Biometrics 000, 000–000 DOI: 000
000 0000

Distributed lag models for hydrological data

Alastair M. Rushworth*

School of Mathematics and Statistics, University of Glasgow, G12 8QQ, UK

**email:* alastair@stats.gla.ac.uk

and

Adrian W. Bowman

School of Mathematics and Statistics, University of Glasgow, UK

and

Mark J. Brewer

Biomathematics and Statistics Scotland, UK

and

Simon J. Langan

The James Hutton Institute, Aberdeen, UK

Summary: The distributed lag model (DLM), used most prominently in air pollution studies, finds application wherever the effect of a covariate is delayed and distributed through time. We explore the use of modified formulations of DLMs to provide flexible varying-coefficient models with smoothness constraints, applicable in any setting in which lagged covariates are regressed on a time-dependent response. The models are applied to simulated flow and rainfall data and to flow data from a Scottish mountain river, with particular emphasis on approximating the relationship between environmental covariates and flow regimes in order to detect the influence of unobserved processes. It was found that under certain rainfall conditions some of the variability in the influence of rainfall on flow arises through a complex interaction between antecedent ground wetness and the time-delay in rainfall. The models are able to identify subtle changes in rainfall response, particularly in the location of peak influence in the lag structure and offer a computationally attractive approach for fitting DLMs.

Key words: P-splines; River flow; Rainfall; Distributed lag; Time series.

1. Introduction

1.1 *Motivation*

Modelling river flow has long been of interest to environmental scientists. In particular, relating river flow to covariates such as hill slope gradient, ground canopy coverage, rainfall and snowmelt has been an important goal, often forming the basis of large catchment-scale models known as distributed models (Beven, 1985). These models commonly make use of rich data sets including high resolution satellite imaging to estimate land usage or snow coverage in discrete areal units. Such data are costly and scarce, and often all that is readily available are average river flows and meteorological data observed at point locations. While large scale distributed models are unavailable in such situations, flexible statistical models may be invaluable in providing simplified approximations to the system of study. Our interest lies in capturing changes in the temporal dependence of river flow on rainfall using approximations based on flexible regression methods when covariates that would have allowed physically-based models to be constructed are absent.

The rainfall-flow relationship is the ensemble of a number of interacting physical processes, most of which are unobserved. River flow is partly generated by a slow 'baseflow' process where infiltration of rainfall from surrounding land seeps out over long periods of time, in a manner which depends on the sponge-like water storage properties of surrounding ground strata (Shaw, 1988). Baseflow accounts for much of the river flow that persists during very dry summer months. In contrast, a faster responding 'runoff' process causes a more instantaneous response of flow to rainfall and accounts for much of the river flow during storms and prolonged rainy periods (Beven, 1984). Fast runoff arises when antecedent soil moisture increases to a level where rainfall can move more quickly near the soil surface without being absorbed, and can result in a more rapid increase in flow over periods of hours.

Baseflow and runoff are in most catchments the two most important drivers of variation in flow levels, with the influence of each determined by physical factors including soil and subsurface composition, evaporation and transpiration.

Accumulation and ablation of transient snow packs also form a key feature in the hydrology of many temperate and high altitude river systems, causing baseflow and runoff to decrease during winter periods and increase suddenly during warmer winter and early spring months. Snow deposition, as well as depth and density, are highly spatially heterogeneous and are less commonly and reliably measured than rainfall data, and in catchments prone to heavy snowfall and accumulation, modellers must be mindful of the increased uncertainty that this presents in flow rates during winter periods.

The dynamics underlying river flow generation are complex and are difficult to capture even in a large physical model, and in addition hydrologists are often interested in identifying when latent unobserved processes are most active, particularly the influence of accumulation and melting of snow. Without detailed covariate data, we proceed by utilising flexible statistical methods with the aim of constructing a framework that allows us to approximate flow generating processes without attempting to identify the individual contributing components, that act over different timescales. The work described here is based on simple point-based rainfall data, but the wider modelling aim is to investigate methods by which complex environmental processes in both space and time can be approximated by semiparametric models.

1.2 *River Dee catchment*

The influence of different flow drivers can be illustrated with graphical summaries of hourly rainfall and flow data collected on the River Dee, whose source is in the Cairngorm Mountains

of Scotland which extends 141km before reaching the North Sea in Aberdeen with a catchment covering 2100km^2 (Baggaley et al., 2009). The River Dee is an important water resource contributing around 50% of the total water supply to 500,000 people for both drinking and industrial purposes, and is also of interest to environmental and conservation scientists with much of the river lying within reserved conservation areas (Langan et al., 1997).

[Figure 1 about here.]

The top left panel of Figure 1 displays a late winter period where little rainfall is observed and there is high flow variability, with some evidence of a daily cycle that might indicate the influence of melting snow. The top right panel displays a summer scenario with sparse rainfall, alongside low levels of river flow that appear to respond sluggishly to intermittent rain storms; this is typical of a period when baseflow dominates. The lower panels display periods when flow and rainfall are at high levels and a strong and immediate response to rainfall impulse is evident; this is a strong indication that runoff dominates during this period. It is evident from Figure 1 that the flow response to rainfall varies throughout the year, in accordance with seasonal changes in rainfall patterns. It is expected that the impact of a unit of rainfall at any point will be delayed and distributed in time, partly due to the distances between the rainfall and flow monitoring sites but also as a result of the influence of rainfall at other locations that is correlated but unobserved.

1.3 *Paper outline*

In Section 2.1, distributed lag models are introduced and some recent developments in their estimation are considered. In Section 2.2 we describe a time-varying distributed lag model for river flow and rainfall data, and in Section 2.3 a general DLM parameterisation is described that allows the incorporation of other covariates into the specification of how rainfall and flow interact. Section 2.4 discusses computational issues around estimation of flexible DLMs.

In Section 3 two different DLMs are applied to hourly rainfall and flow data, and Section 4 concludes with some discussion on model adequacy and suggestions for further work.

2. Modelling with distributed lag models

2.1 *The distributed lag model*

Approaches to modelling the temporal dependence of flow on rainfall often assume that rainfall $r(t)$ and flow $f(t)$ are determined by the convolution

$$f(t) = \int_0^\infty h(s)r(t-s)ds$$

where t is a point in time, s is a lag variable and h is some response function. This is known as the *instantaneous unit hydrograph* (Nash, 1957), describing the impact over time that a unit of rainfall has on flow. Jakeman et al. (1990) suggested filtering rainfall data to first estimate 'effective runoff' before proceeding to estimate h. Direct approaches to modelling rainfall and flow include ARX, NARMAX (Tabrizi et al. 1998) and functional coefficient modelling (Wong et al. 2007). It has been recognised that the shape of h is an important model choice and some authors have implemented polynomial constraints (Tabrizi et al. 1998) on h or used local polynomial smoothers (Wong et al. 2007). Models of the form

$$E(y_t) = \alpha + \beta_0 x_t + \beta_1 x_{t-1} + \ldots + \beta_l x_{t-l}$$

where the impact of one time-dependent variable (x_t) on another (y_t) is spread over time, can be called a distributed lag model (DLM). We refer to the β_is as *lag coefficients*, and these can be considered as forming a discrete estimate, $\hat{h}$, of the underlying function h, which we term the *lag structure*. In many time series settings, multicollinearity emerges when a time-dependent variable is transformed to a set of l lagged covariates and care must be taken in estimation to avoid the highly variable estimates that result from an unconstrained regression. Typically some constraint is applied to the β_ls, a common choice being the Almon lag (Almon, 1965) in which the lag coefficients must lie on a polynomial of order p, $f^p(l)$,

$l \in \{1, \ldots, L\}$, or the Koyck lag (Koyck, 1954) in which the lag coefficients are subject to a geometric decay constraint determined by the lag number.

DLMs have seen much development (Muggeo et al., Zanobetti et al., 2000; Welty et al., 2009) in the context of the delayed impact of urban air pollution on daily mortality counts, the former using a penalised spline approach and the latter a Bayesian approach with parameter constraints determined by carefully chosen priors. In the work by Zanobetti et al., (2000), particular interest lay in 'mortality displacement' or the 'harvesting' effect, a phenomenon characterised by negative coefficients in the tail of the estimated lag structure. More recently, the DLM was extended to allow lag coefficients to change smoothly with a covariate (Gasparrini et al., 2010) in addition to lying on a smooth curve, so that a smooth surface of coefficients results. Gasparrini et al., (2010) did not implement an explicit smoothing parameter, but instead transformed the log of the lag dimension onto a spline basis which has the effect of stronger smoothing at high lags.

Smoothing on model parameters rather than data is an example of "invisible smoothing " where appropriate smoothness levels are not easily judged by inspecting the fitted model. For this reason, a P-splines approach (Eilers and Marx, 1996) is convenient and is adopted here, where a rich set of uniformly spaced B-spline basis functions, together with a roughness penalty on neighbouring basis functions yields a fitted function with the appropriate level of smoothness. The strength of the roughness penalty is typically selected by minimising AIC or GCV. We propose a model specification similar to that of Gasparrini et al., (2010) enforcing smoothness on the lag structure, and allowing the smooth across lags to evolve over time or in response to a covariate.

2.2 *Time varying DLM*

We set up a model for flow at time t, $f(t)$, in terms of preceding upstream rainfall $(r(t-1),\ldots,r(t-L))$ and a parameter vector $(\beta_1,\ldots,\beta_L)$, with the constraint that the β_l lie on a spline constructed from a set of I degree 3 basis functions $\{B_1(\cdot),\ldots,B_I(\cdot)\}$. The form of the model is

$$\begin{aligned} f(t) &= \sum_{l=1}^{L}\beta_l r(t-l)+\epsilon(t) \quad \text{where} \quad \beta_l=\sum_{i=1}^{I}a_iB_i(l) \\ &= \sum_{l=1}^{L}\sum_{i=1}^{I}a_iB_i(l)r(t-l)+\epsilon(t) \end{aligned}$$

where $\epsilon(t)$ is an IID error process. We further allow the relationship between each rainfall lag variable $r(t-l)$ and $f(t)$ to change smoothly with time, the form of which depends on a further set of J B-spline basis functions $\{B_1(\cdot),\ldots,B_J(\cdot)\}$ so that $a_i=\sum_{j=1}^{J}b_{ij}B_j(t)$. This gives the representation

$$f(t)=\sum_{l=1}^{L}\sum_{i=1}^{I}\sum_{j=1}^{J}b_{ij}B_j(t)B_i(l)r(t-l)+\epsilon(t).$$

In matrix notation,

$$\begin{aligned} \mathbf{f} &= \mathbf{X}\theta+\epsilon=\left(f(t_1),\ldots,f(t_n)^T\right) \\ \mathbf{X} &= \mathbf{B}_{\mathrm{J}}\square\mathbf{R}\mathbf{B}_{\mathrm{I}}=(\mathbf{B}_{\mathrm{J}}\otimes\mathbf{1}_{\mathrm{I}}')\odot(\mathbf{1}_{\mathrm{J}}'\otimes\mathbf{R}\mathbf{B}_{\mathrm{I}}) \\ \theta &= (b_{11},b_{21},\ldots,b_{I1},\ldots,b_{1J},b_{2J},\ldots,b_{IJ})^T \end{aligned}$$

Where the i^{th} of $\mathbf{B}_{\mathrm{J}}$ is $\{B_1(t_i),\ldots,B_J(t_i)\}$, i^{th} row of $\mathbf{R}$ is $(r(t_i-1),\ldots,r(t_i-L))$ and $\square$ is the Box product as used by Eilers et al. (2006). The specification omits and intercept term in order to preserve the interpretation of the β_i as an approximation of the transfer function of ... The smooth change in each β_i through time is captured by the same basis set regardless of lag number i, simplifying the model specification considerably. Furthermore, the intercept if included, would represent the flow levels that persist when rain has not recently fallen, which is closely related to baseflow, which in turn varies with time and is not identifiable

with short term variation.

We wish to control the level of smoothness in the fitted coefficients in two ways: by how each rainfall lag variable $r(t-l)$ influences $f(t)$ as t changes; and by how different the influence of $r(t-l)$ and $r(t-l+1)$ is allowed to be at any time t. These constraints will be represented by two different roughness penalties. The first term, $\lambda_1 \mathbf{D}_1^T \mathbf{D}_1$, penalises wiggliness of the β_is through time, and so $\mathbf{D}_1$ is a block diagonal matrix where each block is a quadratic difference matrix $\mathbf{P}_\mathrm{J}$ with J columns so that

$$\mathbf{P}_\mathrm{J}\alpha' = \sum_{j=1}^{J-2} (\alpha_{j+2} - 2\alpha_{j+1} + \alpha_j)^2$$

and, in Kronecker notation, $\mathbf{D}_1 = \mathbf{P}_\mathrm{J} \otimes \mathbf{I}_\mathrm{I}$. The second penalty term, $\lambda_2 \mathbf{D}_2^T \mathbf{D}_2$, controls differences between β_l and β_{l+1}, $l \in \{1, \ldots, L\}$ at any time t and this is achieved similarly by penalising differences between $b_{i,j}$ and $b_{i,j+1}$ for $i \in \{1, \ldots, I\}$ and for all j, so that $\mathbf{D}_2 = \mathbf{I}_\mathrm{J} \otimes \mathbf{P}_\mathrm{I}$. Combining the two penalties, the parameter estimates $\hat{\theta}$ are

$$\hat{\theta} = \mathbf{S}\mathbf{f} = \left(\mathbf{X}^T\mathbf{X} + \lambda_1 \mathbf{D}_1^T \mathbf{D}_1 + \lambda_2 \mathbf{D}_2^T \mathbf{D}_2\right)^{-1} \mathbf{X}^T \mathbf{f}$$

with standard errors given by $\mathbf{se}(\hat{\theta}) = \sqrt{\mathrm{diag}(\mathbf{H}^T\mathbf{H})}$ where $\mathbf{H} = \mathbf{X}\mathbf{S}$

2.3 *General specification*

More generally, DLMs can be specified so that the lag structure varies with any set of covariates. For example, if the β_is are required to change smoothly and non-linearly with one additional covariate, a model matrix with one additional Box product, $\square$, must be constructed.

Let $x_1(t), \ldots, x_r(t)$ be r n-length time-dependent covariates and $\mathbf{J}^1, \ldots, \mathbf{J}^r$ be marginal basis matrices defined on the $x_i()$ such that $\mathbf{J}^i_{m\bullet} = [B_1(x_i(m)), \ldots, B_{J_i}(x_i(m))]$, and J_i is the size of the basis set defined on the ith covariate. A full multidimensional DL model matrix is

then

$$\mathbf{X} = \mathbf{J}_1 \square \mathbf{J}_2 \square \ldots \square \mathbf{J}_r \square \mathbf{RB}_{\mathrm{I}}$$

where $\mathbf{RB}_{\mathrm{I}}$ is defined as in section 2.2 and the corresponding θ is a vector of spline coefficients of length $I \prod_{i=1}^{r} J_i$. Since the coeffients define a smooth estimate in $r+1$ dimensions (r for coefficient bases and 1 for lag structure basis) we require $r+1$ penalty terms that can be expressed as a sequence of kronecker products with identity matrices

$$\mathbf{D}_i = \bigotimes_{j<i} \mathbf{I}_j \otimes \mathbf{P}_i \otimes \bigotimes_{j>i} \mathbf{I}_j$$

where each $\mathbf{D}_i$ corresponds to a penalty on the ith dimension of $\mathbf{A}$.

2.4 *Computational aspects*

The models described in Sections 2.2 and 2.3 potentially require the storage and manipulation of $n \times IJ$ and $IJ \times IJ$ matrices which can be expensive. Currie et al. (2006) describe how, if model matrices arising in tensor-product type models can be factorised so that $\mathbf{X} = \mathbf{X}_1 \otimes \mathbf{X}_2$, then much of the computational and storage overhead can be bypassed. In the present case, $\mathbf{X}$ cannot be so factorised, due to the row-wise tensor product matrix structures. However, significant savings can be made by exploiting the sparseness properties of many of the model objects. Since a set of penalised B-splines is used, all basis matrices are sparse, and additionally they are defined on consecutive sequences of integers (time and lag indices here) and are therefore banded. Therefore $\mathbf{RB}_{\mathrm{I}}$ is a banded sparse matrix, and hence $\mathbf{X} = \mathbf{Z} \square \mathbf{RB}_{\mathbf{I}}$ is banded, and in turn, $\mathbf{X}^T\mathbf{X}$ and $\left(\mathbf{X}^T\mathbf{X} + \lambda_1 \mathbf{D}_1^T \mathbf{D}_1 + \lambda_2 \mathbf{D}_2^T \mathbf{D}_2\right)$ are banded and sparse. Hence we are required only to manipulate a banded sparse matrix objects which is faster than the general sparse case, and drastically reduces the need for storage requirement. The sparseness properties are further enhanced by the zero-inflated distribution of hourly rainfall data. Further computational gains can be made by working with the Cholesky decompositions (see additional materials for `R` implementation). In `R`

sparse matrix algebra is easily performed using the `Matrix` package (Bates and Maechler., (2011)).

As discussed in section 2.1, choice of appropriate smoothing here depends on first choosing a rich basis and then choosing λ that reduces some information criterion. In the present case we proceed iteratively by selecting a basis size, then fitting a model over a small number of diverse roughness penalty strengths.

3. Application

3.1 *Time varying DLM on simulated data*

To first show how well the method described in Section 2 performs in recovering lag structures in noisy time series we first perform some simulations. First of all, simulated flow is generated by convoluting each of three hypothetical lag structures (Figure ...) with our observed rainfall at Braemar in 2006. We then add some white noise to each of these simulated data and attempt to recover the DL curves using the method outlined in section 2.1.

3.2 *Time varying DLM on River Dee data*

We consider river stream discharge data ($m^3 s^{-1}$) collected from Polhollick on the River Dee catchment in the North East of Scotland. In addition, we have hourly rainfall data collected from a monitoring station 10 miles upstream in the same catchment at the Braemar irrigation farm. High resolution data is relatively scarce and what follows has been fitted to 8861 average hourly flows and rainfall for the year 2006 only; ideally several years data would be considered and adjustment made to account for seasonal and interannual variation.

The model in Section 2.2 was fitted to the River Dee data with $J = 100$, $I = 50$ and $L = 100$. A rich basis was first chosen, with K and L selected so that, without penalty terms

(ie. $\lambda_1 = \lambda_2 = 0$), the model overfits the data. The optimal λ_1 and λ_2 were found by searching a logarithmic grid for the values that minimised AICc (Hurvich et al., 1998). AICc is designed to avoid undersmoothing in semiparametric models and is defined as $\log(\hat{\sigma}^2) + 1 + \frac{2(\mathrm{tr}(\mathbf{H})+1)}{n-\mathrm{tr}(\mathbf{H})-2}$. The fit of the model can be examined by inspecting plots of observed and fitted values during different months of 2006 (Figure 2).

[Figure 2 about here.]

In the upper plot of Figure 2 we see a period of very low rainfall and high flow; it is clear that the model performs poorly here. By contrast, the lower plot corresponds to a wet period and the model fits well, despite the extreme levels reached in river flow.

[Figure 3 about here.]

We can also examine the fitted DL curves that yield the fitted values in Figure 2, by examining fitted curves at monthly intervals, as shown in Figure 3. There is clear evidence of differences in the estimated lag structures throughout the year: in summer months, lag structures lie mostly flat, indicating a slow and delayed response; correspondingly DL curves during winter months are sharply peaked and very tightly contained within their 95% confidence intervals. The strong peaks seen particularly in November indicates a fast response that coincides with a period of prolonged wet weather. Lag structure estimates for late winter and early spring are very high, indicating extremely high influence of rainfall up to the most distant lags. The largest lag is 100 hours prior to a flow event; such high estimates are physically unrealistic, and suggest the influence of other unobserved processes.

In the final weeks of observation, a sustained period of heavy rain and an overall increase in flow with progressively more extreme peaks is observed. The lag structures within this period (not shown) progressively increase in size both in overall influence and in the peak influence at approximately 10 lags. Furthermore, the changes in lag structure are consistent

with an increase in ground saturation causing a higher proportion of rainfall to convert to runoff. Subsequent flow levels are highly sensitive to new rainfall and these are reflected with highly peaked DL curves. We proceed to construct a model that attempts to account for temporal variation in lag structures during wet periods using information on long-term ground wetness.

3.3 *A 'ground-wetness' varying DLM on River Dee data*

We now consider introducing a covariate representing unobserved antecedent ground wetness, for which a 30 day moving-window mean of observed hourly rainfall with exponentially decaying weights is constructed as a proxy. We call this proxy $W(t)$, and for the River Dee data, $W(t)$ is shown in the lower right panel of Figure **??**. The choice of 30 days represents the belief that variation in rainfall response is driven by a larger ensemble of precipitation outwith the largest lag of the DLM, particularly during prolonged wet periods. A number of window widths were tried and the resulting model was not found to be sensitive to small changes. An alternative approach might make use of water residence time distributions, if known, to inform the weights and window widths in construction of such a proxy. In what follows, $W(t)$ is assumed to be the only modifying factor of the lag structure and if successful should account for much of the temporal variation in the β_i observed in section 3.1. In similar notation to 2.2 the model is specified as

$$f(t) = \sum_{l=1}^{L} \sum_{j=1}^{J} \sum_{m=1}^{M} c_{jm} B_m(W(t)) B_j(l) r(t-l) + \epsilon(t).$$

Estimation proceeds as in Section 2.2, where the coefficient vector $\theta = (c_{11}, \ldots\ldots, c_{LN})$. The model was fitted using hourly data on the River Dee in 2006 during which rain storms occurred regularly and ground moisture is expected to have changed throughout the observation period. The model parameters were $L = 100$, $M = 50$, $J = 100$ again, representing an overfitted model when the penalty vector $\lambda = \mathbf{0}$. It was found when selecting optimal λ that AICc often preferred undersmooth estimates for variation in the $W(t)$ dimension, hence

it was decided to use the 'optimal' estimate as an upper bound on λ to maintain model simplicity. Interest lies in how the β_i respond to different levels of $W(t)$. The left panels of Figure **??** illustrate the changes in lag structure at different quantiles of the distribution of $W(t)$; at higher levels of $W(t)$ more peaked and overall larger lag structures are visible, particularly at the highest levels of $W(t)$. In the right hand panels of Figure **??**, images illustrating the changes in lag structure across the range of $W(t)$, and through time, are given. An important feature here is the shift in peak influence from later lags to earlier lags which is visible as $W(t)$ increases. It is also notable that less dominant peaks later in the lag structure appear at the lowest and highest levels of $W(t)$.

4. Discussion

We have proposed flexible and computationally attractive distributed lag models with smoothness penalties that have been successful in capturing rainfall and stream flow dynamics. The models were applied to the River Dee data in section 3 and identified a complex and time varying relationship between river flow and rainfall and in particular, the second model presents evidence that some of this variability arises through a complex interaction between slowly changing ground wetness and the time when rain falls. It was also found that the degree and location of peak influence in the lag structure can change dramatically, and that these were persistent features under the use of different strength penalties.

We note a strong and consistent responsiveness in flow levels when rainfall has been heavy or prolonged, with a clear peak in influence that indicates the predominance of fast-moving runoff. At other times less consistent or interpretable response functions are estimated which might be the symptom of a number of possible problems. One problem is that in winter periods, rainfall data collected high on the River Dee is far less reliable, owing to intermittent bouts of snowfall and freezing temperatures so that pieces of snow or ice may be mistakenly

recorded as heavy rainfall when they finally melt. Another more significant factor in the potential 'mismatch' between rainfall and flow is that flow is driven by a *spatial* ensemble of precipitation as well as temporal, and it is highly likely that during spells of localised weather, our data is inadequate, and may result in spurious estimates for lag structures.
The model of Section 3.3 attempts to take account of the impact of rainfall ensembles acting over longer periods of time by assuming a proxy for ground moisture content $W(t)$. The choice of $W(t)$ is based simply on a notional upper lag beyond which rainfall is thought not to impact the current catchment state. A less ad-hoc approach might involve a modified version of the cross-lag interaction DLM of Muggeo et al, (2008). in which the value of a lagged covariate can interact with the impact that more recent lagged covariates can have on flow levels. In particular a heirarchical model in which an additional, long-term transfer function over longer periods of time acts on the coefficients of a shorter range DL curve might better represent the temporal dynamics of flow responsiveness. While more robust to model misspecification, this approach would present significant challenges in estimation and parameterisation.

Despite the application of validation procedures for finding optimal smoothness, there is also a potential need for stronger smoothing on the lag structure during summer months than at other times, and more generally on later parts of the lag structure compared to later parts. A solution to the latter point might be gained by applying ridge penalties to the β_i so that models in which $\beta_i \to 0$ as $i \to \infty$ are preferred; see Muggeo (2008) for an example using health data. For the former we might specify a functional form *a priori* for the smoothing parameters λ so that the penalty term depends on the the lag index i or a seasonal index for example. However, these approaches have not been pursued, as they require that we make assumptions about the underlying response structure of our time series that may be not

be justified given the shortcomings of the data collection process and spatial heterogeneity outlined earlier.

Further work might make use of additional rainfall time series data in order to understand better the impact of the spatial distribution of rainfall, clearly evident in the estimates presented here. Flow data at additional sites on the river might also be incorporated to identify how spatially heterogeneous rainfall affects flow measured at different locations on the river network.

Acknowledgements

The stream discharge data used in this report were provided by Scottish Environmental Protection Agency (SEPA), while the precipitation and temperature data come from the Met Office MIDAS data set and was provided via the British Atmospheric Data Centre website. Thanks to Nikki Baggaley for her help with flow data and advice at an early stage. Alastair Rushworth was funded by an EPSRC CASE studentship. Mark Brewer and Simon Langan were funded by the Rural and Environment Science and Analytical Services division of the Scottish Government.

References

Almon, S. (1965). The distributed lag between capital appropriations and expenditures. *Econometrica* **33**, 178–196

Alves, M.B., Gamerman, D. and Ferreira, M.A.R. (2010). Transfer functions in dynamic generalized linear models. *Statistical Modelling* **10**, 03–40

Baggaley, N.J., Langan, S.J., Futter, M.N., Potts, J.M., and Dunn, S.M. (2009). Long-term trends in hydro-climatology of a major Scottish mountain river. *Science of the Total Environment* **407**, 4633–4641.

Bates, D. and Maechler M. (2011). Matrix: Sparse and Dense Matrix Classes and Methods. R package version 1.0-1. http://CRAN.R-project.org/package=Matrix

Beven, K.J. (1985). Distributed models. *Hydrological forecasting* 405–435.

Beven, K.J. (2004). Rainfall-runoff modelling: the primer *John Wiley and Sons, England.* ISBN 0470866713.

Currie, I.D., Durban, M. and Eilers, P.H.C (2006). Generalised linear array models with applications to multidimensional smoothing. *Journal of the Royal Statistical Society B* **68**, 259–280.

Eilers, P.H.C. Currie, I.D. and Durban, M. (2006). Fast and compact smoothing on large multidimensional grids *Computational Statistics and Data Analysis* **50**, 61–76.

Eilers, P.H.C. and Marx, B.D. (1996). Flexible smoothing with B-splines and penalties. *Statistical Science* **1**, 89–121.

Ferguson, R.I. (1984). Magnitude and modelling of snowmelt runoff in the Cairngorm mountains, Scotland. *Hydrological Sciences Journal* **29**, 49–62.

Gasparrini, A., Armstrong, B. and Kenward, M.G. (2010). Distributed lag non-linear models. *Statistics in medicine* **29**, 2224–2234.

Hurvich, C.M., Simonoff, J.S., and Tsai, CL. (1998). Smoothing parameter selection in non-parametric regression using an improved Akaike information criterion. *JRSSB* **60**, 271–293.

Jakeman, A.J., Littlewood, I.G. and Whitehead, P.G. (1990). Computation of the instantaneous unit hydrograph and identifiable component flows with application to two small upland catchments. *Journal of Hydrology* **117**, 275-300.

Koyck, L.M. (1954). Distributed lags and investment analysis. North-Holland Publishing Company, Amsterdam.

Langan, S.J., Wade, A.J., Smart, R., Edwards, A.C., Soulsby, C.,

Billett, M.F., Jarvie, H.P., Cresser, M.S., Owen, R. and Ferrier, R.C. (1997). The prediction and management of water quality in a relatively unpolluted catchment: current issues and experimental approaches. *The Science of the Total Environment* **194-195**, 419-435.

Muggeo, V.M.R. (2008). Modeling temperature effects on mortality: multiple segmented relationships with common break points *Biostatistics*. **9**, 613-620.

Nash, J.E. (1957). The form of the instantaneous unit hydrograph. *General Assembly of Toronto*. **101**, 114–121.

Shaw, E.M. (1988). Hydrology in Practice. *Van Nostrand Reinhold (International)* ISBN 0748744487

Tabrizi, M.H.N., Said, S.E., Badr, A.W., Mashor, Y. and Billings, S.A. (1998). Nonlinear modeling and prediction of a river flow system. *Journal of the American Water Resources Association* **34**, 1333–1339.

Welty, L.J., Peng, R.D., Zeger, SL. and Dominici, F. (2009). Bayesian distributed lag models: estimating effects of particulate matter air pollution on daily mortality. *Biometrics* **65**, 282–291.

Wong, H., Ip, W., Zhang, R. and Xia, J. (2007). Non-parametric time series models for hydrological forecasting. *Journal of Hydrology* **332**, 337–347.

Zanobetti, A., Wand, M.P., Schwartz, J. and Ryan, L.M. (2000). Generalized additive distributed lag models: quantifying mortality displacement. *Biostatistics* **1**, 279–292.

Figure 1. Rainfall and flow responses from the River Dee 2006. Red lines are flow rates ($m^3 s^{-1}$); black line segments are hourly rainfall levels (mm).

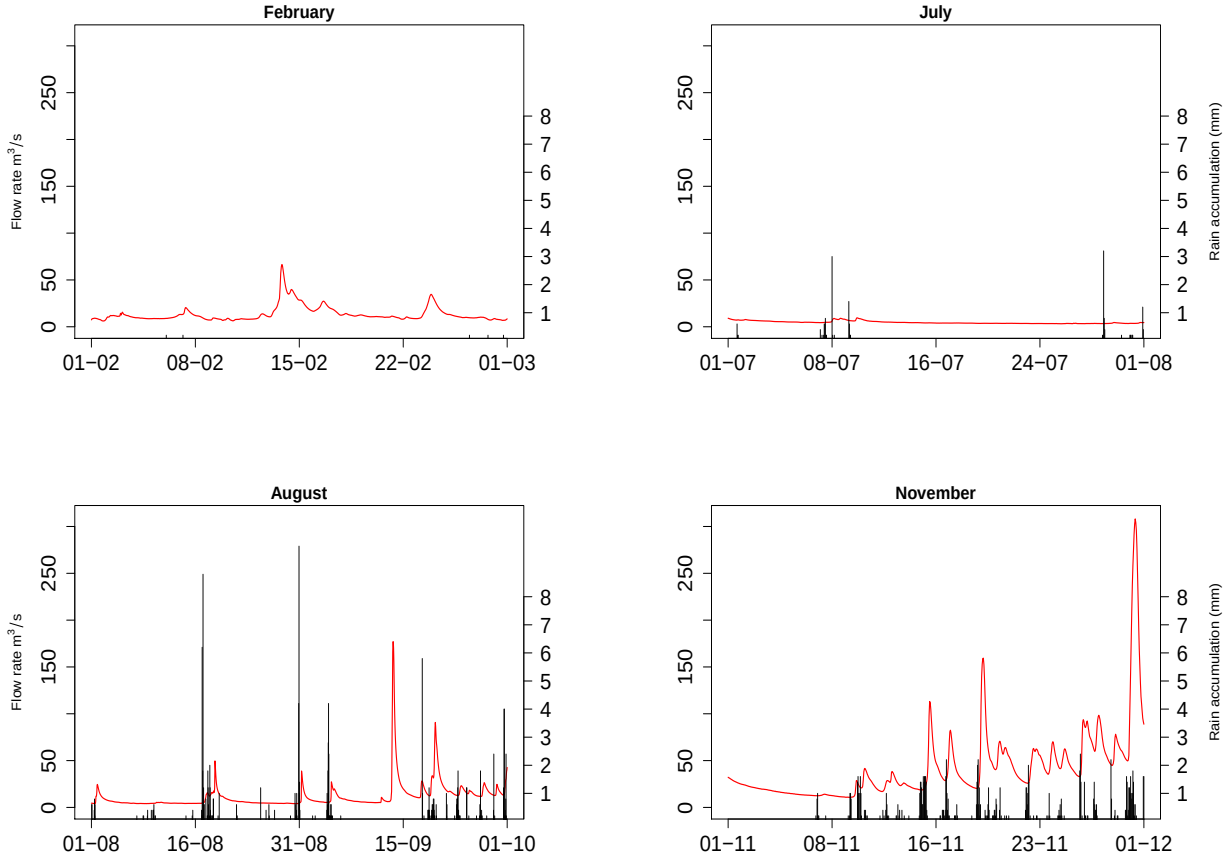
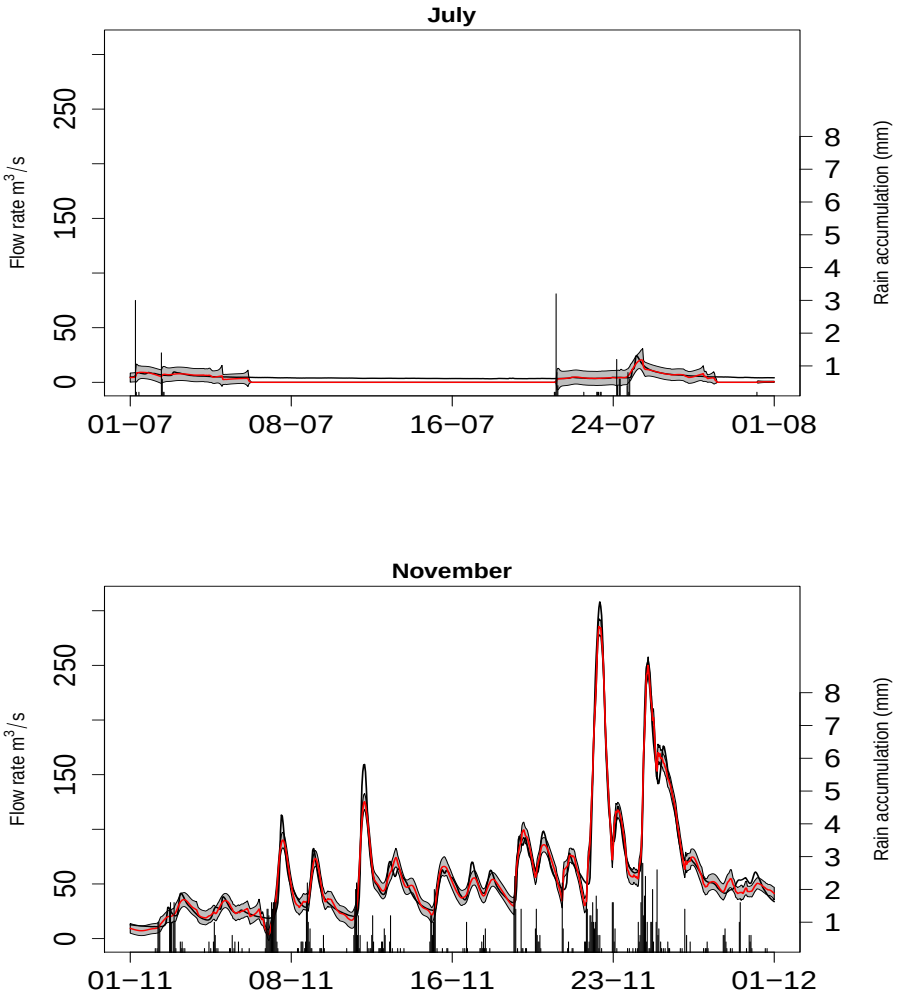


Figure 2. Fitted flows alongside flows observed on the River Dee: black lines represent observed flow levels, red lines represent fitted flows and vertical black line segments are hourly precipitation



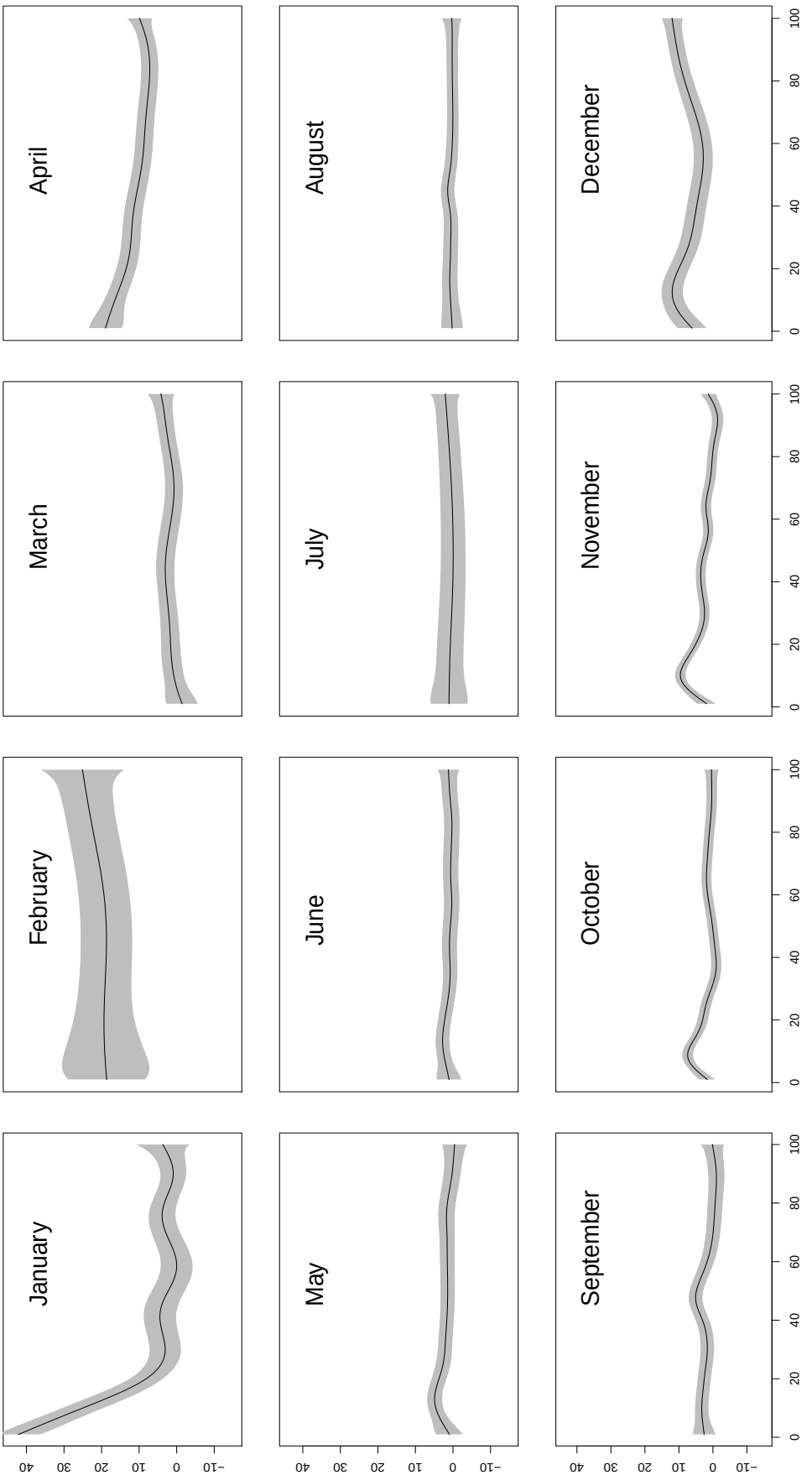


Figure 3. Lag structures with point wise 95% confidence regions estimated from River Dee rainfall and flow data plotted at monthly 'snaphots'; x-axes correspond to the lag numbers (between 1 and 100) of points on the response function