthermore, if the observations have location information, for example groundwater sample locations, the principal component values for each observation location can be plotted spatially to assess spatial relationships.

Supplementary Table Captions

Table S1: Groundwater quality data for the 96 unique sampling locations used in this study. The data for each location are mean values for the years 2000-2016, inclusive, and also include the matched BGS borehole data.

Table S2: The number of samples taken for each groundwater property at each unique sampling location between the years 2000-2016, inclusive.