Flexible regression models over river networks

David O'Donnell, Alastair Rushworth, Adrian W. Bowman*, E. Marian Scott
School of Mathematics and Statistics
The University of Glasgow
U.K.

Mark Hallard
Scottish Environment Protection Agency

February 6, 2013

Summary

Many statistical models are available for spatial data but the vast majority of these assume that spatial separation can be measured by Euclidean distance. Data which are collected over river networks constitute a notable and commonly occurring exception, where distance must be measured along complex paths and, in addition, account must be taken of the relative flows of water into and out of confluences. Suitable models for this type of data have been constructed based on covariance functions. The aim of this paper is to place the focus on underlying spatial trends by adopting a regression formulation and using methods which allow smooth but flexible patterns. Specifically, kernel methods and penalised splines (p-splines) are investigated, with the latter proving more suitable from both computational and modelling perspectives. In addition to their use in a purely spatial setting, p-splines also offer a convenient route to the construction of spatiotemporal models, where data are available over time as well as over space. Models which include main effects and spatiotemporal interactions, as well as seasonal terms and interactions, are constructed for data on nitrate pollution in the River Tweed. The results give valuable insight into the changes in water quality in both space and time.

Keywords: flexible regression; kernels; p-splines; network; smoothing; spatial; spatiotemporal; water quality.

*Corresponding author, email: adrian.bowman@glasgow.ac.uk

1 Introduction

Statistical models for data collected over a spatial region are widely available and heavily used in an enormous range of applications. However, the majority of these models assume that the spatial region of interest is a straightforward subset of $\mathbb{R}^2$ where Euclidean distance is the natural metric. One interesting example of spatial data which does not have these characteristics arises from measurements made over a network consisting of continuous, connected curved line segments. The sample space is intrinsically one-dimensional, although embedded in two-dimensional space. River catchments are a particular, and commonly occurring, example of this. Figure 1 illustrates both the network and a series of point sampling stations for the River Tweed, which spans the border between Scotland and England. (Note that the picture shows some apparently unconnected stream segments. This is simply because some small lochs and other types of water body are not shown.)

Models for this type of spatial data require different constructions. In particular, Euclidean distance needs to be replaced by 'stream distance', defined by Ver Hoef *et al.* (2006) as 'the shortest distance between two locations, where distance is only computed along the stream network'. This approach has been used in geostatistical models for stream networks for some time, for example by Cressie and Majure (1997) and Gardner *et al.* (2003). However, Ver Hoef *et al.* (2006) showed that substituting stream distance for Euclidean distance in standard geostatistical theory does not produce a valid spatial covariance model except when the exponential covariance structure is used. Ver Hoef *et al.* (2006) and Cressie *et al.* (2006) used moving average constructs to define a much broader class of valid spatial covariance

models which use stream distance as well as other information, such as flow volume and the flow-connectedness of locations. One of the defining properties of these models is that they assign a correlation of zero to pairs of locations which are not flow-connected. Ver Hoef and Peterson (2010) developed the theory set out in these earlier papers by defining both 'tail-up' and 'tail-down' moving average constructions, in order to allow for correlation between pairs of locations which are not flow-connected. A variety of applications have subsequently been built on this theoretical structure; see Peterson and Ver Hoef (2010), Peterson and Urquhart (2006), Peterson *et al.* (2006), Garreta *et al.* (2010) for examples.

Covariance functions, and the use of kriging for prediction at locations which have not been monitored, provide a very well established approach to the construction of statistical models for spatial data. However, in some applications the principal focus is on the presence and nature of underlying trends, created by effects such as land use, geological patterns, dominant weather patterns or other influences with a strong systematic component which persists over repeated sampling of the same spatial region. Linear trends can be accommodated easily in covariance function models but in environmental settings trends often take the form of more flexible, non-parametric patterns. An attractive approach is then to place the emphasis on the direct modelling of these trends, using suitable forms of flexible regression, incorporating appropriate forms of spatial errors terms where necessary. This line of thinking is also well established and expressed, for example in the geoadditive models of Kammann and Wand (2003) and the more general semiparametric and additive modelling frameworks described by Ruppert *et al.* (2003) and Wood (2006) among others. Bowman *et al.*

(2009) describe a model of this type for spatiotemporal data.

The aim of the present paper is to develop methods of flexible regression for data over a network. In common with all spatial models, a regression approach allows estimates to be constructed over an entire spatial region from point located data, but it also provides a framework within which spatial, temporal and other covariate effects can be treated simultaneously. Smoothing techniques form the basis of flexible regression methods and these have been applied to a variety of data structures. However, the published literature shows very little evidence of their use in a network setting. The challenge is to devise methods built on the concept of 'borrowing strength' locally, while respecting the specific topology of a network and the additional complications of directionality and size of flow. A key issue in addressing these issues is how to deal with confluence points, where different branches of the network combine. It is shown below that successful treatment of these issues leads to significant improvements over more standard smoothing techniques in this setting. In particular, the estimators exhibit features, such as sharp changes which are often expected at confluence points, but which cannot easily be reproduced by more standard approaches.

Monitoring systems which are designed to collect data spatially also commonly record data over time. In fact, in many applications, the detection of changes over time is equally important as the identification of spatial pattern and this has motivated a large body of research in spatiotemporal modelling. However, very little of this work is directed at a network setting. This provides an opportunity for a successful network flexible regression approach to be extended into the spatiotemporal setting, where spatial, temporal and interaction terms can all be identified in an informative manner.

Different approaches to flexible regression over a network, including local fitting and penalised methods, are discussed in Section 3. A spatiotemporal model, including main effects and interactions, is constructed in Section 4 where a correlated error structure is also considered. Visualisation of the complex nature of the interactions is also discussed here. Throughout the paper, the methods and models are applied to data from the River Tweed. Some final discussion is given in Section 5.

2 The Tweed data

The data that will be used in the analysis here are supplied by the Scottish Environment Protection Agency (SEPA) and refer to the catchment of the River Tweed. SEPA is responsible for the routine collection and evaluation of water quality data from Scotland's lochs, rivers and estuaries. The importance of this is underlined by European Union directives such as the Nitrates Directive, adopted in 1991, and the Water Framework Directive, adopted in 2002, which set targets in terms of water quality and ecological status. EU members have committed to meet these targets and the collection and analysis of monitoring data is therefore essential.

Data on the Tweed catchment are available from January 1987 to August 2011 for eighty three monitoring stations on the river, although the timing of observations varies across the stations. Figure 1 indicates the highly dendritic nature of the Tweed network and illustrates the monitoring stations by plotting nitrate measurements recorded in February 2004. Stations which were not monitored at this particular time point are shown as empty circles. The scale of the map is approximately 70 miles in each direction.

A number of different chemical and biological measurements are avail-

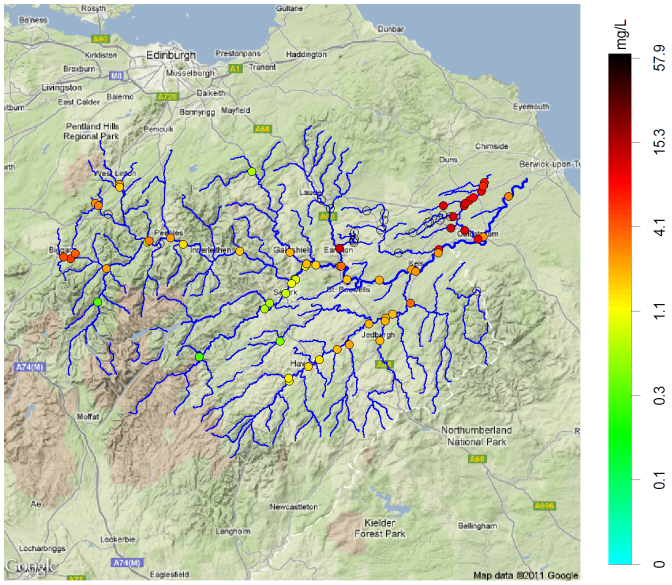

Figure 1: The River Tweed catchment, with sampling stations colour coded by nitrate level recorded in February 2004. Stations where no measurement was available at this time point are indicated by open circles. The map was produced from Google maps, through the `RgoogleMaps` package described by **?**.

able, but the most important one is nitrate. Diffuse pollution such as sewage effluent and runoff from fertilisers are among the largest contributors to nitrate levels, which makes this an appropriate choice for the Tweed, which is surrounded by mainly arable land. Two different types of measurement are available, namely Nitrate (N) and Total Oxidised Nitrate (TON), both measured in milligrams per litre. TON is technically the sum of nitrate and nitrite levels but, since the latter tends to be very small, SEPA regards them as being essentially equivalent and this will be assumed in the analysis. In order to improve normality and stabilise the variance, nitrate level will be analysed on the log scale.

The analysis discussed in later sections of the paper will highlight the importance of measuring water flow. This is only available at a limited

number of locations on the river and it would be impractical to measure this more widely. SEPA therefore uses a hydrological model to estimate flow for each of the 298 separate stream segments. This is an adequate representation because, in later analysis, only relative flow across the stream segments is required. The widths of the stream segments shown in Figure 1 are used to reflect these relative flow volumes. In cases where there is concern about the quality of flow information, an alternative is to follow Ver Hoef *et al.* (2006) by using a surrogate such as 'stream order', which indexes each stream segment by its location in the hierarchy of tributaries to the main river.

The Tweed catchment includes areas of outstanding natural resource which it is of high importance to monitor and preserve. However, the catchment also has considerable diversity of land use, including hill country, farmland and populated areas. The pattern of pollution is therefore likely to evolve differently over time across these land types and this motivates the need for a spatiotemporal description of pollution levels.

3 Network smoothing

There is a wide variety of approaches to the construction of smooth, flexible regression models. Wood (2006) provides a very useful starting point for the large associated literature, with Hastie and Tibshirani (1990), Bowman and Azzalini (1997), Schimek (2000) and Ruppert *et al.* (2003) providing earlier helpful overviews. Of the broad concepts involved, a common approach involves the use of local fitting through kernel functions or other forms of weighting schemes. Another involves the use of basis functions, often in conjunction with a penalty function to control smoothness. Both of these

approaches are developed below for network data.

3.1 Kernel functions

When data $\{(y_i, x_i), i = 1, \ldots, n\}$ are to be modelled by a flexible regression $y_i = m(x_i) + \varepsilon_i$, where m denotes a smooth function and the ε_i denote error terms, a very simple approach to the estimation of m at any point of interest x is to compute a local average in the form

$$\hat{m}(x) = \frac{\sum_i w(x_i - x; h)\ y_i}{\sum_i w(x_i - x; h)},$$

where the weight function w decreases with distance from zero, at a rate determined by the parameter h, and so controls the degree of influence of each observation on the estimate. A normal density function with standard deviation h is a convenient choice for w. It turns out that fitting a local linear regression rather than a local mean has better theoretical properties, as described by Fan and Gijbels (1996) for example. However, where the data are sparse, as occurs in some areas of the Tweed catchment, this can cause stability problems, so the simple local mean is considered here.

One attraction of this approach is that the generality of the underlying idea allows it to be modified to suit data of very different types. In the network setting, these modifications follow the patterns of the spatial models for networks outlined in Section 1. If the locations of the observed values are denoted by s_i, and the location of the point of estimation by s, then the appropriate covariate values x_i and x now refer to 'river distance', using the river mouth as a natural origin. Secondly, the weights are non-zero only where there is a flow path between s_i and s. Thirdly, additional weights should be used to reflect the volume of water flowing in different sections of the river. This is exactly the approach proposed by Ver Hoef *et al.* (2006) but

employed in the context of a regression model rather than spatial prediction through kriging. The flow weighting is derived from a 'tail up' model, using the terminology of Ver Hoef and Peterson (2010), which assumes that points which are not flow-connected are uncorrelated. Specifically, the additional weights are expressed in the function

$$\delta_i(x) = \begin{cases} \Pi_{k \in B_{x,x_i}} \sqrt{\omega_k} & \text{if } s \text{ and } s_i \text{ are flow connected} \\ 0 & \text{if } s \text{ and } s_i \text{ are not flow connected} \end{cases}$$

where B_{s,s_i} is the set of all stream segments between and including the locations s and s_i. Here a stream segment refers to a stretch of water between two neighbouring confluence points. The quantity ω_k denotes the proportion of flow contributed by water stretch k to its subsequent confluence. It contributes on the square-root scale in order to stabilise variance across the contributing stream segments at confluence points, as discussed by Ver Hoef *et al.* (2006).

An estimate of the mean value at s, located at a distance x from the mouth of the river, is then available as

$$\hat{m}(x) = \frac{\sum_i w(x_i - x; h)\delta_i(x)\ y_i}{\sum_i w(x_i - x; h)\delta_i(x)} = \sum_i v_i y_i,$$

where v_i denotes the combined effect of the weighting schemes, incorporating river distance, flow connectedness and flow values. A vector of estimated values can then be written as Sy, where the rows of the smoothing matrix S contain the weights applied to the data vector y in order to construct an estimate at a particular location. Where the estimation points are set to the observed locations, the trace of this 'hat matrix' has the interpretation of an approximate degrees of freedom, as discussed by Hastie and Tibshirani (1990) and many other authors.

Further details of this approach are described in O'Donnell (2012).

3.2 Penalised splines

An alternative approach to the estimation of smooth functions is to use a set of basis functions, $\phi_j(x), j = 1, \ldots, p$, as the components in a regression model, representing the estimate in the form $\hat{m}(x) = \sum_j \beta_j \phi_j(x)$, with coefficients β_j. This has the advantage that the estimate can be expressed in proper functional form simply through the specification of the β_j's and that the dimensionality of the estimation problem can be kept low, independent of the size of the dataset. A convenient choice of basis functions is b-splines, whose construction from polynomial pieces gives them many attractive computational properties, as described by de Boor (1978). The approach to smoothing known as p-splines, proposed by Eilers and Marx (1996) and now used by many other authors, uses a rich b-spline basis but imposes a roughness penalty on the coefficients β_j.

The construction of a basis set over a network faces the difficulty of combining the basis components in a suitable manner at the confluence points. This is possible, but slightly awkward, with the usual pattern of smooth overlapping functions. An attractive alternative is to divide the network into a large number of small pieces within which the function m is likely to change very little. In the river setting, this arises naturally through the identification by SEPA of 'stream units' corresponding to short water stretches which are judged to be relatively constant in terms of environmental conditions. There may be several stream units within each stream segment (the stretch of water between two adjacent confluence points). As the sizes of the stream units are very small compared to the network as a whole, a regression model can be constructed in a piecewise constant manner through a set of mean values $\beta_j, j = 1, \ldots, p$, associated with the p stream units which make

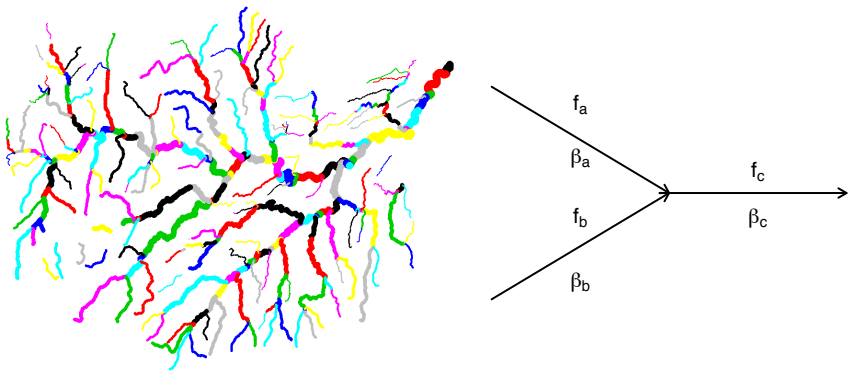

Figure 2: The left hand plot uses different colours to show the decomposition of the river network into a large number of small 'stream units'. The right hand plot gives a schematic representation of a confluence, with model parameters (β_a, β_b), flows (f_a, f_b) and the corresponding outgoing versions (β_c, f_c).

up the network. The estimate at any point of interest s is then simply the estimated value of β_j, where j indexes the stream unit in which s lies. If the number of stream units is large then the loss in resolution through the approximation of m in this piecewise constant manner is very small. This is, in fact, equivalent to the use of b-splines of order 0.

There are likely to be considerably more stream units than observed values and so the estimation process is ill-defined. However, a penalty approach immediately overcomes this difficulty and also provides a means of controlling the smoothness exhibited by the estimate. The 'smoothness' of β-values corresponding to adjacent stream units j and k, with no intervening confluence, can be measured by $(\beta_j - \beta_k)^2$. Where a confluence point is involved, as illustrated in Figure 2, the measure of smoothness needs to reflect the relative levels of flow in the contributing streams a and b. If the flows are denoted by f_a, f_b, f_c, then we expect $f_c = f_a + f_b$ and the

mixing of pollutants to be controlled by the relative flows of the inputs, $\omega_a = f_a/f_c$ and $\omega_b = f_b/f_c$. Following the principle of mass balance, the combined pollution input $\omega_a\beta_a + \omega_b\beta_b$ and the output β_c are identical if $\omega_a(\beta_a - \beta_c) + \omega_b(\beta_b - \beta_c) = 0$. Smoothness across the confluence can therefore be achieved through the penalty

$$\lambda\{\omega_a^2(\beta_a - \beta_c)^2 + \omega_b^2(\beta_b - \beta_c)^2\}, \tag{1}$$

where λ controls smoothness through the weight attached to the penalty. This has the attractive form of combining penalties for smoothness across each flow path of the confluence, with weights determined by the relative flow volumes.

A p-spline model can be formulated as $y = B\beta + \varepsilon$, where y, β, and ε denote the vectors of responses, parameters and errors respectively, while the design matrix B is simply an $n \times p$ indicator matrix whose ith row has the value 1 in the column corresponding to the stream unit of y_i and 0's elsewhere. Following Eilers and Marx (1996), the model is fitted by minimising the penalised sum-of-squares

$$(y - B\beta)^T(y - B\beta) + \lambda\beta^T D^T D\beta$$

with respect to β. Here D is the matrix which generates the differences between β's from adjacent stream units, weighted by flow where this is an intervening confluence point, as in the components of (1). The penalty parameter λ controls the degree of smoothing. The solution to this least squares problem is easily shown to be $\hat{\beta} = (B^T B + \lambda D^T D)^{-1} B^T y$. The linear form of this expression again allows an approximate degrees of freedom to be computed as the trace of the 'hat' matrix.

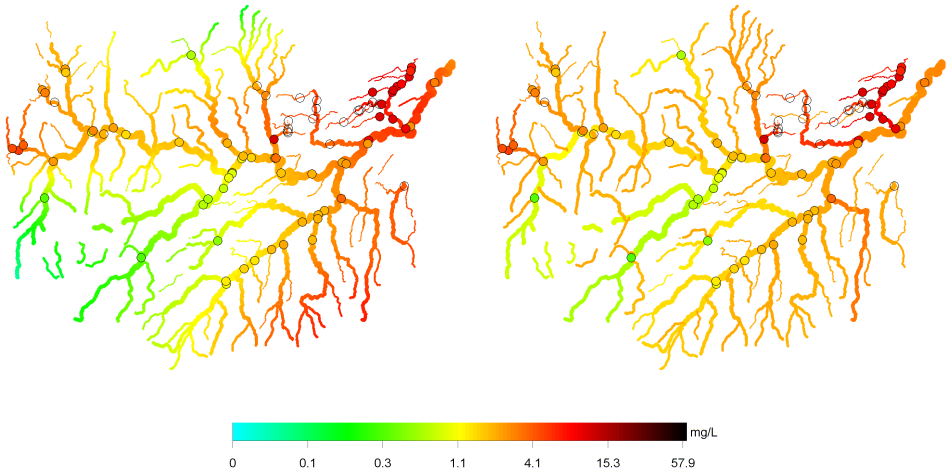

Figure 3: Left plot is a Euclidean distance smooth estimate. Right plot is a network smooth estimate.

Figure 3 shows the effects of smoothing on the average nitrate measurements from February 2004. The right hand panel shows the effects of p-spline smoothing with 12 degrees of freedom. The ability to view the spatial structure over the whole network is a significant benefit, in comparison with plots of the original point-based observations shown in Figure 1. In contrast to the use of standard two-dimensional Euclidean smoothing, shown in the left hand panel of this figure, proper recognition of the network structure gives appropriate weighting to observations from stream segments of different size and allows quite sharp changes in estimated level. For example, the relatively high concentrations of pollution exhibited by some of the tributaries in the northern periphery of the network are not immediately inherited by the larger and relatively unpolluted streams into which they flow, as a result of dilution effects. This behaviour reflects what we believe to be happening in the river but it cannot be captured by methods which ignore the special structures of a network.

3.3 A comparison of the two approaches

The use of kernel functions rather than p-splines produces an estimate which is qualitatively similar to the right hand panel of Figure 3. This confirms the general view that the particular form of construction of smoothing techniques is relatively unimportant - it is the choice of degree of smoothing which matters.

However, a comparison of these two approaches to smoothing over a network indicates that the p-spline method has some advantages. It characterises the levels of pollution over the entire network through a set of parameters, one for every stream unit, which makes it particularly easy and efficient to identify estimated values at any location of interest, simply by identifying the stream unit in which it lies. In contrast, local mean estimation requires a new calculation for each new estimation point. The p-spline estimate is able to handle regions where the data are sparsely located, because the penalty can bridge the gap between data points, while local mean estimation can run into difficulties with very small weights. The p-spline construction is based simply on the relationship between values and flows in neighbouring stream units, so there is no need to define connectedness across the whole network simultaneously or to consider the cumulative effect of flow across distances which span several confluences. There is some potential loss of information with the use of stream units in the p-spline approach because the lengths of the stream units are not used. However, this loss of information is likely to be very small, precisely because the stream units have been defined as homogeneous stretches of water.

In view of these issues, further modelling work will use the p-spline approach. From a computational point of view, the kernel and p-spline meth-

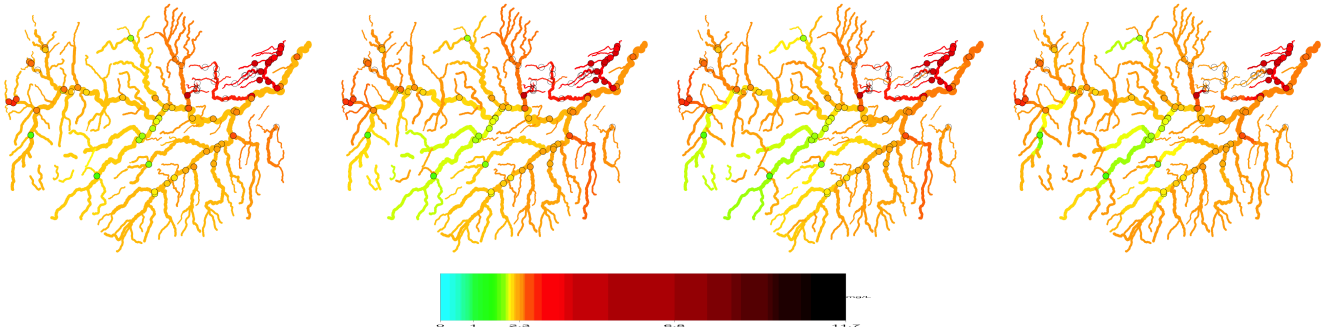

Figure 4: The three panels show the effects of smoothing with 6, 12 and 48 degrees of freedom. Possibly replace with two plots of 6 and 48 df on the original scale, against which the network smooth in the previous figure can be compared. THis figure shows the extremes.

ods are similar in their demands. However, within the context of larger models involving time, seasonal and possibly other covariates, to be discussed below, the direct nature of the estimation process involved in p-splines leads to much greater computational efficiency. This considerably strengthens the case for the use of p-splines for network smoothing.

With any method of flexible regression, the degree of smoothing, which expresses the complexity of the model, is an important choice. There is a very large literature on how this might be done in an automatic manner, using general principles such as cross-validation or an information criterion such as AIC. Hurvich *et al.* (1998) propose a version of AIC where the optimal value of λ is chosen to minimise $AICc = \log(\hat{\sigma}^2) + 1 + \frac{2+2\,dof}{n-dof-2}$ (Hurvich and Tsai, 1989), where dof denotes the approximate degrees of freedom associated with λ. This imposes an additional penalty for large degrees of freedom and is designed to avoid the under-smoothing that might result in such cases when criteria such as GCV or standard AIC are used.

However, there can also be merit in considering the effects of different levels of smoothing, in a sensitivity or multiresolution analysis. Expressing

flexibility through degrees of freedom offers a very convenient scale on which to explore this. It is interesting to note that the qualitative features of the estimates produced by three widely varying degrees of freedom in Figure 4 shows remarkably little change. This suggests that the constraints imposed by the network structure have a strong stabilising effect on the estimation process.

4 Spatiotemporal models for networks

If the aim is to estimate the levels of pollution over a network at a single time point, then there is little to choose between kriging and flexible regression. Indeed, from a single realisation of a spatial process it is not feasible to separate persistent trend from transient spatial variation. However, where observations are also available over time, the nature of spatial and temporal effects and their potential interactions becomes of considerable interest and, in contrast with approaches based on covariance functions, flexible regression methods can be extended to this setting in a relatively straightforward manner.

There are, of course, very many current examples of spatiotemporal models. Cressie and Wikle (2011) describe a wide variety of modelling tools, while many authors describe applications in areas such as air pollution (Guttorp *et al.*, 1994; Shaddick and Wakefield, 2002), rainfall (Brown *et al.*, 2001), snow water (Huang and Cressie, 1996), fish population size (Reyjol *et al.*, 2005) and many more. There are also many examples of river data being modelled through space and time, as described for example by Cressie and Majure (1997), Clement and Thas (2007), Akita *et al.* (2007) and Thorp *et al.* (2006). However these examples do not consider the essential network

features of river distance, flow connectedness and flow weighting. Only a very small number of papers consider a river network structure for data through space and time. Money *et al.* (2009) consider the use of the tail-up structure in space-time models, using a Bayesian/maximum-entropy method of fitting, but their dataset does not make the use of flow-connectedness feasible. Gardner and McGlynn (2009) also use the tail-up model for nitrate data from the Rocky Mountains, but the analysis is primarily based on a spatial analysis for each of a small set of time points over a period of approximately one year, in order to highlight seasonal changes.

There are three principal variables which need to be accommodated in a spatiotemporal model. One is space expressed through the river network locations (s_i), the second is time (t_i) measured on a scale of years to express long term trends, while the third is time within the year (z_i) to express the seasonal changes which are very often exhibited in environmental measurements. Additive models are natural tools to consider, as they provide a framework within which flexible regression can be extended to a wide variety of data structures. Hastie and Tibshirani (1990) provided an early synthesis of this approach which has encouraged wide use of these models in many different application areas. Wood (2006) provided a modern overview which markedly extends the range of the tools available.

In the present setting, a very simple additive model is

$$y_i = \mu + m_s(s_i) + m_t(t_i) + m_z(z_i) + \varepsilon_i \tag{2}$$

where the three functions m_s, m_t, m_z describe spatial, temporal and seasonal trends and ε_i denotes error terms assumed to have a $N(0, \sigma^2)$ distribution marginally. If each of the trend functions is estimated by b-splines then, following the derivation in Section 3.2 above, they can be represented as

$B_s\beta_s$, $B_t\beta_t$, $B_z\beta_z$ where the columns of the design matrices evaluate each basis function at the observed values of the relevant covariate. B-splines of order 0 can be used for the spatial network, while cubic b-splines would be a good choice for the temporal and seasonal effects, as these are defined over more standard sample spaces. The full model can be represented as $y = B\beta+\varepsilon$, where B combines the columns of the individual design matrices, with an initial column of 1's.

It remains to construct suitable penalty terms to induce smoothness on the estimates of the trend functions. First-order differences were the natural choice for the spatial network parameters, as described in Section 3.2 and computed through a difference matrix. For cubic b-splines, second-order differencing of the parameter vector is the more standard choice. The smoothness penalty can then be expressed as $\beta^T P\beta$, where the matrix P has block-diagonal form which combines the individual penalties as $(0, \lambda_s D_s^T D_s, \lambda_t D_t^T D_t, \lambda_z D_z^T D_z)$. Cyclical behaviour in the seasonal term can be induced by requiring the coefficients of the first r basis functions to be identical with the last r basis functions. The penalty $\sum_{k=1}^{r}(\beta_{z,k}-\beta_{z,p+1-k})^2$ achieves this, with $r = 3$ for cubic splines. This can be adopted into the definition of D_z.

In the presence of an overall mean parameter μ in model 2, the identifiability of each additive component can be achieved by the addition of a ridge penalty, as described by Eilers and Marx (2002). This corresponds to a penalty of the form $\beta^T Q\beta$, where Q is a diagonal matrix constructed from the vector $(0, \nu_s 1_s, \nu_t 1_t, \nu_z 1_z)$, with the ridge parameters denoted by ν_s, ν_t, ν_z and with 1_a denoting a vector of 1's whose length is determined by the number of basis functions in the term denoted by a. The fitted model can then

be expressed through the parameter estimates $\hat{\beta} = (B^T B + P + Q)^{-1} B^T y$. Denoting this as Hy, standard errors for $\hat{\beta}$, and so for fitted values, are available from the diagonal elements of HH^T, multiplied by an estimate of the error variance which is constructed as $\hat{\sigma}^2 = RSS/(n - dof)$ where RSS denotes the residual sum-of-squares and dof the approximate degrees of freedom for the model. A penalised spline approach to (generalised) additive modelling is described by Marx and Eilers (1998), and many subsequent authors including Wood (2006), where further details are available.

The additive model (2) is a natural starting point but it is implausible that the spatial pattern of pollution will change in exactly the same way over time, or throughout the year, at every location. It is therefore more appealing to consider an interaction model of the form

$$y_i = \mu + m_s(s_i) + m_t(t_i) + m_z(z_i) + m_{s,t}(s_i, t_i) + m_{s,z}(s_i, z_i) + m_{t,z}(t_i, z_i) + \varepsilon_i, \tag{3}$$

where the functions $m_{s,t}$ and $m_{s,z}$ encapsulate the adjustments required to capture how the time trend and seasonal effects vary over the river network. The term $m_{t,z}$ allows an adjustment to the overall seasonal component, allowing different patterns in different years. The interaction terms can also be conveniently represented in spline basis form, this time using a basis formed by all possible products of the spline basis functions on each separate variable. More precisely, we can write $m_{s,t} = \sum_{j,k} \beta_{jk} \phi_{s,j} \phi_{t,k}$, where $\phi_{s,j}$ and $\phi_{t,k}$ denote b-spline functions for space and time. Since the spatial basis is constructed from b-splines of order 0, this has the simple interpretation that the parameters associated with each stream unit are now allowed to evolve smoothly over time. Corresponding structures and interpretations can be adopted for the space-season and time-season interaction terms. In

matrix notation, the model matrix is

$$B = \left[\begin{array}{cccccccc} \mathbf{1} & B_s & B_t & B_z & B_s \square B_t & B_s \square B_z & B_t \square B_z \end{array} \right]$$

where $\square$ is the row-wise tensor product defined as $A \square Y = (A \otimes \mathbf{1}') \odot (\mathbf{1}' \otimes Y)$ and $\odot$ denotes the Hadamard (elementwise) product; see Eilers *et al.* (2006).

Smoothness in the model terms is induced by applying appropriate penalties, and corresponding penalty parameters λ_i, for each term. In the case of main effects, these are constructed through the difference matrices described above. Penalties for the interaction terms can be constructed by considering the coefficients $\{\beta_{jk}\}$ in matrix form and applying smoothness penalties to both the rows and the columns. For example, space-time smoothness is induced by applying a first-order network penalty to the columns of the matrix $\{\beta_{jk}\}$ and a second order difference penalty over the rows. As described above, identifiability is ensured, and ill-conditioning avoided, by adding a ridge penalty for each term in the model, expressed in a diagonal matrix Q. Other constraints or penalties exist that could have achieved a similar effect, for example constraints that force the mean value of each component to be zero. However, the straightforward specification of the ridge penalty and the subsequent retention of sparseness of model objects makes this a convenient choice, as discussed in the subsection on computational details below. Each of the λ_i is estimated by a short search procedure to find values that minimise the corrected AIC as defined in Section 3.3.

Figure 5 shows the results of fitting this interaction model to the Tweed data, using AICc to select all the penalty parameters. The top left hand plot, together with those in the second row, show estimates of the main effects for space, year and day of the year. This highlights that areas of high pollution are present in the tributaries to the North-East of the River

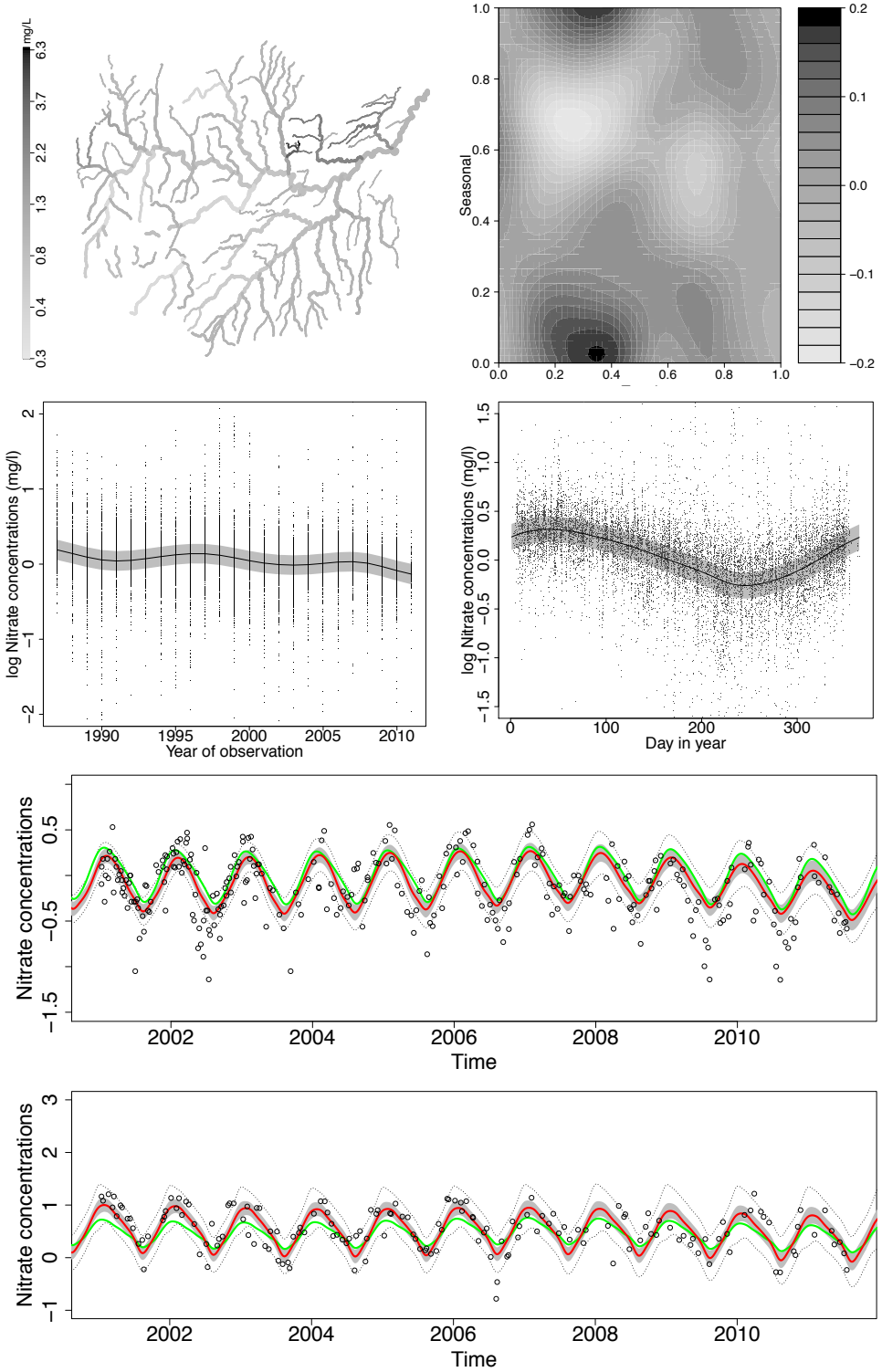

Figure 5: The top four plots show the main effects of space, year and day of the year. The lower two plots show fitted values at two specific spatial locations, namely Gala Water Foot and Norham Gauging Station respectively, including comparison of simple main effects model (green) and the interaction model (red) in each case.

Tweed. Across the years the overall pollution levels are relatively stable, but with some indication of a slight decreasing trend. The overall seasonal effect is strong, as expected, with a gentle decrease from February to August and a sharper rise at the end of the calendar year. The shaded bands in Figure 5 correspond to two standard errors on either side if the estimates and these indicate high precision, as a result of the substantial size of the dataset. The top right hand plot shows the estimate of interaction between year and season. The values of the adjustments plotted here are small, indicating that the change in seasonal pattern over the years is modest. The lower two plots in Figure 5 show fitted values at four specific spatial locations, along with a comparison of a simple main effects model (green) and the interaction model (red). This shows clear improvement at some sites as a result of fitting the interaction terms.

4.1 Computational details

Tensor product spline smooths such as those specified in Section 3 rely on many basis parameters to represent each bivariate interaction and may therefore be intensive to fit. In the present case, the model matrix B is mostly composed of terms involving B_s, an $n \times p_s$ matrix where p_s is the size of the network partition. B_s contains exactly n nonzero entries. If V denotes the $n \times q$ matrix evaluating the bases for the other terms in the model, then $B_s \square V$ is at least $100(1 - \frac{1}{p_s})\%$ sparse and much more so if V is sparse. Sparse matrix algorithms can be used to decrease storage requirements and vastly increase performance. For example, the B that was fitted to the Tweed data, was 16000×5000 where $p_s = 298$ and was 99.8% sparse.

The starting point in solving systems of sparse matrices is to define all

of the model components in a compressed format. The `spam` or `Matrix` packages in `R` (R Development Core Team, 2011) perform this, and allow the fitting of models with large amounts of data, or complex structures. All model summaries and standard errors can then be easily and efficiently calculated using a sequence of operations on the Cholesky factor L of (B^TB+P+Q) to obtain the quadratic form of interest, where $(B^TB+P+Q) = LL^T$. For example, the standard errors of the fitted values $se(\hat{y}_i)$ are the diagonal elements of $BL^{-1}L^{-T}B^T$ and can be calculated efficiently by solving first $AL = B$ and then calculating the summations of each row of A^2.

4.2 Residual correlation

Having established a spatial interaction model for the River Tweed nitrate data that is appropriate for its network structure, it remains to check the assumption of independence made of the residuals. Evidence of residual temporal correlation at short time lags is to be expected, particularly as the model accounts for trends over longer time periods. Under the assumption of independent errors, all standard error estimates are likely to be underestimated when the underlying error process is correlated, so they must be adjusted appropriately. As a conservative measure, it was decided to fit a separable spatiotemporal model to the errors so that

$$\hat{\Sigma}_{ij} = Cov(\epsilon_i, \epsilon_j) = \omega_{ij}\,\sigma^2\,\exp\left\{-\frac{d_{ij}}{\rho} - \frac{|t_i - t_j|}{\psi}\right\}$$

where $\omega_{ij} = \prod_{k \in N} \omega_k$ and k indexes the set of stream units that lie between i and j and on the same flow path as both. The spatial and temporal correlation in the error process is assumed to depend on $t_i - t_j$, the time lag, and d_{ij}, the network separation measured in numbers of stream units. The correlation model was fitted by weighted least squares.

Having obtained an estimate for $\hat{\Sigma}$, the standard errors for the fitted values were then adjusted by

$$\mathsf{s.e.}\{\hat{y}\} = \sqrt{\mathsf{var}\Big\{\hat{H}y\Big\}} = \sqrt{\mathrm{diag}(\hat{H}\hat{\Sigma}\hat{H}^T)}$$

where $\hat{H}$ is the projection or hat matrix given by $B(B^TB+P+Q)^{-1}B^T$. The estimated parameters in the correlation model were $\rho = 8.3$ and $\psi = 27.4$ which represents moderate residual temporal correlation and (after adjusting with weights) weak residual spatial correlation. These parameters refer to a spatial scale in miles relative toa catchment diameter of approximately 70 miles, and a temporal scale in days, relative to a span of 26 years for the whole dataset. The overall variance parameter σ^2 was estimated as 0.1554, which is very close to the estimate under an independence assumption (0.1442). The corresponding adjustments to standard errors are displayed in Figure 5 as dashed lines, from which it is clear that the increases in width over the independence model are not sufficiently large to lead to any substantive change in conclusions.

It would be possible to consider incorporating the correlation structure into the fitting process for the model. This would, however, considerably increase the complexity of the computations, particularly as sparsity would be compromised. The post-fitting adjustment approach combines computational efficiency with an effective first-order approximation to the correlation structure, which has been used to good effect in similar settings, as discussed by Giannitrapani *et al.* (2011).

4.3 Visualisation

While simple spatial terms can be plotted in map or network form, interactions with spatial components are more problematic to view. The lower

panels of Figure 5 show temporal effects at particular point locations. An alternative illustrated in Figure 6 is to display the estimated spatial effects at different time points, here at three different months (January, June and November) in 2005. This helpfully focusses attention on the spatial areas where seasonal change is strongest. However, changes in colour alone can be difficult to assess, especially where those changes are modest. The plots shown in Figure 6 represent the values over the network as 'nodes', plotted approximately in the geographical midpoint of each stream unit. In addition to colour code, each node has radius proportional to the estimated nitrate pollutant level (on the original rather than log scale). This form of display is particularly effective at illustrating changes over time as small changes in size are more easily identifiable than small changes in colour.

A more satisfactory solution involves animation of the spatial pattern across time. This kind of effect can be achieved with graphical tools such as those provided by the `rpanel` package (Bowman *et al.*, 2007) for `R` (R Development Core Team, 2011). This allows the time setting for the spatial display to be controlled through a slider. In a similar manner, sliders can also be used to control the degrees of smoothing through interactive selection of values for teh approximate degrees of freedom. Since visualising and understanding spatiotemporal model fits is challenging from static printed plots, two animations of the fitted models are provided online (`RSS web site to be inserted`). These illustrate spatial and temporal variation in the fitted mean nitrate levels in both network and node form. The effect of the spatial penalty across neighbouring stream units is more evident in the first, while the degree of pollution and change through time is arguably better represented by the second.

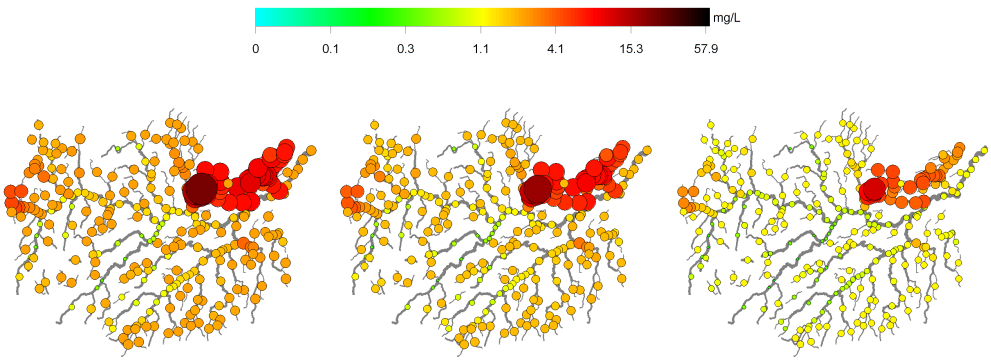

Figure 6: Estimated spatial effects at three different months (January, May and October) in 2005, indicated by colour and scaling of 'nodes' located at the stream units.

5 Discussion

This paper has proposed flexible regression models which respect the unique spatial structure of data which arises from a directed network and which allow appropriate spatially-varying coefficient models for capturing spatial change. In addition, the methods are capable of capturing the complex temporal changes which can occur in environmental pollutant concentrations. The results from the River Tweed data show evidence of a strong overall seasonal effect but only a small degree of overall change in nitrate levels over the 26 year period. Of greater interest is the insight gained from interaction terms, as these allow the inclusion of spatially-varying effects which also respect the nature of the network. It is clear from Figure 5 that local effects account for much of the variation seen at a specific site over time. This may support an argument in favour of the retention of a robust, well designed long term monitoring network which is able to detect changing local environmental pressures.

The penalised models described here use a discrete approximation to what is essentially a continuous spatial process, which in combination with

the assumption of conditional independence structure among neighbouring sites, renders all model components sparse and computationally straightforward to store and manipulate. The sparseness property allows complex network models to be constructed and further covariates to be included without the computational limitations which can hamper other approaches.

The penalised model specification has strong similarity with a Bayesian hierarchical model in which the β_i are treated as random effects with a Gaussian Markov random field (GMRF) prior and a fixed variance, controlling pairwise differences on neighbouring stream units. A Bayesian approach offers some attractive properties, particularly because sampling from the full posterior distribution allows uncertainty associated with smoothing parameters to be integrated out. At present the various smoothing parameters are selected through a grid search over a range of candidate values which can be cumbersome for multidimensional smooths or many covariates. In addition, for MCMC updates, the conditional independence structure of the random effects β_i lends itself to the efficient block updating for GMRFs described by Fahrmeir and Lang (2001). This approach will be the focus of future research.

Attention has focussed on the estimation of model terms and their standard errors, as these give clear and interpretable insight into the structure of the data. If more formal methods of model comparison are required, these can be implemented through approximate F-tests as described by Bowman *et al.* (2009) in the spatiotemporal setting.

Regardless of estimation procedure, average flow values are required for each stream unit in the network partition. Flow data used here was not observed but modelled and supplied by SEPA. Observed flow data would

allow a model to adapt to different flow settings over time, as observed by Cressie and O'Donnell (2010). However, here it is the flow ratios which are the crucial quantities.

The choice of spatial metric, namely river distance in the kernel approach or separation by a number of stream units in the p-splines approach, represents the belief that pollution changes in a slow and consistent way along stream segments. An argument might also be made for a metric with a Euclidean component. This could be valuable where the surrounding land is the source of pollution, rather than point sources, as land characteristics are naturally mapped on a Euclidean scale. It would be possible to use a combination (weighted average) of river distance and Euclidean distance, as in Cressie *et al.* (2006) for spatial prediction. O'Donnell (2012) explores this in the context of kernel methods. Alternatively, land use information could be included as covariates where such data are available.

Acknowledgements

David O'Donnell and Alastair Rushworth gratefully acknowledge support from EPSRC studentships. Some of the research was conducted while Alastair Rushworth was seconded to the Scottish Environment Protection Agency under a project funded by the Scottish Sensors Systems Centre. The work also benefitted from Adrian Bowman's participation in the discussions on spatiotemporal models under project MTM2008-02901 from the Spanish Ministry of Science and Innovation. The comments of a Joint-Editor, and associate editor and two reviewers were very helpful in revising the paper.

References

Akita, Y., Carter, G., and Serre, M. L. (2007). Spatiotemporal nonattainment assessment of surface water tetrachloroethylene in new jersey. *Journal of Environmental Quality 36*(2), 508–520.

Bowman, A. and Azzalini, A. (1997). *Applied Smoothing Techniques for Data Analysis*. Oxford: Oxford University Press.

Bowman, A., Crawford, E., Alexander, G., and Bowman, R. W. (2007). rpanel: Simple interactive controls for r functions using the tcltk package. *Journal of Statistical Software 17*(9).

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

Brown, P., Diggle, P., Lord, M., and Young, P. (2001). Space-time calibration of radar rainfall data. *Journal of the Royal Statistical Society Series C-Applied Statistics 50*, 221–241.

Clement, L. and Thas, O. (2007). Spatio-temporal statistical models for river monitoring networks. *Journal of Agricultural, Biological, and Environmental Statistics 12*(2), 161–176.

Cressie, N., Frey, J., Harch, B., and Smith, M. (2006). Spatial prediction on a river network. *Journal of Agricultural Biological and Environmental Statistics 11*(2), 127–150.

Cressie, N. and Majure, J. J. (1997). Spatio-temporal statistical modeling of livestock waste in streams. *Journal of Agricultural, Biological, and Environmental Statistics 2*, 24–47.

Cressie, N. and O'Donnell, D. (2010). Comment: Statistical dependence in stream networks. *Journal of the American Statistical Association 105*(489), 18–21.

Cressie, N. and Wikle, C. K. (2011). *Statistics for Spatio-Temporal Data*. New York: Wiley.

de Boor, C. (1978). *A Practical Guide to Splines*. New York: Springer-Verlag.

Eilers, P., Currie, I., and Durbán, M. (2006). Fast and compact smoothing on large multidimensional grids. *Computational statistics & data analysis 50*(1), 61–76.

Eilers, P. and Marx, B. (1996). Flexible smoothing with b-splines and penalties. *Statistical Science 11*(2), 89–102.

Eilers, P. and Marx, B. (2002). Generalized linear additive smooth structures. *Journal of Computational and Graphical Statistics 11*(4), 758–783.

Fahrmeir, L. and Lang, S. (2001). Bayesian inference for generalized additive mixed models based on markov random field priors. *Journal of the Royal Statistical Society Series C-Applied Statistics 50*, 201–220.

Fan, J. and Gijbels, I. (1996). *Local polynomial modelling and its applications*. London: Chapman and Hall.

Gardner, B., Sullivan, P., and Lembo, A. (2003). Predicting stream temperatures: geostatistical model comparison using alternative distance metrics. *Canadian Journal of Fisheries and Aquatic Sciences 60*(3), 344–351.

Gardner, K. K. and McGlynn, B. L. (2009). Seasonality in spatial variability and influence of land use/land cover and watershed characteristics on stream water nitrate concentrations in a developing watershed in the rocky mountain west. *Water Resources Research 45*, W08411.

Garreta, V., Monestiez, P., and Hoef, J. M. V. (2010). Spatial modelling and prediction on river networks: up model, down model or hybrid? *Environmetrics 21*(5), 439–456.

Giannitrapani, M., Bowman, A., and Scott, E. (2011). Additive models for correlated data with applications to air pollution monitoring. In R. Chandler and E. Scott (Eds.), *Statistical Methods for Trend Detection and Analysis in the Environmental Sciences*, Chapter 7, pp. 267–282. London: Wiley.

Guttorp, P., Meiring, W., and Sampson, P. (1994). A space-time analysis of ground-level ozone data. *Environmetrics 5*(3), 241–254.

Hastie, T. and Tibshirani, R. (1990). *Generalized Additive Models*. London: Chapman and Hall.

Huang, H. and Cressie, N. (1996). Spatio-temporal prediction of snow water equivalent using the kalman filter. *Computational Statistics & Data Analysis 22*(2), 159–175.

Hurvich, C., Simonoff, J., and Tsai, C. (1998). Smoothing parameter selection in nonparametric regression using an improved akaike information criterion. *Journal of the Royal Statistical Society Series B-Statistical Methodology 60*, 271–293.

Kammann, E. E. and Wand, M. P. (2003). Geoadditive models. *Applied Statistics 52*, 1–18.

Marx, B. and Eilers, P. (1998). Direct generalized additive modeling with penalized likelihood. *Computational Statistics & Data Analysis 28*(2), 193–209.

Money, E., Carter, G. P., and Serre, M. L. (2009). Using river distances in the space/time estimation of dissolved oxygen along two impaired river networks in new jersey. *Water Research 43*(7), 1948–1958.

O'Donnell, D. (2012). *Spatial Prediction and Spatio-Temporal Modelling on River Networks*. Ph. D. thesis, The University of Glasgow.

Peterson, E. E., Merton, A. A., Theobald, D. M., and Urquhart, N. S. (2006). Patterns of spatial autocorrelation in stream water chemistry. *Environmental Monitoring and Assessment 121*(1-3), 571–596.

Peterson, E. E. and Urquhart, N. S. (2006). Predicting water quality impaired stream segments using landscape-scale data and a regional geostatistical model: A case study in maryland. *Environmental Monitoring and Assessment 121*(1-3), 615–638.

Peterson, E. E. and Ver Hoef, J. M. (2010). A mixed-model moving-average approach to geostatistical modeling in stream networks. *Ecology 91*(3), 644–651.

R Development Core Team (2011). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing.

Reyjol, Y., Fischer, P., Lek, S., Rosch, R., and Eckmann, R. (2005). Studying the spatiotemporal variation of the littoral fish community in a large prealpine lake, using self-organizing mapping. *Canadian Journal of Fisheries and Aquatic Sciences 62*(10), 2294–2302.

Ruppert, D., Wand, M. P., and Carroll, R. (2003). *Semiparametric regression*. London: Cambridge University Press.

Schimek, M. G. (Ed.) (2000). *Smoothing and Regression: approaches, computation, and application*. New York: John Wiley.

Shaddick, G. and Wakefield, J. (2002). Modelling daily multivariate pollutant data at multiple sites. *Journal of the Royal Statistical Society Series C-Applied Statistics 51*, 351–372.

Thorp, J., Thoms, M., and Delong, M. (2006). The riverine ecosystem synthesis: Biocomplexity in river networks across space and time. *River Research and Applications 22*(2), 123–147.

Ver Hoef, J. M., Peterson, E., and Theobald, D. (2006). Spatial statistical models that use flow and stream distance. *Environmental and Ecological Statistics 13*(4), 449–464.

Ver Hoef, J. M. and Peterson, E. E. (2010). A moving average approach for spatial statistical models of stream networks. *Journal of the American Statistical Association 105*(489), 6–18.

Wood, S. (2006). *Generalized Additive Models: an introduction with R*. London: Chapman and Hall/CRC.