Data management of river water quality data: A semi-automatic procedure for data validation

L. Clement,[1] O. Thas,[1] J. P. Ottoy,[1] and P. A. Vanrolleghem[1,2]

Received 18 May 2006; revised 12 February 2007; accepted 16 May 2007; published 29 August 2007.

[1] Monitoring networks typically generate large amounts of data. Before the data can be added to the database, they have to be validated. In this paper, a semi-automatic procedure is presented to validate river water quality data. On the basis of historical data, additive models are established to predict new observations and to construct prediction intervals (PI's). A new observation is accepted if it is located in the interval. The coverage of the prediction intervals and its power to detect anomalous data are assessed in a simulation study. The method is illustrated on two case studies in which the method detected abnormal nitrate concentrations in the water body provoked by a dry summer which was followed by an extreme winter period. The case studies also show that similar to classical multivariate outlier detection tools, the semi-automatic procedure allows the detection of suspicious observations lying at the edges as well as observations lying at the center of the univariate distribution of the observations, but, without having to impose linear relationships typically associated with these classical methods.

Citation: Clement, L., O. Thas, J. P. Ottoy, and P. A. Vanrolleghem (2007), Data management of river water quality data: A semi-automatic procedure for data validation, *Water Resour. Res.*, *43*, W08429, doi:10.1029/2006WR005187.

1. Introduction

[2] The European Water Framework Directive (WFD, 22 December 2000) is one of the driving forces in environmental policy in Europe. The WFD's overall environmental objective is the achievement of "good status" for all of Europe's surface- and groundwaters within a 15-year period. The development of monitoring networks is a crucial step in the implementation of the WFD. They are for instance needed for a coherent and comprehensive overview of the water status, to identify pressures on water systems, as a warning system for detecting negative changes in the water quality and to detect trends. Due to the complex nature of water systems and the large amount of data that is collected, there is a clear need for models and Internet and Communication Technology (ICT) tools to assist the implementation of the WFD.

[3] Monitoring networks typically generate large amounts of data. These data have to be validated before they can be added to the database. On the one hand, there can be a problem with the quality of the data, due to errors during the analysis in the laboratory, wrong calibration of the equipment, or to error while entering the data. On the other hand, it is possible that there is a change in the system that causes changes in the water quality. The large amount of water quality data and its complex nature make it difficult for experts to validate all incoming data. Data validation is a clear example where ICT tools could be of great help to assist experts with the maintenance and analysis of monitoring databases compelled by the WFD.

[4] One way to deal with the validation problem is to use models to predict future measurements based on the historical data. In time series literature, this is called forecasting. The use of point forecasts to compare with incoming observations is meaningless if the extent of associated uncertainty is unknown. Interval forecasts should be used instead as they provide more information on future uncertainty and take the sampling variability present in the estimates correctly into account. These intervals, characterized by an upper and lower limit, correspond to a specified coverage probability [*Kim*, 1999; *Chatfield*, 1993]. In time series literature, Autoregressive Moving Average (ARMA) and Autoregressive Integrated Moving Average (ARIMA) models are mainly used. In order to obtain stationarity, trends and seasonal variation have to be eliminated first. Subsequently, the ARMA model is fitted to the stationary residual time series [*Pourahmadi*, 2001]. The models are then used to compute a forecast and a forecast interval. To reduce the assumptions on the distribution of the residuals, bootstrap-based intervals were developed [*Kim*, 1999, 2004; *Clements and Taylor*, 2001; *Chatfield*, 1993].

[5] Another approach is to use techniques from signal processing and statistical outlier detection. In these fields, the terms "fault detection" and "outlier detection" is used. History-based methods that require a large amount of historical data are commonly used. They consist of neural networks or multivariate statistical techniques, mainly based on Mahalanobis distances and principal component analysis (PCA) [*Venkatasubramanian et al.*, 2003; *Penny*, 1996]. The multivariate methods imply the presence of a constant number of variables measured simultaneously. However, in many databases, not all variables are measured at each time

[1]Department of Applied Mathematics, Biometrics and Process Control, Ghent University, Gent, Belgium.
[2]modelEAU, Département de génie civil, Université Laval, Québec, Canada.

0043-1397/07/2006WR005187$09.00

instant. So the use of multivariate methods would only be applicable to a selection of variables.

[6] Moreover, classical techniques used in time series analysis and multivariate statistics are highly parametric and are imposing several assumptions on the data, such as Gaussian residuals, homoscedasticity, and linear relationships between the different variables. Unfortunately, like other environmental data, water quality data typically possess a nonlinear nature [e.g., *Dominici et al.*, 2002; *Wood and Augustin*, 2002; *Cai and Tiwari*, 2000; *McMullan et al.*, 2003]. Hence linear models are not well suited to model the relationships between water quality variables. As an additional difficulty, it is likely that the water quality is susceptible to change due to, for example, the environmental regulation that is becoming more stringent. This implies the use of techniques with an appropriate adaptivity and flexibility to enable data validation under structural changes. When techniques of time series analysis are used, the model structure of the trend, seasonal variation, and ARMA models has to be changed from data set to data set and over time. This leads to an additional investment in time and people since time series modeling requires experience and expert knowledge.

[7] In this paper, a semi-automatic data validation procedure is proposed. A new observation at time $n + 1$ is compared with a prediction interval (PI). If the new observation is included in PI, the observation is accepted. Otherwise, the observation can be passed on to an expert for further evaluation. To deal with the nonlinear character of the data and to enable an appropriate flexibility of the method toward changes in the process, nonparametric additive models (AMs) are proposed. Analytical and bootstrap-based PIs are proposed in this study. In contrast to techniques from time series analysis, the procedure is entirely nonparametric when bootstrapping is used. This reduces the number of assumptions that have to be made considerably.

[8] First, the data will be introduced in section 2. In section 3, the methodology will be presented. Section 4 consists of an illustration of the entire procedure on a real data case, a simulation study to check the coverage of the derived intervals, a power study, and two case studies where two years of data are validated. Finally, the conclusions are summarized in section 5.

2. Description of the Data

[9] In the region of Flanders (Belgium), the Flemish Environmental Agency (VMM) established several monitoring networks. The physico-chemical monitoring network covers 1425 sampling locations distributed over the different catchments. Each sampling location is evaluated 12 to 26 times a year on a basic spectrum of physico-chemical variables: water temperature, dissolved oxygen (DO), pH, chemical oxygen demand (COD), nitrogen compounds, phosphorus, chloride, and conductivity. All these data are stored in a database, which is also managed by the VMM. Data can be classified according to their catchment and the Yzer basin is considered in this study. On a monthly basis, grab samples are taken at every sampling location. The nitrate measurements of the Yzer in 2003 and 2004 will be validated in section 4.

[10] The Yzer Valley is a typical lowland river, located in a polder area. A map of the Yzer catchment indicating the sampling locations maintained by the VMM is given in Figure 1. The total area of the catchment is 1101 km^2. The stream length is 76 km and 44 km of it is located in Belgium. At the French border, the river is relatively small, between 8 and 10 m. The river gets gradually wider to reach a width of 20 to 25 m near its mouth at Newport, Belgium. The river enters the North Sea by a complex of sluices. In Belgium, the river can be subdivided in three major parts. Part I is an area where the river is more or less in its original state. In part II, the river is straightened and has marshes to its right side. In part III, the river has artificial dammed banks [*De Rycke et al.*, 2001].

[11] The river is subjected to eutrophication due to the high nitrate and phosphate concentrations. A major source of nutrient pollution is the intensive agricultural activity at the river banks. Besides the agricultural pollution, other sources are from an industrial origin and from untreated sewage discharged by households.

3. Methods

[12] In order to validate new observations, the historical data are used to fit a nonparametric additive model presented in section 3.1. This model is subsequently used to construct prediction intervals. These intervals give the boundaries of the region where new measurements can be expected with a pre-specified probability. Section 3.2 deals with the derivation of the prediction intervals. Finally, section 3.3 describes diagnostic plots to detect possible causes of data rejection by our validation procedure. The methodology is illustrated by examples on data originating from sampling location 913000 along the Yzer River. Measurements between January 1990 and December 2002 are considered as historical data. A scatterplot of the evolution of nitrate over time is given in Figure 2.

3.1. Additive Models for Water Quality Data

[13] Suppose we have n observations of the response Y, which are denoted by $\mathbf{y} = (y_1, \ldots, y_n)^T$ and that measured simultaneously with the p predictor vectors $\mathbf{x_j} = (x_{1j}, \ldots, x_{nj})^T$, $j = 1, \ldots, p$, then a typical water quality data set $\mathbf{D}$ is represented by $\mathbf{D} = (\mathbf{x}_1, \ldots, \mathbf{x}_p, \mathbf{y})$. A general framework to model the relationships between the mean of Y and its predictors X can be written in the following form,

$$Y = m(X_1, \ldots, X_p) + \epsilon, \tag{1}$$

where m is the unknown regression function and ϵ is a zero mean random term. The data analyst now has to choose a certain structural form to approximate the conditional mean $m(X_1, \ldots, X_p)$. This can be done in a parametric, nonparametric, or semiparametric way. A well-known example of a fully parametric model is the standard multiple linear regression model. Because the relationships between the response and the predictors are assumed to be linear, equation (1) can be written as

$$Y = \alpha + \beta_1 X_1 + \cdots + \beta_p X_p + \epsilon = \alpha + \sum_{j=1}^{p} \beta_j X_j + \epsilon, \tag{2}$$

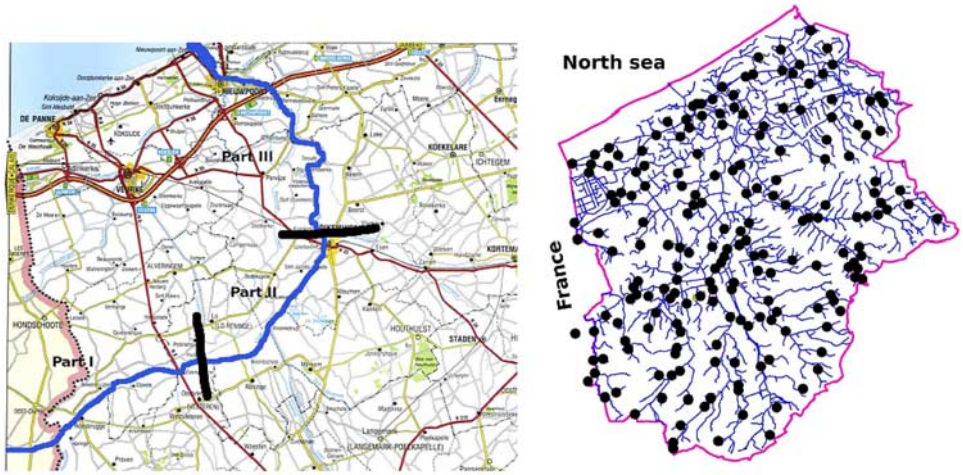

Figure 1. The Yzer catchment: in the left panel, the main river is shown; in the right panel, the entire catchment is shown along with the sampling locations of the VMM (indicated with black circles).

with the parameters α and $\beta^T = (\beta_1, \ldots, \beta_p)$. To fit the model to the data, the parameters have to be tuned so that the fitted values $\hat{\mathbf{y}} = \hat{\alpha} + \sum_{j=1}^{p} \hat{\beta}_j \mathbf{x}_j$ are in some sense as close as possible to the observed values $\mathbf{y}$ (i.e., by the use of least squares). The popularity of linear models is largely due to their simplicity and ease of interpretation. However, the model depends on a strong assumption of linearity between the predictors and the response. Unfortunately, water quality data typically possess a nonlinear nature [e.g., *Cai and Tiwari*, 2000; *McMullan et al.*, 2003]. Therefore it would be better to let the data drive the specification of the functional relation between the predictor variables and the response. This is exactly what scatterplot smoothers do for the two-dimensional case (Y, X_1). They model Y as $Y = f_1(X_1) + \epsilon$, where $f_1(X_1)$ is a smooth function used to approximate the underlying function $m(X_1)$ without imposing a rigid parametric relationship such as in the linear model. A principle used by many smoothers is to estimate the regression surface locally instead of globally. The fit at a certain predictor value x_i is only based on the data that lies in a certain neighborhood of x_i. This adds much more flexibility to the estimation of the underlying function. An example is the loess smoother [*Cleveland and Devlin*, 1988] (Figure 2), which indicates an increase in the nitrate level between January 1990 and December 1997, and a steady decrease in the average nitrate concentration afterward. The linear regression line remains more or less constant over the entire temporal domain because it cannot handle slope changes.

[14] Scatterplot smoothing can be easily extended to the p-dimensional case [e.g., *Cleveland and Devlin*, 1988; *Cleveland and Grosse*, 1991; *Loader*, 1999] where $m(X_1, \ldots, X_p)$ is approximated by a p-dimensional smoother $f_{1,\ldots,p}(X_1, \ldots, X_p)$. Note that the number of dimensions equals the number of regressors. There are unfortunately some problems related to multidimensional smoothers. In particular,

[15] 1. *Buja et al.* [1989] showed that most multidimensional extensions of univariate smoothers are not attractive from a computational point of view.

[16] 2. Due to their multivariate nature, multivariate smoothers also suffer from "the curse of dimensionality". These problems are mainly triggered by the multidimensional neighborhoods which have to be defined. *Hastie et al.* [2001] illustrated that the neighborhoods are less local when the number of predictors increases. Another issue is related to the data sparseness in a high dimensional setting where more data ends up in the boundary region. Since smoother estimates are known to be more biased in the boundary regions, the boundary problem is more dominant in a multidimensional setting.

[17] 3. It is difficult to define a sensible metric for the multidimensional neighborhoods, because the predictors are often measured in different units.

[18] 4. The visualization of multivariate smoothers is less obvious. Especially when the number of predictors is greater than 2. In order to study the effects of the individual predictors, projections from the hypersurface can be made on a lower dimensional space, but this projection depends on the fixed values of the remaining predictors and thus they are rather noisy.

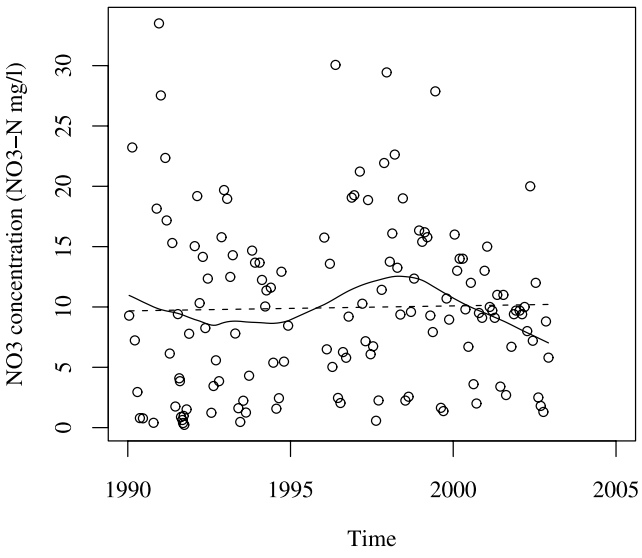

Figure 2. Scatterplot of nitrate concentration as a function of time together with a least squares regression line (dashed line) and a loess smoother (solid line).

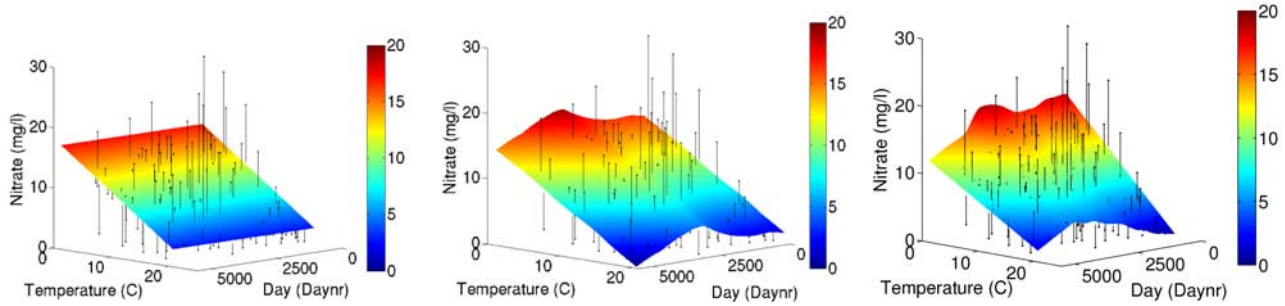

Figure 3. Nitrate concentration as a function of time (day number) and temperature (°C). In the left panel, nitrate is modeled using a linear model; in the middle panel, an additive model is presented built up by two univariate local linear regression smoothers; and in the right panel, a two-dimensional local linear regression smoother is used.

[19] To overcome the abovementioned problems, *Buja et al.* [1989] proposed an alternative approach. They suggested one-dimensional smoothers as additive building blocks of the model, resulting in a more restricted class of nonparametric regression models, also referred to as additive models. Additive models extend standard linear models and approximate the unknown regression surface $m(X_1, \ldots, X_p)$ by,

$$Y = \alpha + \sum_{j=1}^{p} f_j(X_j) + \epsilon, \tag{3}$$

where f_j can be any function; however, in most cases, smoothers are used. Similar to linear models, the effect of a predictor on the fitted response surface does not depend on the values of the other predictors. But, additive models are much more flexible because they are not necessary linearly additive in the covariates X.

[20] Thus, the contribution of each predictor can still be studied individually. This enables the user to decompose the model in each of its smooth functions, which can be graphically depicted. Figure 3 shows the difference between a linear model $Y = \alpha + \beta_1 X_1 + \beta_2 X_2 + \epsilon$, an additive model $Y = \alpha + f_1(X_1) + f_2(X_2) + \varepsilon$ and a multivariate regression smoother $Y = f_{12}(X_1, X_2) + \epsilon$, where Y is the nitrate concentration, X_1 represents the time, and X_2 is the temperature. Due to the additivity assumption, the bump at low temperatures and at intermediate dates is less high for the additive model than in the multivariate smoother model. However, the bump is situated in a data-sparse region and might be a boundary effect from the multivariate smoother. Apart from this feature, the fits by the additive model and the multivariate smoother look similar. The additive model however enables the analyst to look at the contribution of each predictor separately. This is illustrated in Figure 4. By the end of 1992, the contribution of the long-term trend shows a steep incline and reaches a maximum at the beginning of 1998 and it decreases afterward. The inverse relation between temperature and nitrate is also obvious. Because local polynomial smoothers are used as the basic building block of the additive models in this paper, a brief review on local polynomial smoothing is needed before we can move on to model fitting and selection.

3.1.1. Local Polynomial Smoothing

[21] *Hastie and Tibshirani* [1990] defined a smoother as a tool for summarizing the trend of a response Y as a function of one or more predictors $X_1, \ldots, X_p$. It produces an estimator which is less variable than Y itself and can be used for several purposes. In this paper, the focus is on its use to estimate a regression surface, without resorting to a parametric class of functions. Because only univariate smoothers are used in additive models, the overview on local polynomial regression is restricted to the univariate case. Since they only contain one predictor, the model can be written as

$$Y = m(X) + \epsilon. \tag{4}$$

The idea of local polynomial regression can easily be motivated by approximating the regression function m in a neighborhood of x_0 by a Taylor expansion,

$$m(x) \approx f(x) = m(x_0) + \sum_{k=1}^{q} \frac{m^{(k)}(x_0)}{k!}(x - x_0)^k, \tag{5}$$

where $m^{(k)}(x_0) = \dfrac{\partial^k m}{\partial x^k}(x_0)$. Local weighted least squares can be used to fit this polynomial minimizing

$$\sum_{i=1}^{n} \left[y_i - \sum_{k=0}^{q} \beta_k (x_i - x_0)^k \right]^2 K\left(\frac{x_i - x_0}{h}\right), \tag{6}$$

where $K(\cdot)$ is a kernel function which will be introduced later on and h is the bandwidth which defines the size of the neighborhood $(x_0 - h, x_0 + h)$. The kernel function assigns

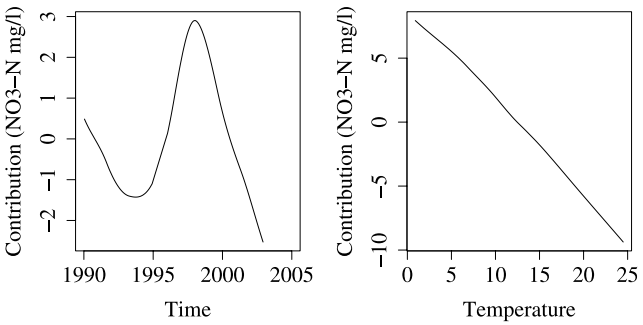

Figure 4. Contribution of (left) long-term trend and (right) temperature to the nitrate concentration predicted by the additive model in Figure 3.

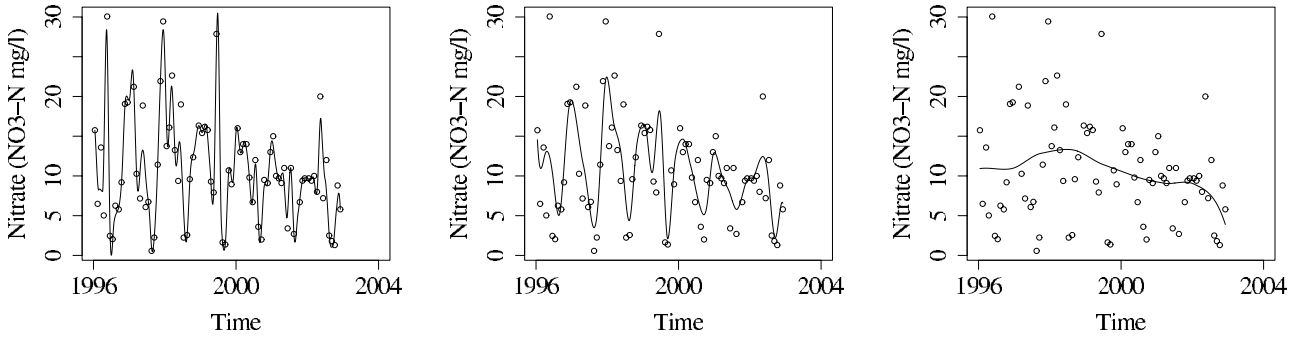

Figure 5. Fit of the nitrate data with local linear regression with bandwidth equal to (left) 2 months, (middle) 4 months, and (right panel) 2 years.

weights to each observation. The solution to this local weighted least squares problem is

$$\hat{\beta}_{\mathbf{0}} = \left(\mathbf{x}_{\mathbf{c}}^{\mathbf{T}}\mathbf{W_0}\mathbf{x_c}\right)^{-1}\mathbf{x}_{\mathbf{c}}^{\mathbf{T}}\mathbf{W_0}\mathbf{y}, \tag{7}$$

where $\mathbf{x_c} = (\mathbf{1}, \mathbf{x_{vc}}, \ldots, \mathbf{x_{vc}^q})$, $\mathbf{1} = (1, \ldots, 1)^T$, $\mathbf{x_{vc}} = (x_1 - x_0, \ldots, x_n - x_0)^T$, and $\mathbf{W_0}$ is a diagonal matrix build up by the kernel weights [*Fan and Gijbels*, 1996]. The response y_0 corresponding to x_0, is then estimated by

$$\begin{aligned}\hat{y}_0 &= [1 \quad 0 \quad \ldots \quad 0]\hat{\beta}_{\mathbf{0}} \\ &= [1 \quad 0 \quad \ldots \quad 0]\left(\mathbf{x}_{\mathbf{c}}^{\mathbf{T}}\mathbf{W_0}\mathbf{x_c}\right)^{-1}\mathbf{x}_{\mathbf{c}}^{\mathbf{T}}\mathbf{W_0}\mathbf{y} \\ &= \mathbf{S_0}\mathbf{y},\end{aligned} \tag{8}$$

where the centered vector of x_0 is

$$[1 (x_0 - x_0) \cdots (x_0 - x_0)^q] = [1\ 0\ \ldots\ 0]. \tag{9}$$

Hence the fit of local polynomial smoothers is a linear combinations of the response. If this procedure is performed for all observations (x_i, y_i) belonging to the data set $(\mathbf{x},\mathbf{y})$, the fit $\hat{\mathbf{y}}$ can be written as

$$\hat{\mathbf{y}} = \mathbf{S}\mathbf{y}, \tag{10}$$

where $\mathbf{S}$ is the smoother matrix. Smoothers that can be written as a linear combination of the response belong to the class of linear smoothers. When similar assumptions are made as in the parametric regression framework, linear smoothers inherit a whole set of inference procedures known from the classical parametric regression context such as the construction of confidence intervals [*Cleveland and Devlin*, 1988]. Two important examples of such assumptions are Gaussian residuals and an unbiased estimation of m by the function $\hat{f}(x)$.

[22] Several important choices have to be made before local polynomial regression can be used. The size of the bandwidth has to be selected, but a practical procedure for bandwidth selection is kept for the next paragraph. The degree of the polynomial has to be set. Since the bias is mainly controlled by the bandwidth, the choice of the degree of the local polynomial is less important. However, for a fixed bandwidth, increasing the degree will reduce the bias, but this is at the expense of an increasing variance of the fit and of a higher computational cost. *Fan and Gijbels* [1996] proved that the variance of the fit does not increase by going from an even order polynomial fit to an odd order polynomial fit. The extra parameter can however reduce the bias significantly. They also argue that even order fits suffer from serious boundary effects, in contrast with odd order fits which have nice adaptive boundary properties. From this point of view, Fan and Gijbels recommended to use the lowest possible odd order for the polynomial fit. Hence here the degree is set to 1. Another question to be addressed is the choice of the kernel function K. The choice of the kernel is not that important from a practical point of view. However, Fan and Gijbels have shown that the Epanechnikov kernel, $K(u) = 3/4\ (1 - u^2)$ for $-1 < u < 1$ and 0 for u outside that range, is asymptotically optimal for the interior of the domain. This kernel is used in the remainder.

[23] The bandwidth and the kernel function K control the size of the local neighborhood. Therefore the choice of the bandwidth in local polynomial regression is a crucial one. When taking the bandwidth close to zero, the data is interpolated, resulting in an overparametrized model. A bandwidth taken arbitrarily large, results in a polynomial of degree p which is fitted globally. Hence the bandwidth is a key element in controlling the complexity of the smoother. The smaller the bandwidth, the more degrees of freedom that can be used for controlling the bias. But this reduction in bias does not come for free. Smaller bandwidths also lead to an increase of the variance associated with the estimates. Too small bandwidths typically result in more wiggly curves and this can conceal the main features which are present in the data. Too large bandwidths, on the other hand, tend to oversmooth the data and can introduce a substantial bias. This is illustrated in Figure 5 where nitrate data are modeled using a local linear smoother and three different bandwidths. When a bandwidth of two months is taken, the curve is very wiggly and highlights features which may be inherent to the sampling variability. A bandwidth of 4 months still highlights a cyclic pattern in the nitrate concentrations but produces a smoother fit. A large bandwidth is sensitive to oversmoothing, leading to an estimate which can miss certain features of the curve. A bandwidth of 2 years, for example, loses the ability to pick up the cyclic behavior of the nitrate concentration. The size of the bandwidth can be chosen to be constant over the domain of X, or can be variable. An example of a variable bandwidth with a very simple nature is the nearest-neighbor bandwidth. The selector requires that a fixed percentage of the data is included in the neighborhood. This percentage is

referred to as the span s. It automatically adapts the amount of smoothing to the local situation, using small bandwidths in a dense design region and large bandwidths in sparse regions [*Altman*, 1992; *Fan and Gijbels*, 1996; *Loader*, 1999]. This method thus prevents that the regression in sparse data regions is based on only a limited number of points. The span s is used in this study to control the size of the bandwidth h. To select a good value for the span s, a criterion that quantifies the trade off between the amount of bias and the associated variance of the estimator is needed. An example of such a criterion is the generalized cross validation (GCV)

$$\mathrm{GCV}(s) = 1/n \sum_{i=1}^{n} \left\{ \frac{y_i - \hat{f}_s(x_i)}{1 - \mathrm{tr}(\mathbf{S})/n} \right\}^2, \tag{11}$$

where $\sum_{i=1}^{n} \left\{ y_i - \hat{f}_s(x_i) \right\}^2$ is a measure for the amount of bias and $\mathrm{tr}(\mathbf{S})$ is a measure for the degrees of freedom used by the model. Both an increased bias and an increase in model complexity lead to a higher GCV. Thus the model with the lowest GCV has to be selected in order to control the bias without using to many degrees of freedom.

[24] Literature on the attractiveness and advantages of local linear regression smoothers can be found in the works of *Cleveland* [1979], *Cleveland and Devlin* [1988], *Fan* [1992, 1993], *Hasti and Loader* [1993], *Fan and Gijbels* [1996], and *Loader* [1999].

3.1.2. Fitting Additive Models

[25] In practice, the backfitting algorithm proposed by *Buja et al.* [1989] is the most widely used method to estimate the additive components. From equation (3), it is obvious that each function can be written as:

$$f_j(X_j) = Y - \alpha - \sum_{k \neq j} f_k(X_k) + \epsilon. \tag{12}$$

When f_j is a linear smoother with smoother matrix $\mathbf{S_j}$ and in the hypothetical case that the other predictor terms are known, f_j can be estimated as

$$\hat{\mathbf{f}}_\mathbf{j} = \mathbf{S_j} \left\{ \mathbf{y} - \alpha - \sum_{k \neq j} \mathbf{f_k} \right\}, \tag{13}$$

where $\mathbf{f_k}$ is the vector $(f_k(x_{1k}), \ldots, f_k(x_{nk}))^T$ and $\boldsymbol{\alpha}$ is a vector $(\alpha, \ldots, \alpha)^T$. When only linear smoothers are used in the model, a similar expression can be used for each smoother. By combining all these expressions, the following set of equations has to be solved,

$$\begin{bmatrix} \mathbf{I_n} & \mathbf{S_1} & \mathbf{S_1} & \cdots & \mathbf{S_1} & \mathbf{1} \\ \mathbf{S_2} & \mathbf{I_n} & \mathbf{S_2} & \cdots & \mathbf{S_2} & \mathbf{1} \\ \vdots & \vdots & \vdots & & \vdots & \vdots \\ \mathbf{S_p} & \mathbf{S_p} & \mathbf{S_p} & \cdots & \mathbf{I_n} & \mathbf{1} \\ \mathbf{1}/n & \mathbf{1}/n & \mathbf{1}/n & \cdots & \mathbf{1}/n & 1 \end{bmatrix} \begin{bmatrix} \mathbf{f_1} \\ \mathbf{f_2} \\ \vdots \\ \mathbf{f_p} \\ \alpha \end{bmatrix} = \begin{bmatrix} \mathbf{S_1} \\ \mathbf{S_2} \\ \vdots \\ \mathbf{S_p} \\ \mathbf{1}/n \end{bmatrix} \mathbf{y} \tag{14}$$

where $\mathbf{1}$ is the vector $(1, \ldots, 1)^T$. The backfitting algorithm solves this set of equations iteratively. In the lth iteration, $\mathbf{f_j^{(l-1)}}$ is updated by

$$\mathbf{f_j^{(l)}} = \mathbf{S_j} \left(\mathbf{y} - \alpha - \sum_{k<j} \mathbf{f_k^{(l)}} - \sum_{k>j} \mathbf{f_k^{(l-1)}} \right). \tag{15}$$

In order to make each function identifiable, an additional constrained has to be introduced, $\sum_{i=1}^{n} f_j(x_{ij}) = 0$. This is simply done by replacing each $\mathbf{S_j}$ in equations (13) (14) (15) by the centered smoother matrix $\mathbf{S_j^*} = (\mathbf{I_n} - \mathbf{1}\mathbf{1}^\mathbf{T}/n)\mathbf{S_j}$. This also forces α to be estimated by the sample mean $\bar{\mathbf{y}}$. In the next section the model uncertainty will be assessed by providing a formula for variance and confidence interval estimation.

3.1.3. Variance and Pointwise Confidence Intervals

[26] In classical parametric statistics, a variance estimate is the key element for statistical inference. Similar to linear regression, the residual sum of squares (RSS) can be used for variance estimation. Here RSS is equal to $\mathrm{RSS} = (\mathbf{y} - \hat{\mathbf{y}})^T (\mathbf{y} - \hat{\mathbf{y}})$. In linear regression with p predictors, the variance estimate then simply becomes $\hat{\sigma}^2 = \mathrm{RSS}/df$, where its degrees of freedom (df) is equal to $n - p - 1$. In the context of linear regression smoothers, we have already used the trace of the smoother matrix, $\mathrm{tr}(\mathbf{S})$, as a definition of the degrees of freedom. *Hastie and Tibshirani* [1990], however, showed that it is better to use another definition for the degrees of freedom of the RSS. In the next paragraph, their definition is explained in some more detail.

[27] When all components of the AM are linear or linear smoothers, there is a projection matrix $\mathbf{H}$ so that $\hat{\mathbf{y}} = \mathbf{H}\mathbf{y}$. For nonparametric AMs using linear smoothers, the additive component functions can be solved by a set of normal equations presented in equation (14). Equation (14), which can thus also be written as

$$\hat{\mathbf{P}}\mathbf{f} = \hat{\mathbf{Q}}\mathbf{y}, \tag{16}$$

But, as *Opsomer* [2000] mentioned, it is possible, at least conceptually, to write the estimators directly as

$$\hat{\mathbf{f}} = \hat{\mathbf{P}}^{-1}\hat{\mathbf{Q}}\mathbf{y}, \tag{17}$$

and after obtaining $\hat{\mathbf{P}}^{-1}$, $\hat{\mathbf{y}}$ can be written as

$$\begin{aligned} \hat{\mathbf{y}} &= [\mathbf{I_n} \quad \mathbf{I_n} \quad \mathbf{I_n} \quad \cdots \quad \mathbf{I_n} \quad \mathbf{1}]\,\hat{\mathbf{f}} \\ &= [\mathbf{I_n} \quad \mathbf{I_n} \quad \mathbf{I_n} \quad \cdots \quad \mathbf{I_n} \quad \mathbf{1}]\,\hat{\mathbf{P}}^{-1}\hat{\mathbf{Q}}\mathbf{y} \\ &= \mathbf{H}\mathbf{y}, \end{aligned} \tag{18}$$

where $\mathbf{I_n}$ is the $n \times n$ identity matrix. From this derivation, it is clear that an additive model using linear smoothers can be considered as a linear smoother with a projection matrix $\mathbf{H}$. If such a projection matrix exists, it can be shown that the RSS has the expectation $E[\mathrm{RSS}] = \{n - \mathrm{tr}(2\mathbf{H} - \mathbf{H}\mathbf{H}^\mathbf{T})\}\sigma^2 + \mathbf{b}^\mathbf{T}\mathbf{b}$, where $\mathbf{b}$ is the bias [*Hastie and Tibshirani*, 1990].

Thus, when the bias is negligible, the variance can be estimated by

$$\hat{\sigma}^2 = \frac{\mathrm{RSS}}{n - \mathrm{tr}(2\mathbf{H} - \mathbf{H}\mathbf{H}^{\mathbf{T}})}, \tag{19}$$

where, in analogy with linear regression, the degrees of freedom of the errors can be defined as $df^{\mathrm{err}} = n - \mathrm{tr}(2\mathbf{H} - \mathbf{H}\mathbf{H}^{\mathrm{T}})$.

[28] When the residuals are i.i.d, the estimate of variance-covariance matrix of $\hat{\mathbf{y}}$ can be calculated as

$$\boldsymbol{\Sigma}_{\hat{\mathbf{y}}} = \mathbf{H}\mathbf{H}^{\mathbf{T}}\hat{\sigma^2}. \tag{20}$$

Similar to $\hat{\mathbf{y}}$, a projection matrix $\mathbf{H_j}$ can be defined for each component $\hat{\mathbf{f}}_{\mathbf{j}} = \mathbf{H_j}\mathbf{y}$. The variance-covariance matrix of each component is simply obtained by replacing $\mathbf{H}$ in equation (20) by $\mathbf{H_j}$.

[29] The calculation of $\hat{\mathbf{P}}^{-1}$ is computationally unattractive since it involves inverting a $(np) \times (np)$ matrix. Moreover, the inverse of $\hat{\mathbf{P}}$ does not always exists. Recently, Giannitrapani et al. [Additive models for correlated data with applications to air pollution monitoring, submitted to *Biometrics*, 2005] provided a simple method to keep track of the important projection matrices while the backfitting algorithm proceeds. When using linear smoothers, in the lth iteration step the estimate of each component $\mathbf{f}_{\mathbf{j}}^{(\mathbf{l})}$ can be written as $\mathbf{f}_{\mathbf{j}}^{(\mathbf{l})} = \mathbf{H}_{\mathbf{j}}^{(\mathbf{l})}\mathbf{y}$. Hence the backfitting scheme can be expressed as

$$\mathbf{H}_{\mathbf{j}}^{(\mathbf{l})} = \mathbf{S}_{\mathbf{j}}^{*}\left(\mathbf{I_n} - \sum_{k<j}\mathbf{H}_{\mathbf{k}}^{(\mathbf{l})} - \sum_{k>j}\mathbf{H}_{\mathbf{k}}^{(\mathbf{l-1})}\right), \tag{21}$$

where $\mathbf{I_n}$ is the $n \times n$ identity matrix. At each stage, the updated projection matrix $\mathbf{H}_{\mathbf{j}}^{(\mathbf{l})}$ remains independent of $\mathbf{y}$. When the backfitting algorithm has converged, a set of projection matrices $\{\mathbf{H_j}, j = 1, \ldots, p\}$ is obtained. They can be used to estimate the individual components $\mathbf{f_j} = \mathbf{H_j}\mathbf{y}$ and the fitted values $\hat{\mathbf{y}} = \mathbf{H}\mathbf{y}$, where $\mathbf{H} = \mathbf{1}\mathbf{1}^{\mathbf{T}}/n + \sum_{j=1}^{n}\mathbf{H_j}$.

[30] The variance estimates of the estimators $\hat{\mathbf{y}}$ can now be used for construction of approximate $(1 - \alpha)$ confidence intervals. Here the interval will be only given explicitly for the estimator $\hat{y}_i$,

$$\left[\hat{y}_i - z_{\left(1-\frac{\alpha}{2}\right)}\hat{\sigma}_{y_i},\ \hat{y}_i + z_{\left(1-\frac{\alpha}{2}\right)}\hat{\sigma}_{y_i}\right], \tag{22}$$

where $z_{\left(1-\frac{\alpha}{2}\right)}$ is the $\left(1-\frac{\alpha}{2}\right)$ percentile of the standard normal distribution and $\hat{\sigma}^{y_i}$ is the square root of the ith diagonal element of $\hat{\mathbf{S}}_{\mathbf{y}}$. The formulation of confidence bands for the component functions $f_j(x_{ij})$ is trivial. We still have to keep in mind that the intervals are only correct when the bias is negligible. When this is not the case, the additive model fit $\hat{\mathbf{y}}$ is a fit for $\mathbf{Hm}$ rather than for the true underlying surface $\mathbf{m}$ evaluated at the design points [*Hastie and Tibshirani*, 1990]. The coverage of the confidence intervals of equation (22) depends upon the normality assumptions. To relax these assumptions, the bootstrap can be used as a nonparametric method for obtaining the confidence intervals. The bootstrap method will be introduced in section 3.2.2.

3.1.4. Model Selection

[31] The model selection can be performed in two stages: (1) span selection of the smoothing parameters $(s_1, \ldots, s_p)$ and (2) selection of variables in the model. As mentioned in section 3.2.1, nearest-neighborhood bandwidths are used for the local polynomial smoothers in the model. The spans are typically tuned by using classical criteria as the GCV. In principle, these methods require a multidimensional search to determine the optimal span for each of the smoothers conditional on the spans of the other components in the model. When the number of smoothers p rises, there is an exponential increase in the number of AMs to be evaluated. This procedure further has to be embedded in a procedure to select the number of predictors used. To ensure that the appropriate smoothing parameters are used, the smoothing parameters of each of the candidate models should be determined. When p gets large and a dense grid is used for the selection of the smoothing parameters, this approach quickly gets computationally demanding.

[32] *Hastie and Tibshirani* [1990] have introduced the BRUTO algorithm as a pragmatic solution to keep the computational burden limited. The algorithm is an adaptation of the backfitting algorithm so that it combines model fitting, smoothing parameter selection and model selection. To avoid computational problems, Hastie and Tibshirani adjusted the GCV criterion

$$\mathrm{GCV}(s_1, \ldots, s_p) = \frac{\sum_{i=1}^{n}\hat{\epsilon}_i^2}{n\left(1 - \mathrm{tr}\left(\mathbf{H}(s_1, \ldots, s_p)\right)/n\right)^2}, \tag{23}$$

to the modified GCV criterion

$$\mathrm{GCV}^b(s_1, \ldots, s_p) = \frac{\sum_{i=1}^{n}\hat{\epsilon}_i^2}{n\left(1 - \left[1 + \sum_{j=1}^{p}\left\{\mathrm{tr}\left(\mathbf{S_j}(s_j)\right) - 1\right\}\right]/n\right)^2}. \tag{24}$$

In this way, the computational difficulties associated with $\mathrm{tr}(\mathbf{H}(s_1, \ldots, s_p))$ are circumvented. But, as shown in the previous section, Giannitrapani et al. [Additive models for correlated data with applications to air pollution monitoring, submitted to *Biometrics*, 2005] provided a very simple method to keep track of the important projection matrices. Hence a modification of the GCV is not required since $\mathbf{H}^{(l)}$ is known at each step. Therefore we incorporate the original GCV criterion in the BRUTO algorithm.

[33] The BRUTO algorithm starts with the projection matrices $\mathbf{H_j} = 0$. In each iteration, one parameter s_j is selected. Hence the span selection is performed one smoothing parameter at a time while the other smoothing parameters remain unchanged. In particular, the s_j is adjusted which minimizes the global GCV. In the cycle (l), this is applied by using the appropriate smoothing parameter $s_j^{(l)}$ to update the projection matrix $\mathbf{H}_{\mathbf{j}}^{(\mathbf{l})}\left(s_j^{(l)}\right) = \mathbf{S}_{\mathbf{j}}^{*}\left(s_j^{(l)}\right)\left(I_n - \sum_{k\neq j}\mathbf{H}_{\mathbf{k}}^{(\mathbf{l-1})}\left(s_k^{(l-1)}\right)\right)$ while the other projection matrices are left unaltered. Hence each iteration only provides for an update of one smoothing parameter s_j and its corresponding projection matrix $\mathbf{H_j}(s_j)$.

The BRUTO algorithm is continued until the GCV converges. The convergence is guaranteed, because each iteration produces a decrease in the criterion. The BRUTO algorithm can easily be extended to incorporate model selection. When the GCV is allowed to be optimized by the selection of the null fit, $\mathbf{H_j} = \mathbf{0}$, it enables the removal of the associated explanatory variable from the model. Hence a particular variable can be included at a certain stage and it can be omitted from the model later on.

[34] For data validation purposes, the model should be able to adapt to changes in the system. Therefore model selection has to be performed on-the-fly. The optimal model at each sampling location is obtained by using a separate model for every sampling location and the model structure is also allowed to change over time by executing the BRUTO algorithm as soon as a new observation is added to the database.

3.2. Prediction Intervals

[35] In section 3.1, the method to model the historical data was presented. To validate new data, a prediction interval (PI) is needed. A PI, however, differs from the pointwise confidence intervals for the fitted values $\hat{\mathbf{y}}$ derived in section 3.1.3. A confidence interval reflects how accurate the mean is estimated. The data validation procedure, however, requires an interval estimate associated with the location of a single observation. Under the normality assumption, the conditional distribution of an observation given the covariates is $N(m(\mathbf{x}),\sigma^2)$. Hence the prediction interval has to incorporate the model uncertainty due to the estimation of $m(\mathbf{x})$ and the additional variability associated with individual observations fluctuating around the mean.

[36] Two different approaches are presented to derive prediction intervals: an analytical procedure which only works for AMs with linear smoothers and assumes the errors to be Gaussian, and more general double bootstrap procedures. The latter are fully nonparametric and they can cope with any type of AM and non-Gaussian errors. Both methods assume that the residuals are independent. The data used in this study is based on monthly grab samples. When the water quality data is sampled at intervals larger than 2 weeks, its dependency is known to be only related to seasonality and trend [*Van Belle and Hughes*, 1984]. In case these dependencies are modeled accurately, the data can be assumed to be independent. Another assumption is that the bias of the estimator is negligible. However, in the presence of bias, the variance estimate is inflated and results in conservative interval estimates [Giannitrapani et al., Additive models for correlated data with applications to air pollution monitoring, submitted to *Biometrics*, 2005].

3.2.1. Analytical Prediction Intervals

[37] Before the analytical PIs can be constructed, an estimator of the variance of a new prediction is needed. As shown in section 3.1.3, a projection matrix exists when the AM is built up by linear smoothers. Thus the prediction by the smoother at a certain predictor value is always a linear combination of the observed values of the response. For the kth local linear smoother (first order polynomial), the prediction at time $n + 1$ is $[1\ 0]\ (\mathbf{x}_{\mathbf{k,c}}^{\mathbf{T}}\mathbf{W}_{\mathbf{k,n+1}}\mathbf{x}_{\mathbf{k,c}})^{-1}\mathbf{x}_{\mathbf{k,c}}^{\mathbf{T}}\mathbf{W}_{\mathbf{k,n+1}}\mathbf{y}$. Thus its corresponding (row)smoothing vector can be written as $S_{k,n+1} = [1\ 0](\mathbf{x}_{\mathbf{k,c}}^{\mathbf{T}}\mathbf{W}_{\mathbf{k,n+1}}\mathbf{x}_{\mathbf{k,c}})^{-1}\mathbf{x}_{\mathbf{k,c}}^{\mathbf{T}}\mathbf{W}_{\mathbf{k,n+1}}$. The centered smoothing (row)vector for the kth predictor at time $n + 1$ can be written as $\mathbf{S}_{\mathbf{k,n+1}}^{*} = \mathbf{S}_{\mathbf{k,n+1}} - \mathbf{1}^{\mathbf{T}}\mathbf{S}_{\mathbf{k}}/n$. Similar to equation (13), an estimate of the contribution of the kth predictor function $\hat{f}_{k,n+1}$ of the additive model is given by

$$\begin{aligned}\hat{f}_{k,n+1} &= \mathbf{S}_{\mathbf{k,n+1}}^{*}\left(\mathbf{y} - \alpha - \sum_{k\neq j}\hat{\mathbf{f}}_{\mathbf{j}}\right)\\ &= \mathbf{S}_{\mathbf{k,n+1}}^{*}\left(\mathbf{I_n} - \sum_{k\neq j}\mathbf{H_j}\right)\mathbf{y}\\ &= \mathbf{H_{k,n+1}}\mathbf{y}.\end{aligned} \tag{25}$$

The estimate of the mean response at time $n + 1$, $\hat{y}_{n+1}$, then becomes

$$\begin{aligned}\hat{y}_{n+1} &= \alpha + \sum_{j=1}^{p}\hat{f}_{j,n+1}\\ &= \left(\mathbf{1}^{T}/n + \sum_{j=1}^{p}\mathbf{H_{j,n+1}}\right)\mathbf{y}\\ &= \mathbf{H_{n+1}}\mathbf{y},\end{aligned} \tag{26}$$

and its variance is

$$\sigma_{\hat{y}_{n+1}}^{2} = \mathbf{H_{n+1}}\mathbf{H}_{\mathbf{n+1}}^{\mathbf{T}}\sigma^{2}. \tag{27}$$

This variance refers to the uncertainty associated with prediction of the mean of new observations at time $n + 1$, and not to the variance of a single observation which is typically fluctuating around the model mean. Hence the variance needed for the construction of a PI is decomposed into a part related to the uncertainty of the modeled mean, $\sigma_{\hat{y}_{n+1}}^{2}$ and into the part due to residual variance, σ^2. Thus, the variance needed for calculating a PI becomes

$$\sigma_{y_{n+1}}^{2} = \left(\mathbf{H_{n+1}}\mathbf{H}_{\mathbf{n+1}}^{\mathbf{T}} + 1\right)\sigma^{2}, \tag{28}$$

and σ^2 is estimated as in equation (19). After plugging this into equation (28), an approximate $1 - \alpha$ PI is given by

$$\left[\hat{y}_{n+1} - z_{\left(1-\frac{\alpha}{2}\right)}\hat{\sigma}_{y_{n+1}},\ \hat{y}_{n+1} + z_{\left(1-\frac{\alpha}{2}\right)}\hat{\sigma}_{y_{n+1}}\right], \tag{29}$$

and $z_{\left(1-\frac{\alpha}{2}\right)}$ is the critical value from the normal distribution. In the remainder of this paper, this PI is referred to as aPI.

3.2.2. Bootstrap Intervals

[38] In general, additive models do not have an analytical solution and the errors can deviate from normality. The calculation of the analytical intervals as described in section 3.2.1 only exists when linear smoothers are used as building blocks and their coverages are only correct when the errors are Gaussian. In this section, a procedure is proposed for the construction of the prediction intervals that can cope with additive models in general. The procedure does not impose any parametric assumptions of the underlying distribution of the errors. Therefore an analytical derivation does not exist for the PI. This implies the use of computational intensive methods for variance estimation such as bootstrapping.

[39] The bootstrap is a statistical inference technique that relies on only some weak distributional assumptions. Bootstrapping consists of resampling from a sample $\mathbf{D} = (\boldsymbol{D}_1, \ldots,$

$\boldsymbol{D_n}$), with replacement, to generate bootstrap replicates $\mathbf{D}^*(\mathbf{b})$, $b = 1, \ldots, B$, of the same size n. The bootstrap replicates are then used to simulate B estimates of a given statistic, resulting in an empirical probability distribution of the statistic. Suppose one wishes to estimate the empirical cumulative distribution function G^* of a statistic $\theta = t(\mathbf{D})$ which is estimated from a given sample $\mathbf{D} = (\mathbf{x_1}, \ldots, \mathbf{x_p}, \mathbf{y})$. Each observation $\mathbf{D_i}$ is sampled with replacement and with an equal probability of $1/n$. Sample $\mathbf{D}$ is resampled with replacement B times, until B bootstrap replicates $\mathbf{D}^*(\mathbf{b})$, $b = 1, \ldots, B$, are generated. With each bootstrap replicate $\mathbf{D}^*(\mathbf{b})$, the statistic θ can be evaluated, yielding B bootstrap estimates $\hat{\theta}^*(\mathrm{b})$. The acquired empirical distribution G^* can also be used to calculate for instance the variance or confidence intervals on $\hat{\theta}$.

[40] When applying the bootstrap in a regression context, there are two common approaches for generating bootstrap samples (1) by resampling the cases $\mathbf{D_i} = (x_{i1}, \ldots, x_{ip}, y_i)$ or (2) by resampling the errors ($\hat{\epsilon}_i$). The use of resampling cases is not really an option since it changes the sample design. Water quality data are gathered over time and so the time covariate is not sampled at random. The water quality data are sampled at intervals larger than 2 weeks. Therefore their dependencies are only related to seasonality and trend, and the residuals can be assumed to be independent after modeling these dependencies [*Van Belle and Hughes*, 1984]. These considerations provide a strong argument in favor of resampling residuals. In this case, bootstrap samples are generated by simply resampling from the empirical distribution of the residuals $\hat{F}$ and creating bootstrapped responses by

$$\mathbf{y}^*(\mathbf{b}) = \hat{\mathbf{y}} + \mathbf{e}^*(\mathbf{b}), \tag{30}$$

where $\mathbf{e}^*(\mathbf{b})$ is a bootstrap replicate of the residuals. A bootstrap data set is then constructed as follows $\mathbf{D}^*(\mathbf{b}) = (\mathbf{x_1}, \ldots, \mathbf{x_p}, \mathbf{y}^*(\mathbf{b}))$. The most straightforward method to obtain $\mathbf{e}^*(\mathbf{b})$ is to resample the crude errors $\hat{\epsilon}_i$. However, when a projection matrix $\mathbf{H}$ exists for the models, *Davison and Hinkley* [1997] suggest sampling the residuals from the distribution of the centered adjusted residuals $r_i - \bar{r}$, where r_i is defined as

$$r_i = \frac{\hat{\epsilon}_i}{\sqrt{1 - h_{ii}}}, \tag{31}$$

where h_{ii} is the ith diagonal element of the projection matrix $\mathbf{H}$ and $\bar{r}$ is the average of the r_i. For linear smoothers, it can be shown that the variance of the estimated residuals e_i is equal to $\sigma^2(1 - h_{ii})$. Hence resampling from the distribution of the centered adjusted residuals is preferred because they have the same variance as the true errors ϵ_i. Now that the bootstrap is introduced in the regression context, it can be applied to the data validation problem.

[41] The aim is to construct a prediction interval for new observations. The point estimate of a new observation, $\hat{\theta} = t(\mathbf{x})$, is a prediction from the additive model. Two sources of variability are involved in the derivation of the PI: the uncertainty due to the model prediction and the variability of the residuals. Therefore a double bootstrap procedure is needed. The main loop takes the variability of the model estimator into account. The second loop adds the additional variability that is associated with a single observation. Two types of bootstrap intervals are derived: percentile-based PIs and prediction error-based PIs, where the prediction error δ is defined by $\delta = \hat{y}_{n+1} - y_{n+1}$.

[42] The percentile method proceeds as

[43] 1. Fit the additive model to the historical data set $\mathbf{D}$

[44] 2. Use the fitted model to calculate the prediction $\hat{\theta} = t(\mathbf{x})$

[45] 3. Extract the empirical distribution $\hat{F}$ of the residuals

[46] 4. First bootstrap loop: For $b_1 = 1, \ldots, B_1$

[47] (i) Take a bootstrap sample of the residuals $\mathbf{e}^*(b_1)$ and construct a bootstrapped response $\mathbf{y}^*(b_1)$ by adding these residuals to the fitted values of the AM ($\hat{\mathbf{y}}$), $\mathbf{y}^*(b_1) = \hat{\mathbf{y}} + \mathbf{e}^*(b_1)$. The bootstrapped data set $\mathbf{D}^*(b_1)$ now becomes $\mathbf{D}^*(b_1) = (\mathbf{x_1}, \ldots, \mathbf{x_p}, \mathbf{y}^*(b_1))$.

[48] (ii) Fit a AM model to $\mathbf{D}^*(b_1)$, and compute the prediction $t(\mathbf{D}^*(b_1))$.

[49] (iii) Second bootstrap loop: For $b_2 = 1, \ldots, B_2$

[50] a. Sample at random a residual $e^*(b_2)$ from the empirical distribution of the residuals ($\hat{F}$).

[51] b. The bootstrap estimate $\hat{\theta}^*(b_1, b_2)$ for the new observation is given by $\hat{\theta}^*(b_1, b_2) = t\big(\mathbf{D}^*(b_1)\big) + e^*(b_2)$.

[52] 5. $1 - \alpha$ confidence intervals are calculated from the bootstrap distribution G^*. First, the $\hat{\theta}^*$'s are ordered so that $\hat{\theta}^*_{(1)} \leq \cdots \leq \theta^*_{(B_1B_2)}$. The interval is obtained by taking the $\alpha/2$ and $1 - \alpha/2$ percentiles of G^* [*Efron and Tibshirani*, 1993] which is denoted as

$$\left[\hat{\theta}^*_{\left(\left[B_1B_2\frac{\alpha}{2}\right]\right)},\ \hat{\theta}^*_{\left(\left[B_1B_2\left(1-\frac{\alpha}{2}\right)\right]+1\right)}\right]. \tag{32}$$

This PI is referred to as the bPI.

[53] *Davison and Hinkley* [1997] showed for linear models that the PI can also be estimated by computing the bootstrap distribution of the studentized predictions errors, $z = \delta/\hat{\sigma}$, mimicking the standard normal theory, where the prediction error $\delta = \hat{y}_{n+1} - y_{n+1}$ and $\hat{\sigma} = \sqrt{(\mathrm{RSS}/df^{\mathrm{err}})}$. This idea can easily be adopted to additive models and require steps 4 and 5 of the main bootstrap loop to be replaced by

[54] 4. First bootstrap loop: For $b_1 = 1, \ldots, B_1$

[55] (i) Take a bootstrap sample of the residuals $\mathbf{e}^*(b_1)$ and construct a bootstrapped response $\mathbf{y}^*(b_1)$ by adding this residuals to the fitted values of the AM ($\hat{\mathbf{y}}$). $\mathbf{y}^*(b_1) = \hat{\mathbf{y}} + \mathbf{e}^*(b_1)$. The bootstrapped data set $\mathbf{D}^*(b_1)$ now becomes $\mathbf{D}^*(b_1) = (\mathbf{x_1}, \ldots, \mathbf{x_p}, \mathbf{y}^*(b_1))$.

[56] (ii) Fit an AM model to $\mathbf{D}^*(b_1)$, and compute the prediction $t(\mathbf{D}^*(b_1))$ and the standard deviation of the residuals, $\hat{\sigma}^*(b_1)$.

[57] (iii) Second bootstrap loop: For $b_2 = 1, \ldots, B_2$

[58] a. Sample at random a residual $e^*(b_2)$ from the empirical distribution of the residuals ($\hat{F}$).

[59] b. Compute the standardized prediction error $z^*(b_1b_2) = \delta^*(b_1b_2)/\hat{\sigma}^*(b_1)$ with $\delta^*(b_1b_2) = \hat{y}^*_{n+1} - \big(\hat{y}_{n+1} + e^*(b_2)\big)$.

[60] 5. The bootstrap prediction interval, after ranking the z^*'s to $z^*_{(1)} \leq \ldots \leq z^*_{(B_1B_2)}$ is given by

$$\left[\hat{y}_{n+1} - \hat{\sigma}z^*_{\left(\left[(B_1B_2+1)\left(1-\frac{\alpha}{2}\right)\right]\right)},\ y_{n+1} - \hat{\sigma}z^*_{\left(\left[(B_1B_2+1)\frac{\alpha}{2}\right]\right)}\right]. \tag{33}$$

This interval is referred to as spbPI.

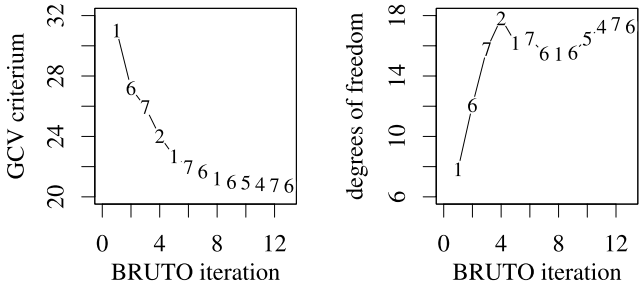

Figure 6. Left: Convergence of the GCV criterion when BRUTO is applied to the data of sampling location 913000 along the Yzer River. Right: The evolution of the total degrees of freedom in the model as a function of the iteration number. The numbers along the curve indicate which of the seven predictors is updated.

[61] In case a projection matrix exists for the additive model, the computational cost of the bootstrap procedures can be reduced significantly. Since the bootstrap procedure only alters the response, the structure of the predictors remains the same. Moreover, the projection matrix only has to be calculated once because its calculation only involves the predictors. Therefore the calculation of the prediction in step 4(ii) of the bootstrap procedure reduces to $t(\mathbf{D}^*(b_1)) = \mathbf{H_n}\mathbf{y}^*(b_1)$ instead of having to perform the full backfitting procedure. This leads to considerable savings in computational costs. Remember that it is also better in this case to sample the residuals from the distribution of the centered adjusted residuals.

3.3. Diagnostic Plots

[62] There are several possible causes for the rejection of incoming data, such as changes in the system, illegal spills, errors during the analysis in the laboratory, wrong calibration of the equipment, outliers in the predictor variables, and so on. Since other physico-chemical variables are present in the model as predictor variables, it is possible that an outlier in one of these variables results in a false rejection of the incoming data: A predictor has an additive effect on the outcome of the model, and outliers can result in an extreme value of the predictor function enhancing a shift in the PI. At first sight, this looks like an anomaly of our methodology. However, such shifts can be detected by simply leaving the predictor out of the model: If the prediction was performed at an outlying observation in a particular predictor variable, the interval will shift back when this predictor variable is omitted.

[63] The following strategy is proposed: all predictor variables are left out of the model one by one. Then the PI is calculated with each of these new models. If the new observation now lies in the PI, the observed deviation is possibly due to an outlier in the predictor which has been left out of the model.

4. Results and Discussion

[64] Here the entire methodology is illustrated on a real data case. The results of this case are then used to generated synthetic data for a simulation study and a power study. These studies are needed to check the coverage and the performance of the derived prediction intervals. Finally, the method is applied to two case studies to validate the nitrate data of the Yzer River measured in 2003 and 2004. In a first case, 2 years of data are validated at one sampling location. In a second case, the data validation is applied to 2 years of data on all sampling locations of the river basin that contain enough data to fit the AM models.

4.1. Illustration of the Methodology on a Real Data Case

[65] The methodology is illustrated on the data of sampling location 913000 which belongs to the physico-chemical monitoring network. The sampling location is located along the Yzer River. The data set consists of eight variables: (1) day number throughout the year, (2) date, (3) temperature, (4) dissolved oxygen (DO), (5) nitrite (NO_2^-), (6) chemical oxygen demand (COD), (7) pH, and (8) nitrate (NO_3^-). First, the additive model is built and a residual analysis is performed. Then the AM is used to validate a new observation by using the different PIs.

4.1.1. Procedure to Build the Additive Model

[66] The nitrate concentration is modeled using an additive model. The building blocks of the model are local linear smoothers resulting in a fully nonparametric model. The first seven variables are allowed to be included in the final model. Variable 1 codes for the seasonal effect and variable 2 models a potential long-term trend. The BRUTO algorithm is used for model selection. The evolution of the algorithm is presented in Figure 6. The numbers in the plot indicate which of the predictors was adjusted in each cycle. During the first four cycles, predictors 1, 6, 7 and 2 are included in the model. From the 5th up to the 9th cycle, the spans of the selected predictors are adjusted. During cycles 10 and 11, predictors 5 and 4 are selected, respectively. Finally, the last two cycles adjust the spans of predictors 7 and 6, respectively. The final model includes predictors 1, 2, 4, 5, 6, and 7. Notice that the 3rd predictor is never included in the model. At first, the GCV decrease is steep. This is due to the take up of extra predictors in the model and is also reflected in the steep increase of the associated degrees of freedom.

[67] The resulting model is presented in Figure 7. To enable a graphical representation of the high dimensional regression surface, we have chosen to represent the fit as a function of the temporal dimension (Figure 7, top). The effect of each of the predictors is given in Figure 7 in the remaining panels. All fits are accompanied by 95% pointwise confidence intervals. A fitted value is equal to the sum of the general mean and each of the effects at the corresponding predictor values. Once the model is fitted, one can predict the mean response for a new observation by simply adding the individual effects for each of the predictor variables observed at time $n + 1$. In this way, a new nitrate value can be calculated, given its day number, date, DO, NO_2^-, COD, and pH values measured for the particular sample under validation. The figure shows a clear seasonal pattern with low contributions in summer and high contributions in winter, and an increasing contribution of the temporal trend until 1998 and decreasing trend from 1999 onward. Low DO concentrations seem to have a negative contribution on the nitrate concentration, while high DO concentrations have a positive contribution. The contribution of COD is inversely related to the nitrate concentration and levels off at high COD concentrations.

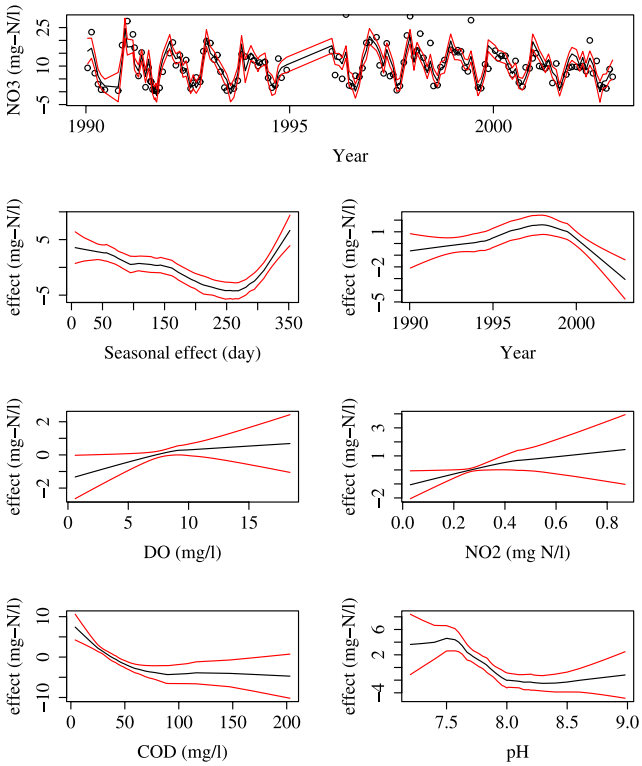

Figure 7. AM for nitrate at sampling location 913000 at the Yzer River. Nitrate is modeled by a long-term trend (date), a seasonal effect, temperature, DO, COD, and pH. The top panel shows the data and the lower panels show the effect of each predictor.

The contributions of DO and COD can be explained from microbiology. Low dissolved oxygen concentrations inhibit the nitrification process which converts ammonium to nitrate. Such oxygen levels are typically occurring at high COD levels. Additionally, in anoxic conditions (in the absence of oxygen and the presence of nitrate), certain microorganisms can use nitrate as their electron acceptor and in the presence of organic matter they convert nitrate to nitrogen gas which eventually escapes from the water phase. The contribution of nitrite seems to be approximately

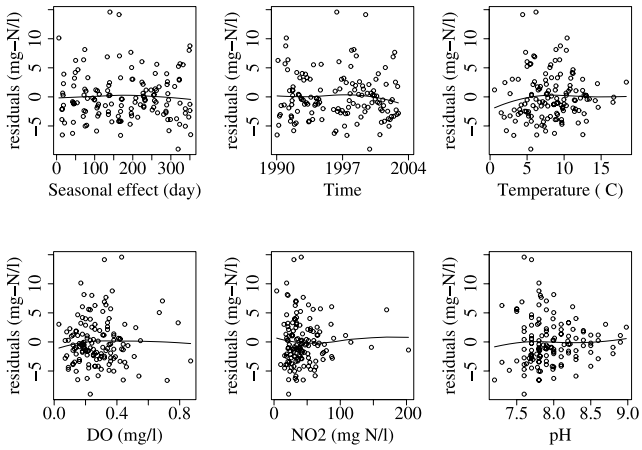

Figure 8. Residual plots from the additive model in Figure 7. A residual smoother is added to each plot.

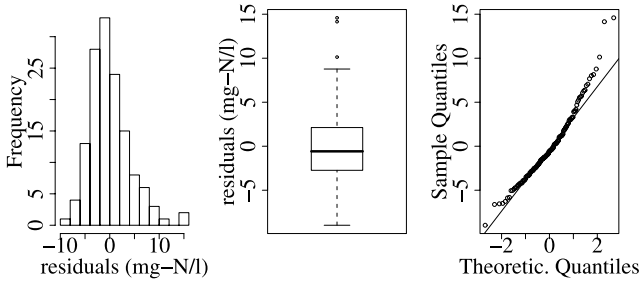

Figure 9. Histogram, box plot, and QQ plot of the residuals from the additive in Figure 7.

proportional to the actual nitrate concentration. In Figure 7, it can be seen that the model is sufficiently flexible to model a large part of the variation of the original data series.

[68] The model quality is checked in a residual analysis. Residual plots for each predictor are given in Figure 8. From the residual plots the data seems more or less homoscedastic. The variance estimate of the residuals is $\hat{\alpha}^2_{913000} = 18.7$. The smoothers added to the residual plots show that the mean of the residuals is centered around zero, except in data-sparse regions at the endpoints, but this is likely to be a boundary effect of the smoother. At the boundaries, the data is sparse and a few residuals can have a large influence on the fit of the smoother used in the residual plot. In Figure 9, the histogram and the QQ plot of the residuals indicate deviations from normality in the upper tail and suggest that the residuals are distributed with a slight tail to the right. The box plot also shows some outliers. When the outliers are removed, the residuals appear to be almost Gaussian (results not shown). However, these nitrate levels cannot be removed because they might be extreme events which are characteristic for the data-generating process. The presence of serial correlation in the residuals is checked using the runs test and by making a variogram of the residuals. The runs test is a nonparametric test that checks the randomness hypothesis of a data sequence [see, e.g., *McWilliams*, 1990]. The run test on the residuals gave a p-value of 0.78, which clearly accepts the null hypothesis of randomness. A variogram is a tool to represent autocorrelation in unequally spaced observations. To construct the variogram, first the differences $d(ij) = y_i - y_j$ and the time differences $\Delta_t(ij) = t_i - t_j$ are calculated for all observations i and j. According to their time difference $\Delta_t(ij)$, all differences $d(ij)$ were classified in time distance classes with mean time distance $\Delta_{t,k}$. The distance classes were taken to be equal in size and the bin was taken at 30 days. For each distance class k, the semivariance is estimated as $\rho_k = \sum_{i=1}^{n_k} d_i^2/(2n_k)$. The semivariance ρ_k is then plotted against $\Delta_{t,k}$. The left panel of Figure 10 represents the variogram for the original data series and the right panel displays the variogram for the residuals of the AM. The grey lines in the background are variograms obtained when white noise was created with the same variance as the derived variograms of interest. The original nitrate measurements are clearly autocorrelated and the seasonal pattern is very obvious. After the AM was fitted, the autocorrelation is completely removed and the variogram behaves similar to white noise. Both the runs test and the variogram support the hypothesis of *Van Belle and Hughes* [1984] that water quality data measured at intervals

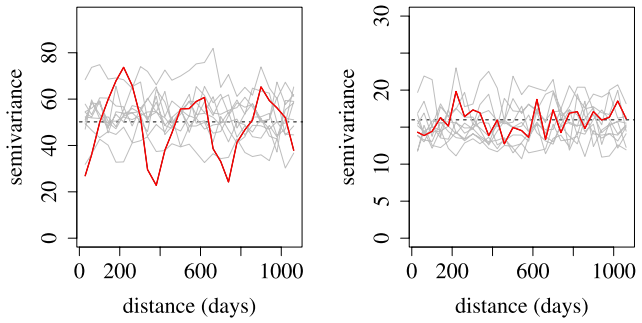

Figure 10. Variogram of the (left) original nitrate series and of the (right) residuals after fitting the AM from Figure 7 are plotted (thick solid line). Ten variograms generated by white noise with the same variance are added to the plot (thin grey lines).

which are larger than 2 weeks can be assumed to be independent when seasonality and trend are removed.

[69] The additive model for the historical data is established and the residuals are shown to be independent. The model can now be used to construct a prediction interval for new observations. In the next section, the validation is performed using the three different PIs derived in section 3.2.

4.1.2. Validation of a new Observation by the use of Prediction Intervals

[70] In the previous section an additive model was established using the data before 01 January 2003. The first new observation is acquired on 14 January 2003 and will be validated. The AM is used to perform a prediction of the fitted response $\hat{y}_{n+1,\,913000} = 12.3$. The variance corresponding to this prediction is $\sigma^2_{\hat{y}_n,913000} = 2.6$. The prediction interval for nitrate on 14 January 2003 is given in Figure 11. Instead of creating a two-sided interval, we prefer to use one-sided interval by concentrating all the uncertainty in the upper tail. Low nitrate concentrations are not harmful for the environment, so it is more interesting to focus on a faster detection of abnormal high nitrate concentrations. In the right panel the historical data is presented together with the optimal fitted model. In the left panel, the new observation is represented by a dot and the upper limit of the bootstrap interval is given using the three different methods. The new observation lies in all intervals and is thus accepted. In the double bootstrap procedure, 1000 bootstraps are calculated for each bootstrap loop (B_1 and B_2) resulting in 1 million bootstrap replicates (B_1B_2). The bPI seems to be slightly higher than the aPI and the spbPI.

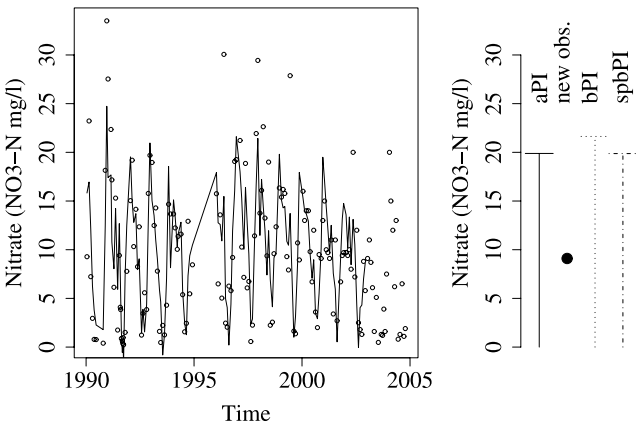

Figure 11. Prediction interval for the nitrate concentration on 14 January 2003 at sampling location 913000 along the Yzer River. Left: Historical data with model fit. Right: The new observation (dot) is accepted by all one-sided prediction intervals (aPI (solid line), bPI (dotted line), and dashed dotted line (spbPI)). The new observation is accepted by all intervals.

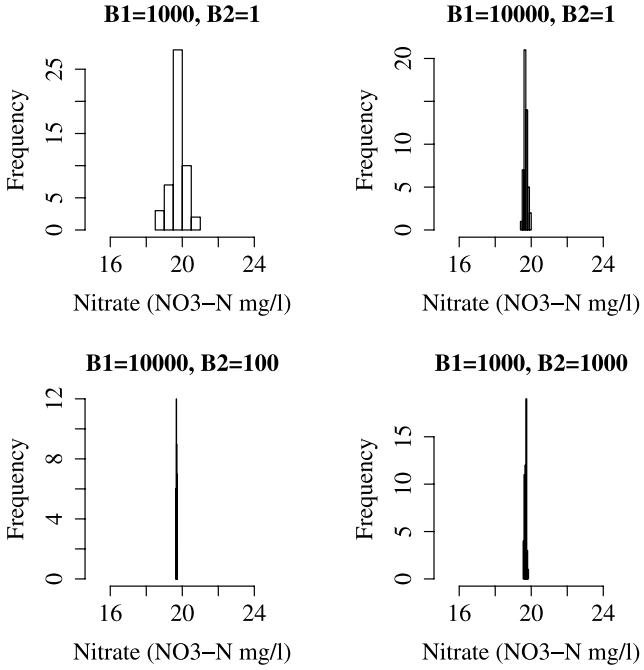

Figure 12. Effect of the number of bootstraps in the first and second loops on the bootstrap resampling variability of one-sided 95% spbPI. Each histogram is based on 50 PIs; B1 is the number of bootstraps in the main bootstrap loop and B2 is the number of bootstraps in the second bootstrap loop.

[71] In this study, B_1 and B_2 are chosen to be 1000, resulting in 1 million bootstrap replicates (B_1B_2). In the ideal case, however, the number of bootstrap replicates should be taken to be ∞. In practice, this is not feasible and the number of bootstrap replicates is set at a large value. This leads to a bootstrap resampling variability. Thus, when the calculation of the bootstrap PI is repeated on the same data, the obtained PI will be slightly different. To stabilize the bootstrap resampling variability, the number of bootstrap replicates should be taken large enough. In a double bootstrap procedure, the bootstrap resampling variability is introduced in both loops. To control the bootstrap resampling variability due to the first loop, the size of B_1 should be appropriate. The bootstrap resampling variability caused by the second loop is controlled by B_1B_2. Hence stable intervals are obtained by taking B_1 and B_1B_2 large enough. The latter can be obtained by taking the number B_1 very large and by taking $B_2 = 1$ or by using moderate values for both B_1 and B_2. In a practical implementation, the computational complexity associated with both bootstrap loops also has to be taken into account. Here the computational load of the second loop is negligible

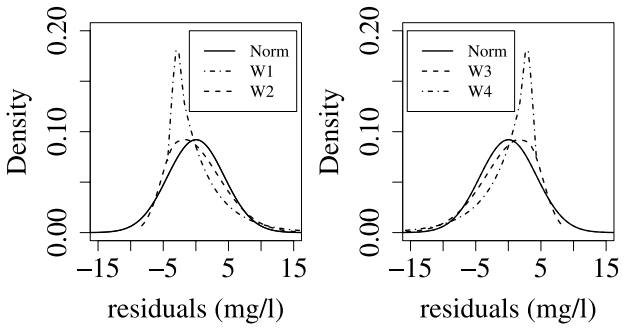

Figure 13. Density functions of the residuals used to generate the data for the simulation study.

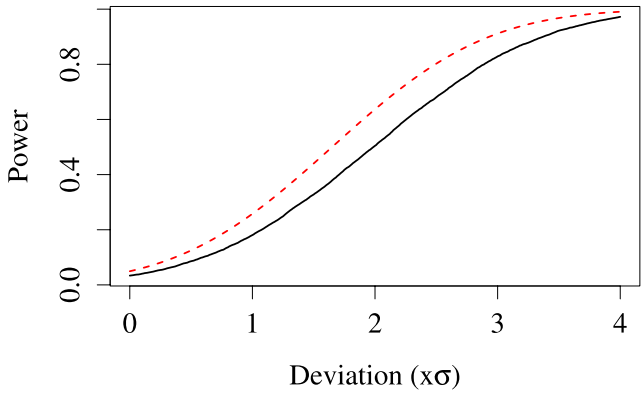

Figure 14. Power curve for the detection of deviations in validated data: solid line: empirical power; dashed line: theoretical power when the model uncertainty is neglected. The size of the deviations ranges between 0 and 4 times $\hat{\sigma}_{913000}$.

compared to the first loop. Hence it is interesting to take B_1 as small as possible in order to reduce the computational power. The impact of the sizes of B_1 and B_2 is assessed in Figure 12. One-sided intervals were calculated to validate nitrate measurements. For the same data set, 50 bootstrap intervals are calculated for (1) B_1 = 1000, B_2 = 1, (2) B_1 = 10,000, B_2 = 1, (3) B_1 = 10,000, B_2 = 100 and (4) B_1 = 1000, B_2 = 1000. For cases (1) and (4), the time needed to calculate the intervals was almost equal because the computational complexity associated with the calculation of 1000 AMs in the first bootstrap loop is much larger than the complexity needed for the second step. For cases (2) and (3), however, 10 times more computational time was needed because the first loop was executed 10 times more. The figure clearly illustrates that, for case (4), the one-sided interval is estimated much more accurately than in case (1) where there is still a considerable amount variability. The stability of the intervals in (4) was slightly better than in case (2). This is because the second loop was only executed 10,000 times for case (2) compared to 1,000,000 times for case (4). In case (3), a small gain in accuracy can be observed compared to case (4). In both cases, the second loop is assessed 1,000,000 times. Hence the bootstrap resampling variability induced by the second loop is controlled at the same level. In case (3), the first loop is executed 10 times as much as in case (4) and therefore a slight reduction of the bootstrap resampling variability is established. But this is at the expense of an increase in the computational time by a factor of 10. In order to reach an acceptable accuracy while keeping the computational time limited, we decided to use B_1 = 1000 and B_2 = 1000.

4.2. Evaluation of the Coverage of the PI's in a Simulation Study

[72] The coverages of the three prediction intervals derived in section 3.2 are evaluated for five different cases:

Table 1. Coverage (in %) of 95% PI's for Data Originating From Different Distributions

	Analytical	Bootstrap	
Distribution	aPI	%bPI	spbPI
Gaussian	96.4	97.2	95
Right-Tailed, W1	94.1	96	94.5
Moderately Right-Tailed, W2	95.5	96.6	94.8
Moderately Left-Tailed, W3	98.8	98.5	95.2
Left-Tailed, W4	99.8	99.9	96.6

normal residuals, two types of residuals originating from right-tailed distributions, and two types of residuals originating from left-tailed distributions. The results of the nitrate data set at location 913000 in section 4.1 are used for constructing the data for the simulation study. The model fitted in Figure 7 is used to construct simulated data sets. For the right-tailed distributions, Weibull distributions with shape factors of 1 and 2 are considered. The scale parameter can be chosen arbitrarily because the simulated residuals are standardized and multiplied with the standard deviation $\hat{\sigma}_{913000}$ of the residuals obtained from the fitted model in Figure 7. The residuals from the left-tailed distributions are generated by changing the sign of the residuals from the right-tailed distributions. Plots of the distribution functions used in the simulation study for each of the different distributions are given in Figure 13. For the normal residuals, we will sample from a normal distribution with mean 0 and variance $\hat{\sigma}^2_{913000}$.

[73] Now that new residuals with the same variance as the original data can be generated, simulated data sets are constructed. First, residuals are simulated as explained above, denoted by ϵ^*. The simulated data sets $\mathbf{D}^*$ then consist of the original predictors ($\mathbf{x_1}$, …, $\mathbf{x_p}$) and the simulated response $\mathbf{y}^* = \hat{\mathbf{y}} + \epsilon^*$. For the simulated data sets, the values of the true underlying function $m(\mathbf{X})$ evaluated at the predictor points $\mathbf{x}$ and the observation under validation $\mathbf{x_{n+1}}$ are known. They are presented as the $\hat{\mathbf{y}}_{913000}$ and $\hat{y}_{n+1,913000}$ in Figure 7, respectively.

[74] For each distribution, 5000 data sets were constructed. As the values at time n + 1 have the true underlying mean m, 95% PIs should accept 95% of the validated data. The coverage for the different intervals are given in Table 1. The aPIs seem to be slightly too large for the Gaussian case. The coverage of the aPIs decreases when the data is right-tailed and increases when the data is left-tailed. This effect is more apparent when the distribution becomes more asymmetric. The bPI seems to have the tendency to be too large, the results for the different distributions are all above 95%. Only the spbPI seems to reach the correct coverage and is robust toward deviations from normality. The coverage of bPI is known to be problematic [*Efron and Tibshirani*, 1993; *Davison and*

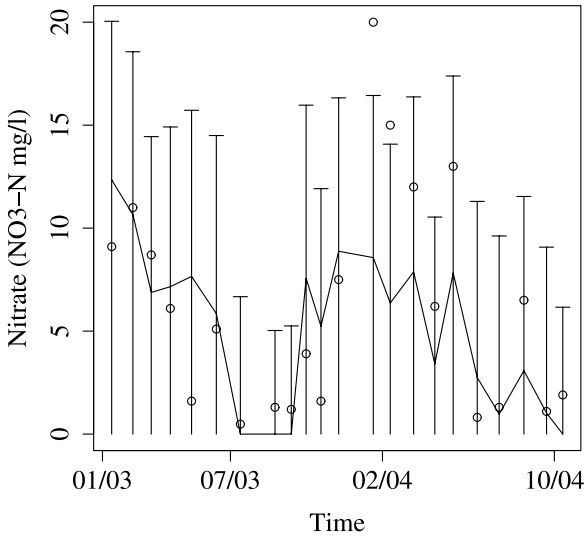

Figure 15. Validation of nitrate at sampling location 913000 of the Yzer monitoring network. Nitrate concentrations in January and February 2004 are considered as anomalous by the automatic validation procedure.

Hinkley, 1997]. Corrections for percentile-based intervals exist, for instance *Efron and Tibshirani* [1993] suggested bias- and acceleration-corrected intervals. But the methods they suggested cannot be constructed for the double bootstrap procedure because the second loop consists of adding a random residual. For the semi-automatic data validation procedure, aPIs are preferred from a computational point of view. However, their coverage can behave poorly, particularly for the combination of upper bounded one-sided PIs and residuals that follow a left-tailed distribution. The studentized prediction error-based bootstrap PIs (spbPI), however, are rather robust toward the distribution of the residuals and, therefore, we suggest to use this PI for data validation purposes.

4.3. Evaluation of the Power

[75] The model fitted in Figure 7 is used to construct simulated data sets. First, residuals ϵ^* are simulated from the normal distribution $N\left(0,\ \hat{\sigma}^2_{913000}\right)$. The simulated data sets $\mathbf{D}^*$ then consist of the original predictors ($\mathbf{x_1}$, …, $\mathbf{x_p}$) and the simulated response $\mathbf{y}^* = \hat{\mathbf{y}}_{913000} + \epsilon^*$. Thus for the simulated data sets, the values of the true underlying function $m(\mathbf{X})$ evaluated at the predictor points $\mathbf{x}$ and $\mathbf{x_{n+1}}$ are $\hat{\mathbf{y}}_{\mathbf{913000}}$ and $\hat{y}_{n+1,913000}$. Now a systematic deviation is introduced in the simulated data $(\mathbf{x}_{n+1},\ y^*_{n+1})$ which has to be validated. Instead of validating $y^*_{n+1} = \hat{y}_{n+1,913000} + \epsilon^*$, $y^*_{n+1} = \hat{y}_{n+1,913000} + \epsilon^* + l\hat{\sigma}_{913000}$ is used and the corresponding power to detect this deviation is calculated. To derive a complete power curve, different values for l are taken which range between 0 and 4. For each l, 5000 data sets are generated to calculate the empirical power. The resulting power curve is displayed in Figure 14 (thick black line). In the same figure, a theoretical power curve is represented. The theoretical power was derived under the assumption that the uncertainty due to the estimation of the model could be neglected. When the model uncertainty can be neglected, the model prediction $\hat{y}^*_{n+1}$ follows a normal distribution $N\left(\hat{y}_{n+1,913000},\ \hat{\sigma}^2_{913000}\right)$. The validated observation y^*_{n+1}, however, follows a normal distribution $N\left(\hat{y}_{n+1,913000} + l\sigma_{913000},\ \hat{\sigma}^2_{913000}\right)$. Hence the power to detect the deviation in y^*_{n+1} is established by using the distribution function $N\left(\hat{y}_{n+1} + l\sigma_{913000},\ \hat{\sigma}^2_{913000}\right)$ to calculate the probability $P\left(y^*_{n+1} > \hat{y}_{n+1,913000} + z_{1-\alpha}\hat{\sigma}_{913000}\right)$. This theoretical power cannot be exceeded because model uncertainty is always present in practical applications. At the beginning, when $l = 0$ both curves start at 5%. This is due to the definition of 95% PIs. For moderate l, the empirical power curve is lower than the theoretical one, but the empirical power remains remarkable high. This suggests that our method is well suited for data validation purposes.

4.4. Case Studies

4.4.1. Validation at one Sampling Location

[76] The data of sampling location 913000 along the Yzer River over the years 2003 and 2004 are validated. The data set at this location contains eight variables: day number, date, temperature (t), dissolved oxygen concentration (DO), nitrite concentration (NO_2^-), chemical oxygen demand (COD), pH, and nitrate concentration (NO_3^-). The time series starts at April 1990 and ends in December 2004. All eight variables are measured on a monthly basis. The data from 1990 until December 2002 are considered as historical data. The nitrate data from 2003 and 2004 are validated in chronological order. In particular, if a new observation lays within the 95% PI, then the measurement is accepted and considered as historical data for the validation of the next observation.

[77] The results of the data validation are presented in Figure 15. All data from 2003 are accepted. The observations in January and February of 2004 are rejected. Diagnostic plots for these observations are given in Figures 16 and 17, respectively. From the diagnostic plots, possible explanations for the rejection of the data may become clear. The measurement in January was only accepted when the trend was omitted from the model, giving a strong indication that this measurement did not follow the expected long-term time trend in the data. The measurement in February

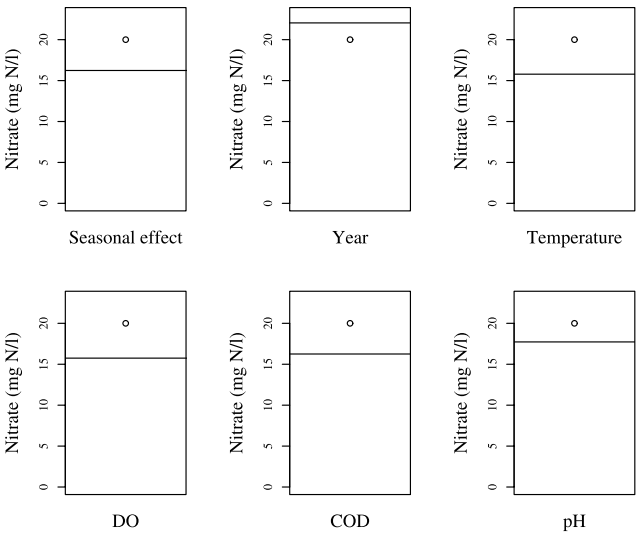

Figure 16. Diagnostic plots for rejected nitrate concentration of January 2004 at sampling location 913000 of the Yzer monitoring network.

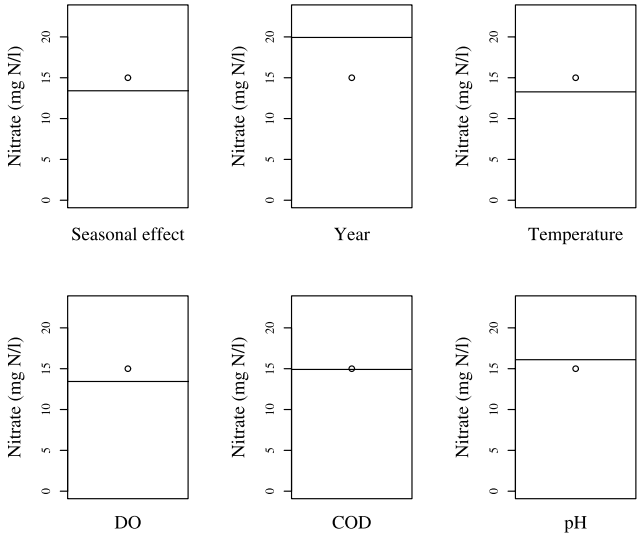

Figure 17. Diagnostic plots for rejected nitrate concentration of February 2004 at sampling location 913000 of the Yzer monitoring network.

was accepted when the trend or pH were omitted from the model. The effect of omitting the pH on the size of the interval was only limited. The effect of omitting the trend from the model was much more pronounced. This indicates again that the main cause of the deviation was related to the trend. The nitrate concentrations in the beginning of 2004 are known to be high [*Vlaamse Milieumaatschappij*, 2005]. The Yzer River is located in the countryside and 2003 was a dry year, which resulted in an accumulation of nitrate in agricultural soils in the summer. The dry summer of 2003 had a beneficial effect on the nitrate concentration, since there was a limited amount of nitrate washed to the water course by the rain. Hence the nitrate accumulated in the soil and was washed out in the winter period. Moreover, January 2004 was recognized to be extremely wet by the Belgian Royal Meteorological Institute (KMI). This means that this phenomenon at most happens once in 100 years. The dry summer combined with an extreme wet winter provoked high nitrate concentrations in receiving river.

4.4.2. Validation of an Entire Basin

[78] The data from 2003 and 2004 for all sampling locations of the entire Yzer River, containing enough data to fit the models, is validated. The data set at each location has information on eight variables: day number, date, t, DO, NO_2^-, COD, pH, and NO_3^-. All eight variables are measured on a monthly basis. The data from 1990 until December 2002 are considered as historical data. The nitrate data from 2003 and 2004 are validated in chronological order. If a new observation lies within the PI, then the measurement is accepted and considered as historical data for the validation of the next observation. The data validation is carried out using 95% spbPIs. The empirical coverage of the intervals in a certain period is calculated by dividing the number of accepted observations in this period by the total number of validated observations in this period. The coverage of the intervals for the whole validation period, was 91%. However, the coverage for the 2003 data of the spbPIs was 94.7% and was close to what is expected from theory when no deviations are present. In 2004, the coverage was only 80%, indicating the presence of a considerable amount of anomalous data. In Figure 18, the results of the data

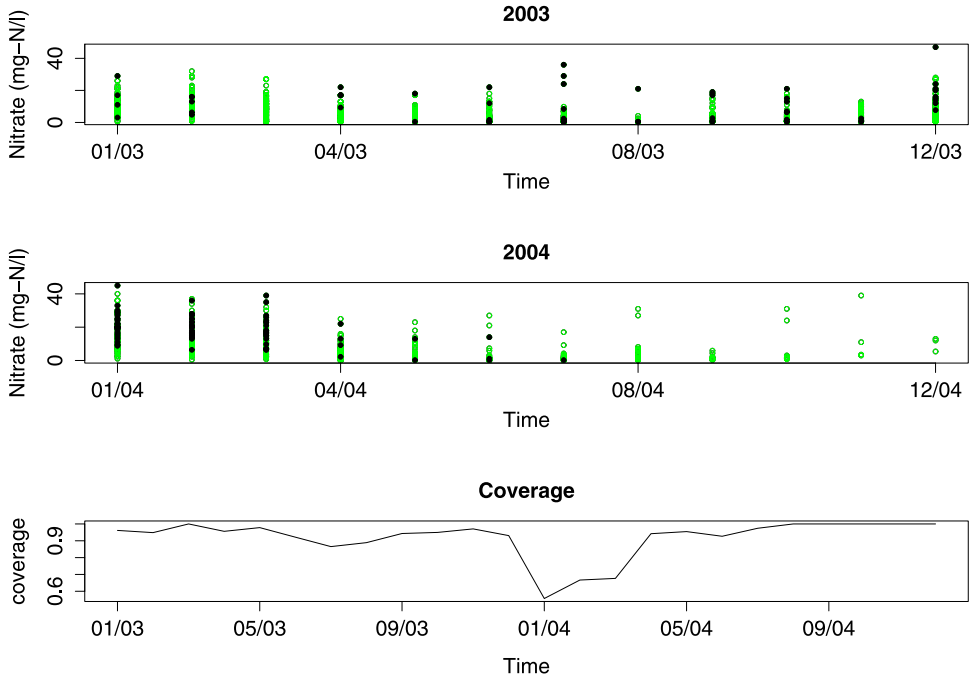

Figure 18. Validation of nitrate at all sampling locations of Yzer monitoring network. The top panel shows the results of the validation in 2003; the middle panel shows the results of 2004; and the bottom panel shows the evolution of the coverage of the PIs during the whole validation period. Accepted data are indicated with an open circle and the rejected data are indicated with a dot. The coverage of the 95% prediction intervals clearly drops in January, February, and March 2004.

validation based on the spbPIs are presented. The top panel shows the results of the validation in 2003, the middle panel shows the results of 2004, and the bottom panel shows the evolution of the coverage of the spbPIs during the whole validation period. Accepted data are indicated with open circles and the rejected data are presented by dots. From the middle panel of Figure 18, it can be seen that a considerable amount of data is rejected in the period of January to March 2004. This is even more obvious in the results presented in the bottom panel. The bottom panel shows the evolution of the empirical coverage in each month. In 2003, the coverage is more or less stable at 95%. In the beginning of 2004, a clear drop of the coverages of the PIs is observed (January 56%, February 66%, and March 67%) indicating that there was a change in the system during the first months of 2004.

[79] A more general feature can be derived from Figure 18: similar to multivariate techniques, our method also detects observations lying in the center of the univariate distribution of the nitrate concentration as outlying observations. Hence our methodology combines the interesting features of multivariate outlier detection without imposing restricted assumptions on the relationship between the response and the predictor variables.

5. Conclusions

[80] A method for the validation of river water quality data is proposed. Based on the historical data, an additive model is fitted, which is subsequently used to construct prediction intervals for future observations.

[81] Our study indicates that the additive models are clearly able to catch the cyclic pattern present in the data and could model the nonlinear behavior and relationships typically associated with river water quality data. As an interesting feature, the observed associations between the response and the predictors reflect well-known physical and biological relationships. Since the model selection is carried out at each time step, the models succeed to adapt to changes in the processes of the underlying river.

[82] From the different prediction intervals which were derived, the studentized prediction error-based bootstrap PIs (spbPIs) are most interesting to be used in practice. The coverages of the 95% spbPIs have been assessed in a simulation study and, in comparison with analytical intervals, which assume the residuals to be Gaussian, they appear to be much more robust against deviations from normality. The power of the method was also shown to be adequate.

[83] The case studies have illustrated that our method could detect anomalous events, such as an abnormal high nitrate release due to a dry summer, which was followed with an extreme wet winter period. The diagnostic plots were also useful to assist the operator for an explanation of the anomalous measurement: they indicated that the rejection was related to the trend. In the case studies, the semi-automatic procedure detected suspicious observations lying at the edges as well as observations lying in the center of the univariate distribution of the observations. Hence it combines the interesting features of classical multivariate outlier detection tools without having to impose linear relationships typically associated with these methods.

[84] An ICT tool based on this methodology could be of great value to analyze and maintain environmental databases originating from monitoring networks such as the ones which are implied by the WFD. It can be used to check the quality of the data and it can also detect abnormal changes in the water quality.

[85] **Acknowledgments.** The results presented in this paper have been elaborated in the frame of research projects VMM/AMO/ADM/TWO/200000901 and VMM/AMO/ADM/TWO/20021106 of the Flemish Environmental Agency (VMM) and the EU project Harmoni-CA, contract no. EVK1-CT-2002-20003. The program is organized within the Energy, Environment, and Sustainable Development Programme in the 5th Framework Programme for Science Research and Technological Development of the European Commission. The Flemish Environmental Agency also provided the data for the case study.

References

Altman, N. S. (1992), An introduction to kernel and nearest-neighbor nonparametric regression, *Am. Stat.*, *46*(3), 175–185.

Buja, A., T. Hastie, and R. Tibshirani (1989), Linear smoothers and additive models, *Ann. Stat.*, *17*(2), 453–510.

Cai, Z., and R. C. Tiwari (2000), Application of a local linear autoregressive model to bod time series, *Environmetrics*, *11*, 341–350.

Chatfield, C. (1993), Calculating interval forecasts, *J. Bus. Econ. Stat.*, *11*(2), 121–134.

Clements, M. P., and N. Taylor (2001), Bootstrapping prediction intervals for autoregressive models, *Int. J. Forecast.*, *17*(2), 247–267.

Cleveland, W. S. (1979), Robust locally weighted regression and smoothing scatterplots, *J. Am. Stat. Assoc.*, *74*(368), 829–836.

Cleveland, W. S., and S. J. Devlin (1988), Locally weighted regression: An approach to regression analysis by local fitting, *J. Am. Stat. Assoc.*, *83*(403), 596–610.

Cleveland, W. S., and E. Grosse (1991), Computational methods for local regression, *Stat. Comput.*, *1*, 47–62.

Davison, A., and D. Hinkley (1997), Bootstrap methods and their applications, in: [*Cambridge Series on Statistical and Probabilistic Mathematics*], first ed., Cambridge Univ. Press, New York.

De Rycke, A., K. Devos, and K. Decleer (2001), *Verkennende ecologische gebiedsvisie voor de IJzervallei. Rapport Instituut voor Natuurbehoud*, 123 p., Instituut voor Natuurbehoud, Brussel.

Dominici, F., A. McDermott, S. L. Zeger, and J. M. Samet (2002), On the use of generalized additive models in time-series studies of air pollution, *Am. J. Epidemiol.*, *156*(3), 193–203.

Efron, B., and R. J. Tibshirani (1993), An introduction to the bootstrap, *Monographs on Statistics and Applied Probability*, first ed., CRC Press, Boca Raton, Fla.

Fan, J. (1992), Design-adaptive nonparametric regression, *J. Am. Stat. Assoc.*, *87*(420), 998–1004.

Fan, J. (1993), Local linear regression smoothers and their minimax efficiency, *Ann. Stat.*, *21*(1), 196–216.

Fan, J., and I. Gijbels (1996), Local polynomial modelling and its applications, *Monographs on Statistics and Applied Probability*, first ed., CRC Press, Boca Raton, Fla.

Hasti, T., and C. Loader (1993), Local regression: Automatic kernel carpentry, *Stat. Sci.*, *8*(2), 120–129.

Hastie, T. J., and R. J. Tibshirani (1990), Generalized additive models, *Monographs on Statistics and Applied Probability*, first ed., CRC Press, Boca Raton, Fla.

Hastie, T. J., R. J. Tibshirani, and J. H. Friedman (2001), The elements of statistical learning, *Springer Series in Statistics*, first ed., 552 p., Springer, New York.

Kim, J. H. (1999), Asymptotic and bootstrap prediction regions for vector autoregression, *Int. J. Forecast.*, *15*(4), 393–403.

Kim, J. H. (2004), Bootstrap prediction intervals for autoregression using asymptotically mean-unbiased estimators, *Int. J. Forecast.*, *20*(1), 85–97.

Loader, C. (1999), Local regression and likelihood, *Statistics and Computing*, first ed., 304 p., Springer, New York.

McMullan, A., A. W. Bowman, and E. Scott (2003), Non-linear and nonparametric modelling of seasonal environmental data, *Comput. Stat.*, *18*, 167–183.

McWilliams, T. P. (1990), A distribution-free test for symmetry based on a runs statistic, *J. Am. Stat. Assoc.*, *85*(412), 1130–1133.

Opsomer, J. D. (2000), Asymptotic properties of backfitting estimators, *J. Multivar. Anal.*, *74*, 166–179.

Penny, K. (1996), Appropriate critical values when testing for a single multivariate outlier by using the Mahalanobis distance, *J. R. Stat. Soc., Ser. C, Appl. Stat.*, *45*(1), 73–81.

Pourahmadi, M. (2001), Foundations of time series analysis and prediction theory, *Wiley Series in Probability and Statistics*,first ed., John Wiley, Hoboken, N. J.

Van Belle, G., and J. P. Hughes (1984), Nonparametric tests for trend in water quality, *Water Resour. Res.*, *20*(1), 127–136.

Venkatasubramanian, V., R. Rengaswamy, S. N. Kavuri, and K. Yin (2003), A review of process fault detection and diagnosis. part III: Process history based methods, *Comput. Chem. Eng.*, *27*(3), 327–346.

Vlaamse Milieumaatschappij (2005), *Water- & waterbodemkwaliteit - Lozingen in het water-Evaluatie saneringsinfrastructuur 2004.*, 87 p., Vlaamse Milieumaatschappij, Aalst.

Wood, S., and N. Augustin (2002), Gams with integrated model selection using penalized regression splines and applications to environmental modelling, *Ecol. Modell.*, *157*, 157–177.

L. Clement, Department of Applied Mathematics, Biometrics and Process Control, Ghent University, Coupure Links 653, B-9000, Gent, Belgium. (lieven.clement@ugent.be)

J. P. Ottoy, O. Thas, and P. A. Vanrolleghem, Department of Applied Mathematics, Biometrics and Process Control, Ghent University, Gent, Belgium.