Distributed lag models for hydrological data

Alastair Rushworth
with Adrian Bowman, Mark Brewer and Simon Langan

26th International Workshop on Statistical Modelling
13th July 2011

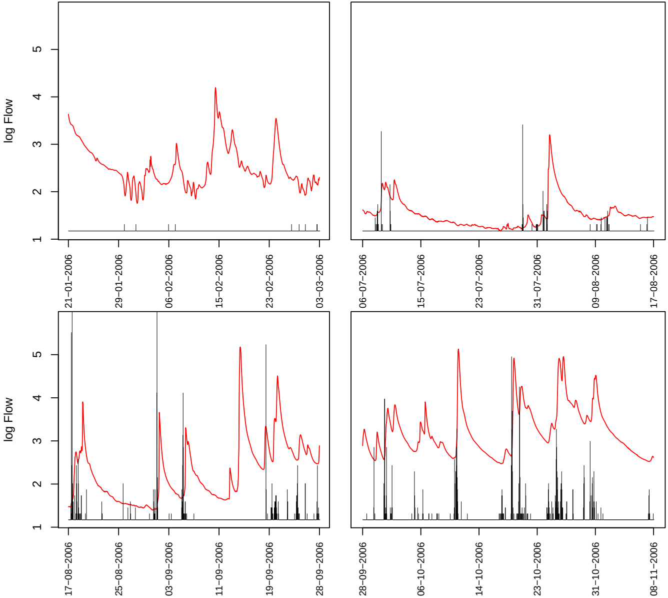
Introduction

Motivated by a need to understand stream flow and its relationship with rainfall

Much hydrological modelling focussed on data and/or assumption rich process based approaches.

What is the best that can be done with only rainfall and stream flow data?

Rainfall scenarios



Rainfall response

$\mathrm{Flow}(t)$ often thought of as the convolution

$$\mathrm{Flow}(t) = \int_a^b \mathrm{Rain}(t-s)h(s)ds$$

sometimes referred to as a transfer function; h often called *impulse response*.

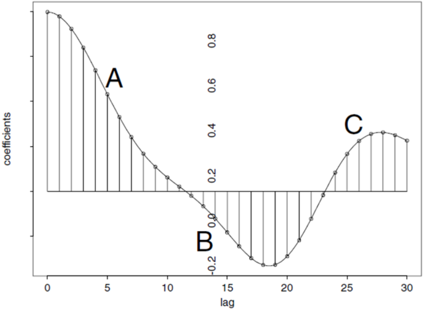
How best to estimate h? Distributed lag model offers an attractive approach

$$\mathrm{Flow}(t) = \beta_0 + \beta_1 \mathrm{Rain}(t-1) + \ldots + \beta_p \mathrm{Rain}(t-p) + \epsilon_t$$

Betas form a coefficient curve, a discrete estimate of $h() \ldots$

Use in air pollution

Same type of problem: impact of recent air pollution exposure on mortality now?



$$E[g(y_t)] = \sum_{i=0}^{30} \beta_i x_{t-i} + \sum_{j}^{p} \gamma_i z_i$$

y_t	mortality on day i
x_{t-i}	pollution on day $t-i$
z_i	other covariates
β_i	$\sum_{l=1}^{k} a_j B_j(i)$

Lag structure corresponding to the mortality displacement effect (taken from Zanobetti et al., 2000).

DLM with splines

Similar to Zanobetti et al., but we would like $\beta_i = \beta_i(t)$

$$\begin{aligned} f_t &= \sum_{i=1}^{p} \beta_i(t) r_{t-i} + \epsilon_t \\ &= \sum_{i=1}^{p} \sum_{j=1}^{k} a_{ij} B_j(t) r_{t-i} + \epsilon_t \end{aligned}$$

In matrix notation, $\mathbf{f} = \mathbf{X}\mathbf{a}$ where

$$\begin{aligned} \mathbf{f} &= (f(1), \ldots, f(n)) \\ \mathbf{X}_{i\bullet} &= (r_{i-1}, \ldots, r_{i-p-1}) \otimes (B_1(t), \ldots, B_k(t)) \\ \mathbf{a} &= (a_{11}, a_{21}, \ldots, a_{p1}, \ldots, a_{1k}, a_{2k}, \ldots, a_{pk}) \\ &= vec(\mathbf{A}), \text{ where } \mathbf{A} \text{ is the matrix of } a_{ij}\text{'s} \end{aligned}$$

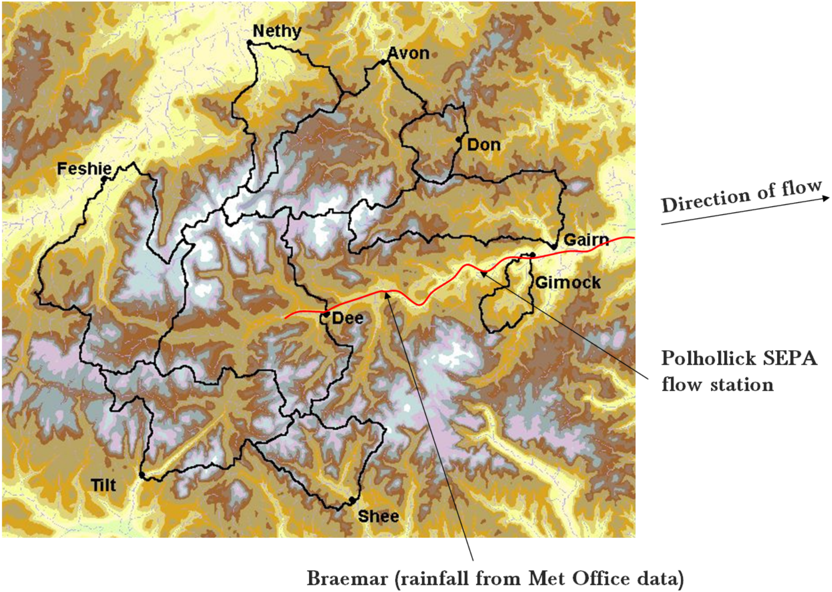
Penalties

Want two penalties: one on the smoothness of each β_i through time; and on the smoothness of the lag structure at any time t. Structure of $\mathbf{X}$ means that penalties are straightforward:

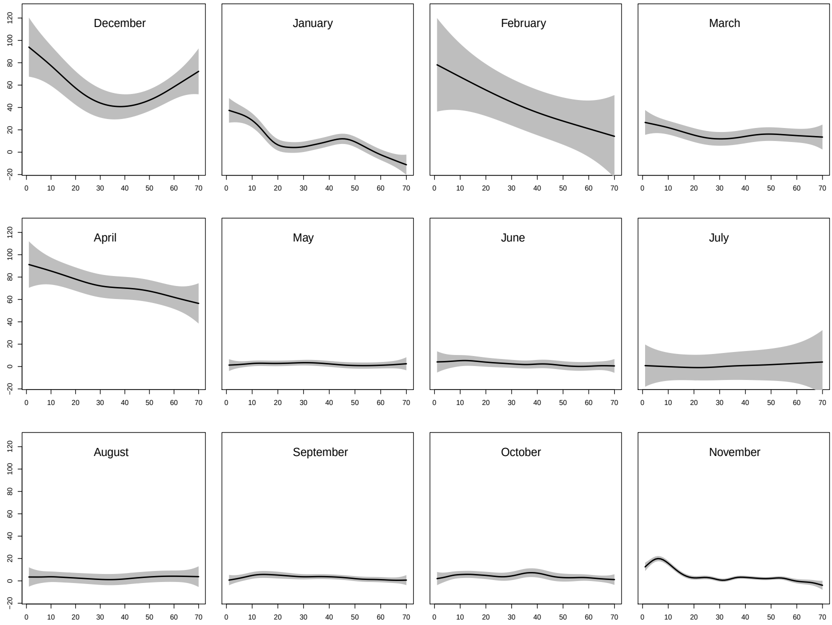
- Time penalty: $\lambda_1^2 \mathbf{D}_1^T \mathbf{D}_1$ where $\mathbf{D}_1 = \mathbf{I}_k \otimes \mathbf{P}_p$
- Lag penalty: $\lambda_2^2 \mathbf{D}_2^T \mathbf{D}_2$ where $\mathbf{D}_2 = \mathbf{P}_k \otimes \mathbf{I}_p$

where $\mathbf{P}_k$ is a quadratic difference matrix with k columns. Parameter estimates are then achieved by

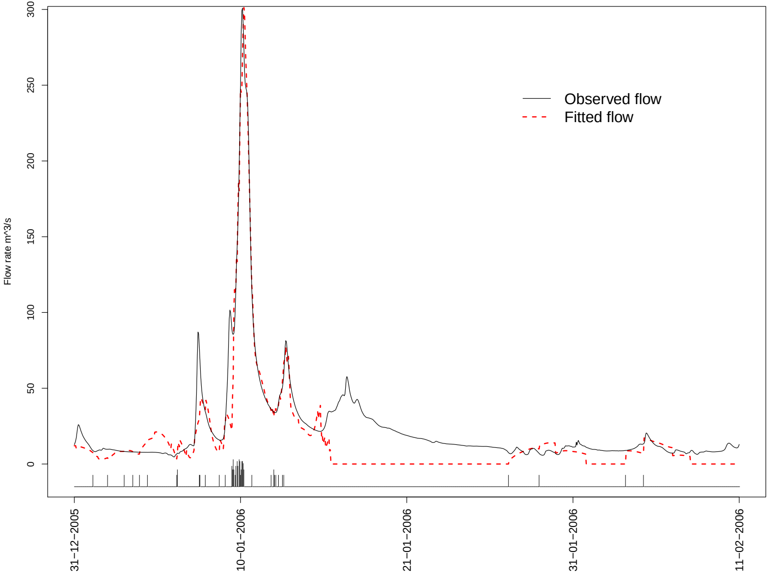
$$\mathbf{a} = \left(\mathbf{X}^T \mathbf{X} + \lambda_1^2 \mathbf{D}_1^T \mathbf{D}_1 + \lambda_2^2 \mathbf{D}_2^T \mathbf{D}_2\right)^{-1} \mathbf{X}^T \mathbf{f}$$


Nethy
Avon
Don
Feshie
Gairn
Girnock
Dee
Tilt
Shee
Direction of flow
Polhollick SEPA flow station
Braemar (rainfall from Met Office data)

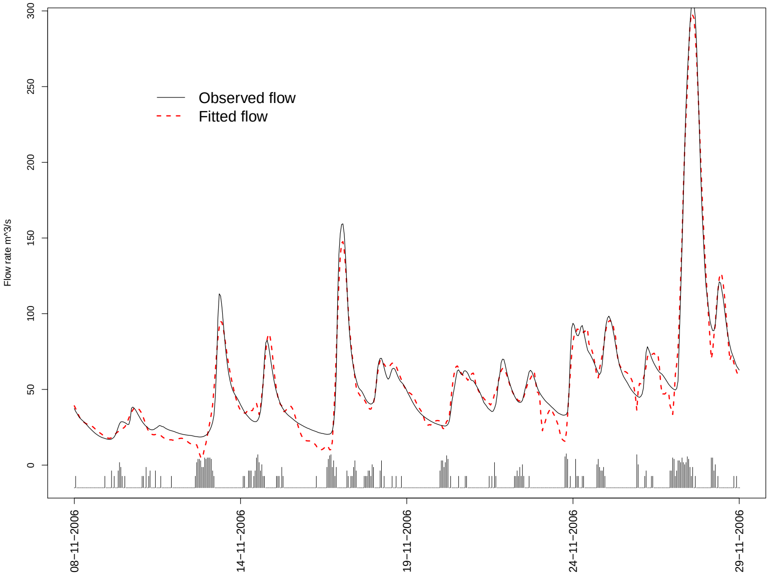
Estimated lag structures for Dee data



Fitted vs observed flows (Jan-Feb 2005)



Fitted vs observed flows (Nov 2006)



Comments

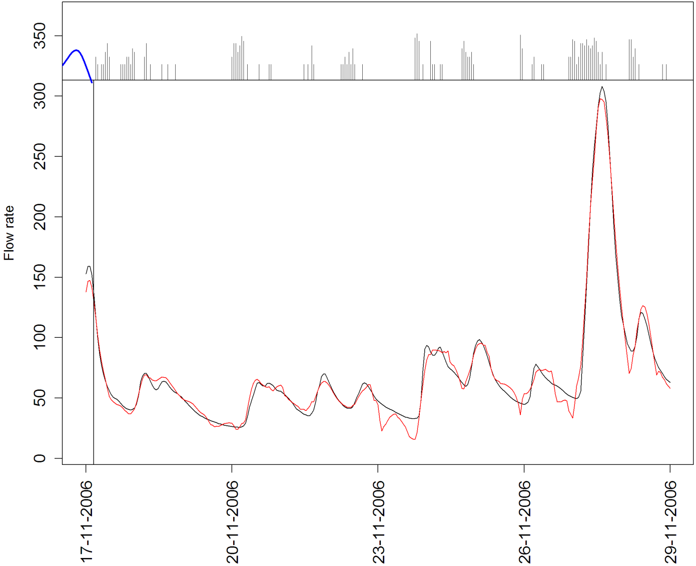
Performance

- poor in cold and/or dry periods
- good during wettest weather.

Features

- Clear differences in lag structure at different times of year
- Changes appear in lag structure appear to be non-linear
- Unlikely looking wiggliness & high estimates
- Interpretation can be unclear

Transfer function in action....

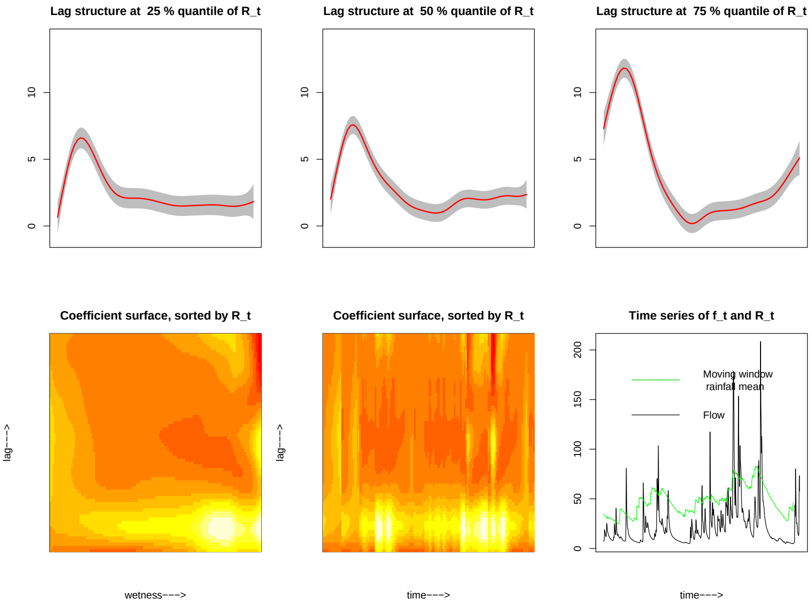


More recent work....

Suppose antecedent rain can be estimated by a 30 day exponential rolling-mean variable R_t. Then we could fit

$$f(t) = \sum_{i=1}^{p} \sum_{j=1}^{L} \sum_{m=1}^{N} c_{jm} B_m(R_t) B_j(i) r(t-i) + \epsilon(t).$$

We can examine the model by looking at the c_{jm}s



Lag structure at 25 % quantile of R_t
Lag structure at 50 % quantile of R_t
Lag structure at 75 % quantile of R_t
0
5
10
Coefficient surface, sorted by R_t
Coefficient surface, sorted by R_t
Time series of f_t and R_t
lag--->
wetness--->
time--->
time--->
Moving window rainfall mean
Flow
0
50
100
150
200

Future work

Refining penalty structure to suit application

- Is it realistic to apply the same penalty across the whole response function?
- Other more drastic possibilities: monotonic ‘tail’; adaptive splines
- ... depends on what we are trying to estimate

Get more rain data: likely explanation for unrealistic wiggliness is unobserved rainfall.

Thanks for listening!

A few references...

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(16), 46334641.

Eilers, P.H.C. and Marx, B.D. (1996). Flexible smoothing with B-splines and penalties. Statistical Science. 89-102.

Zanobetti, A. and Wand, M.P., and Schwartz, J. and Ryan, L.M. (2000). Generalized additive distributed lag models: quantifying mortality displacement. Biostatistics. 1(3), 279.