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 specify modified formulations of DLMs to provide computationally attractive, flexible varying-coefficient models that are applicable in any setting in which lagged covariates are regressed on a time-dependent response. We investigate the application of such models to rainfall and river flow and in particular their role in understanding the impact of hidden variables at work in river systems. We apply two models to data from a Scottish mountain river, and we fit to some simulated data to check the efficacy of our model approach. During heavy rainfall conditions, changes in the influence of rainfall on flow arises through a complex interaction between antecedent ground wetness and a time-delay in rainfall. The models identify subtle changes in responsiveness to rainfall, particularly in the location of peak influence in the lag structure.
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, surrounding land usage, 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 rainfall-flow relationships during winter periods.

The dynamics underlying river flow generation are complex and are difficult to capture in detailed physical models, and in addition, hydrologists are often interested in identifying when latent 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 data*

The influence of different flow drivers is best illustrated with graphical summaries of hourly rainfall and flow data collected on the River Dee (that is later used in modelling): hourly rainfall accumulations (mm) collected to the nearest 0.1mm at Braemar and river discharge

data (m^3s^{-1}) are collected from Polhollick, both located in the North East of Scotland. Previous work on the River Dee showed that hourly flows are the highest resolution necessary to identify peak flow levels (Baggaley et al., 2009). The source of the River Dee is in the Cairngorm Mountains of Scotland, and it extends 141km before reaching the North Sea in Aberdeen with a total catchment area 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 a November period in which flow and rainfall are at high levels and a strong and immediate response to rainfall impulse is evident - a strong indication that runoff dominates during this period. The nature of the responsiveness is more easily seen in the bottom right panel of Figure 1 which shows a single week from its lefthand neighbour. 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 also clear that the influence of a unit of rainfall is delayed and spread over time, caused by spatial separation (and intervening ground conditions) of rainfall across the catchment and flow gauges.

1.3 *Paper outline*

In Section 2.1, distributed lag models (DLMs) are introduced and recent developments in DLM methodology is reviewed. In Section 2.2 we construct a time-varying distributed lag model for river flow and rainfall data, and in Section 2.3 a general DLM formulation 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 simulated data and to hourly rainfall and flow data from the River Dee. 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. 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 (Zanobetti et al., 2000; Muggeo et al., 2008; Welty et al., 2009; Gasparrini et al., 2010) in the context of the delayed impact of urban air pollution on daily mortality counts. In this setting interest lies in specifying plausible shapes for DL curves and in particular the 'mortality displacement' effect, a phenomenon characterised by negative coefficients in the tail of the estimated lag structure. Zanobetti et al. (2000) propose a generalised model taking a penalised spline approach to modelling DL curves while Welty et el. (2009) discuss a Bayesian approach with penalties on parameters determined by carefully chosen priors. Others (Muggeo et al., 2008; Gasparrini et al., 2010) allow lag coefficients to change with temperature in addition to lying on a smooth curve, so that a surface of lag coefficients results. Muggeo et al. (2008) present a framework where dependance of the DL curve on temperature is piecewise linear, with unknown breakpoints. Gasparrini et al. (2010) propose DL curves that lie on a bivariate surface parameterised by splines defined on lag index and temperature values.

Smoothing on model parameters rather than data, as is the case with DL curve estimation, is

a situation where appropriate smoothness levels are not easily judged by visual inspection of 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 some information criterion. We proceed to construct a DL model for rainfall and river flow rates, relaxing the assumption of a fixed lag structure; the use of P-splines are found to facilitate specification of flexible models while maintaining a high level of computational efficiency by taking advantage of sparse model objects.

2.2 *Time varying DLM*

We set up a model for flow at time t, $f(t)$, in terms of a weighted sum of preceding upstream rainfall $(r(t-1),\ldots,r(t-L))$ with weights $(\beta_1,\ldots,\beta_L)$, subject to 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

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

where α is an intercept term and $\epsilon(t)$ 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

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

In matrix notation,

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

Where the i^{th} row 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 intercept included in the specification represents flow rates after rainfall has not been observed for over L lags.

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 the 'wiggliness' of the β_is through time, and so $\mathbf{D}_1$ is a block matrix where each block is a quadratic difference matrix $\mathbf{P}_{\mathrm{J}}$ with J columns so that

$$\mathbf{P}_{\mathrm{J}}\mathbf{b} = \sum_{i=1}^{I} \sum_{j=1}^{J-2} (b_{i,j+2} - 2b_{i,j+1} + b_{i,j})$$

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}, $l \in \{1, \ldots, L-1\}$ at any time t and this is achieved similarly by penalising differences between $b_{i,j}$, $b_{i,j+1}$ and $b_{i,j+2}$ for $i \in \{1, \ldots, I\}$ and $j \in \{1, \ldots, J-2\}$, so that $\mathbf{D}_2 = \mathbf{I}_{\mathrm{J}} \otimes \mathbf{P}_{\mathrm{I}}$. Combining the two penalties, the parameter estimates $\hat{\boldsymbol{\theta}}$ are obtained by penalised least squares by

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

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

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, $\Box$, 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 $x_i()$ so that the m^{th} row of $\mathbf{J}^i$ is $[B_1(x_i(m)), \ldots, B_{J_i}(x_i(m))]$, and J_i is the size of the basis set defined on the ith covariate. A general multidimensional DLM model matrix is defined as

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

where $\mathbf{RB}_{\mathrm{I}}$ is defined as in Section 2.2 and the corresponding $\boldsymbol{\theta}$ is a vector of spline coefficients and an intercept with length $1 + I \prod_{i=1}^{r} J_i$. Since the model sets up a smooth in $r+1$ dimensions (r for coefficient bases and 1 for lag structure basis) we require $r+1$ penalty terms, these can be expressed as a sequence of Kronecker products with identity matrices

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

where $\bigotimes_{j<i} \mathbf{I}_j = \mathbf{I}_1 \otimes \ldots \otimes \mathbf{I}_{i-1}$, and each $\mathbf{D}_i$ corresponds to a roughness penalty on the ith dimension of the tensor smooth defined by $\mathbf{X}$.

2.4 *Computational aspects*

The models described in Sections 2.2 and 2.3 potentially require the storage and manipulation of $n \times (1 + IJ)$ and $(1 + IJ) \times (1 + IJ)$ 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 their rows are defined on consecutive sequences of integers (time and lag indices here) and are therefore banded. Therefore $\mathbf{RB_I}$ is a banded sparse matrix, and hence $\mathbf{X} = \mathbf{B_J} \square \mathbf{RB_I}$ is banded, and in turn, $\mathbf{Z}^T\mathbf{Z}$ and $\left(\mathbf{Z}^T\mathbf{Z} + \lambda_1\mathbf{D}_1^T\mathbf{D}_1 + \lambda_2\mathbf{D}_2^T\mathbf{D}_2\right)$ are banded and sparse. Hence we are required only to manipulate a banded sparse matrix object which is faster than the general sparse case, and dramatically reduces storage requirements. 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)).

In order to choose an appropriate 'rich' basis size when applying models of Sections 2.2 and 2.3, a short iterative process is required to determine the minimal basis size so that on application of different strength penalties, a broad range of model smoothnesses result and when $\boldsymbol{\lambda} = \mathbf{0}$, the model overfits the data. Having selected a rich basis, the optimal penalties λ_i are 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 $\log(\hat{\sigma}^2) + 1 + \frac{2(\mathrm{tr}(\mathbf{H})+1)}{n-\mathrm{tr}(\mathbf{H})-2}$.

3. Application

3.1 *Time varying DLM on simulated data*

Before applying the models in Section 2 to river flow data, we first apply the methods to simulated data to examine how well the model captures different lag structures in the presence of different error processes, and to test the performance of AICc in selecting the

optimal level of smoothness for the fitted DL curves.

Three time series of flow data were constructed by convoluting the 2006 Braemar hourly rainfall data described in Section 1.2, with three different DL curves defined up to 50 lags so that,

$$f_{ij}(t) = \sum_{l=1}^{L} \beta_{li} r(t-l) + \epsilon_j$$

where β_{L1} and β_{L2} are time invariant and are based on Gamma distribution functions (shown in Figure 2), and β_{L3} varies smoothly over time between the shapes of β_{L1} and β_{L2}. Furthermore, $\epsilon_1 \sim N(0, 0.04)$, $\epsilon_2 \sim N(0, 0.16)$ and ϵ_3 follows a normal random walk process through time with $\sigma = 0.01$ so that nine possible scenarios result. We then simulated 200 times from each of the nine scenarios and fit the time varying model of Section 2.2, with $L = 50$ lags and moderate basis sizes of $I = J = 20$. Since our main interest is in recovering the underlying DL structures, the fitted DL curves can be compared against the true curve using 95% pointwise simulation envelopes and the root mean squared error (RMSE), both shown in Figure 2.

[Figure 2 about here.]

From Figure 2 it can be seen in all cases that the DLM model recovers the underlying lag structure well, with the 95% envelope functions lying close to the true curve. Some detail is lost in some of the fitted functions at the 'peak' of the estimates, particularly where the peak is pointed, which is likely to be a result of the choice of basis size being slightly too small. The random walk process is included as an example of a strongly correlated error process that environmental data often exhibit, and while the model performs less favourably than with independent errors, it still succeeds in recovering the shape of the underlying function.

3.2 *Time varying DLM on River Dee data*

The River Dee data describe in Section 1.2 is now considered with the model of section 2.2. High resolution data is relatively scarce and what follows has been fitted to the 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.

A rich basis was first chosen, with $I = 50$ and $J = 100$ selected so that, without penalty terms (ie. $\lambda_1 = \lambda_2 = 0$), the model overfits the data. A large number ($L = 100$) of lags were chosen and the optimal λ_1 and λ_2 were found by the method described in Section 2.4. The fit of the model can be examined by inspecting plots of observed and fitted values during different parts of 2006, shown in Figure 3.

[Figure 3 about here.]

In the upper plot of Figure 3 we see a period of very low rainfall and low flow; the model performs poorly where rainfall has not been observed for more than $Ls = 100$ hours, with the intercept $\alpha = 12.7$ left to account for the remaining recession of flow levels. By contrast, the lower plot corresponds to a wet period and the model fits well, despite the extreme levels reached in river flow.

[Figure 4 about here.]

We can also examine the fitted DL curves which are shown in Figure 4. 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, and during wet autumn months are sharply peaked and very tightly contained within their 95% confidence intervals. A strong and consistent responsiveness in flow levels when rainfall has been heavy or prolonged is visible, for example during November 2006, with a clear peak in lagged influence that most likely indicates the predominance of fast-moving runoff. At other times

less consistent or interpretable response functions are estimated, for example in January 2006, shown in Figure 4, responses appear very high, suggesting extremely high influence of rainfall up to the most distant lags which is unlikely to be the case as snow is the most likely flow driver at this time. During periods in which freezing temperatures are common rainfall data can be unreliable as snow and ice accumulate in the measuring device until they melt, often much later. We therefore interpret the estimates for January and winter months with caution, and note that they indicate the presence of some effect yet to be accounted for.

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) gradually increase in height, particularly in the 'peak' of influence at approximately a 10 hour lag. Such changes in lag structure are consistent with an increase in ground saturation causing a higher proportion of rainfall to convert to runoff, with flow levels subsequently appearing to be highly sensitive to new rainfall. It is therefore desirable 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, which we now call $W(t)$. 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 of Section 3.2, 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 catchment-specific 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 is intended to account for much of the temporal variation in the β_i observed in section 3.1 during wet periods. In similar notation to Section 2.2 the model is specified as

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

Estimation proceeds as in Section 2.2, where the coefficient vector $\boldsymbol{\theta} = (\alpha, c_{11}, \ldots\ldots, c_{JM})$. The model parameters were $L = 100$, $M = 50$, $J = 100$ again, representing an overfitted model when the penalty vector $\boldsymbol{\lambda} = \mathbf{0}$. It was found when selecting optimal $\boldsymbol{\lambda}$ that AICc often preferred undersmooth estimates for variation in the $W(t)$ dimension, hence it was decided to use the 'optimal' estimate as a lower bound on $\boldsymbol{\lambda}$ and select stronger penalties to maintain simplicity of the model. The intercept term was similar to that in Section 2.2 with an estimate of $\alpha = 12.2$. Interest lies in how the β_i respond to different levels of $W(t)$. The top panels of Figure 5 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 bottom panels of Figure 5, 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 $W(t)$.

[Figure 5 about here.]

4. Discussion

We have proposed flexible and computationally attractive DLMs with roughness penalties that are successful in capturing the dependence between river flow and a sequence of preceding rainfall measurements. In Section 3.2, a complex and time varying relationship between

river flow and rainfall was identified, with Section 3.3 uncovering 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.

From Figure 3 it is clear from the fitted values that the assumption of independent errors is not always justified, with some residual correlation present. Although the simulations in Section 3.1 show that the underlying DL curves could be recovered effectively under strongly correlated errors, associated standard errors are likely to be underestimated. A possible adjustment might involve fitting a model to the residuals, and adjusting the hat matrix $\mathbf{H}$ by the estimated residuals variance matrix as Bowman et al. (2009) did in the context of spatio-temporal modelling. If required, approximate hypothesis tests could then be constructed, as in Bowman et al. (2009), to assess and compare aspects of competing models, in particular to determine whether a time varying model is required over a fixed lag structure.

The issue of biased estimation may arise if the temporal extent and influence of baseflow is not adequately characterised. In the current context an intercept term is all that accounts for the decay in flow rates after rainfall has been absent for L or more hours. More sophisticated approaches might treat L as unknown, or assume a very large L in order to fully incorporate baseflow response into the DL specification. Both approaches may require more structured penalties, for example ismposing stronger penalties at higher lags than shorter ones so that models in which $\beta_i \to 0$ as $i \to \infty$ are preferred; see Muggeo (2008) for an application using health data. A related issue was the tendency to undersmooth that resulted from automatic smoothing parameter selection using AICc, in which case the selected parameter was treated

as a lower bound to maintain a realistic level of wiggliness. While this undersmoothing may result partially from misspecification as described, an important component of further work will be to ensure that automatic smoothness selection achieves believably smooth estimates.

It is likely that bias and autocorrelation are induced primarily by spatio-temporal heterogeneity in the rainfall process that can not be well represented by point-location rainfall data. Bias in lag structures can arise when the underlying weather is dynamic, for example, when rain storms occur near to the rain gauge but are not recorded. It is therefore the intention in future research to represent river flow as *both* a temporal and spatial ensemble of rainfall, using data containing information on the spatial position of rainfall events.

Acknowledgements

We would like to thank two anonymous reviewers and an associate editor for their comments that greatly improved the manuscript. The stream discharge data used in this report were provided by Scottish Environmental Protection Agency (SEPA), while the precipitation 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.

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.

Bowman, A.W., Giannitrapani, M., and Marian Scott, E. (2009). Spatiotemporal smoothing and sulphur dioxide trends over Europe. *Journal of the Royal Statistical Society: Series C (Applied Statistics)* **58**, 737-752.

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: Series B (Statistical Methodology)* **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, C.L. (1998). Smoothing parameter selection in non-parametric regression using an improved Akaike information criterion. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **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. *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) Co. Ltd., London. 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, S.L., 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 for four selected months in 2006. Continuous lines are flow rates ($m^3 s^{-1}$); vertical line segments are hourly rainfall levels (mm).

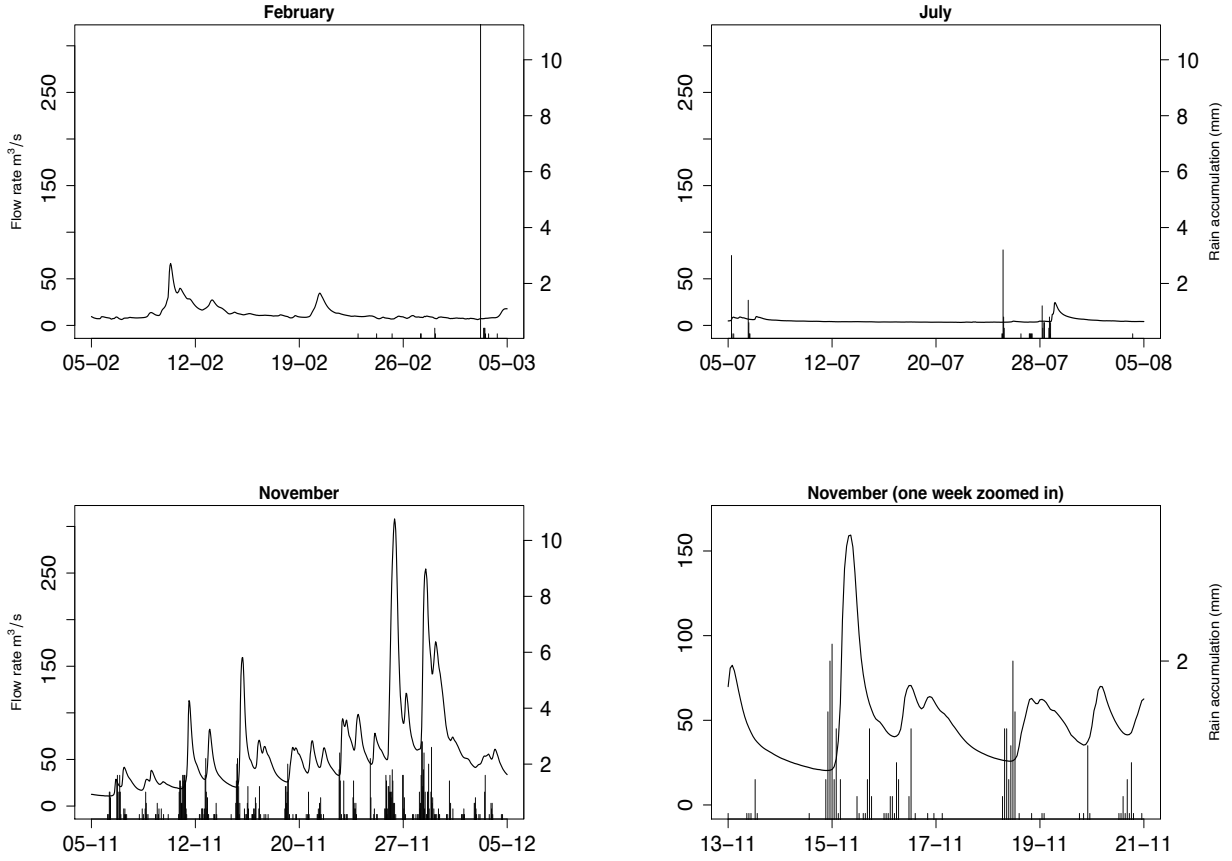


Figure 2. True DL curves with pointwise 95% simulation envelopes plotted in grey. Each row corresponds to simulations under differing DL shapes shown in black, and each column to differing error structures ϵ_1, ϵ_2 and ϵ_3, respectively. The third row shows a mid-July snapshot of the true DL surface and 95% simulation envelopes under the time varying DL scenario. Mean RMSE values are quoted at the top of each panel.

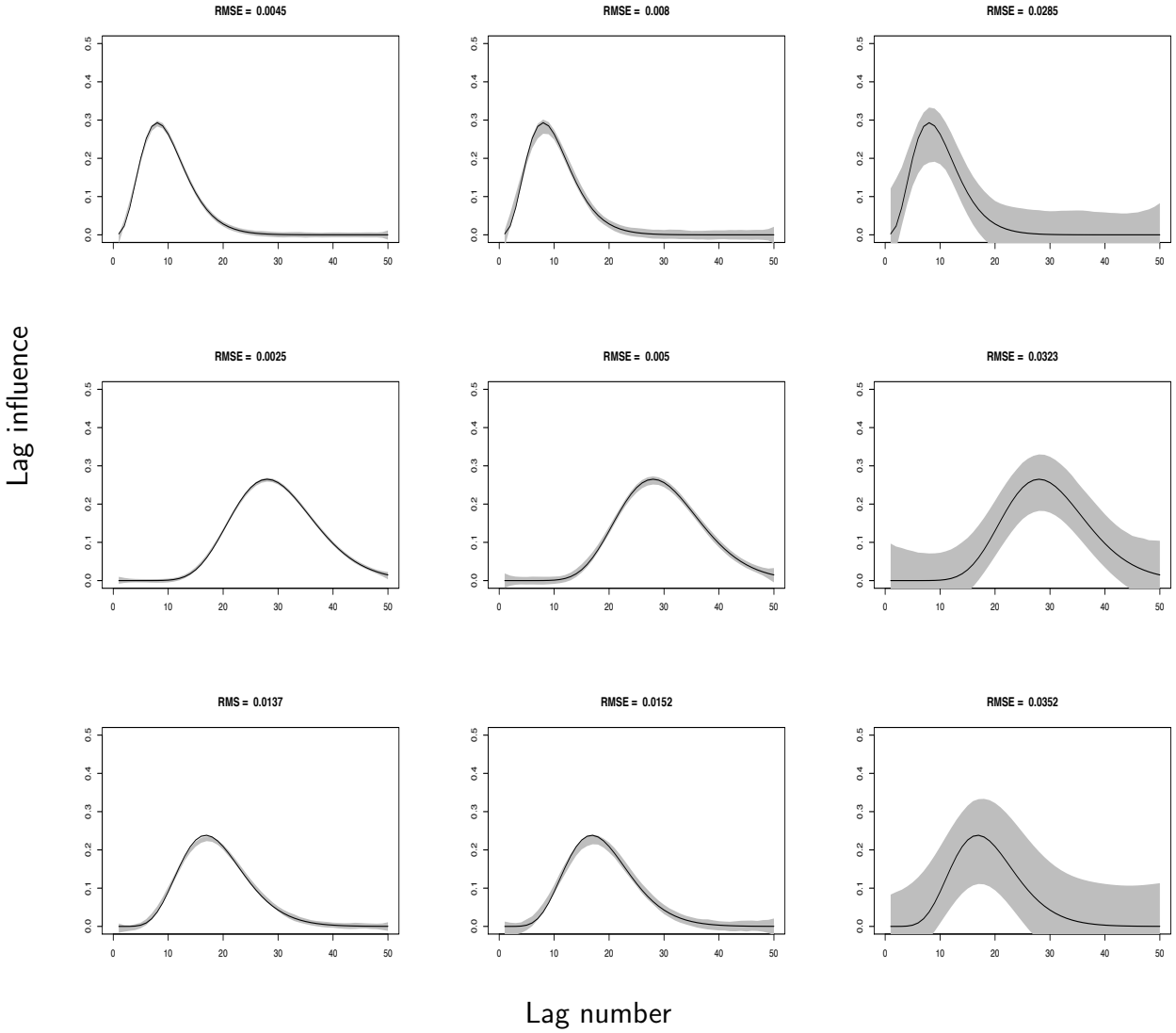
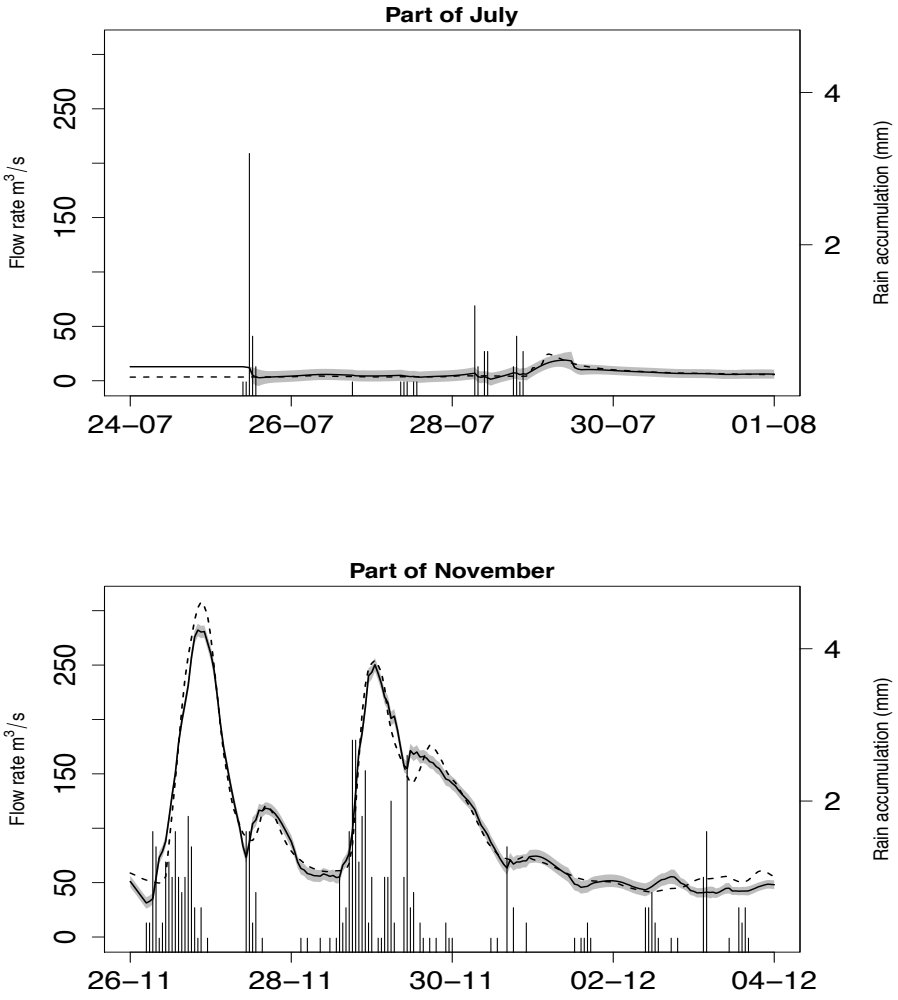


Figure 3. Fitted flows alongside flows observed on the River Dee: continuous lines represent observed flow levels, dashed lines represent fitted flows and vertical line segments are hourly precipitation and grey shaded areas are 95% confidence regions.



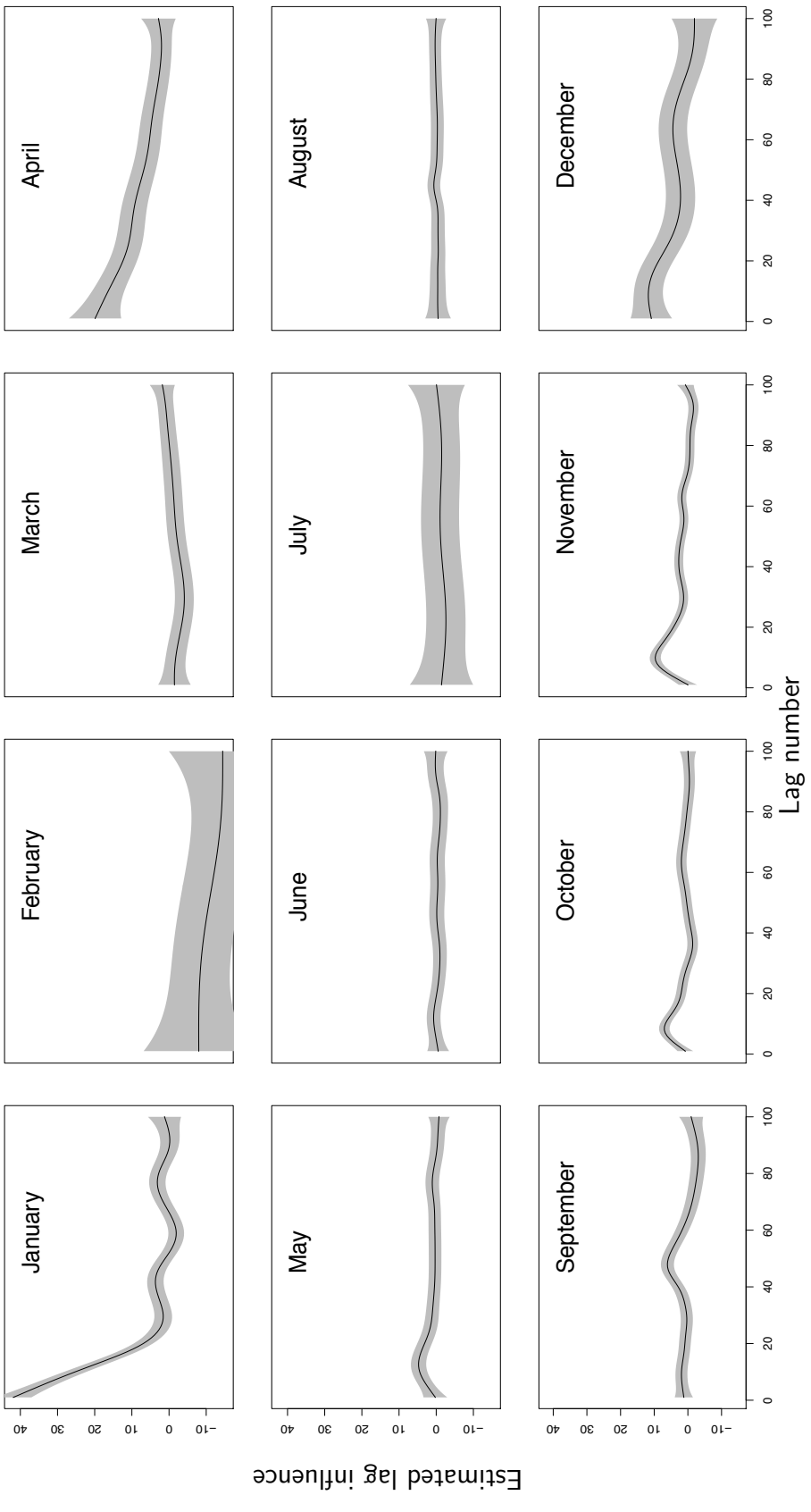


Figure 4. Estimated mean DL curves with pointwise 95% confidence regions estimated from River Dee rainfall and flow data plotted at monthly 'snaphots'; x-axes correspond to the lag numbers (between 1 and 100) of points on the response function

Figure 5. Top: Lag structure at quantiles of $W(t)$, with grey band representing pointwise 95% confidence intervals. Bottom (from left) lag structures estimates smoothly varying with $W(t)$; standard errors in lag structure estimates, smoothly varying with $W(t)$