Assessing and comparing the performance of models for stream network data

Alastair Rushworth
with Adrian Bowman, Erin Peterson and Jay Ver Hoef

Thursday 13th June 2013

University of Glasgow

Some important features of river network data

- Confluences and 'connectedness'
- Data and networks can be *large*
- Measurements often very *sparse* in space

Methods and software available

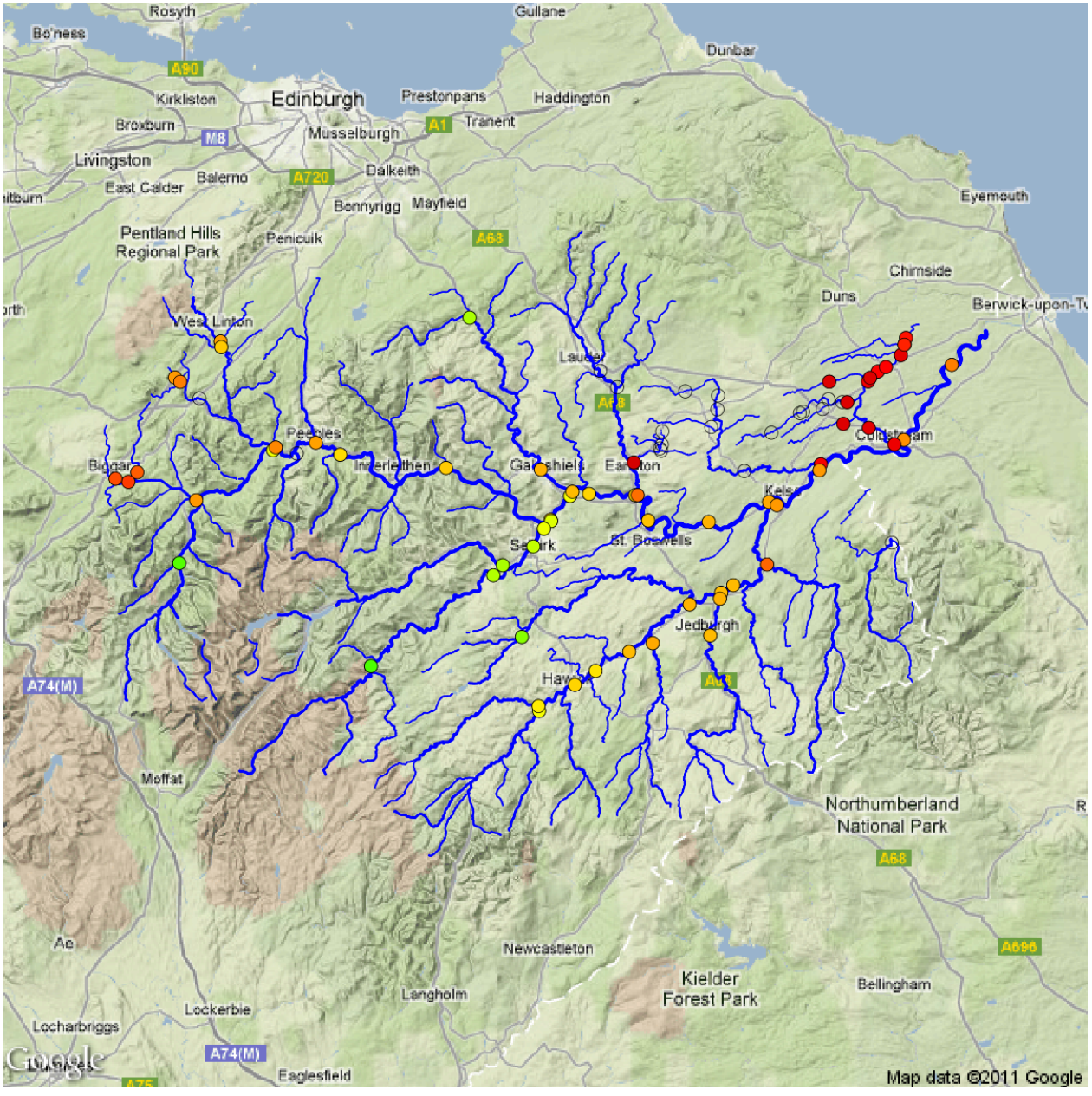
SSN

- Based on spatial covariance functions
- Ver Hoef et al. (JASA; 2010)
- R package: http://cran.r-project.org/web/packages/SSN/

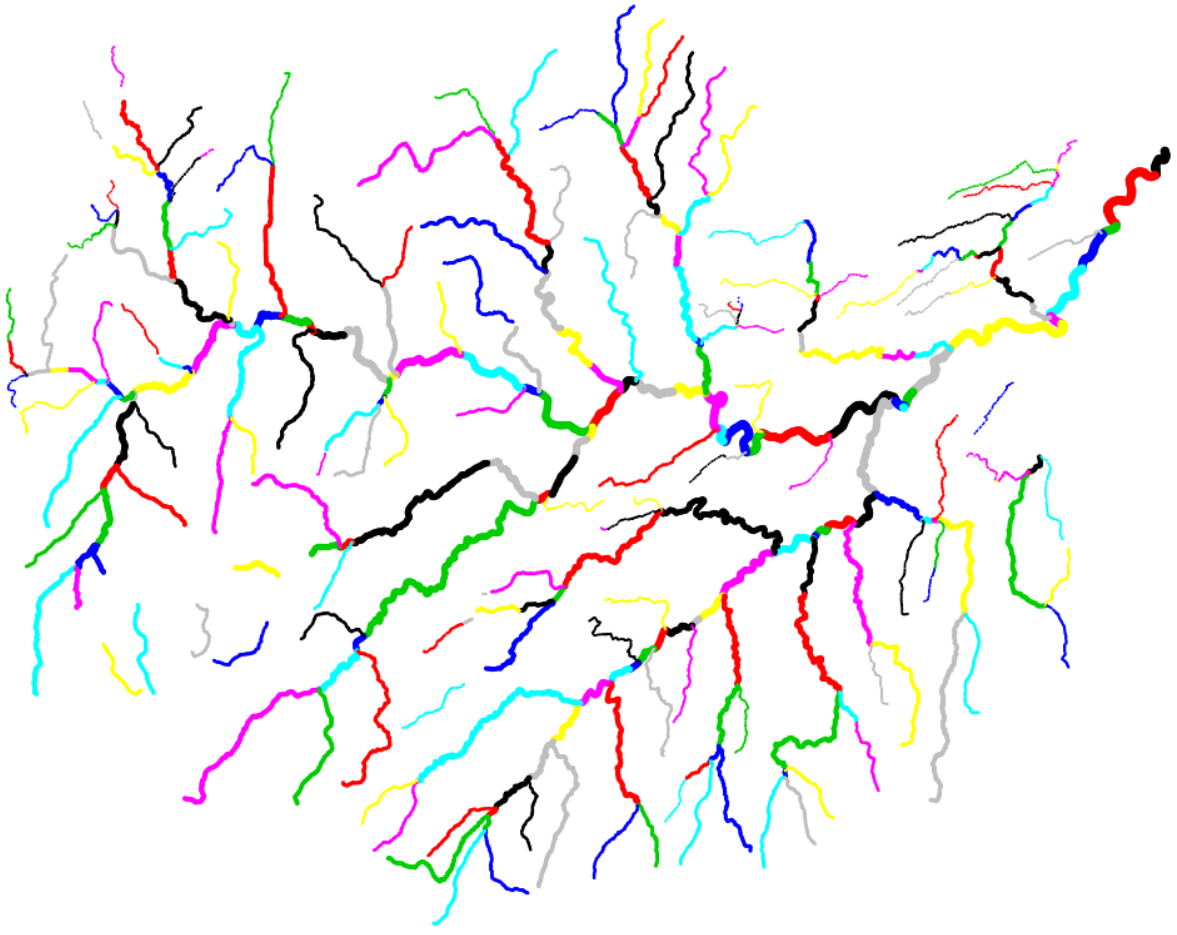
SmoothNetwork

- Based on P-splines and penalised least squares
- O' Donnell et al. (JRSSC (*in press.*; 2013))
- R package: http://alastairrushworth.wordpress.com

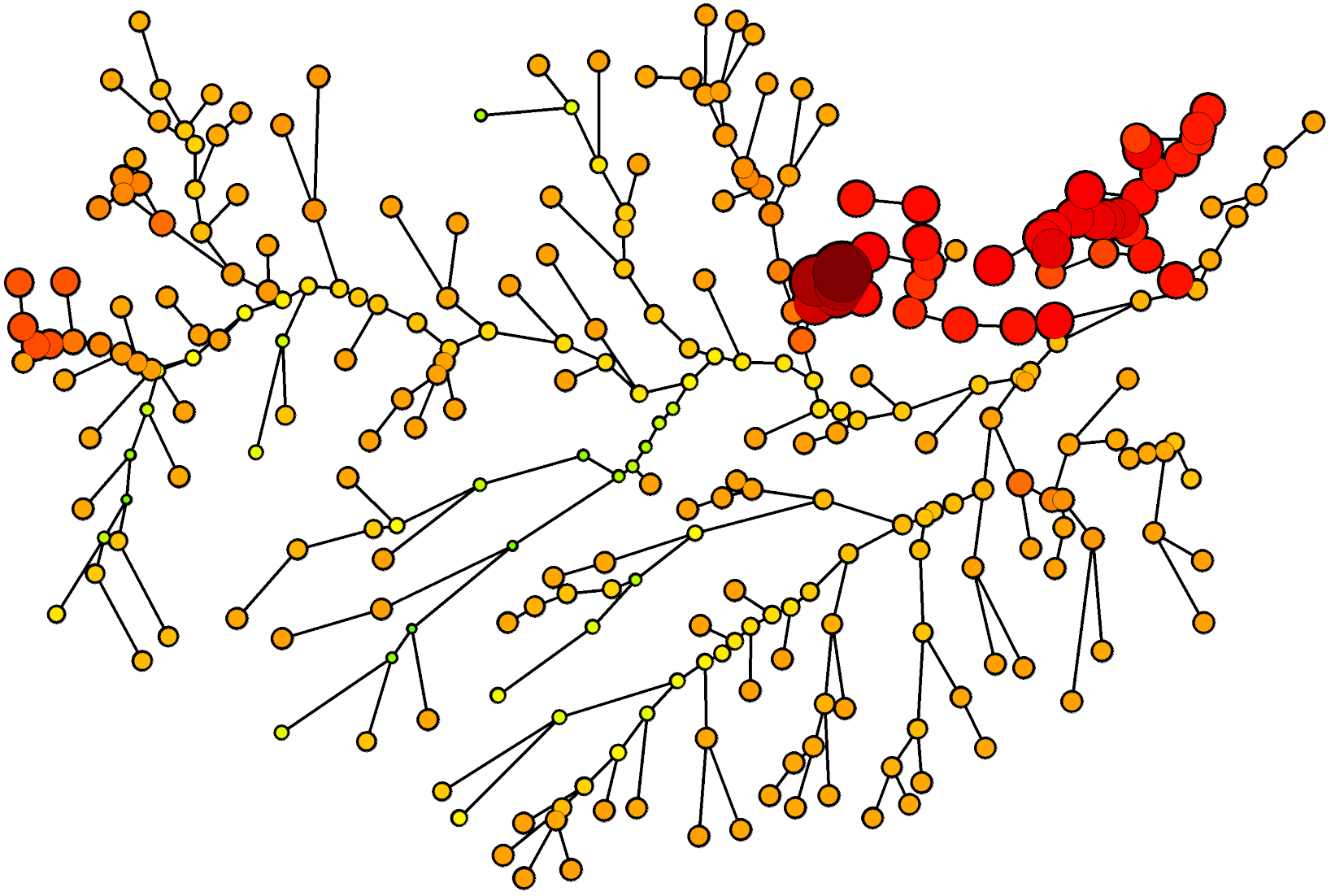
Example: River Tweed



Example: River Tweed



Example: River Tweed



A simple model for segments

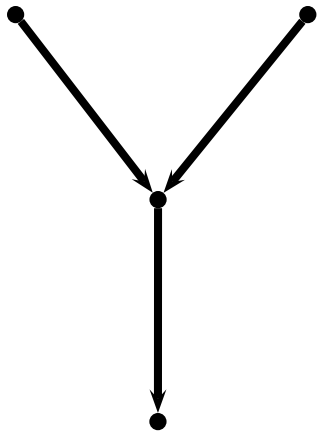
Suppose we have set of n responses each observed on one of a set of p nodes of a graph, and $\mathbf{x}$ is a covariate vector. Then

$$E(\mathbf{y}) = \underbrace{\boldsymbol{\alpha}}_{\text{intercept}} + \underbrace{\mathbf{Mb}}_{\text{Dummy variable with } p \text{ levels}}$$

where

$$\mathbf{M} = \begin{pmatrix} 0 & 0 & 1 & 0 & \dots \\ 0 & 0 & 0 & 1 & \dots \\ 0 & 1 & 0 & 0 & \dots \\ 0 & 0 & 0 & 0 & \dots \\ \vdots & \vdots & \vdots & \vdots & \vdots \end{pmatrix}$$

Confluence penalty



An idealised confluence with stream units 1, 2 **and** 3, **associated mean pollutant levels** b_1, b_2 **and** b_3.

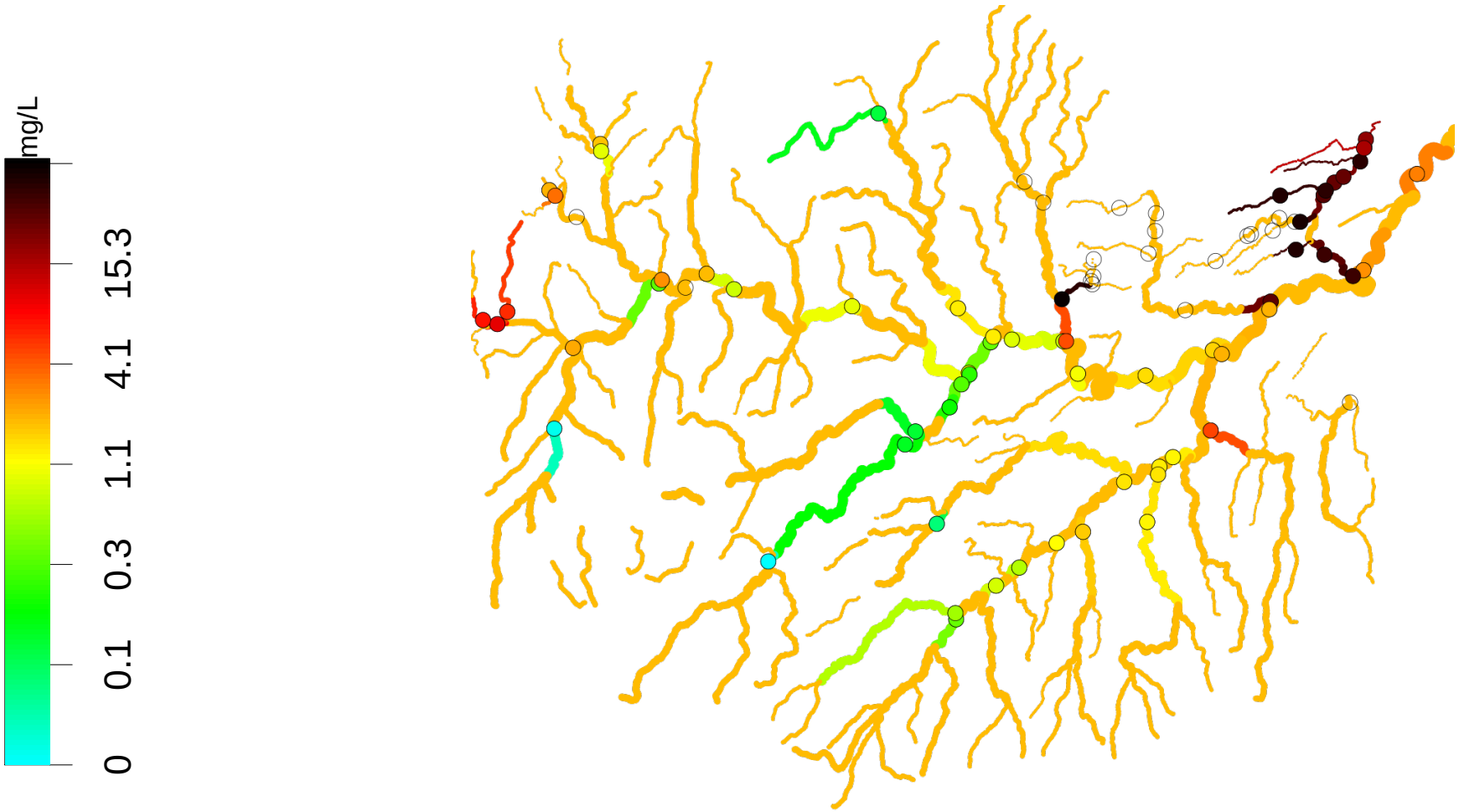
Values upstream
should be similar to those directly downstream
subject to their respective flow contributions:

$$\frac{f_1}{f_3} b_1 + \frac{f_2}{f_3} b_2 \approx b_3$$

$$\min_{\mathbf{b}} \left[(\mathbf{y} - \mathbf{M}\mathbf{b})^T (\mathbf{y} - \mathbf{M}\mathbf{b}) + \sum_{i,j \sim k} \lambda \left(\frac{f_i^2}{f_k^2} (b_i - b_k)^2 + \frac{f_j^2}{f_k^2} (b_j - b_k)^2 \right) \right]$$

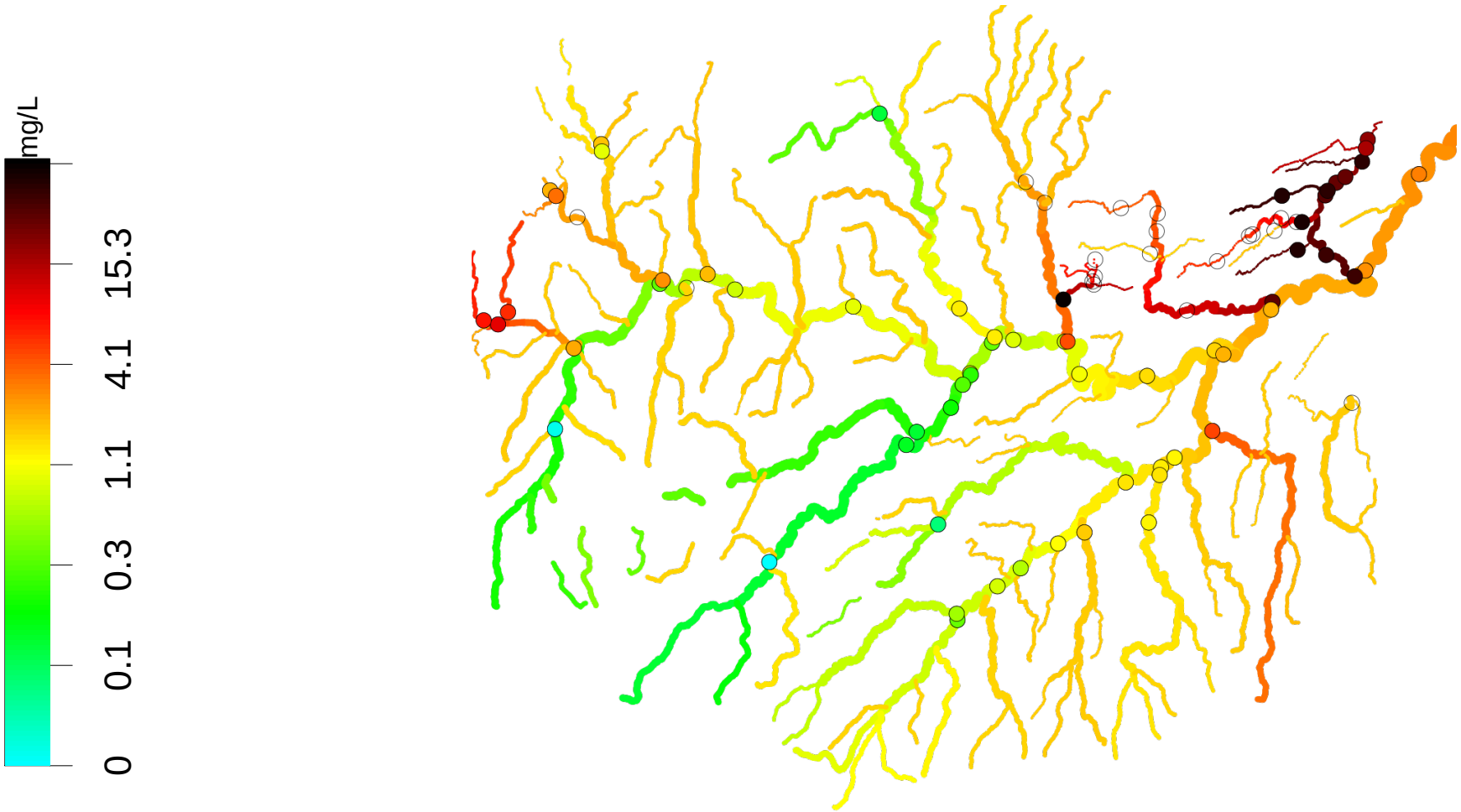
So what happens if you vary λ ?

Effect of varying the confluence penalty; $\lambda = 0$



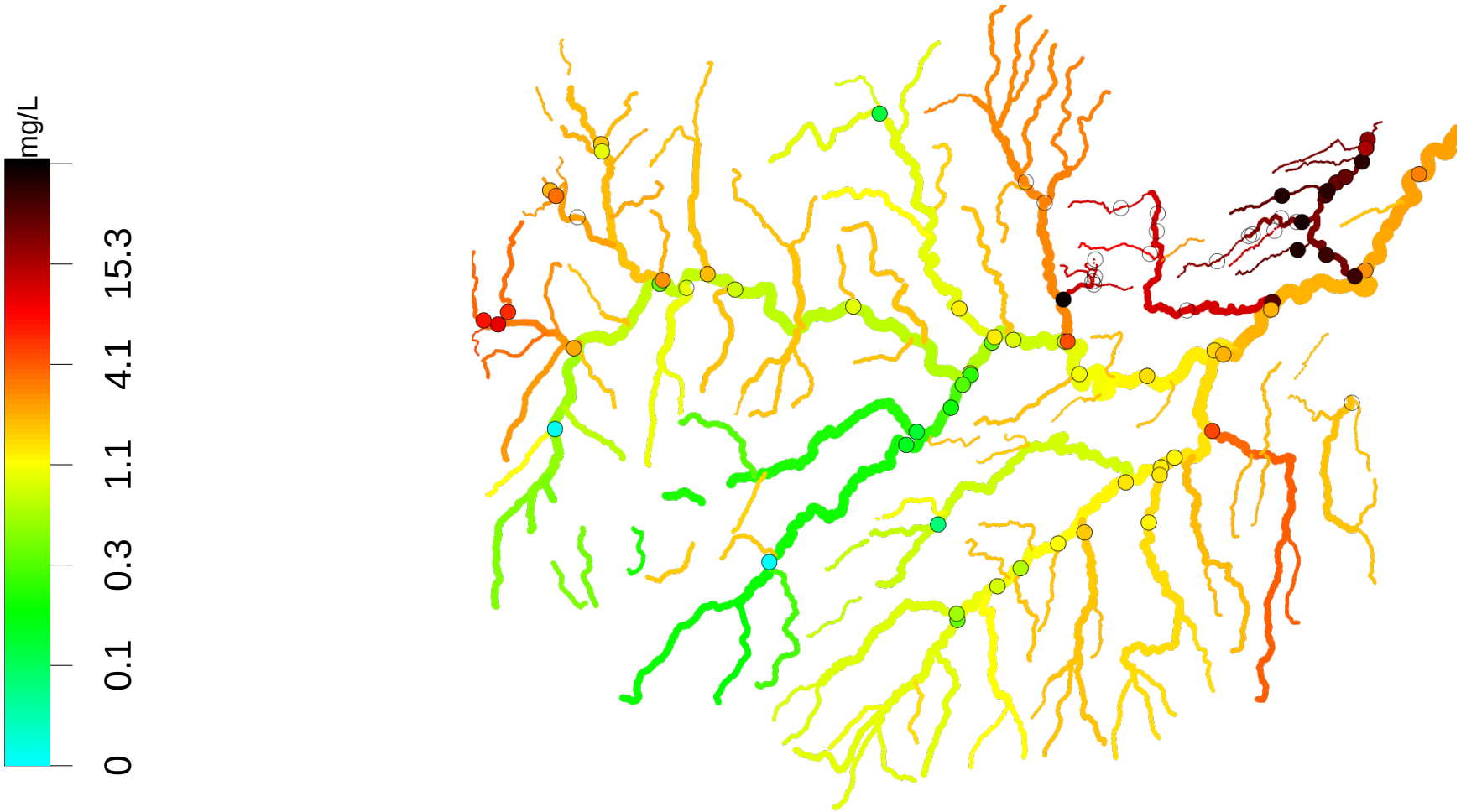
A weaker penalty allows the spatial smooth to interpolate the observed data

Effect of varying the confluence penalty; $\lambda = +$



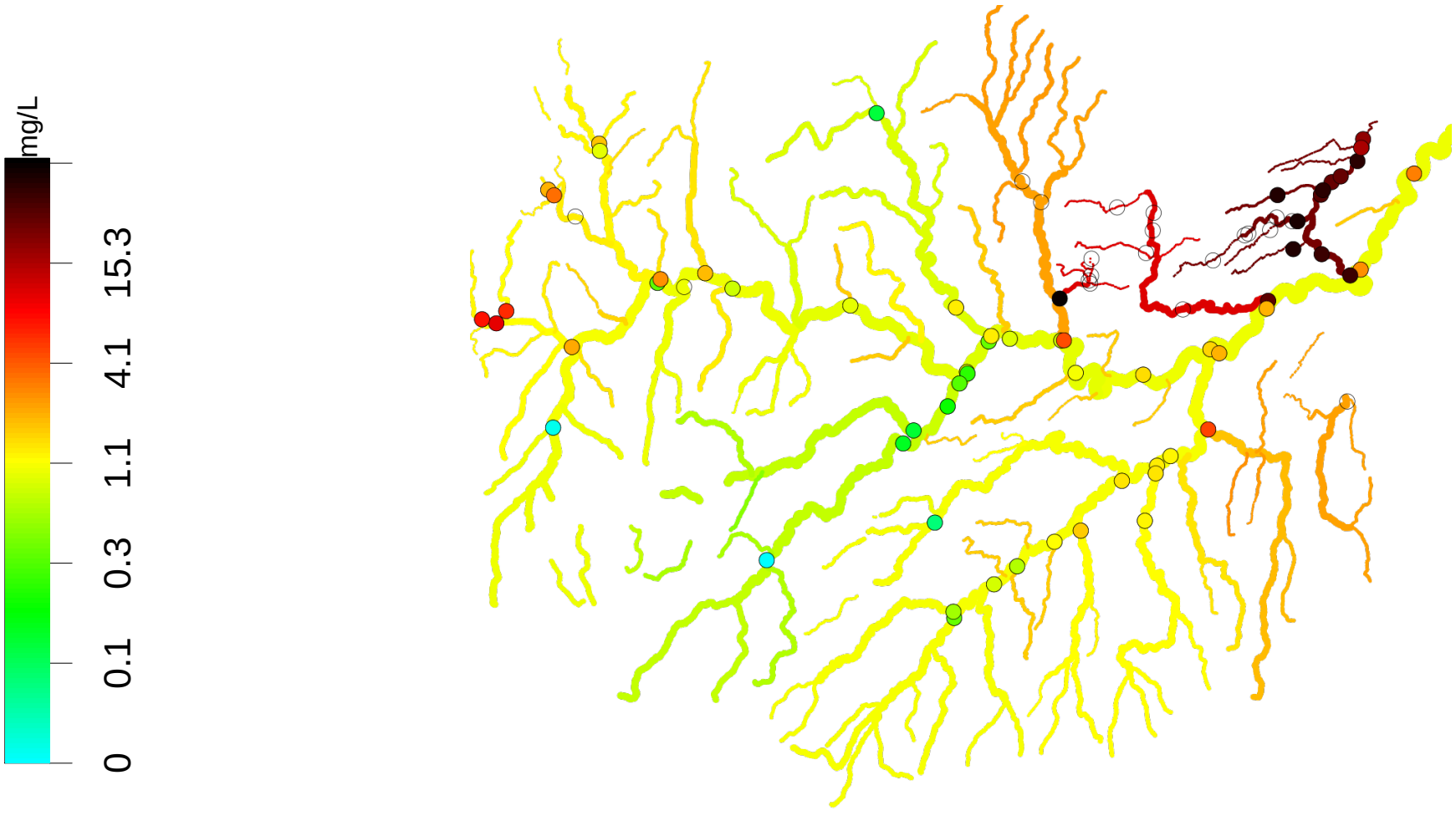
A weaker penalty allows the spatial smooth to interpolate the observed data

Effect of varying the confluence penalty; $\lambda = ++$



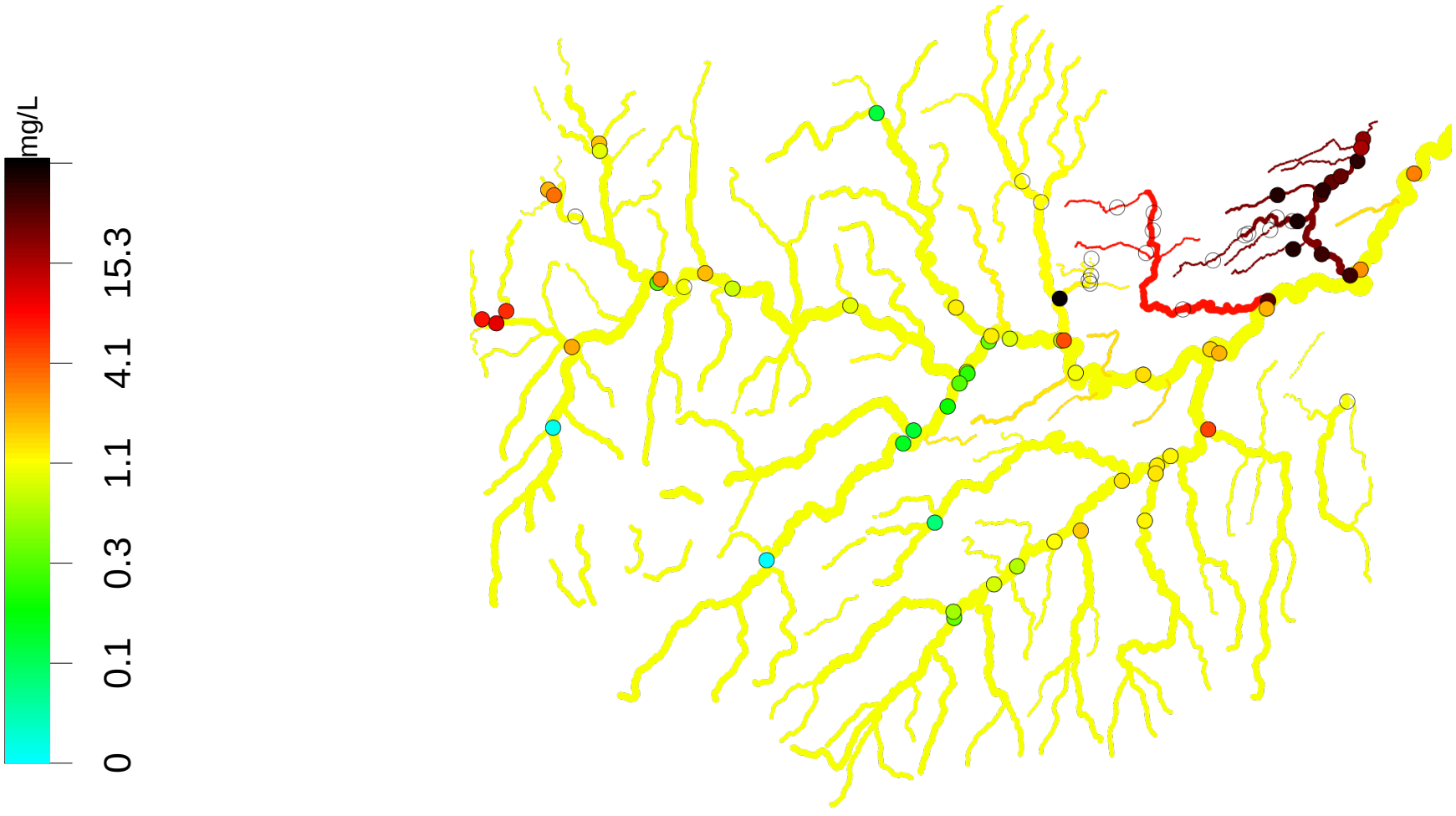
A weaker penalty allows the spatial smooth to interpolate the observed data

Effect of varying the confluence penalty; $\lambda = +++$



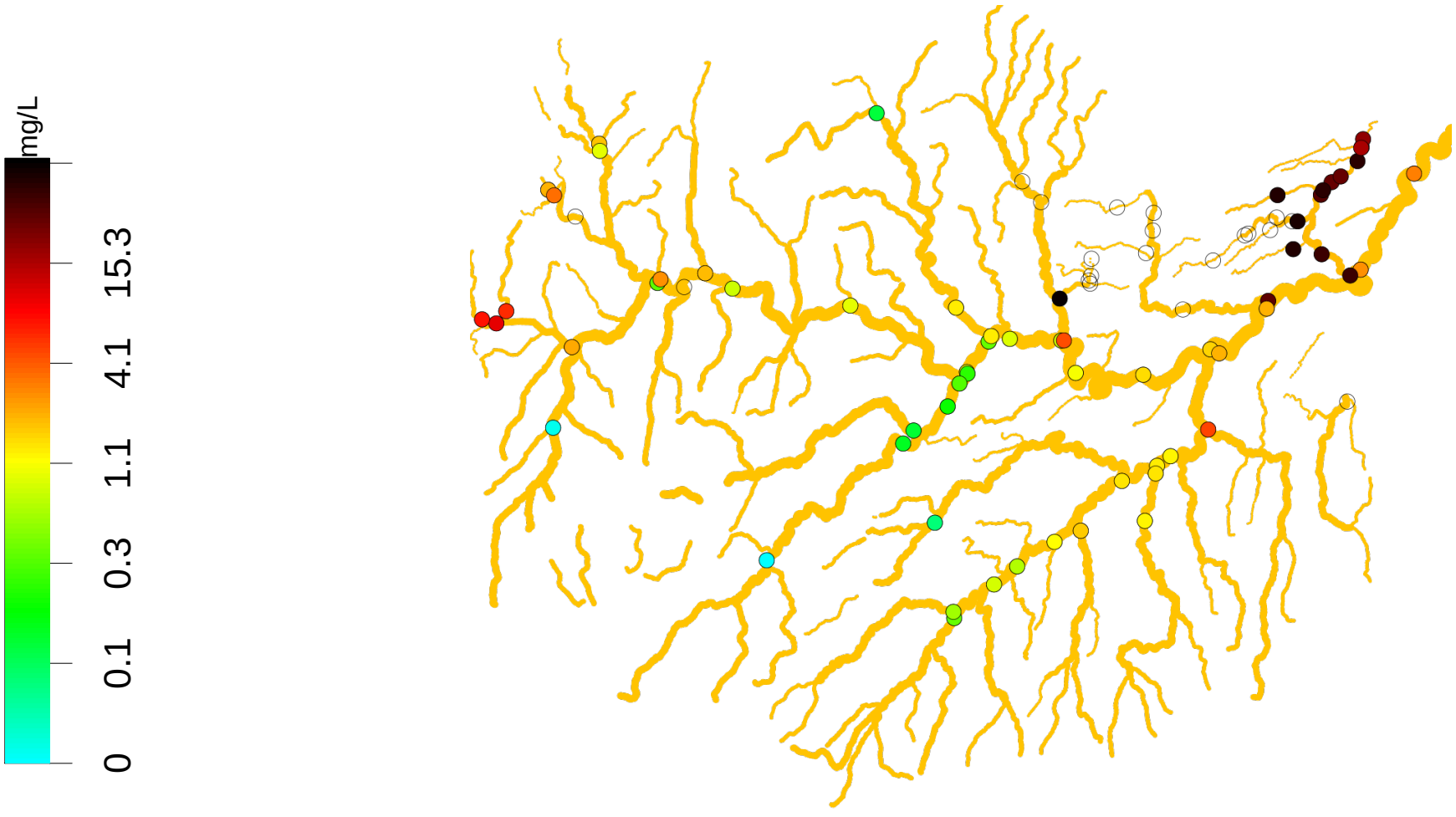
A stronger penalty tends to bring stream unit means nearer the *overall* mean

Effect of varying the confluence penalty; $\lambda = ++++$



A stronger penalty tends to bring stream unit means nearer the *overall* mean

Effect of varying the confluence penalty; $\lambda \to \infty$



A stronger penalty tends to bring stream unit means nearer the *overall* mean

Models based on P-splines: summary

- Piecewise constants for small stream units
- Estimation by PLS; very fast because of matrix sparsity
- Model easily to extend with additional smooth components

River network modelling with spatial covariance functions

Given data $\mathbf{Y} = (Y_1, \ldots, Y_n)$, covariates $\mathbf{X} = (\mathbf{X}_1, \ldots, \mathbf{X}_p)$ let

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{z} + \boldsymbol{\epsilon}$$

where $\mathbf{z} \sim N(\mathbf{0}, \boldsymbol{\Sigma})$ and $\boldsymbol{\epsilon} \sim N(\mathbf{0}, \sigma^2\mathbf{I})$.

Cressie et al. (2006) and Ver Hoef et al. (2006) describe ways that are valid for stream networks to populate $\boldsymbol{\Sigma}$ such that $\boldsymbol{\Sigma} = \boldsymbol{\Sigma}(h|\boldsymbol{\theta})$ and where

- h is the stream distance separating two observations
- $\boldsymbol{\theta}$ is a vector of additional parameters

River network modelling with spatial covariance functions

Example: The covariance function for the 'tail up' exponential model is defined as

$$C_u(r_i, s_j|\boldsymbol{\theta}) = \begin{cases} \pi_{i,j} C_t(h|\boldsymbol{\theta}) & \text{if } r_i < s_j \text{ are flow-connected} \\ 0 & \text{if } r_i \text{ and } s_j \text{ are flow-unconnected} \end{cases}$$

where

$$C_t(h|\boldsymbol{\theta}) = \theta_v \exp(-h/\theta_r)$$

and $\pi_{i,j}$ is a weight function to account for the impact of confluences intervening two points.

River network modelling with spatial covariance functions

Key features, compared

- Very flexible: broad class of spatial structures can be captured with choice of covariance functions
- Non-Gaussian responses
- Estimation by REML; spatial effects are variance components, straightforward to add extra pieces - eg. Euclidean components.

Comparison

- Aim to answer the question

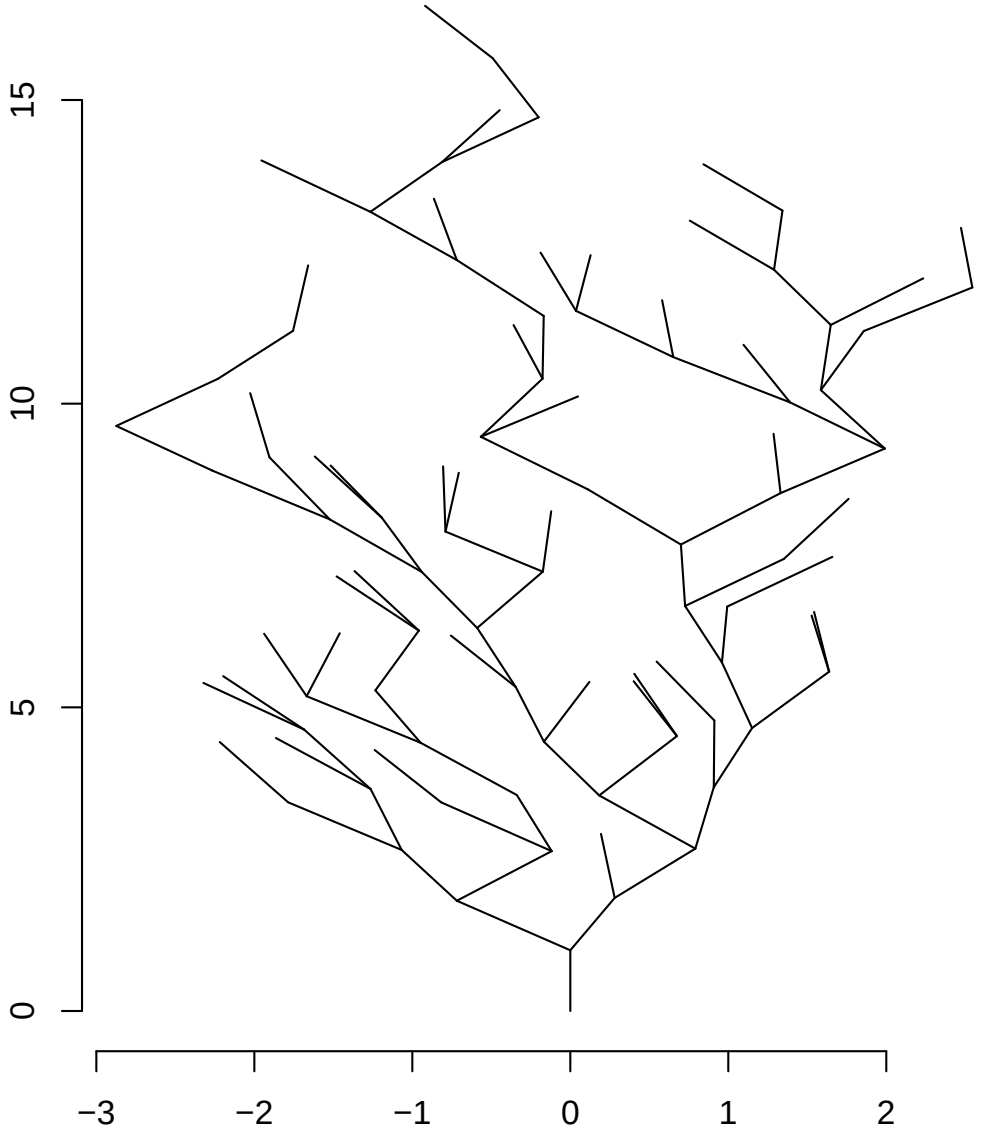
  which technique is most appropriate and when?

- Performance measured by
  - Predictions: RMSPE, Prediction interval coverage, bias...
  - Fixed effects: RMSE, Confidence region coverage, bias...
  - Time to fit model

- Need some mechanism for generating data...

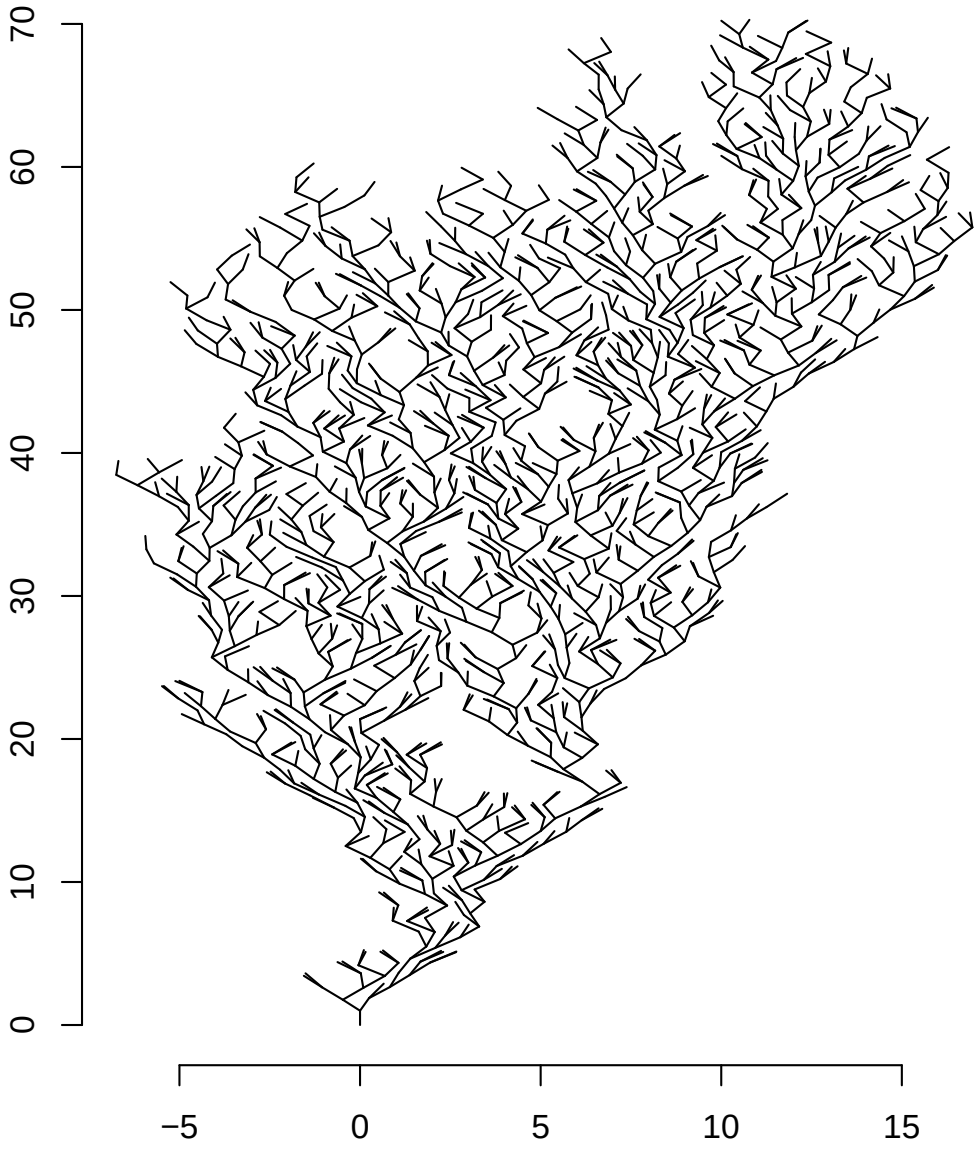
Comparison study - generating data

From the R package **SSN**, `createSSN()`, 100 stream segments



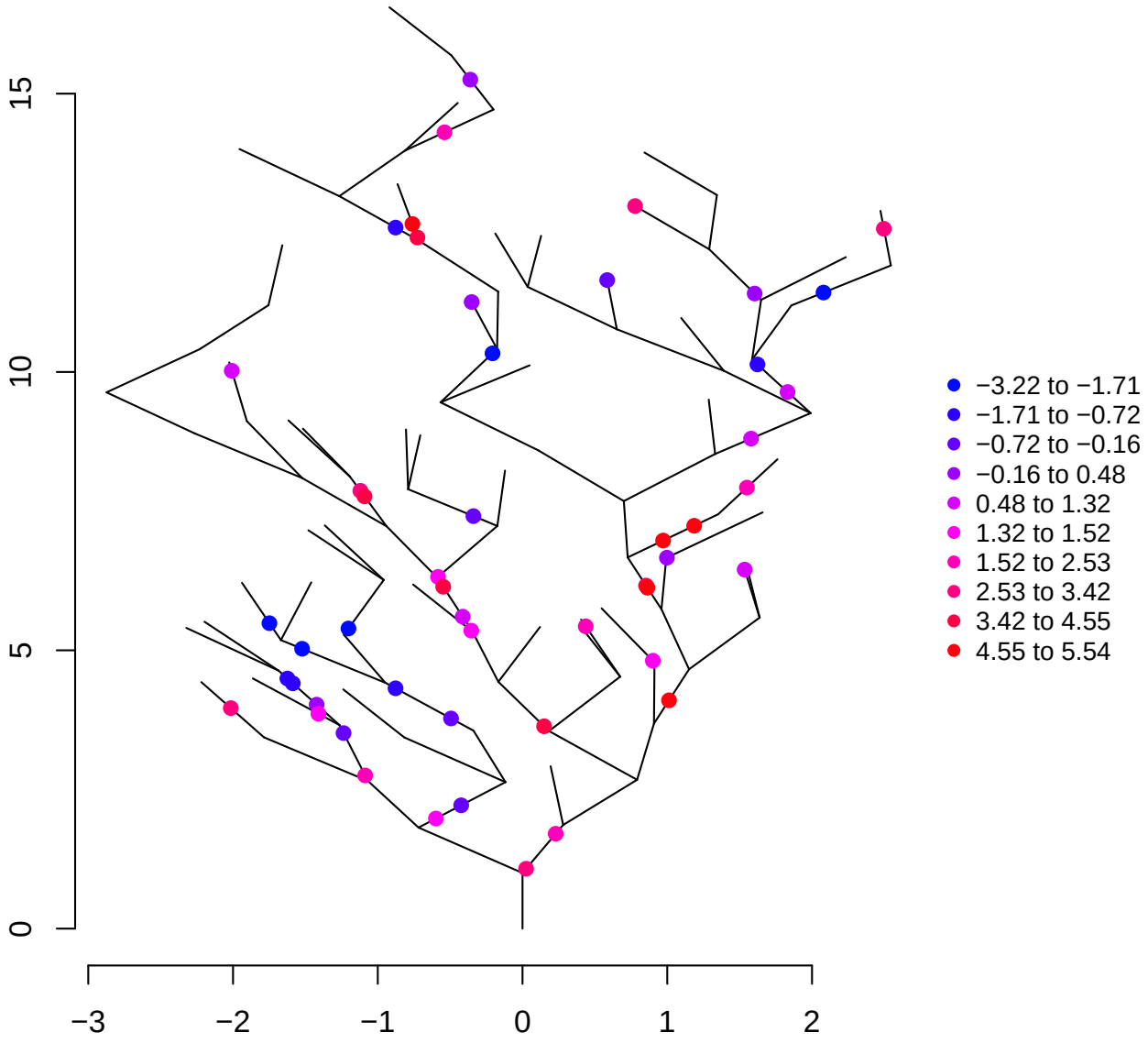
Comparison study - generating data

From the R package **SSN**, `createSSN()`, 2000 stream segments



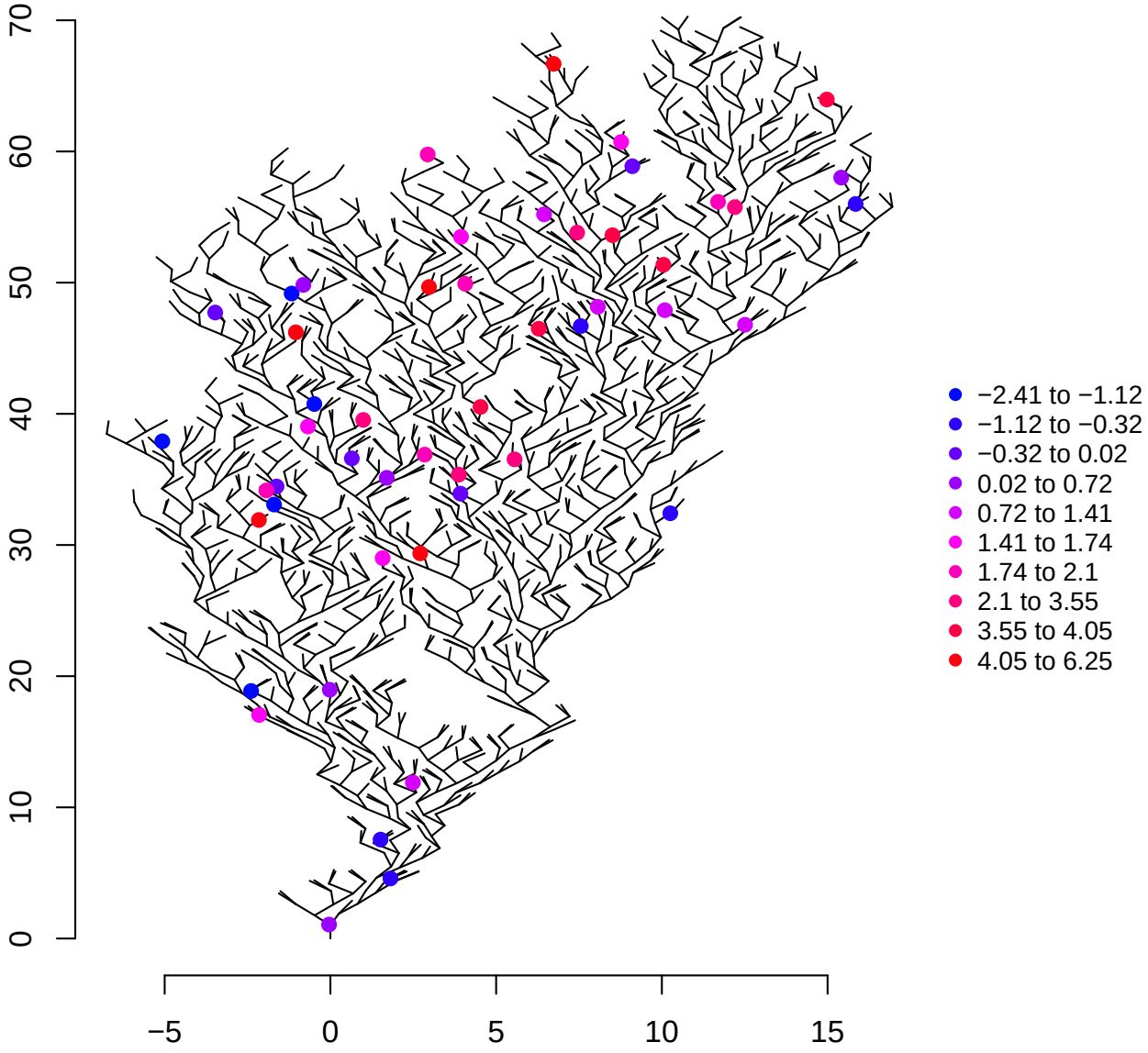
Comparison study- generating data

From the R package **SSN**, `SimulateOnSSN()`, 100 stream segments



Comparison study- generating data

From the R package **SSN**, `SimulateOnSSN()`, 2000 stream segments



Comparison study - design

Sparsity and abundance of data

- Network size: small (100 stream segments) vs. large (2000 stream segments)
- 50 data points vs. 500

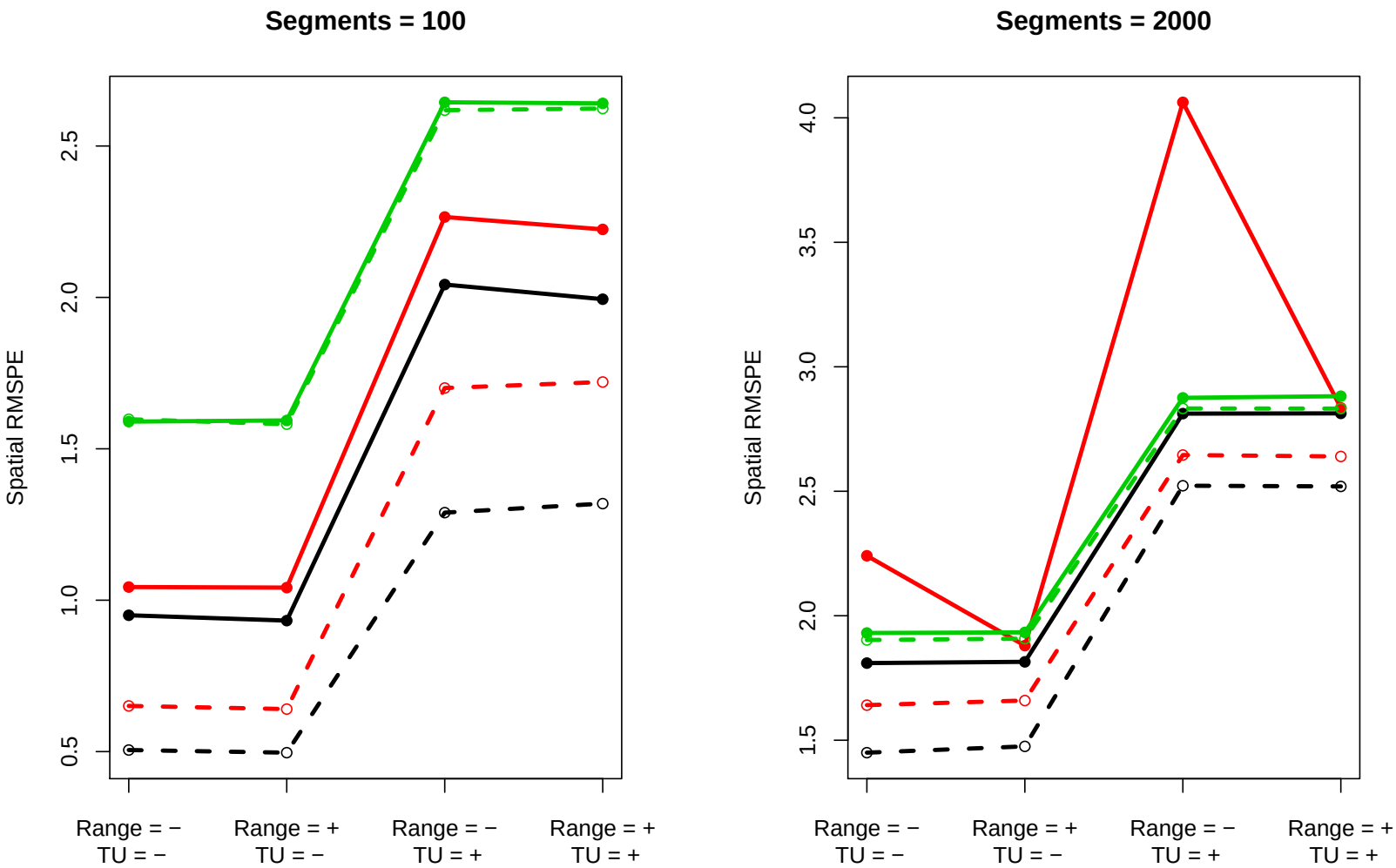
Data spatial structure

- Large or small range parameter
- Different levels of spatial strength (relative to fixed effects)

Fixed effects

- One strong and observed, one strong and unobserved and one weak and observed.

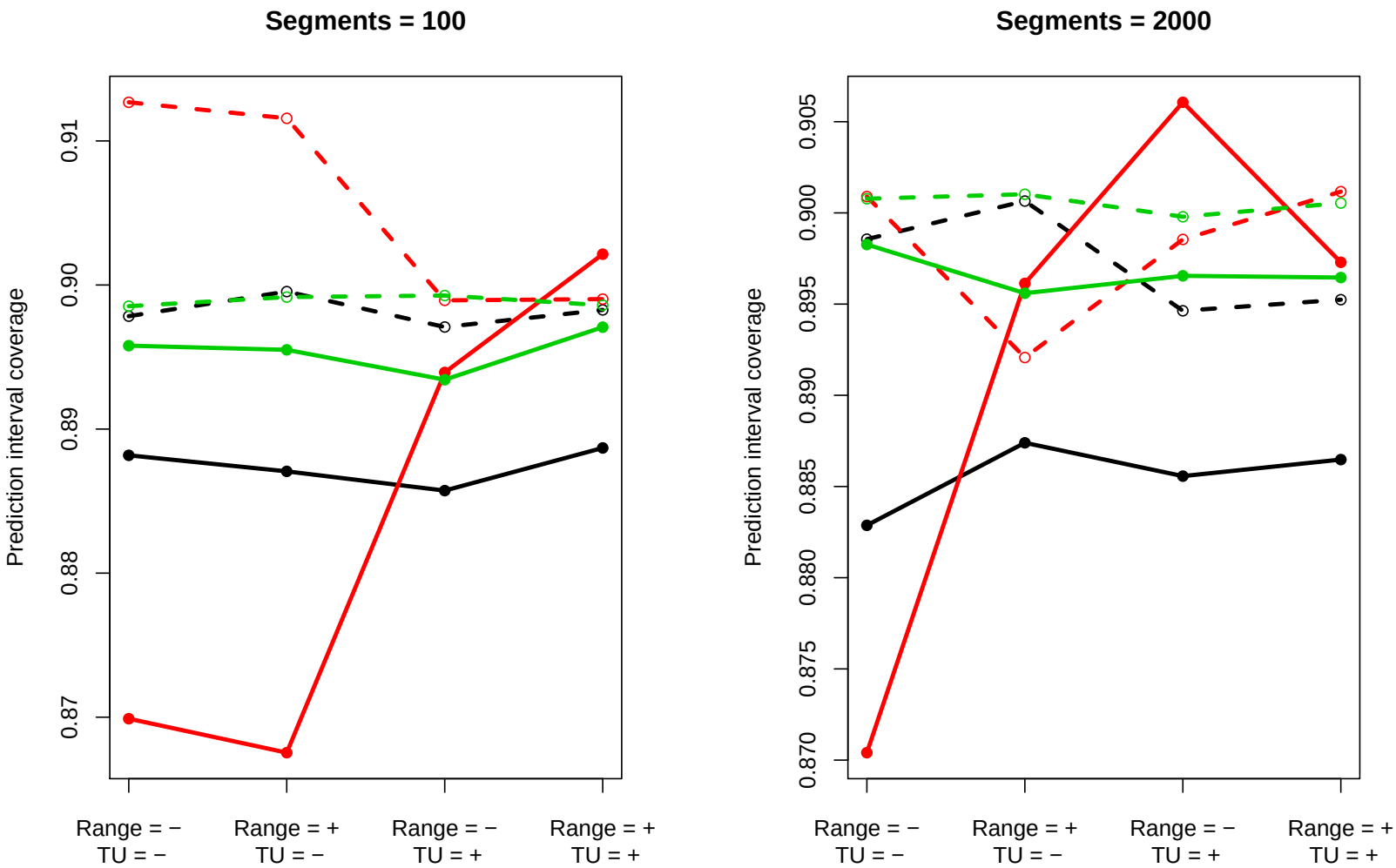
Prediction Error



Dashed lines: $n = 500$; solid lines: $n = 50$.

Black: SSN model; Red: spline model; Green: Linear model.

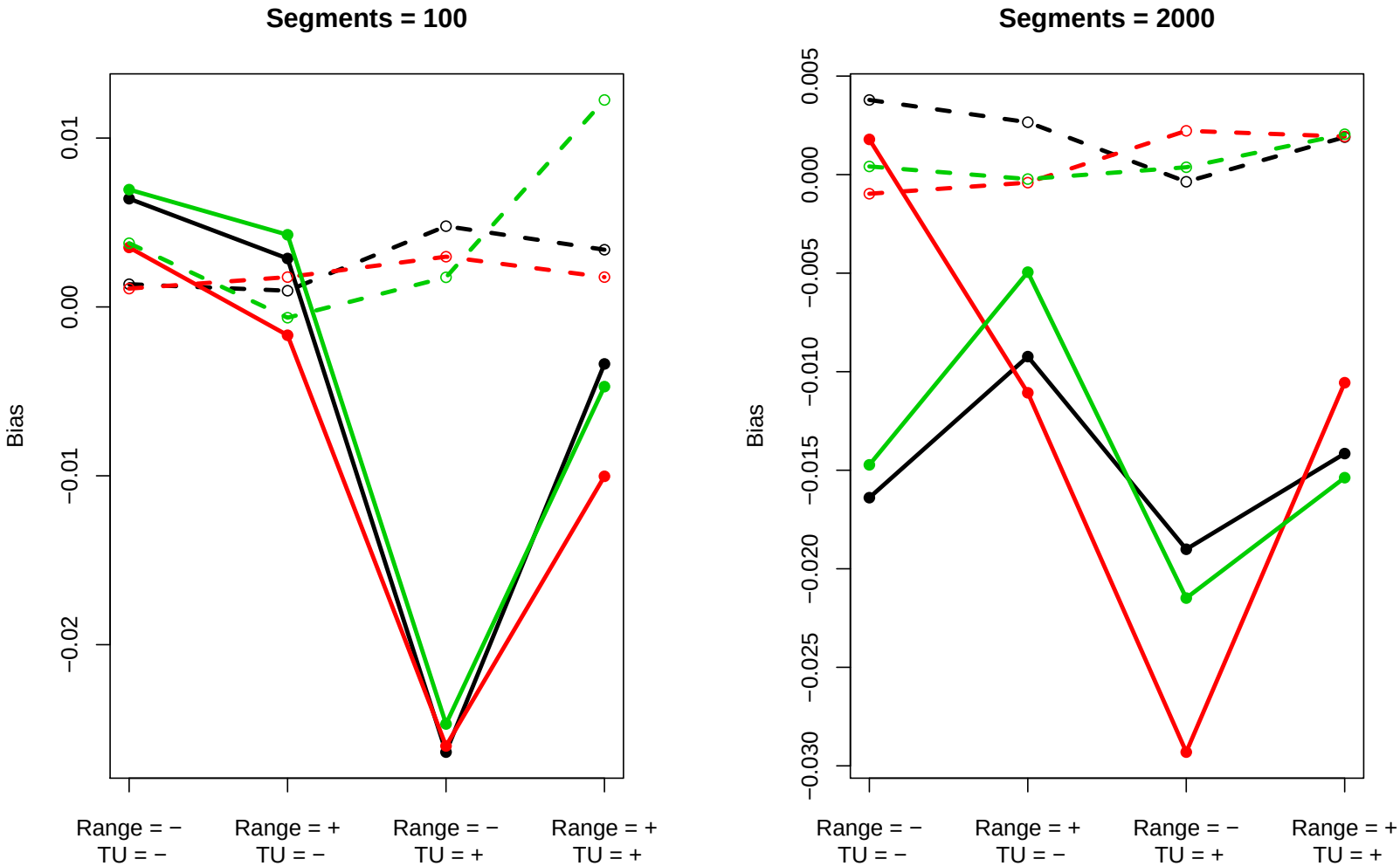
Prediction Interval coverage



Dashed lines: $n = 500$; solid lines: $n = 50$.

Black: SSN model; Red: spline model; Green: Linear model.

Prediction bias



Dashed lines: $n = 500$; solid lines: $n = 50$.

Black: SSN model; Red: spline model; Green: Linear model.

Summary

Prediction errors are smallest under SSN model

- Interval coverage is relatively poor for spline model, best for SSN; improves for larger n
- Strong dependance at small scales is the most challenging to capture well

Comparison is a work in progress

- Compare performance with other types of simulated spatial structures eg. Euclidean
- Sensitivity of SSN to misspecification
- Spatio-temporal data; smooth covariate effects eg. altitude

Extending the spline approach...

λ **selection**: Uncertainty not accounted for: Bayesian approach?

Further idea: Smooth basis functions *on* the network, rather than piecewise constant

- Natural extension, suggested first by Cressie and O'Donnell (2010).
- Sparsity and computational feasibilty should be preserved

Treat flow rates as parameters

- Bayesian analysis inevitable; uncertainty accounted for
- Priors based on proxies/ partial data

Thanks for your attention!