Conditional parametric models for storm sewer runoff

H. Jonsdottir,[1,2] H. Aa Nielsen,[1] H. Madsen,[1] J. Eliasson,[2] O. P. Palsson,[2] and M. K. Nielsen[3]

Received 12 August 2005; revised 12 July 2006; accepted 6 November 2006; published 31 May 2007.

[1] The method of conditional parametric modeling is introduced for flow prediction in a sewage system. It is a well-known fact that in hydrological modeling the response (runoff) to input (precipitation) varies depending on soil moisture and several other factors. Consequently, nonlinear input-output models are needed. The model formulation described in this paper is similar to the traditional linear models like final impulse response (FIR) and autoregressive exogenous (ARX) except that the parameters vary as a function of some external variables. The parameter variation is modeled by local lines, using kernels for local linear regression. As such, the method might be referred to as a nearest neighbor method. The results achieved in this study were compared to results from the conventional linear methods, FIR and ARX. The increase in the coefficient of determination is substantial. Furthermore, the new approach conserves the mass balance better. Hence this new approach looks promising for various hydrological models and analysis.

Citation: Jonsdottir, H., H. A. Nielsen, H. Madsen, J. Eliasson, O. P. Palsson, and M. K. Nielsen (2007), Conditional parametric models for storm sewer runoff, *Water Resour. Res.*, *43*, W05443, doi:10.1029/2005WR004500.

1. Introduction

[2] Hydrology is one of the oldest fields of interest in science and has been studied on both small and large scales for about 6000 years. The goal of the present work is to achieve good predictions of flow in a sewage system. Black box models have been providing good prediction results, often much better than conceptual or physical models, depending on how well the actual system is known. *Carstensen et al.* [1998] showed that data driven models are more reliable for online applications in sewers than stationary deterministic models.

[3] Black box models have been used in hydrology for decades; *Sherman* [1932] presented the first black box model by introducing the theory of unit hydrograph. The unit hydrograph is an impulse response function and as such is estimated directly as a FIR model, i.e., the flow is modeled as lagged values of precipitation. The unit hydrograph describes the relation between effective precipitation and quick flow. Hence, for the flow data, a base flow separation must be performed and the effective precipitation must be calculated from the precipitation data. Quite often, physical equations are used for effective precipitation calculations, e.g., Horton's infiltration formula [*Horton*, 1935] or Philip's equation [*Philip*, 1969]. Effective rain identification can also be incorporated in the hydrograph modeling process itself [e.g., *Hsu et al.*, 2002].

[4] For the purpose of flow predictions, ARX and ARMAX (autoregressive moving average exogenous) models are in most cases more successful than FIR models. This means that the flow is modeled not only as a function of precipitation, but also by using past flow values and in that case all the available information is applied. *Todini* [1978] used an ARMAX model for online flow predictions, and *Novotny and Zheng* [1990] used an ARMAX model for deriving watershed response function and their paper provides an overview of how ARMAX models, transfer functions, Green's functions and the Muskingum routing method are related.

[5] Both the FIR models and the ARMAX models are linear time-invariant models. These models are simple and easy to use and in many cases provide acceptable results, particularly when the volume of the flood is large compared to the infiltrated volume. Nevertheless, the rainfall-runoff process is believed to be highly nonlinear, time-varying and spatially distributed [e.g., *Singh*, 1964; *Chiu and Huang*, 1970; *Pilgrim*, 1976]. With increased computer power, nonlinear models have become increasingly popular. *Capkun et al.* [2001] handle the nonlinearity by using an ARX model and by modeling the variance as a function of past rainfall. Bayesian methods have also been applied; *Campbell et al.* [1999] used such a procedure for parameter estimation in their nonlinear flood event model. *Iorgulescu and Beven* [2004] used nonparametric techniques for the identification of rainfall-runoff relationship using direct mapping from the input space to the output space with good results. During the last decade neural networks have been popular as in the work by *Hsu et al.* [1995] and *Shamseldin* [1997] and, more recently, the SOLO-ANN model by *Hsu et al.* [2002]. *Karlson and Yakowitz* [1987] used a nonparametric regression method, which they refer to as the nearest neighbor method. They compare FIR, ARMAX and nearest neighbor models. Their results favor the nearest neighbor and the ARMAX models; however, they do not distinguish between the ARMAX and the

[1]Department of Informatics and Mathematical Modeling, Technical University of Denmark, Lyngby, Denmark.
[2]Faculty of Engineering, University of Iceland, Reykjavik, Iceland.
[3]Wastewater Control ApS, Virum, Denmark.

0043-1397/07/2005WR004500

nearest neighbor models. *Porporato and Ridolfi* [1996] used a nearest neighbor model and found that the local linear model with small neighborhoods gave the best results. *Porporato and Ridolfi* [1997] detected strong nonlinear deterministic components in the discharge series. They used noise reduction techniques specifically proposed for the field of chaos theory to preserve the delicate nonlinear interactions, and then they used nonlinear prediction (NLP) with good results. *Porporato and Ridolfi* [2001] followed up on these methodologies for multivariate systems. *Prevdi and Lovera* [2004] tackle the nonlinearity by using time-varying ARX models, which they refer to as nonlinear parameter varying models (NLPV). The parameter variation is defined as an output of a nonlinear function and the optimization is performed by using Neural Networks. *Young et al.* [2001] considered the time variable parameters to be state dependent and the method is thus referred to as the SDP approach. For nonlinear phenomena this approach results in a two-stage approach, called the data-based mechanistic approach (DBM) [*Young*, 2003]. In recent years fuzzy methods have been tested for flood forecasting [e.g., *Chang et al.*, 2005; *Nayak et al.*, 2005].

[6] In the present paper conditional parametric models are used to develop models for flow predictions in a sewage system. A conditional parametric model is a linear regression model where the parameters vary as a smooth function of some explanatory variable. Thus the method presented here is in a line with the SDP and the NLPV methodologies. The name conditional parametric model originates from the fact that if the argument of the functions is fixed then the model is an ordinary linear model [*Hastie and Tibshirani*, 1993; *Cleveland*, 1994]. In the models presented here, the parameters vary locally as polynomials of external variables, as described by *Nielsen et al.* [1997]. In contrast to linear methods like FIR and ARX, this methodology allows fixed input to provide different output depending on external circumstances.

[7] This paper is organized as follows: In section 2 the models are described, followed by section 3 with a description of the parameter estimation method. Section 4 contains results, and in section 5, online prediction and control in sewage systems are discussed. Finally, in section 6 conclusions are drawn.

2. Models

[8] In the present paper the excess outflow is modeled as a function of total precipitation (the base flow in the sewage system does not originate in rainfall). To avoid the calculation of infiltration it was decided to use the total precipitation as measured online. This is very convenient, particulary since the infiltration rate depends on several physical factors and no perfectly quantified general formula exists [*Viessman and Lewis*, 1996]. Some of the more recently developed models identify the effective precipitation along with the hydrograph [e.g., *Nalbantis et al.*, 1995]. The goal is to predict flow in the sewage system as a function of measured precipitation; consequently division of the precipitation into effective rain and infiltration/evaporation is not important. For the purpose of flow prediction, conditional parametric models are applied [*Nielsen et al.*, 1997]. These models are an extension of the well known linear regression model where the parameters vary as functions of some external variable. In this research two types of models were tested: conditional parametric FIR models and conditional parametric ARX models. The models are formulated as

FIR

$$y_t = \sum_{i=0}^{q_1} h_i(\mathbf{x}_{t-m}) z_{t-i} + e_t \quad e_t \in N\left(0, \sigma^2_{\mathrm{FIR}}\right) \tag{1}$$

ARX

$$y_t = \sum_{i=1}^{p} a_i(\mathbf{x}_{t-m}) y_{t-i} + \sum_{i=0}^{q_2} b_i(\mathbf{x}_{t-m}) z_{t-i} + e_t \quad e_t \in N\left(0, \sigma^2_{\mathrm{ARX}}\right) \tag{2}$$

where y_t is the output (flow), z_t is the input (precipitation), $\mathbf{x}_t$ is the explanatory variable and m is the time delay if any. Here the explanatory variable is season and/or threshold (see section 4). The order of the FIR model in equation (1) is denoted q_1 and the order of the ARX model in equation (2) is denoted (p, q_2).

[9] In the FIR model the function **h**, represented by the coefficients $h_i(\mathbf{x}_{t-m})$ $i = 1,\ldots q_1$ is known as the impulse response function. It demonstrates how the system responds to the input. In the ARX model $A_{\mathbf{x}_{t-m}}(q^{-1})$ is defined as the pth-order polynomial operator

$$A_{\mathbf{x}_{t-m}}\left(q^{-1}\right) = a_1(\mathbf{x}_{t-m})q^{-1} + \ldots + a_p(\mathbf{x}_{t-m})q^{-p} \tag{3}$$

where q^{-1} is the backward shift operator. Similarly $B_{\mathbf{x}_{t-m}}(q^{-1})$ is defined as:

$$B_{\mathbf{x}_{t-m}}\left(q^{-1}\right) = b_0(\mathbf{x}_{t-m}) + b_1(\mathbf{x}_{t-m})q^{-1} + \ldots + b_{q_2}(\mathbf{x}_{t-m})q^{-q_2} \tag{4}$$

as the q_2th-order polynomial operator. Then for a fixed $\mathbf{x}_{t-m}$ the impulse response function can be derived from the transfer function $A_{\mathbf{x}_{t-m}}(q^{-1})/B_{\mathbf{x}_{t-m}}(q^{-1})$ as described, for example, by *Ljung* [1987]. In this case the impulse response function includes coefficients h_0, h_1, ... up to infinity. However, in practice, only the first n coefficients are used.

[10] *Ashan and O'Connor* [1994] define the gain factor G of a unit hydrograph as

$$G = \frac{1}{A}\sum_{0}^{n} h_i \tag{5}$$

where A is the area of the watershed, n is the order of the model, the coefficients h_i are the coefficients in a unit hydrograph, where the input is effective rain and the output is excess flow. In an ideal situation the gain factor is one, but *Høybye and Rosbjerg* [1999] state that such a linear relationship does not exist. Furthermore, *Ashan and O'Connor* [1994] state that the overall model efficiency is in general very sensitive to the magnitude of the gain factor. In this paper the input is total precipitation; however, the value in equation (5), will be referred to as the gain factor. The gain factor will not be one. However, the gain provides valuable information about the system, it provides the fraction of the total precipitation that becomes excess

rainfall and thus also the fraction that infiltrates into the ground. Furthermore, the change in the gain factor as the external variable $\mathbf{x}$ changes provides a valuable information for understanding the system.

3. Estimation Method

[11] The models used are locally linear regression. In order to describe the ARX and FIR models together the notation is changed to the notation of a linear regression

$$y_t = \mathbf{z}_t^T \theta(\mathbf{x}_t) + e_t; \quad t = 1, \ldots, N, \quad e_t \in N\left(0, \sigma^2\right) \tag{6}$$

where the output or the response, y_t is a stochastic variable; $\mathbf{z}_t \in \mathcal{R}^k$ is the input; $\mathbf{x}_t \in \mathcal{R}^r$ is an explanatory variable. The parameter vector $\theta(\cdot) \in \mathcal{R}^k$ is a vector of smooth functions of $\mathbf{x}_t$, and $t = 1,\ldots, N$ are observation numbers. In the case of a FIR model, the variable $\mathbf{z}_t$ is the lagged values of the precipitation and $\theta(\mathbf{x_t})$ are the coefficients in a hydrograph. In the case of an ARX model, the $\mathbf{z}_t$ consist of the lagged values of precipitation and the lagged values of flow, where $\theta(\mathbf{x_t})$ consists of the corresponding parameters. If $\mathbf{x}_t$ is constant across all the observations, the model reduces to a traditional linear regression model, hence the name. The estimation of $\theta(\cdot)$ is accomplished by estimating the functions at a number of distinct values of $\mathbf{x}$. Given a point $\mathbf{x}$, each $\theta_{\mathbf{j}}$, $j = 1,\ldots k$ is approximated by a local linear function

$$\theta_j(\mathbf{x}_t) = \theta_{j0} + \theta_{j1}^T \mathbf{x}_t \qquad j = 1, \ldots, k. \tag{7}$$

The coefficients θ_{j0} and θ_{j1}^T are estimated by using weighted least squares (by using kernels).

[12] If $\mathbf{x}_t$ is two-dimensional $\theta_j(\mathbf{x}_t)$ can be written as

$$\theta_j(\mathbf{x}_t) = \theta_{j0} + \theta_{j1} x_{1t} + \theta_{j2} x_{2t} \qquad j = 1, \ldots, k \tag{8}$$

hence

$$\begin{aligned} y_t = {} & z_{1t}\theta_{10} + z_{1t}\theta_{11}x_{1t} + z_{1t}\theta_{12}x_{2t} \\ & + \ldots + z_{kt}\theta_{k0} + z_{kt}\theta_{k1}x_{1t} + z_{kt}\theta_{k2}x_{2t} + e_t. \end{aligned} \tag{9}$$

Then a row in a new design matrix can be defined as

$$\mathbf{u}_t^T = \left[z_{1t}, z_{1t}x_{1t}, z_{1t}x_{2t}, \ldots, z_{jt}, z_{jt}x_{1t}, z_{jt}x_{2t}, \ldots, z_{kt}, z_{kt}x_{1t}, z_{kt}x_{2t}\right] \tag{10}$$

and by defining the column vector

$$\theta_{j\mathbf{x}} = \left[\theta_{j0}, \theta_{j1}, \theta_{j2}\right] \tag{11}$$

and

$$\theta_{\mathbf{x}} = \left[\theta_{1\mathbf{x}}^T, \ldots, \theta_{j\mathbf{x}}^T, \ldots, \theta_{k\mathbf{x}}^T\right]^T \tag{12}$$

the flow vector $\mathbf{y_t}$ can be written as

$$y_t = \mathbf{u}_t^T \theta_{\mathbf{x}} + e_t \qquad t = 1, \ldots, N. \tag{13}$$

[13] The parameter vector $\theta_{\mathbf{x}}$ is fitted locally to $\mathbf{x}$. This is accomplished by using the traditional weighted least squares, where the weight on observation t is related to the distance from $\mathbf{x}$ to $\mathbf{x}_t$, so that

$$w_t(\mathbf{x}) = W(||\mathbf{x}_t - \mathbf{x}||/d(\mathbf{x})), \tag{14}$$

where $||\mathbf{x}_t - \mathbf{x}||$ is the Euclidean distance between $\mathbf{x}_t$ and $\mathbf{x}$. The function W: $\mathcal{R} \rightarrow \mathcal{R}$ is a nowhere increasing function. In this paper the tricube function

$$W(v) = \begin{cases} (1 - v^3)^3, & v \in [0; 1) \\ 0, & v \in [1; \infty) \end{cases} \tag{15}$$

is used. The scalar $d(\mathbf{x}) > 0$ is called the bandwidth. If $d(\mathbf{x})$ is constant for all values of $\mathbf{x}$, it is denoted a fixed bandwidth. On the other hand, if $d(\mathbf{x})$ is chosen so that a certain fraction of the observations is within the bandwidth, it is denoted as nearest neighbor bandwidth. The advantage of the tricube weighting function is that it is a smooth function like the Gauss bell, but unlike the Gauss bell the tricube function is zero outside the bandwidth, which makes the computational effort smaller. The choice of weighting function or kernel does not have a large impact [see *Silverman*, 1986].

[14] In general, if $\mathbf{x}$ has a dimension of two or larger, scaling of the individual elements of $\mathbf{x}$ before applying the method should be considered [e.g., *Cleveland and Develin*, 1988]. A rotation of the coordinate system, in which $\mathbf{x}$ is measured, could also be relevant. When the local estimate in equation (13) $\widehat{\theta}_{\mathbf{x}}$ is obtained, the elements of $\hat{\theta}(\mathbf{x})$ in equation (6) are calculated as

$$\hat{\theta}_j(\mathbf{x}_t) = [1, x_{1t}, x_{2t}]\widehat{\theta}_{j\mathbf{x}} \qquad (j = 1, \ldots k). \tag{16}$$

[15] When $\mathbf{z}_j = 1$ for all j this method is almost identical to the method introduced by *Cleveland and Develin* [1988]. Furthermore, if $\theta(\cdot)$ is a local constant, then the method of estimation reduces to determining the scalar $\hat{\theta}_j(\mathbf{x})$ so that $\sum_{t=1}^{n} w_t(\mathbf{x})(y_t - \hat{\theta}(\mathbf{x}))^2$ is minimized, i.e., the method reduces to traditional kernel estimation [see also *Härdle*, 1990; *Hastie and Loader*, 1993]. Furthermore, it is worth mentioning that, as for traditional linear regression, the fitted values $\widehat{y}_i$, $i = 1,\ldots, N$ are linear combinations of the observations [see *Nielsen et al.*, 1997].

[16] As noted earlier, the method of conditional parametric modeling has certain similarities to the SDP method [e.g., *Young et al.*, 2001], as well as the NLPV as used by *Prevdi and Lovera* [2004]. All these models are time varying ARMAX type of models. In the NLPV approach, the nonlinear optimization is by use of neural network, which is completely black box oriented. The nonlinear SDP methodology results in the two stage DBM approach [*Young*, 2003]. In the first stage an appropriate model structure is identified by considering a class of linear transfer function models whose parameters are allowed to vary over time. In the second stage any identified (significant) parameter variation is modeled using a parametric approach, and the parameters of the resulting parametric nonlinear model are estimated using nonlinear least squares or maximum likelihood estimation. Hence the

resulting model is a nonlinear model with fixed parameters. *Young* [2005] provides a fine overview and comparison of the SDP and NLPV modeling approaches.

[17] The suggested method is typically a one-stage approach. The resulting model is a conditional parametric model, where the total parametrization is a combined parametric and nonparametric model. Consequently, in every neighborhood, there exists an approximately linear parametric model. Furthermore, by studying the values of the parameter $\theta(\mathbf{x})$ as the external variable $\mathbf{x}$ changes, a complete parameterized model might be developed if that is desired. A complete parameterized model has both advantages and disadvantages: It often provides a better "physical" understanding of the system. However, the parameters are under all circumstances estimated by use of existing data and if the external circumstances change, this involves extrapolation. Local estimates (estimates in a neighborhood, kernel estimation) will adapt to new circumstances quickly. In this paper the nonlinearity is described directly without any use of recursive/adaptive estimation. In the case of time-varying models, adaptive estimation, as described by *Nielsen et al.* [2000], can be superimposed on the method. Hence the approach makes it possible to track time variation in a nonlinear model, this extension is however, not the focus of the present paper.

4. Results

4.1. Description of the Data and the Circumstances

[18] The data originate from the company Wastewater Control ApS in Denmark and consist of 68 rain events which occurred in the period 1st January 2003 to 4th May 2004. The 68 rain events cover many types of rain, of a varying intensity and length. The data consist of pairs of measured precipitation [mm/(6 min)] and excess flow [m^3/(6 min)]. The sampling time is, as indicated, 6 min. The excess flow is calculated from the total flow by subtracting the system's base flow. The base flow, or the dry weather flow, in the sewage system does not originate from rain and is defined as a constant plus a daily variation [see *Carstensen et al.*, 1998]. The sewage system is built up in the traditional manner as a net, with pumping stations located at some of the node points. During heavy rain events the volume of water entering the pipes can exceed the pumping station's capacity causing some kind of saturation/threshold in the system.

[19] The area of the watershed is 10.89 km^2. The impermeable area is dominated by urban area and the soil is mostly clay. The data contained 3 heavy rain events where the threshold/saturation phenomena can be seen. A water balance study was performed which showed a yearly variation in the water balance. This seasonality is mostly due to the soil moisture content in the root zone and variation in groundwater level; the soil is much drier during the summer than during the winter.

4.2. Model Construction

[20] An analysis of the data using linear models with nonvarying coefficients showed that the time delay from input to output is 2 lags, i.e., 12 min. It has been found that at most 19 lags are needed in a FIR model, as in equation (1). In the ARX models, as in equation (2), the "best" linear model, using AIC criteria, is *ARX*(2, 6) with a time delay of 2, i.e., the output y_t is a function of y_{t-1}, y_{t-2}, z_{t-2},...z_{t-7}. For the sake of convenience these model degrees were used in the whole study, i.e., the same number of lags was used in the conditional parametric models.

[21] The numerator in the ARX models is quite high compared to what is often seen in hydrology. However, most rainfall-runoff studies are on a daily basis. In this project the sampling time is 6 min, consequently the numerator needs to be higher. The linear model order was chosen by use of AIC/BIC criteria (the AIC and BIC indicated the same model order) and use of some other criteria might have led to lower orders. However, *Porporato and Ridolfi* [1996] indicate that the degree of the numerator should not be less than the basin concentration time, and in order to capture the entire subsequent runoff the numerator should be even greater. The basin concentration time in the sewage system, using the 6 min sampling time is about 6 lags (the basin lag is estimated to be 4 lags and, referring to *Singh* [1988], the time of concentration is 1.42 times the basin lag time, which is close to 6 lags). The order of the numerator in the ARX model is 6. Evidently the FIR model has a larger model degree than the ARX model.

[22] As mentioned earlier the nonlinear effects are mostly due to seasonal variations and the saturation/threshold effects in the pipes. The seasonal variation is modeled as the first term in a Fourier series, i.e., a sinus wave

$$x_t^s = C\sin(\omega t + \phi) \tag{17}$$

where x_t^s is the explanatory variable due to season. The water balance study showed the largest response to precipitation in February and the smallest in August. Consequently the parameters ω and ϕ are chosen such that x_t^s peaks in mid February. The parameter C is set to 100 which is a necessary scaling in the two-dimensional model presented later in this section. In practice the seasonal variation is not as regular as a sinus wave. However, since only 16 months of data are available it is not possible to estimate a seasonality function without restrictions as in equation (17). The seasonality in the parameters can most likely be modeled globally as in a PARMA model [e.g., *Rasmussen et al.*, 1996].

[23] The saturation/threshold effect is modeled either as a function of the rain intensity or as a function of the flow, depending on the model type. In a FIR model the conditional variable representing the saturation/threshold is precipitation intensity, x_t^p and set as

$$x_t^p = (u_{t-2} + u_{t-3} + u_{t-4} + u_{t-5} + u_{t-6})/5 \tag{18}$$

i.e., the average rain intensity in lags 2 to 6. This choice is based on the facts that the time delay is 2 lags, and the time of concentration is 6 lags.

[24] In the ARX models the saturation/threshold is modeled as a function of the flow itself instead of the rain intensity. This is in fact more physically correct because the threshold occurs because there is more water in the pipes than the pumps in the node points can serve, even though all this water is caused by heavy rain. Hence the explanatory variable is defined as

$$x_t^f = y_{t-1}. \tag{19}$$

Table 1. Coefficient of Determination R^2 for Various Models and Different Conditions[a]

	Condition	All	Winter (8 Events)	Spring (13 Events)	Summer (13 Events)	Fall (12 Events)	Heavy Rain (3 Events)
FIR	linear	0.79	0.64	0.73	0.68	0.90	0.83
FIR	seasonal	0.84	0.70	0.78	0.85	0.92	0.89
FIR	threshold/saturation	0.82	0.63	0.75	0.79	0.93	0.93
ARX	linear	0.94	0.91	0.94	0.91	0.96	0.93
ARX	seasonal	0.95	0.92	0.94	0.93	0.96	0.94
ARX	threshold/saturation	0.96	0.93	0.95	0.94	0.97	0.97
ARX	seasonal × threshold/saturation	0.97	0.95	0.96	0.97	0.98	0.99

[a]The unit is m^3/6 min as the sampling time is 6 min. The R^2 calculations are one-step prediction, performed using overall data. For winter, spring, summer, and fall, only 2 months were used for better seasonal distinction. Finally, R^2 is calculated for the three heaviest rain events; those events are excluded in the seasonal calculations.

There were only 3 heavy rain events during this period and since 19 coefficients need to be identified in the FIR model, the 3 events with heavy rainfall were not quite enough to identify the 19 coefficients within an acceptable confidence level, meaning that several combinations of solutions might be possible. However, some solutions were found and those were used for prediction. As a consequence of this sparse data it was not possible to identify a FIR model where the coefficients varied both with the season and the threshold. On the other hand in the ARX models the constants are rather well identified and it was possible to identify coefficients depending on two variables, season and flow.

[25] The local estimation requires bandwidth decisions. The bandwidth determines the smoothness of the estimate. If the bandwidth is small the variance is large and the bias is small. If the bandwidth is large the variance is small but the bias increases. An "optimal" bandwidth is a bandwidth which is a compromise of these two factors. In the traditional kernel estimation, as by *Härdle* [1990], the estimates are local constant; here the estimates are local lines. This allows a larger bandwidth without the cost of a bias problem. The bandwidths are different, depending on the model types. In each case the bandwidths were found by manual optimization, the bandwidth needs to be small enough to detect differences in the conditional variable. However, the larger it is, the less variation in the estimates. For example in the ARX model where the conditional variable is the season, it was possible to use a large bandwidth. The seasonal variable is almost evenly distributed, and the optimal bandwidth included 2400 data points, which is about 65% of the data. On the other hand in the FIR model with rain intensity as a conditional variable, the bandwidth included only 55 data points, which is about 1.5% of the data. This is because there are few events with heavy rain, and thus, by using a larger bandwidth, the few data points no longer have an effect.

[26] The calculations are performed by using a program named LFLM (locally weighted fitting of linear models) which is an S-PLUS/R library package. For a description, see *Nielsen* [1997].

4.3. Modeling Results

[27] Model validation demands some measure of the model's quality. This measure is not a single number which can be used for each and every model and in each and every situation. In this project the main goal was to achieve accurate predictions. Thus the optimization (model calibration) is a least squares method and as such the model's performance is validated with respect to that. In hydrology several other factors might be of higher importance, like the overall water balance, the timing of the peak flow or other things.

[28] The coefficient of determination R^2, often referred to as the Nash efficiency, is a widely used model criterion in hydrology and it is a fine measure of the model's efficiency with respect to the least squares minimization. The residuals used for model validation are one step prediction errors, using the calibration data series. A cross validation would have been adequate. However, the model's parameters were estimated locally and will thus adapt to the data in use; hence cross validation does not have the same meaning as when the parameters are estimated globally. Table 1 shows the R^2 for the conditional parametric models. As the mass balance is an important dimension in hydrology, the mean value of the error was also calculated. The mean value of the error demonstrates the mass balance on average. If the mass balance is well conserved, the mean value of the error will be zero. Table 2 shows the mean value of the error. Tables 1 and 2 show the overall performance and also the seasonal performance. For each season the seasonal calculations are based two months in each season for better distinction between the seasons. For the seasonal calculations the 3 most heavy rain events were excluded. These were events, where the saturation/threshold effect exists, those were grouped together and the calculations were performed for them separately in order to measure how the models perform in that situation. As a reference R^2 was also calculated for the corresponding well known linear models FIR and ARX. It is well known that the coefficient of determination does not penalize overparameterizations. However, both AIC and BIC studies led to this model order as does a physical study like the basin concentration time as discussed in section 4.2. Specifically the conditional parameter models are compared to a linear model of same orders. Thus the FIR models are comparable, and the ARX models are as well. The estimation was based on all the events, and Tables 1 and 2 show total results for all the 68 rain events.

[29] In a comparison of the three FIR models, the seasonal FIR model has the best performance both with respect to the R^2 and to the mass balance. The seasonal FIR is clearly an improvement of the traditional FIR model with constant parameters. Even in a situation with heavy rain, the seasonal FIR outperforms the traditional FIR. The threshold

Table 2. Mean Value of the Error[a]

	Condition	All	Winter (8 Events)	Spring (13 Events)	Summer (13 Events)	Fall (12 Events)	Heavy Rain (3 Events)
FIR	linear	1.13	11.18	19.54	−22.80	−2.4	−37.22
FIR	seasonal	0.23	0.76	15.79	−1.14	−6.8	−10.23
FIR	threshold/saturation	3.40	14.13	20.47	−17.12	−2.3	−17.50
ARX	linear	0.05	1.67	2.79	−3.84	−0.48	−7.58
ARX	seasonal	−0.20	0.01	2.72	−0.27	−1.45	−2.58
ARX	threshold/saturation	0.12	1.77	3.3	2.85	−1.29	−4.74
ARX	seasonal × threshold/saturation	−0.01	0.02	2.37	0.19	1.48	−1.76

[a]Values are in m^3/6 min.

FIR is the best during heavy rain events, as a result of its design, however the bias is quite large.

[30] All the conditional parametric ARX models outperform the traditional ARX, both the one-dimensional models and the two-dimensional model, which is the best both with respect to the Nash efficiency and the bias.

[31] In Table 2 it can be seen how the traditional linear models underestimate the runoff during the winter and overestimate it during the summer, especially the FIR models. The FIR models have a larger bias, and even the seasonal FIR is not quite acceptable in all seasons, especially in the spring season. Thus the FIR models are not acceptable for predictions.

[32] As expected, the ARX models outperform the FIR models. However, it must be stressed that the FIR models and the ARX models need different inputs for prediction. Using a FIR model, one-step prediction demands past precipitation which is also required when using an ARX model, but past values of the flow are additionally required. For k-step prediction both the FIR and the ARX models need past and present values of the precipitation and also $(k - 2)$ prediction of the precipitation. Additionally the ARX models demand past, present and $(k - 1)$ step prediction of the flow. It must be mentioned though that using predicted values of precipitation as input will never be quite as reliable as using measured values since the predicted values have much larger variance; this is also true for the predicted values of the flow. Moreover, the parameter estimates are performed with the assumption that the input is measured, not predicted.

[33] For visual comparison two events were chosen. These are a 'typical' event in winter time and an event with heavy rain, showing the theshold/saturation. Note that even though single events are shown in Figures 1–3, the estimate is based on all the events. A single conditional ARX model and a single conditional FIR model will be drawn along with the traditional linear models. The figures show the best conditional FIR model, the one-dimensional seasonal FIR and the best conditional ARX model, the two-dimensional ARX along with the linear FIR and ARX, for comparison.

[34] Figure 1 shows an event in the winter time; the duration of the event is about 6.5 hours. Note that the linear

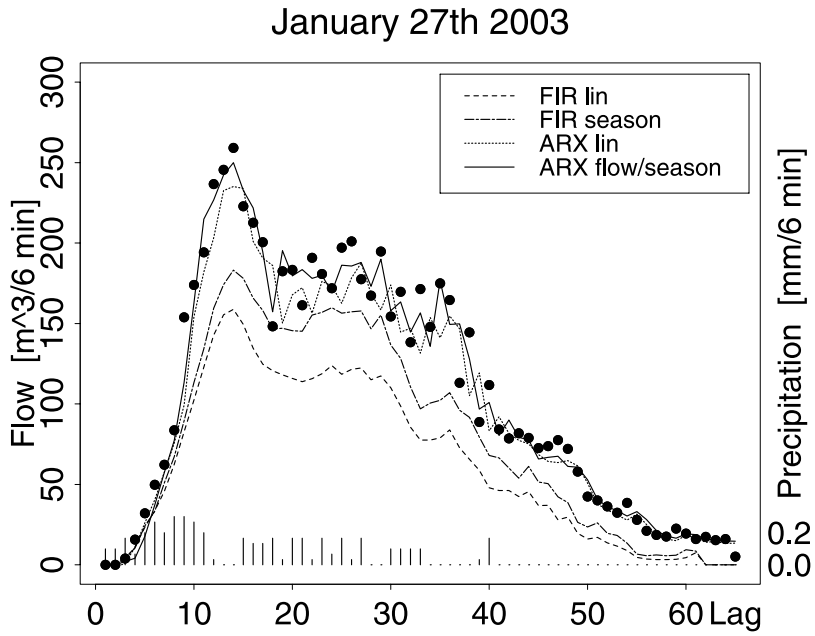

Figure 1. Event in winter. The data are shown as points, and the precipitation is shown as bars, with the scale on the right axis. The time lag is 6 min, and a conditional ARX model, where the parameters depend on season and flow, and a conditional FIR model, where the parameters depend on season, are shown. For comparison the conventional time-invariant linear models FIR and ARX are also shown.

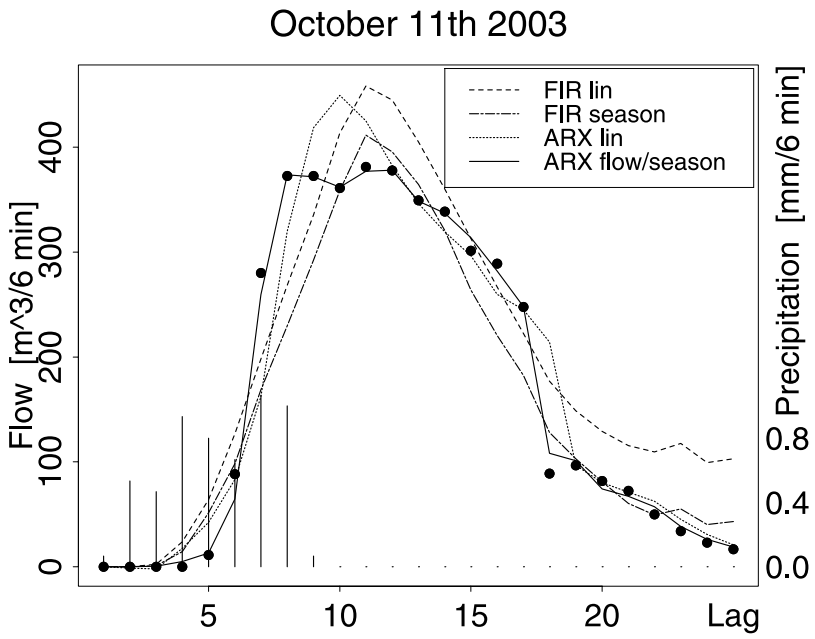

Figure 2. Event in autumn. The data are shown as points, and the precipitation is shown as bars, with the scale on the right axis. The time lag is 6 min. A conditional ARX model, where the parameters depend on season and flow, and a conditional FIR model, where the parameters depend on season, are shown. For comparison the conventional time-invariant linear models FIR and ARX are also shown.

FIR model underestimates the runoff as demonstrated in Table 2, and a seasonal FIR model is clearly an improvement on the traditional linear FIR. The linear ARX model is better, but not as good as the conditional parametric ARX. Figure 2 shows the same for an event with heavy rain and thus the threshold/saturation effect; the duration of this event is about 2.5 hours. In this case both the linear FIR and the linear ARX overestimate the flow peak, as does a seasonal FIR, while the conditional ARX nicely captures the flat and long peak.

[35] It might be argued that in real applications a confidence interval for the predictions would be required. This is indeed true; confidence intervals for the predicted output are valuable. However, since the models are nonlinear, it is believed that prediction intervals should be estimated by

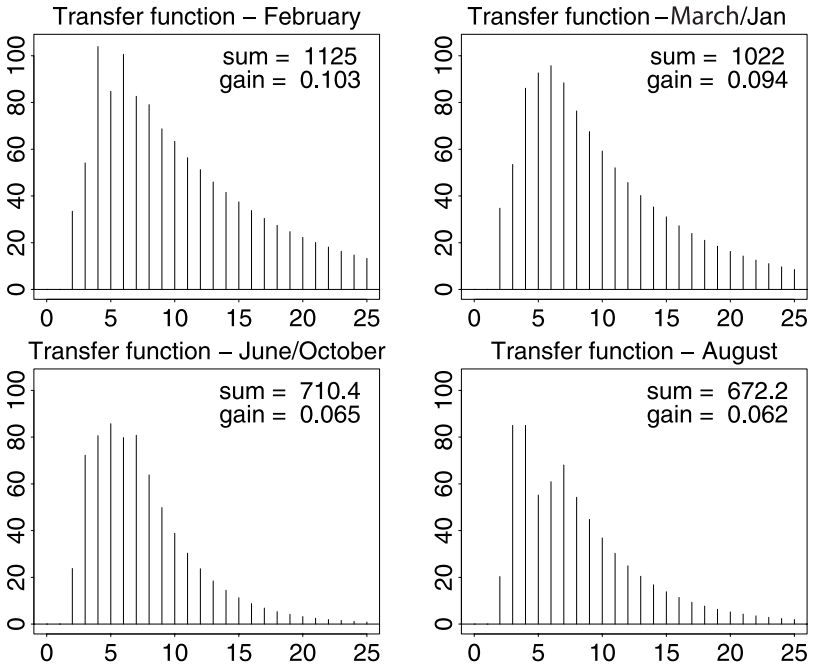

Figure 3. Impulse response function estimates for four different seasons calculated using a seasonal ARX model.

Table 3. Local Parameter Estimates in a Seasonal ARX Model[a]

Season	a_1	a_2	b_2	b_3	b_4	b_5	b_6	b_7
S_1	0.34	0.51	33.45	42.75	68.56	21.98	19.13	5.48
S_2	0.62	0.20	32.66	50.67	48.33	23.57	21.12	5.23
S_3	0.82	−0.03	23.80	52.73	22.09	21.88	11.95	18.04
S_4	0.73	0.08	20.36	70.23	21.59	−13.30	14.13	19.44

[a] S_1 = February, S_2 = April/December, S_3 = June/October, and S_4 = August.

methods like quintile regression as by *Nielsen et al.* [2006]. This is not covered in this paper.

[36] Finally, conditional parametric models with local estimates can also be used to study the circumstances of the watershed and thus provide a useful information for developing a nonlinear global parametric model if wanted.

[37] For example, the seasonality can be studied. For this purpose a seasonal ARX model is used. In this study the conditional variable representing the seasonality is a sinus wave and the parameters are estimated as local lines, depending on the values of the sinus. Table 3 shows estimated coefficients in the ARX for fixed values of the season. Because of symmetry, it is not possible to distinguish between spring and autumn. A comparison of the autoregressive parameters a_1 and a_2, shows that a_2 is larger than a_1 during the winter while a_2 is close to zero during the summer and a_1 is the dominating autoregressive parameter. The negative value of the parameter b_5 in August is physically incorrect, and this is probably due to sparse data, since there is only one summer season and August is close to and on the boundary of the seasonal variation parameter.

[38] Using the estimated parameters, the impulse response function can be calculated, as shown in Figure 3, which also shows the sum of the coefficients and the corresponding gain factor, calculated by equation (5). Note that the impulse response function has the longest tail during winter and shortest tail during summer, and it also reaches larger values during winter than summer. Consequently, during winter about 10% of the total water reaches the sewage system, while during summer about 6% enters the sewage system. A similar analysis has been performed for the flow dependence of the impulse response function and it mostly shows that when the flow is large the impulse response function is flatter, it peaks later, the values are smaller, and the tail is longer.

5. Discussion

[39] The FIR models provide two-step prediction, i.e., information 12 min ahead, since the time delay between precipitation and flow is 2 lags. The ARX models provide one-step prediction, since flow at time $t - 1$ is used for prediction. For real time online prediction and automatic control it might be necessary to achieve information with longer time horizon, say 30 min, i.e., five-step prediction. For both the FIR and the ARX models a five-step prediction requires three-step prediction of precipitation, i.e., online weather forecast. However, since it is only a question of a couple of minutes, it might be possible to use online precipitation measurements a bit further from the treatment plant, i.e., a weather station capturing the frontal rain a little bit earlier. Evidently this depends on the wind and frontier movement direction although in many cases the wind during rain is from the (south) west, which is the dominating wind direction. The wind direction might also be a conditional variable in the model if enough data are available. On the other hand, for the ARX model the situation is a little bit more complicated because the flow y_{t-1} is a conditional variable. The most practical thing would be to provide online flow data from a couple of node points in the sewage system net, node points which are distributed geographically in the sewage system. The flow in the node points is naturally delayed compared to the flow in the wastewater treatment plant, and obviously the delay is different depending on the geographical localization. However, if data from the node points are available in general, it would be most convenient to use the flow in the node points as an input in a model for online prediction and control, at the wastewater treatment plant, and thereby remove much of the unaccountable rain distribution.

6. Summary and Conclusion

[40] Conditional parametric models have been developed and tested for rainfall-runoff modeling in a sewage system. The models are FIR and ARX models with the coefficients varying as a function of external variables. The input of the models is the total precipitation as measured online, and the output is the excess flow prediction. The base flow is separated by using simple equations since the base flow in the sewage system does not originate in rainfall.

[41] Both the conditional parametric FIR and the conditional parametric ARX provide results which are significantly superior to results from conventional linear models. As expected the ARX models provide the best one-step predictions.

[42] In this study the conditional variables are used to capture seasonal fluctuations and threshold/saturation due to the limited capacity of the system pumps and pipes.

[43] Use of this modeling approach has a good potential for developing good prediction models. Furthermore, the method can also be used for sensitivity analysis while constructing a physical model of the system flow. The method of conditional modeling is a useful contribution to the tools of nonlinear modeling techniques used in hydrology.

[44] **Acknowledgments.** The authors wish to thank the company Wastewater Control ApS in Denmark for delivering the data and information about the system. Furthermore, the authors wish to thank the Hydrological Service at the National Energy Authority of Iceland for lending their facilities.

References

Ashan, M., and K. M. O'Connor (1994), A simple non-linear rainfall-runoff model with a variable gain factor, *J. Hydrol.*, *155*, 151–183.

Campbell, E. P., D. R. Fox, and B. C. Bates (1999), A Bayesian approach to parameter estimation and pooling in nonlinear flood event models, *Water Resour. Res.*, *35*(1), 211–220.

Capkun, G., A. Davison, and A. Musy (2001), A robust rainfall-runoff transfer model, *Water Resour. Res.*, *37*(12), 3207–3216.

Carstensen, J., M. K. Nielsen, and H. Strandbæk (1998), Prediction of hydraulic load for urban storm control of municipal WWTP, *Water Sci. Technol.*, *37*(12), 363–370.

Chang, L.-C., F.-J. Chang, and Y.-H. Tsai (2005), Fuzzy exemplar-based inference system for flood forecasting, *Water Resour. Res.*, *41*, W03002, doi:10.1029/2004WR003514.

Chiu, C. L., and J. T. Huang (1970), Nonlinear time-varying model of rainfall runoff relation, *Water Resour. Res.*, *6*(1), 1277–1286.

Cleveland, W. S. (1994), Coplots, nonparametric regression, and conditionally parametric fits, in *Multivariate Analysis and Its Applications*, edited by T. W. Anderson, K. T. Fang, and I. Olkin, pp. 21–36, Inst. of Math. Stat., Hayward, Calif.

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

Härdle, W. (1990), *Applied Nonparametric Regression*, Cambridge Univ. Press, New York.

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

Hastie, T., and R. Tibshirani (1993), Varying-coefficient models, *J. R. Stat. Soc.*, *55*, 757–796.

Horton, R. E. (1935), Surface runoff phenomena: Part I. Analysis of the hydrograph, technical report, *Horton Hydrol. Lab. Publ. 101*, Edwards Bros., Ann Arbor, Mich.

Høybye, J., and D. Rosbjerg (1999), Effect of input and parameter uncertainties in rainfall-runoff simulations, *J. Hydrol. Eng.*, *4*, 214–224, doi:10.1061/ (ASCE)1084-0699 [1999]4:3 (214).

Hsu, K., H. V. Gupta, and S. Sorooshian (1995), Artificial neural network modeling of the rainfall-runoff process, *Water Resour. Res.*, *31*(10), 2517–2530.

Hsu, K., H. V. Gupta, Z. Gao, and S. S. B. Imam (2002), Self-organizing linear output mat (SOLO): An artificial neural network suitable for hydrologic modeling and analysis, *Water Resour. Res.*, *38*(12), 1302, doi:10.1029/2001WR000795.

Iorgulescu, I., and K. Beven (2004), Nonparametric direct mapping of rainfall-runoff relationships: An alternative approach to data analysis and modelling, *Water Resour. Res.*, *40*(8), W08403, doi:10.1029/2004WR003094.

Karlson, M., and S. Yakowitz (1987), Nearest-neighbor methods for nonparametric rainfall-ruoff forecast, *Water Resour. Res.*, *27*(7), 1300–1308.

Ljung, L. (1987), *System Identification: Theory for the User*, Prentice-Hall, Upper Saddle River, N. J.

Nalbantis, I., C. Obled, and J. Rodriguez (1995), Unit hydrograph and effective precipitation identification, *J. Hydrol.*, *168*, 127–157.

Nayak, P., K. P. Sudheer, D. Rangan, and K. S. Ramasastri (2005), Short-term flood forecasting with neurofuzzy model, *Water Resour. Res.*, *41*, W04004, doi:10.1029/2004WR003562.

Nielsen, H. A. (1997), LFLM version 1.0: An SPLUS/R library for locally weighted fitting of linear models, *Tech. Rep. 22*, Dep. of Math. Modell., Tech. Univ. of Denmark, Lyngby. (Available at: http://www.imm.dtu.dk/~han/software.html)

Nielsen, H. A., T. S. Nielsen, and H. Madsen (1997), Conditional parametric ARX-models, in *11th IFAC Symposium on System Identification*, vol. 2, pp. 475–480, Elsevier, New York.

Nielsen, H. A., T. S. Nielsen, A. K. Joensen, H. Madsen, and J. Holst (2000), Technical note: Tracking time-varying-coefficient function, *Int. J. Adaptive Control Signal Process.*, *14*, 813–827, doi:10.1002/1099-1115 (200012).

Nielsen, H. A., H. Madsen, and T. S. Nielsen (2006), Using quantile regression to extend an existing wind forecasting system with probabilistic forecast, *Wind Energy*, *9*, 95–108, doi:10.1002/we.180.

Novotny, V., and S. Zheng (1990), Rainfall-runoff transfer function by ARMA modeling, *J. Hydraul. Eng.*, *115*, 1386–1400.

Philip, J. R. (1969), Theory of infiltration, *Adv. Hydrosci.*, *5*, 215296.

Pilgrim, D. H. (1976), Travel times and nonlinearity of flood runoff from tracer measurements on a small watershed, *Water Resour. Res.*, *31*, 2517–2530.

Porporato, A., and L. Ridolfi (1996), Clues to the existence of deterministic chaos in river flow, *Int. J. Mod. Phys. B*, *10*(15), 1821–1862.

Porporato, A., and L. Ridolfi (1997), Nonlinear analysis of river flow time sequences, *Water Resour. Res.*, *33*(6), 1353–1367.

Porporato, A., and L. Ridolfi (2001), Multivariate nonlinear prediction of river flows, *J. Hydrol.*, *248*, 109–122.

Prevdi, F., and M. Lovera (2004), Identification of non-linear parametrically varying models using separable least squares, *Int. J. Control*, *77*(16), 1382–1392, doi:10.1080/00207170412333188863.

Rasmussen, P. F., J. D. Salas, L. Fagherazzi, J.-C. Rassam, and B. Bobe (1996), Estimation and validation of conteporaneous PARMA models for streamflow simulation, *Water Resour. Res.*, *32*(10), 3151–3160.

Shamseldin, A. Y. (1997), Application of a neural network technique to rainfall-runoff modelling, *J. Hydrol.*, *199*, 272–294.

Sherman, L. K. (1932), Streamflow from rainfall by the unit-graph method, *Eng. News Rec.*, *108*, 501–505.

Silverman, B. (1986), *Density Estimation for Statistics and Data Analysis*, CRC Press, Boca Raton, Fla.

Singh, V. P. (1964), Nonlinear instantaneous unit hydrograph theory, *J. Hydraul. Div. Am. Soc. Civ. Eng.*, *90*(HY2), 313–347.

Singh, V. P. (1988), *Hydrologic Systems*, vol. 1, *Rainfall-Runoff Modelling*, Prentice-Hall, Upper Saddle River, N. J.

Todini, E. (1978), Using a desk-top computer for an on-line flood warning system, *IBM J. Res. Dev.*, *22*, 464–471.

Viessman, W., and G. L. Lewis (1996), *Introduction to Hydrology*, Harper-Collins, New York.

Young, P. C. (2003), Top-down and data-based mechanistic modelling of rainfall-flow dynamics at the catchment scale, *Hydrol. Processes*, *17*, 2195–2217.

Young, P. C. (2005), Comments on identification of non-linear parametrically varying models using separable least squares by F. Previdi and M. Lovera: Black box or open box, *Int. J. Control*, *78*(2), 122–127, doi:10.1080/00207170500073772.

Young, P. C., P. McKenna, and J. Bruun (2001), Identification of non-linear stochastic systems by state dependent parameter estimation, *Int. J. Control*, *74*(18), 1837–1857, doi:10.1080/00207170110089824.

J. Eliasson, H. Jonsdottir, and O. P. Palsson, Faculty of Engineering, University of Iceland, Hjardarhaga 2-6, 107 Reykjavik, Iceland. (halloharpa@gmail.com)

H. Madsen and H. A. Nielsen, Department of Informatics and Mathematical Modeling, Technical University of Denmark, Building 321, DTU, DK-2800 Lyngby, Denmark.

M. K. Nielsen, Wastewater Control ApS, Kollemosevej 47, DK-2830 Virum, Denmark.