Principal filter analysis for luminescence excitation-emission data

Chris Brunsdon

Department of Geography, University of Newcastle upon Tyne, UK

Andy Baker

Centre for Land Use and Water Resources Research (CLUWRR), University of Newcastle-upon-Tyne, UK

Received 18 August 2002; revised 29 August 2002; accepted 23 October 2002; published 19 December 2002.

[1] A new method (termed Principal Filter Analysis (PFA)) for analysing large time series of luminescence excitation-emission matrices (EEMs) is proposed, based on the idea of identifying 'filters' that detect time periods where interesting variations in the EEMs occur. A mathematical exposition of the technique is supplied, followed by a discusion of how it may be implemented in practice. The method is applied to EEMs taken from a stalagmite in Crag Cave, W. Ireland resulting in three distinct time periods of luminescence properties being identified. *INDEX TERMS:* 0649 Electromagnetics: Optics; 0644 Electromagnetics: Numerical methods; 3299 Mathematical Geophysics: General or miscellaneous; 1055 Geochemistry: Organic geochemistry. **Citation:** Brunsdon, C., and A. Baker, Principal filter analysis for luminescence excitation-emission data, *Geophys. Res. Lett.*, *29*(24), 2156, doi:10.1029/2002GL015977, 2002.

1. Introduction

[2] Using luminescence spectrophotometry, it is now possible to generate large excitation-emission matrices (EEMs) for hydrological and geological studies *Baker et al.* [2000]. A single EEM typically consists of a matrix of luminescence intensity values arranged in a rectangular grid-typically containing 10^3 to 10^5 values. The rows and columns of the matrix represent the excitation and emission wavelengths. Analysis and visualisation of these matrices can provide geological and hydrological information essentials for a number of applications, including speleothem and peat derived records of humifcation and the identification of organic carbon fractions in marine and terrestrial waters and groundwaters, including sewage and farm wastes *del Castillo et al.*, [1999]; *Jiji et al.* [1999]; *le Coupannec et al.*, [2000]; *Matthews et al.* [1996]; *Mayer et al.* [1999]; *Mobed et al.* [1996]; *Parlanti et al.* [2000]; *Wu and Tanoue* [2001]; *Yan et al.*, [2000]; *Patel-Sorrentino et al.*, [in press]. However, single EEMs are rarely considered in isolation. They are most usefully considered as a time series-for example based on 365 daily hydrological samples, or much longer series for geological samples *Coble* [1996]; *McGarry and Baker* [2000]; *Baker et al.*, [1998]; *Caseldine et al.* [2000]; *Newson et al.* [2001]; *Baker* [2001, 2002]. To analyze a series as a whole, we must regard the series of EEMs as a 'data cube' with a third time dimension in addition to those of excitation and emission described above.

[3] A data model of this type presents a number of problems for visualisation and analysis. Firstly, the data is essentially four dimensional, as any intensity measure also has associated values of time, excitation and emission wavelength. Clearly, four-dimensional data may not be visualised directly. Secondly, there is a large amount of data. Even in the hydrological case, where the time series are relatively short, there may be 365 EEMs to consider. Analysis of EEMs generally consists of identifying 'features' such as peaks or ridges of intensity which correspond to the presence of certain substances. Identifying EEMs in the time series where such features change markedly helps to identify periods of environmental change. However, with very large numbers of EEMs the task of finding 'interesting' ones is non-trivial. In an earlier paper *Baker et al.* [2000], we offered an isoline-based method for viewing the data cube as a four-dimensional entity. Here, we offer an alternative approach, which more directly tackles the problem of identifying 'interesting' points in the EEM time series. This is achieved using the technique we term *Principal Filter Analysis* PFA, which we outline in this paper.

2. Principal Filters: An Outline

[4] For a single point in time we may consider the luminescence intensity (I) to be a function of the excitation wavelength ν_{ex} and the emission wavelength, ν_{em}. Thus we write $I = I(\nu_{\text{ex}}, \nu_{\text{em}})$. An EEM may therefore be considered as a set of discretely sampled I-values using regularly spaced values of ν_{ex} and ν_{em}. We may extend this to consider every point in time t in the continuous period spanning the time series of EEMs, and write $I = I(\nu_{\text{ex}}, \nu_{\text{em}}, t)$. The time series of EEMs (the data cube) may then be considered as as a set of discretely sampled I-values as above, but with an added dimension of discrete sampling points in time. A further refinement is to standardise I by subtracting the time averaged value of I at each ($\nu_{\text{ex}}, \nu_{\text{em}}$) and dividing this quantity by the time-based RMS value. This has the effect of highlighting *relative* changes in the EEM matrix over time, rather than *absolute* changes.

[5] Suppose we know that a certain period in time, say from t_1 to t_2 is interesting. We could define a measurement of overall intensity during this period for an excitation-emission pair ($\nu_{\text{ex}}, \nu_{\text{em}}$) as

$$I^*(\nu_{\text{ex}}, \nu_{\text{em}}) = \int_T I(\nu_{\text{ex}}, \nu_{\text{em}}, t) f(t) dt \qquad (1)$$

where

$$f(t) = \begin{cases} 1 & \text{if} \qquad t_1 \le t \le t_2 \\ 0 & \text{Otherwise} \end{cases}$$

and T is the entire time span of the EEM time series.

0094-8276/02/2002GL015977

[6] The function f may be thought of as a *filter* selecting out points in time that are interesting in some prescribed way. The function $I^*(\nu_{ex}, \nu_{em})$ may then be plotted using contours or three-dimensional surface plots to identify features of the excitation-emission intensities in the time range. However, f need not be confined to a binary 0/1 switching function. By allowing f to vary continuously it is possible to arrive at a function I^* which applies relative weights of importance to different points in time. Furthermore, by allowing f to take negative values for certain time periods, it is possible to create an I^* function which gives higher values for certain (ν_{ex}, ν_{em}) pairs which do *not* have high luminescence intensities during these time periods.

[7] In practice we do not have prior knowledge of the 'interesting' time periods, and hence 'interesting' choices of f. Here we propose a method for making such choices on the basis of $I = I(\nu_{ex}, \nu_{em}, t)$, or more precisely, on the basis of the data cube as a discrete sample of this function. The underlying idea is to find f giving the 'most interesting' I^* function. We define 'most interesting' to mean the I^* exhibiting the most variability, $\mathrm{V}(I^*)$, over the sampled ranges of ν_{ex} and ν_{em}, defined by

$$\mathrm{V}(I^*) = \int\int_{2} (I^*(\nu_{ex}, \nu_{em}) - \mathrm{M}(I^*))^2 \, d\nu_{ex} d\nu_{em} \qquad (2)$$

where $\boldsymbol{\nu}^2$ is used as a shorthand to denote the region spanned by ν_{ex} and ν_{em}, and $\mathrm{M}(I^*)$ is the mean value of the function I^* over this same region, defined by

$$\mathrm{M}(I^*) = \frac{\int\int I^*(\nu_{ex}, \nu_{em}) \, d\nu_{ex} d\nu_{em}}{\int\int d\nu_{ex} d\nu_{em}}$$

Thus, the problem of finding a suitable f may be stated as

$$\begin{gathered}\text{Find a function } f \text{ minimising } \mathrm{V}(I^*) \\ \text{where} \\ I^*(\nu_{ex}, \nu_{em}) = \int_T I(\nu_{ex}, \nu_{em}, t) f(t) \, dt\end{gathered} \qquad (3)$$

[8] Note that multiplying f by a constant, or adding a constant to f would allow $\mathrm{V}(I^*)$ to increase without bound, so we subject the above problem to the two constraints: $\int_T f(t) \, dt = 0$ and $\int_T (f(t))^2 = 1$.

[9] Having found f according to the conditions above, we may then plot f against t to identify 'interesting' time periods, and plot the associated $I^*(\nu_{ex},\nu_{em})$ to identify excitation-emission wavelength pairs associated with the time periods. The f found in this way is referred to as a *Principal Filter.*

[10] Suppose we now refer to this function as f_1. It is now possible to consider other 'interesting' filters, f_2, f_3 and so on. Firstly consider f_2. To identify different features from f_1, we solve equation 3, imposing a further constraint of orthogonality-that is

$$\int_T f_1(t) f_2(t) \, dt = 0 \qquad (4)$$

[11] This ensures that f_2 will identify a different time pattern from f_1-the 'integrating to zero' property ensures that the two filters cannot both have the same sign for very large time periods in T. We may then go on to find the associated I^* function-which we will denote by I^*_2.

[12] This process may be continued indefinitely, initially by finding f_3 such that it solves equation 3, with the additional constraints that it is orthogonal to both f_1 and f_2, and more generally by finding f_k such that it is orthogonal to $f_1, f_2, \ldots f_{k-1}$.

3. Computational Issues

[13] In practice we do not work with the continuous function I, but with the discrete 'data cube.' Thus, we work with discrete approximations for all of the functions in the last section. To do this, we re-arrange the data cube into a matrix $\mathbf{X}$, whose rows are the layers of the excitation-emission levels, and whose columns correspond to the time intervals. That is, if the data cube has l excitation levels, m emission levels and k time intervals, the matrix $\mathbf{X}$ will have lm rows and k columns. This being done, the function f in equation (1), is recast as a row vector with k elements, say $\mathbf{a} = (a_1, a_2, \ldots a_k)$ and the whole expression is recast as a matrix multiplication $\mathbf{x} = \mathbf{Xa}$. The summations in the matrix multiplication replace the integration in the original expression, and the result, $\mathbf{x}$, is a column vector with lm elements, replacing $I^*(\nu_{ex},\nu_{em})$ in equation (1). Note that we may work with standardised intensities as suggested in the previous section. In this case, $\mathbf{X}$ is transformed by subtracting the mean from each row and then by dividing each row by its standard deviation.

[14] Equation (2) is then replaced by the variance of the vector $\mathbf{x}$, that is:

$$\mathrm{V}(\mathbf{x}) = \frac{1}{lm} \sum (x - \bar{x})^2 \qquad (5)$$

[15] Having translated the expressions in the previous section into discrete form, the problem stated in (3) can also be stated in discrete form:

$$\begin{gathered}\text{Find a vector } \mathbf{a} \text{ minimising } V(\mathbf{x}) \\ \text{where} \\ \mathbf{x} = \mathbf{Xa}\end{gathered} \qquad (6)$$

[16] As before we need to add two further constraints, since adding a constant multiplied by a vector of ones to $\mathbf{a}$, or multiplying $\mathbf{a}$ by a constant allows $\mathrm{V}(\mathbf{x})$ to increase without bounds. The discrete form of these constraints are $\Sigma a_i = 0$ and $\Sigma a_i^2 = 1$. Finally, it is possible to define a series of $\mathbf{a}$-vectors, say $\{a_1, a_2, \ldots\}$ using the orthogonality constraint. In discrete form, this constraint is expressed in terms of the vector dot product:

$$\mathbf{a}_i \cdot \mathbf{a}_j = 0 \text{ if } i \neq j \qquad (7)$$

[17] The utility of re-expressing the problem in discrete form using matrix algebra is that the the solution to problem (6) is well known. This is discussed, for example, in *Maxwell*

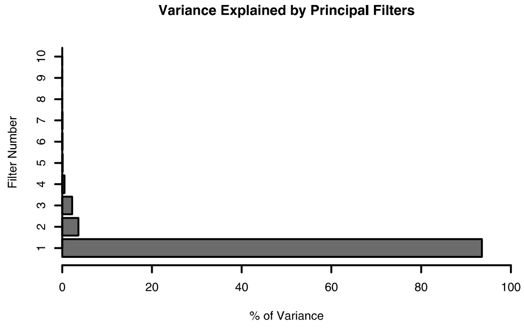

Figure 1. Variance explained by principal filtering.

[1977]. A similar approach was used in *Compagnucci et al.* [2001] to investigate evolving spatial patterns in atmospheric systems. The values of $\mathbf{a}$ solving (6) are the eigenvectors of $\mathbf{X_D}^T\mathbf{X_D}$, where $\mathbf{X_D}$ denotes the matrix $\mathbf{X}$ with the mean value of each row subtracted from that row. Note that when working with standardised intensities, $\mathbf{X} = \mathbf{X_D}$. Thus, we have an explicit form for $\mathbf{a}_1$, $\mathbf{a}_2$. We may then treat $\mathbf{a}_i$'s as approximate solutions for $f_i(t)$. Each $\mathbf{a}_i$ has an associated $\mathbf{x}$-value—call this $\mathbf{x}_i$. As noted earlier, this has lm elements, and is a discrete approximation of $I^*(\nu_{\mathrm{ex}}, \nu_{\mathrm{em}})$. To interpret this correctly, $\mathbf{x}_i$ should be re-shaped into a l by m element rectangular array, which may be used by a contour-drawing or surface-plotting package to visualise $I^*(\nu_{\mathrm{ex}}, \nu_{\mathrm{em}})$.

[18] Finally, the eigenvalues of $\mathbf{X}$ also have an interpretation. Call these $(\lambda_1, \lambda_2, \ldots)$, sorted in order of magnitude. Then λ_i is proportional to the variance of the elements of $\mathbf{x}_i$. Since the $\mathbf{a}_i$ vectors form an orthogonal set, note that the total variance of all of the elements of X is proportional to $\Sigma\lambda_i$. Thus, the proportion of the total variance 'explained' by $\mathbf{x}_i$ is equal to $\lambda_i/\Sigma\lambda_i$. Also, recalling that we order the index i according to the magnitude of λ_i, define $\phi_i = \Sigma_{i=1,j}\lambda_j/\Sigma\lambda_i$. This indicator is useful for determining the success of the first few $(\mathbf{a}_i, \mathbf{x}_i)$ pairs-values of ϕ_1, ϕ_2 and other low-indexed ϕ_i's close to 1 suggest that much of the variability in the whole data cube is explained by the first few $\mathbf{x}_i$'s.

4. An Example

[19] In the following example, data was obtained from a stalagmite sample from Crag Cave, W. Ireland, that has already undergone extensive research in the form of isotope and crystal structure variations *Mc Dermott et al.* [1999] and our previous research into visualising luminescence EEMs *Baker et al.* [2000]. The luminescence excitation-emission matrix timeseries comprises 440 data points covering the period 10,000 years bp to present (giving an effective mean resolution of 2.5 yrs/EEM). *Baker et al.* [2000] demonstrate three periods of distinct luminescence properties: (1) 0–4,000 BP (0–75 mm from top), (2) 4,000–9,600 BP (75–420 mm from top), and (3) Before 9,600 BP (420 mm to base). A principal filter analysis was carried out using standardised intensities. The values of ψ_i for $i = 1\ldots 10$ are plotted in Figure 1. This shows that much of the variance is explained by the first three principal filters. Here we consider the first two. PF1 (Figure 2) increases towards the top of the sample, and exhibits a clear period of change to high values at 80 mm and a possible transition to low values at the base of the sample. The former clearly

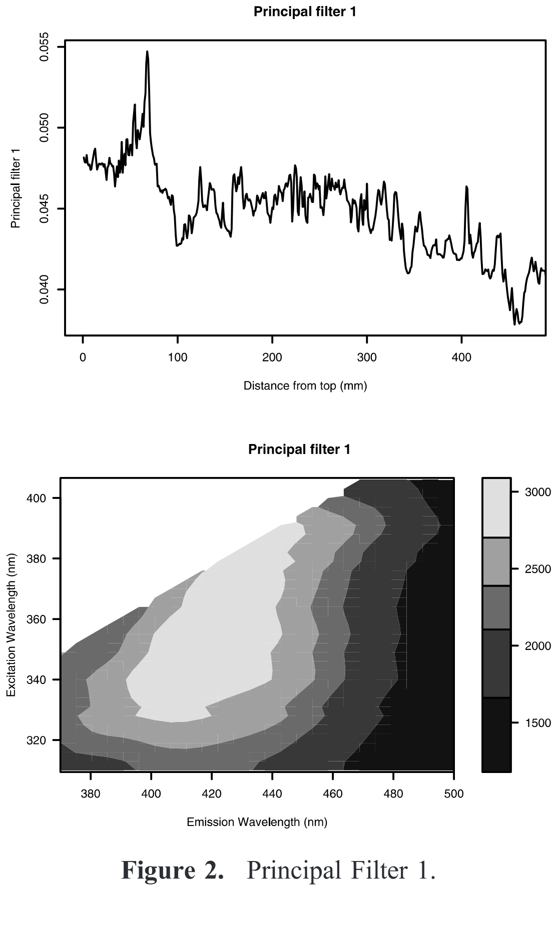

Figure 2. Principal Filter 1.

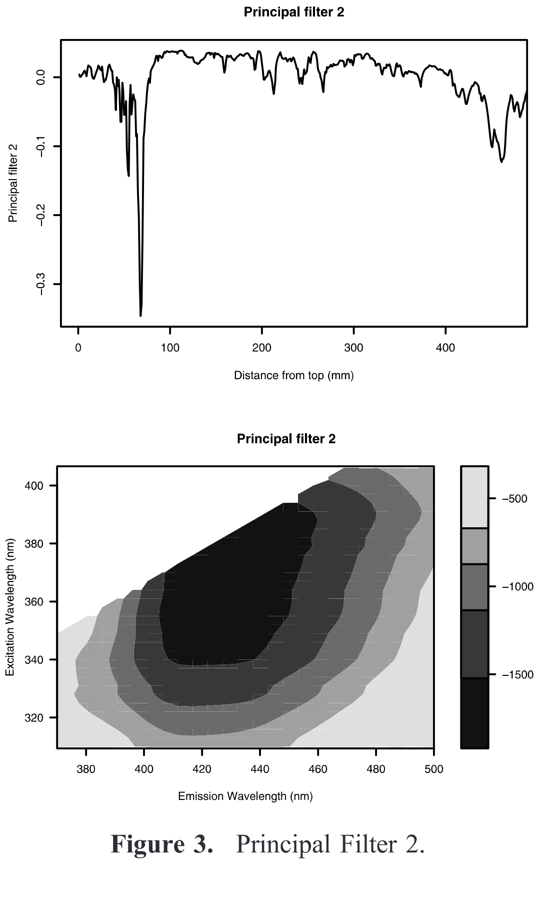

Figure 3. Principal Filter 2.

identifies the change in luminescence properties identified in *Baker et al.* [2000] at about 4000 BP in this sample. There is a notable shift in the relative distribution of intensities in the EEMs. This corresponds to a luminescence shape that has both high and low wavelength luminescence peaks; a high score occurs only when both peaks are present and a low score when one or both are absent. The importance of this factor as the first PF here is that it was not readily observed visually through individual analysis of EEMs and thus demonstrates the importance of the PFA technique as a diagnostic tool. PF2 (Figure 3) identifies change in luminescence EEM properties at higher wavelengths than PF1 and is therefore indicative of increased humic-like fluorescence. Again, the greatest change is in the top 80 mm of the stalagmite, and in *Baker et al.* [2000] an increase in the fluorescence emission wavelength at this time is similarly observed. Hence we are confident that PF2 is identifying organic matter preserved in the stalagmite which is more humic in nature; this includes a weaker transition to such material in the base of the sample, and three events between 200 and 300 mm from top which were also identified by *Baker et al.* [2000].

5. Conclusion

[20] Luminescence excitation emission wavelength timeseries can provide high-resolution palaeoclimate, hydrological and environmental records. However, due to the vast amount of inter-related data that can be obtained in four dimensions (luminescence excitation and emission wavelengths, luminescence intensity, time), interpretation of these data in terms of climate or environmental change is not always straightforward. PFA on the luminescence EEM timeseries from the Crag Cave stalagmite has shown the utility of this technique.

[21] An important aspect of this technique is an emphasis on the interaction between the computational and visual approaches. Clearly producing visual representations of EEMs is key to interpretation, by a very large number of images are produced in any given analysis. This technique uses a computational approach to finding 'interesting' filters which highlight a small number of patterns explaining nearly all of the variability in the EEM data matrix, making interpretation of the key trends in the data easily identifiable. Ongoing research by the authors addresses integrating PFA with existing visualisation software on order to provide a customisable user-friendly interface, making the technique available to a broader range of users. Also, research is underway to develop statistical tests of whether shifts or peeks in the PFA curves are due to genuine processes or artifacts of residual random noise in the data. Recent developments in fibre-optic probe technology have decreased spot-size available, and the increasing speed of spectrophotometers over recent years has greatly increased the rate at which data may be collected. Therefore increasing quantities of optical data at increasing optical and temporal/spatial resolution will require increasingly sophisticated statistical and visualisation techniques such as the PFA methodology outlined here.

References

Baker, A., Fluorescence excitation-emission matrix characterisa-tion of some sewage impacted rivers, *Environmental Science and Technology*, *35*, 948–953, 2001.

Baker, A., Fluorescence properties of some farm wastes: Implications for water quality monitoring, *Water Research*, *36*, 189–194, 2002.

Baker, A., D. Genty, and P. L. Smart, High-resolution records of soil humification and palaeoclimate change from speleothem luminescence excitation-emission wavelength variations, *Geology*, *26*, 903–906, 1998.

Baker, A., C. Brunsdon, M. Charlton, and F. McDermott, Visualisation of luminescence excitation-emission timeseries: Palaeoclimate implications from a 10,000 year stalagmite record from Ireland, *Geophysical Research Letters*, *27*, 2145–2148, 2000.

Caseldine, C., A. Baker, D. Charman, and D. Hendon, A comparative study of optical properties of NaOH peat extracts: Palaeoenvironmental implications for humification studies, *The Holocene*, *10*, 649–658, 2000.

Coble, P., Characterisation of marine and terrestrial dissolved organic matter in seawater using excitation-emission matrix spectroscopy, *Marine Chemistry*, *51*, 325–346, 1996.

Compagnucci, R., D. Araneo, and P. Canziani, Principal sequence pattern analysis: A new approach to classifying the evolution of atmospheric systems, *International Journal of Climatology*, *21*, 197–218, 2001.

del Castillo, C., P. Coble, J. Morell, J. Lopez, and J. Corredor, Analysis of the optical properties of the orinoco river plume by absorption and fluorescence spectroscopy, *Marine Chemistry*, *66*, 35–51, 1999.

Jiji, R., G. Cooper, and K. Booksh, Excitation-emission matrix fluorescence based determination of carbanate pesticides and polycyclic aromatic hydrocarbons, *Analytica Chimica Acta*, *397*, 61–72, 1999.

le Coupannec, F., D. Morin, O. Sire, and J. Peron, Characterization of dissolved organic matter (DOM) in landfill leachates using fluorescence excitation-emission matrix, *Environmental Technology*, *21*, 515–524, 2000.

Matthews, B., A. Jones, N. Theodorou, and A. Tudhope, Excitation-emission-matrix fluorescence spectroscopy applied to humic acid bands in coral reefs, *Marine Chemistry*, *55*, 317–322, 1996.

Maxwell, A., *Multivariate Analysis in Behavioural Research*, Chapman and Hall, 1977.

Mayer, L., L. Schick, and T. Loder, Dissolved protein fluorescence in two maine estuaries, *Marine Chemistry*, *64*, 171–179, 1999.

McDermott, F., et al., Holocene climate variability in europe: evidence from δ^{18}O, textural and extension-rate variations in three speleothems, *Quaternary Science Reviews*, *18*, 1021–1038, 1999.

McGarry, S., and A. Baker, Organic acid fluorescence: applica-tions to speleothem palaeoclimate reconstruction, *Quaternary Science Reviews*, *19*, 1087–1101, 2000.

Mobed, J., S. Ingsen, J. Autry, and L. McGown, Fluorescence characterisation of ihss humic substances: Total luminescence spectra with absorbence correction, *Environmental Science and Technology*, *30*, 3061–3066, 1996.

Newson, M., A. Baker, and S. Mounsey, The potential role of freshwater luminescence measurements in exploring runoff pathways in upland catchments, *Hydrological Processes*, *15*, 989–1002, 2001.

Parlanti, E., K. Worz, L. Geoffroy, and M. Lamotte, Dissolved organic matter fluorescence spectroscopy as a tool to estimate biological activity in a coastal zone submitted to anthropogenic inputs, *Organic Chemistry*, *31*, 1756–1781, 2000.

Patel-Sorrentino, N., S. Mounier, and J. Benaim, Excitation-emission fluorescence matrix to study ph influence on organic matter fluorescence in the amazon basin rivers, *Water Research*, in press.

Wu, F., and E. Tanoue, Molecular mass distribution and fluorescence characteristics of dissolved organic ligands for copper (II) in lake biwa, japan, *Organic Geochemistry*, *32*, 11–20, 2001.

Yan, Y., H. Li, and M. Myrick, Fluorescence fingerprint of waters: Excitation-emission matrix spectroscopy as a tracking tool, *Applied Spectroscopy*, *54*, 1539–1542, 2000.

C. F. Brunsdon, Department of Geography, University of Newcastle-upon-Tyne, NE1 7RU, UK. (chris.brunsdon@ncl.ac.uk)

A. Baker, Centre for Land Use and Water Resources Research (CLUWRR), University of Newcastle-upon-Tyne, NE1 7RU, UK. (andy.baker@ncl.ac.uk)