Big BAM: Detecting non-stationarity in peak river flows

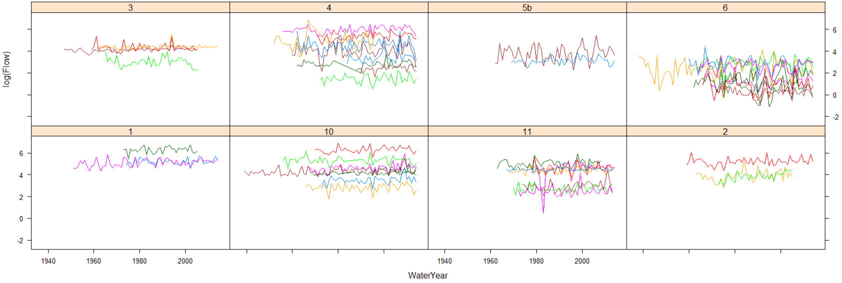
Emiko Dupont, Robbie Peck, Tom Smith,
Ilaria Prosdocimi, Nicole Augustin

ITT7

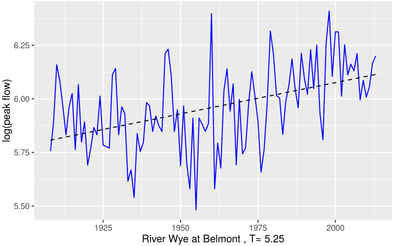
31 January 2017

Data

- 29,687 obs. of 42 variables.
- Time series at 675 stations.



Linear model at each station



Model:

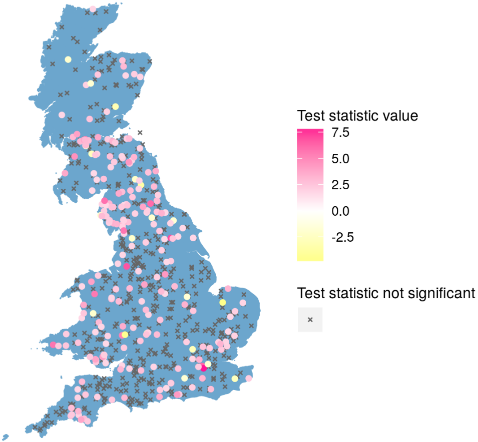
$$y_i = \log(\text{flow}_i) = \beta_0 + \beta_1 t_i + \varepsilon_i \quad \text{where} \quad \varepsilon_i \sim N(0, \sigma^2) \text{ iid}$$

Hypothesis test: $H_0 : \beta_1 = 0$ against $H_1 : \beta_1 \neq 0$

$$T = \frac{\hat{\beta}_1}{\hat{\sigma}_{\hat{\beta}_1}} \sim_{\text{approx}} N(0,1) \text{ under } H_0$$

Reject H_0 if $|T| > 1.64$ (time dependence is significant)

Test statistic for $\alpha = 0.1$



Time dependence significant at around 20% of stations

What can we do to improve this?

Improvement 1: Build a spatial model for the test statistic

- Pooling data improves **power** of statistical test

Approach 1: Mixed effects model based on hydrometric areas:
For station i in area j

$$T_i = \mu + b_j + \varepsilon_i \quad \text{where} \quad b_j \sim N(0, \sigma_b^2), \quad \varepsilon_i \sim N(0, 1)$$

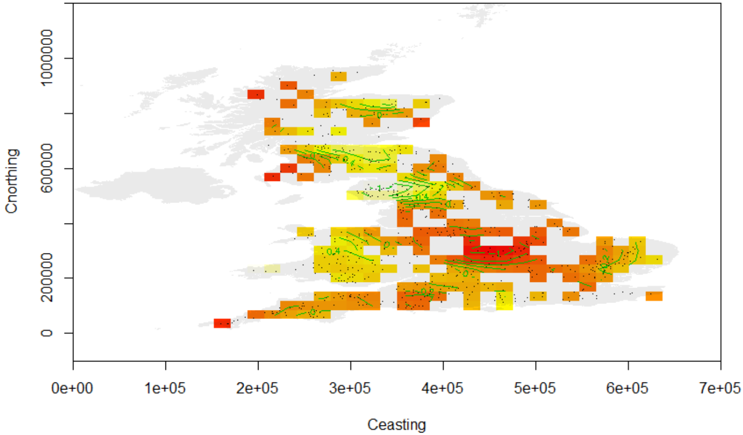
Results:

- Reproduced results from Emikos project.
- 80% of tests now significant at level 0.1.

Approach 2: GAM with spatial random effect

$$T_i = \mu + f(\text{c-east}_i, \text{c-north}_i) + \varepsilon_i \quad \text{where} \quad \varepsilon_i \sim N(0, \sigma^2)$$

- Use of centroids of catchment as locations.



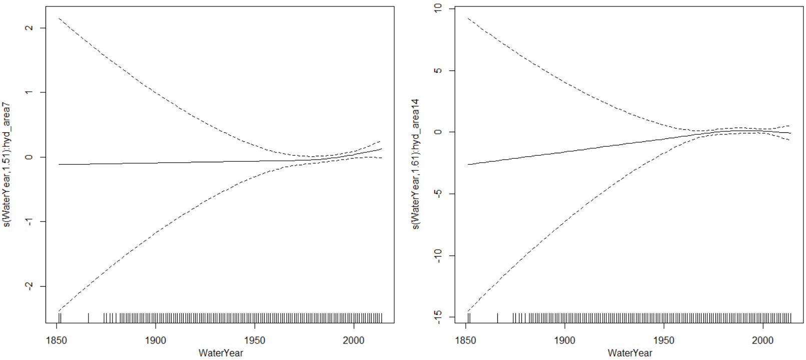
Improvement 2:
Approach 1:
Investigate shape of time dependence (GAMM)
For station i in hydrometric area j at time t_k

$$\log(\text{flow}_{ik}) = f_j(t_k) + b_{0,i} + b_{1,i} t_k + \varepsilon_{ik}$$

$\text{b}_{0,i} \sim N(0, \sigma^2_{b_0}), b_{1,i} \sim N(0, \sigma^2_{b_1}), \varepsilon_i \sim N(0, \sigma^2)$

- Standardise the flow
- f_j for each hydrometric area.
- 

Plots of two f_j's



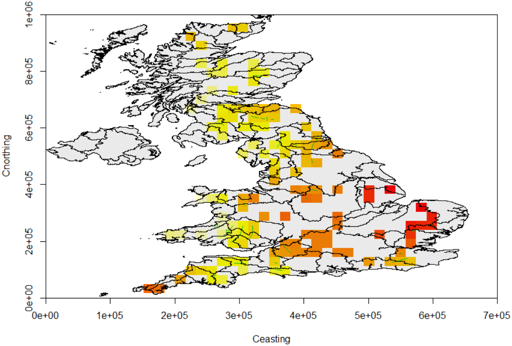
Approach 2:

Varying coefficient model based on centralised spatial coordinates. For station i, hydrometric area j and water year k,

$$\log(\text{Standardised Flow}_{ik}) = f((\text{c-north}, \text{c-east})_i)t_k + b_j + \epsilon_{ik}$$

, where $b_j\ N(0, \sigma_b^2)$.

- Reflects that stations nearer are more similar.
- Shows spatial pattern of time dependence.



Future Work

- Can we obtain "estimated time dependence" from test statistic results?
- From the varying coefficient model, how do we assess whether slopes are significant?
- Improve measure of similarity of stations.
- Model AMAX using GEV distribution.