Proceedings of the 17th World Congress
The International Federation of Automatic Control
Seoul, Korea, July 6-11, 2008

Flood forecasting for heteroscedastic streamflow processes

Francesca Pianosi * **Luciano Raso** **

* *Dipartimento di Elettronica e Informazione, Politecnico di Milano, Milan, Italy (Tel: +39-2-2399.9630; e-mail: pianosi@elet.polimi.it).*
** *Politecnico di Milano, Milan, Italy (e-mail: raso@asp-poli.it).*

Abstract: The paper presents a nonlinear heteroscedastic model for flow forecasting. The model is composed of two submodels: the former provides the expected value of the flow, conditional on available information, e.g. past flow and precipitation records; the latter provides the variance of the prediction error as a function of past values of the prediction error itself and precipitation measures. The proposed model is tested on a real world case study, the inflow to Lake Verbano, Italy, where the inflow forecast is used for optimizing release decisions from the lake. Results are discussed and compared with those obtained with conventional modelling approach, where the flow is estimated based on a linear model of the flow logarithm, and the variance is not given a dynamical description but is assumed to be a time-varying parameter.

1. INTRODUCTION

Most of the methods in conventional prediction theory are based on the assumption that the model residual has a constant variance. As pointed out by Engle (1982), mainly with regards to econometric applications, often this assumption is not satisfied: on the contrary, large errors are more likely to be followed by large errors and small errors by small errors. In brief, an autoregressive effect can be revealed in the variance of the model residual. These considerations led Engle to propose an AutoRegressive Conditional Heteroscedasticity (ARCH) model for a discrete-time stochastic process $y = \{y_0, y_1, \ldots\}$. According to the ARCH model, the expected value $E[y_{t+1}|I_t]$, conditional on information I_t at time t, is a linear combination of past records of the variable itself and possibly of other measures, as in conventional regression analysis, while the conditional variance $V[y_{y+1}|I_t]$ is not constant but it is given by a linear combination of past values of the squared prediction error. Bollerslev (1986) generalized this approach by using an ARMA model to describe the conditional variance, and called the global model a Generalized AutoRegressive Conditional Heteroscedasticity (GARCH) model.

In environmental systems modelling, a typical example of stochastic process that shows non-uniform conditional variance is the outflow process from an uncontrolled catchment. However, to the author's knowledge, few attention is paid to this specificity in the literature, even if, to the purpose of flood forecasting and control, improving the accuracy in the estimation of the conditional variance is almost as crucial as improving the prediction ability. In the case study that will be discussed in this paper, the goal is to identify an inflow model to be used in an on-line control scheme for the management of a regulated lake. Since the control scheme is based on stochastic optimization, the performances of the closed-loop system depend on the accuracy of the estimate of both the expected value and the variance of the inflow.

The paper is organized as follows. The next section is devoted to a review of the conventional approach for modelling flow processes and accounting for heteroscedasticity. The limit of this approach when exogenous input (e.g. precipitation) is introduced in the model is discussed in section 3. To overcome the difficulty, a nonlinear model is introduced and its prediction ability tested on to a real world case study. Then, in section 4 a dynamical model of the variance of the prediction error is presented and compared with conventional static models. Final remarks and issues for future research are presented in the last section.

2. MODELLING AUTONOMOUS FLOW PROCESSES

The outflow process q from an uncontrolled catchment can be modelled as a discrete-time stochastic process. The probability distribution of the process, i.e. the distribution of q_t at each time t, can be approximately assumed to be uniform and log-normal. As a consequence, the process y, where $y_t = \log(q_t)$ for each t, is normal and it can be suitably modelled as an autoregressive process. In general, q (and thus y) is cyclostationary, due to exogenous forcing inputs that are periodic, e.g. the contribution of snow melt in late spring and summer. Periodicity is usually accounted for in two ways: 1) the normal process y is first standardized through its periodic mean μ_t and standard deviation σ_t, and then the normal standard process $\tilde{y}$ so obtained is modelled as an autoregressive process (deseasonalized-ARMA model); 2) the process y is directly modelled as a periodic autoregressive model, i.e. with periodic parameters (PARMA model) Hipel and McLeod (1994). In the example of the snow melt contribution, the first approach seems more suitable, since it recognizes the origin of periodicity from exogenous inputs and not from a change in the system dynamics. Once a model of the process y is available, at each time t the prediction $y_{t+1|t}$ can be obtained based

⋆ Corresponding author F. Pianosi. Tel. +39-2-2399.9630. Fax +39-2-2399.9611.

on the available information $I_t = |\log(q_t), \log(q_{t-1}), ...|$. The prediction $E[q_{t+1}|I_t]$ of the original variable and its conditional variance $V[q_{t+1}|I_t]$ are then given by

$$E[q_{t+1}|I_t] = \exp(2y_{t+1|t} + \sigma^2)/2 \quad (1a)$$

$$V[q_{t+1}|I_t] = \exp(2y_{t+1|t} + 2\sigma^2) - \exp(2y_{t+1|t} + \sigma^2) \quad (1b)$$

where σ^2 is the (constant) conditional variance of $y_{t+1|t}$. From equation (1b), it follows that the conditional variance of q_{t+1} changes over time and increases with the value of the prediction, i.e. that the model is able to reproduce heteroscedasticity of the flow process.

Wang et al. (2005) consider daily and monthly flow time series of upper Yellow River, China, and show that the residual of the deseasonalized-ARMA model is not a stationary homoscedastic process. Instead, the residual shows, depending on the sample time (daily or monthly), either cyclo-stationarity or both cyclo-stationarity and heteroscedasticity. For the latter case, they propose an ARMA-GARCH model for describing the deseasonalized process $\tilde{y}$, which results in a fully dynamical description of the residual variance. To the authors' knowledge, here the time-variability and heteroscedasticity of the residual process can not be given any physical interpretation, while it must be interpreted as an evidence that the standardization of y does not fully capture its seasonality, and that the ARMA model is not completely adequate to modelling $\tilde{y}$.

3. MODELLING FLOW PROCESSES WITH EXOGENOUS INPUT

The approach based on logarithmic transformation of the flow values presents two main limits. First, the log-transformation results in a bias of the flow prediction towards small values, which can be unsatisfactory, especially if the model is to be used for flood forecasting (see discussion in Romanowicz (2006)). Second, the model is suitable for describing the flow formation process as an autonomous process, while the introduction of exogenous inputs can be critical. For example, assume that precipitation measures are added as an exogenous information in the model of process y, i.e. consider the following deseasonalized-ARMAX model

$$y_t = \left(\log(q_t) - \mu_t^{\log q}\right)/\sigma_t^{\log q} \quad (2a)$$

$$y_t = \frac{B(z^{-1})}{A(z^{-1})}p_t + \frac{C(z^{-1})}{A(z^{-1})}e_t \quad (2b)$$

where $\mu_t^{\log q}$ and $\sigma_t^{\log q}$ are the a priori estimates of the mean and standard deviation of the flow logarithm, for time t (in general, they are time-varying periodic); p_t is the measure of average precipitation over the catchment for the time interval $[t-1, t)$, e_t is a zero mean white noise, and $A(z^{-1})$, $B(z^{-1})$ and $C(z^{-1})$ are polynomials in the backward shift operator z^{-1} (such that $z^{-1}q_t = q_{t-1}$), of order n_a, n_b and n_c respectively, i.e.

$$A(z^{-1}) = 1 - a_1 z^{-1} - \ldots - a_{n_a} z^{-n_a}$$
$$B(z^{-1}) = b_1 z^{-1} + \ldots + b_{n_b} z^{-n_b}$$
$$C(z^{-1}) = 1 + c_1 z^{-1} + \ldots + c_{n_c} z^{-n_c}$$

If the prediction of y_{t+1}, based on model (2), is used to derive the prediction of q_{t+1} through equation (1a), then q_{t+1} turns out to be exponentially related to precipitation p_t. This is in contrast with the physical knowledge of the flow formation process, which shows an approximately linear relation among precipitation and flow, and worsens the prediction ability of the model, since floods due to precipitation peaks are overestimated. In particular, when using the model outside of the estimation data set, flood peaks can be dramatically overestimated whenever the precipitation exceeds the maximum value recorded in the estimation data set, as it will be shown when discussing the case study.

3.1 Nonlinear model of the flow

In order to overcome the difficulties pointed out in the previous section, the following model will be discussed here

$$y_t = \frac{1}{A(z^{-1})}e_t \quad (3a)$$

$$q_t = \exp\{\sigma_t^{\log q} y_t + \mu_t^{\log q}\} + B(z^{-1})p_t + C(z^{-1})\varepsilon_t \quad (3b)$$

where all variables are defined as in model (2) and ε_t is the model residual, i.e. the difference between observed and predicted flow, assumed to be a zero mean white noise.
The rational behind model (3) is the following. If no exogenous input enters the system, the flow shows a decrease, which is accounted for by the autoregressive model of the flow's logarithm. The presence of the logarithm ensures a description of the recession phase that is consistent with both hydrological theory and the trend revealed by data. The residual from this autonomous logarithmic model is linearly related to precipitation. The underlying idea is that what can not be explained by autocorrelation, must be due to the exogenous forcing input, i.e. the precipitation. Finally, the flow prediction can be corrected based on past errors. This moving average component can prove to be useful for correcting the lag in the peaks prediction, if any, as discussed below.

Model (3) is nonlinear, therefore traditional parameter estimation techniques can not be directly applied. The identification strategy that we propose is the following. An initial estimate of the model parameters, namely the coefficients of the polynomials $A(z^{-1})$, $B(z^{-1})$ and $C(z^{-1})$, is obtained through a sequential identification based on ordinary least squares: first, the coefficients of the polynomial $A(z^{-1})$ are estimated, based on a sample of the deseasonalized process y; least squares are then used again to fit a linear regression among the residual $q_t - \exp\{\sigma_t^{\log q}(1 - A(z^{-1}))\log(q_t) + \mu_t^{\log q}\}$ and precipitation, thus obtaining the coefficient of polynomial $B(z^{-1})$; the same is repeated with the residual from the autoregressive-exogenous model, to derive the coefficient of $C(z^{-1})$. This initial parameter estimate is sub-optimal because 1) it is obtained sequentially while all parameters should be estimated at once; 2) the coefficient of the moving average component have been estimated over the deterministic component residual, not the residual of the global model. Therefore, this estimate is used to initialize a global optimization of the parameter estimates performed with Levenberg-Marquardt algorithm.

3.2 Application results

The prediction ability of model (3) is tested on a real world case study. The output variable is the net inflow

n_a	n_b	n_c	R^2 est.	R^2 val.	
2	1	0	0.29	-2.19	deseasonalized-ARMAX (2)
			0.68	0.73	model (3)
2	2	0	0.63	-0.01	deseasonalized-ARMAX (2)
			0.69	0.74	model (3)

Table 1. Coefficient of determination R^2 over estimation and validation data set for the deseasonalized-ARMAX model and model (3), with different choices of the model order. Results obtained with higher order models (in particular $n_c > 0$) are not reported because no significant improvement is obtained.

q_t to the Lake Verbano, a multipurpose regulated lake in northern Italy. This variable can not be measured but it is derived from level and release measures by inverting the reservoir's mass balance equation. The input is the total precipitation (both rainfall and snowfall undistinguished) over the catchment. This is obtained by averaging the values relevant to 18 meteorological stations. Time series of these variables are available for the period 1/Jan/1992 - 30/Nov/2000. Data relevant to the period 1992-1997 are used for the model identification, while data from the beginning of 1998 to the end of the sample are used for validation.

Since the lake release decision is taken daily, and the inflow prediction is going to be used to optimize decisions, the inflow model must be time-discrete with a sample time Δ equal to 24 hours. As it shall be seen in the following, the travel time of the catchment is lower than Δ thus the model will show a systematic error during flood events. This is unavoidable since the decision time step can not be changed.

Table 1 provides a brief comparison of the performances obtained with a deseasonalized-ARMAX model of form (2) (where an exponential relation exists among precipitation and flow) and model (3), for two different choices of the model order. Poor performances of the deseasonalized-ARMAX model are justified by the fact that the parameter estimation is strongly conditioned by the greatest flood event in the estimation data set, while all other events are misinterpreted. It must be concluded that the deseasonalized-ARMAX model is not robust: bad performances on the validation data set are due to the presence of a big flood event (142 mm of precipitation on the 14/Oct/2000, against a maximum value of 100 mm over the estimation data set) where the model, exponential in the precipitation, highly overestimates the flow (figure 1). Model (3), instead, provides satisfactory results. Its performances can be further improved by introducing time-varying periodic parameters for the exogenous component, in order to account for different effect of precipitation on the inflow in different seasons (usually due to different conditions of the ground). To this end, each coefficient of the polynomials $B(z^{-1})$ is replaced by a Fourier series whose coefficients are estimated through linear least squares. With this structure, good results are obtained also with a low order model: with $n_a = 1$, $n_b = 2$, $n_c = 1$ and a 2nd order Fourier series for the two coefficients of the polynomial $B(z^{-1})$, the coefficient of determination R^2 is equal to 0.74 and 0.77 over the estimation and validation

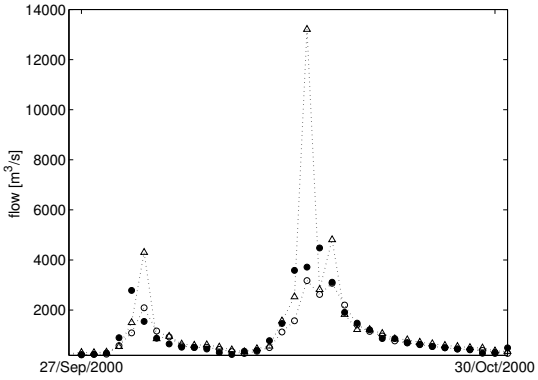

Fig. 1. Observed flow (black dots), prediction of the deseasonalized-ARMAX model (triangle) and of the model (3) (circle) - both with $n_a = 2$, $n_b = 2$ and $n_c = 0$ - over a flood event in the validation data set.

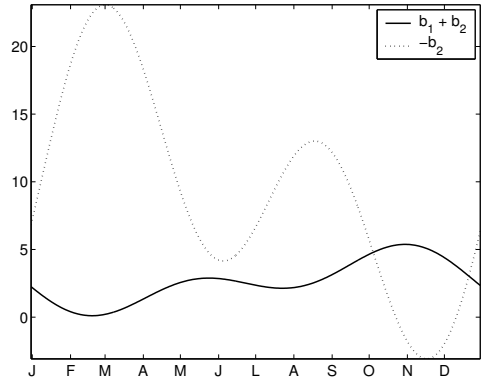

Fig. 2. Time-varying periodic parameters (period = 1 year) of the exogenous component of (3). Since $n_b = 2$, the precipitation contribution to the prediction $q_{t+1|t}$ is given by $(b_1 + b_2)p_t - b_2(p_t - p_{t-1})$.

data set respectively. The trajectory of the parameter estimates relevant to the exogenous input are consistent with physical interpretation of the flow-formation process: in particular, the sum $b_1 + b_2$, which weights the contribution of the last available precipitation measure p_t, is higher in the late-spring period and in autumn, when soil moisture, and thus the runoff, are higher (see figure 2). Finally, as for the moving average component, the estimated value of c_1 is negative (it is equal to -0.6390 with a variance of 0.0013). This result can be interpreted as follows. Since the travel time in the catchment is lower than Δ (it is approximately 12-18 hours), flow peaks due to abrupt precipitation are predicted with one-step lag; this systematic error of the 'deterministic prediction' can be partially corrected when the noise term $c_1\varepsilon_t$ is introduced: its contribution, in fact, is negative after a flow underestimation, since by definition $\varepsilon_t = q_t - q_{t|t-1}$ (figure 3). Intuition is confirmed by the fact that if this moving-average component is removed ($n_c = 0$), the model performances decrease to $R^2 = 0.72$ over the estimation data set and $R^2 = 0.74$ over the validation data set.

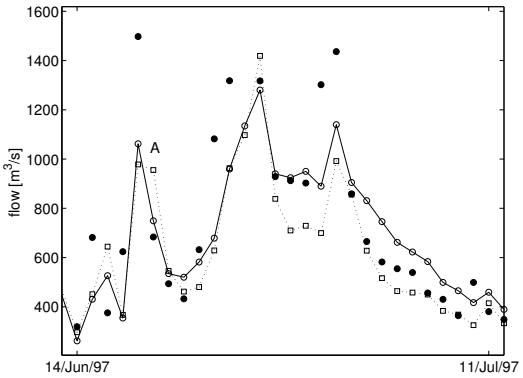

Fig. 3. Observed flow (black dots), prediction of model (3) with $n_c = 0$ (square) and $n_c = 1$ (circle), over a portion of the estimation data set. Letter A indicates a case where the MA component helps correcting overestimated prediction after an underestimation due to abrupt precipitation.

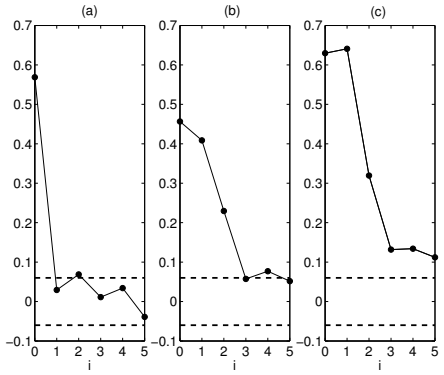

Fig. 4. Cross-correlation between model residual ε_{t+1} and precipitation p_{t+1-i} (a), ε^2_{t+1} and p_{t+1-i} (b), and $|\varepsilon_{t+1}|$ and p_{t+1-i} (c), for different values of i. Dashed lines are 95% confidence bounds.

4. ESTIMATION OF THE VARIANCE OF THE PREDICTION ERROR

The prediction error ε_t of the nonlinear model (3) is approximately Gaussian. Its distribution can not be considered exactly Gaussian since it shows a positive skew: extreme underestimation errors, in fact, are larger than extreme overestimation errors. This is because the maximum unpredictability is connected to flood peaks caused by abrupt precipitation (when the flow q_t is strongly influenced by the precipitation p_t in the same interval), which are systematically underestimated. In the following, however, we will neglect this skew, assume that ε be generated by a Gaussian process and exploit this assumption to derive confidence bounds for the flow prediction.

Further analysis reveals that the prediction error can be assumed zero mean and white noise. However the squared error is autocorrelated, which is an evidence of the heteroscedasticity of the prediction error. Therefore, the variance σ^2_t of ε_t will be described by means of a dynamical model. In order to derive the model of the error variance, let us first develop some considerations. Unpredictability is mainly concentrated during flood events. Figure 4.a shows the cross-correlation function between precipitation and prediction error. It can be noticed that the error ε_{t+1} in the prediction made at time t is highly correlated with the precipitation p_{t+1}, because the prediction error is often due to an 'unconsidered' precipitation event occurring in the time interval of the prediction; instead, the correlation between ε_{t+1} and past precipitation records p_{t+1-i}, with $i > 0$, is almost zero, thus confirming that all the information available at time t is correctly exploited by the model. However, as it can be seen from figure 4.b, the cross-correlation between precipitation p_{t+1-i} and squared error ε^2_{t+1} is high also for $i > 0$. Thus it can be concluded that while the error value does not depend on past precipitation, the error variance does. Finally, figure 4.c shows the cross-correlation function between precipitation and the absolute value of the error, which is even stronger than between precipitation and squared error. The analysis of the autocorrelation function revealed that also the autocorrelation of process $|\varepsilon|$ is higher than the autocorrelation of process ε^2. From the above, the following conclusions can be drawn:

- precipitation measures can be helpfully used as exogenous input of the variance model, to account for the increased unpredictability of the flow during flood events;
- the time series of the absolute value of the residual provides more information for identifying the variance model than the time series of the squared residual.

Therefore the idea is to identify a dynamical model for the residual absolute value, use it to compute the conditional expected value of $|\varepsilon_{t+1}|$ and from this derive an estimate of σ_{t+1}. This is possible since, under the assumption that ε_{t+1} be normally distributed for any t, its standard deviation is given by

$$\sigma_{t+1} = \sqrt{2\pi}/2 \cdot E[|\varepsilon_{t+1}|] \tag{4a}$$

(see Appendix for proof). The following linear model has been identified for predicting the absolute value of the error

$$E[|\varepsilon_{t+1}|] = \alpha + \sum_{i=1}^{n_\beta} \beta_i |\varepsilon_{t+1-i}| + \sum_{i=1}^{n_\gamma} \gamma_i p_{t+1-i} \tag{4b}$$

Parameters α, β_i (with $i = 1, \ldots, n_\beta$) and γ_j (with $j = 1, \ldots, n_\gamma$) are all taken as positive, thus guaranteeing positivity of $E[|\varepsilon_{t+1}|]$, and they can be estimated through constrained nonlinear least squares. Different model orders were tested and good results were obtained also for low orders. Results that will be shown in the following refer to the case $n_\alpha = 3$ and $n_\beta = 1$.

4.1 Application results

A constant and a periodic model of the prediction error variance will be used as benchmarks. The former assumes that the standard deviation of the prediction error be constant and equal to the sample standard deviation

$$\sigma_{t+1} = \sqrt{\frac{1}{N-1} \sum_{t=0}^{N-1} \varepsilon^2_{t+1}} \tag{5}$$

where N is the number of data in the estimation data set. The latter assumes that the standard deviation be periodic and derives it from the variance σ^2_{t+1}, which is given by a fifth order Fourier series

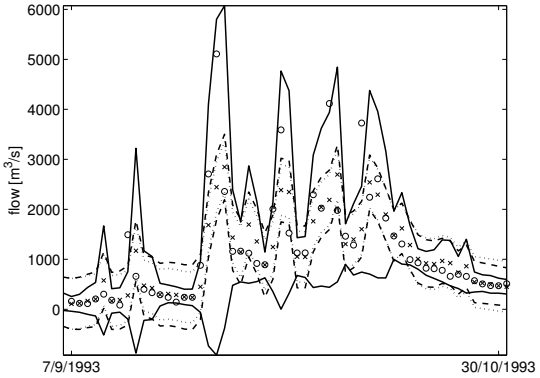

Fig. 5. Measured flow (circle), predicted flow (cross) and 99% confidence bounds assuming that the prediction error be Gaussian and its standard deviation be constant (dotted line), periodic (dashed) or dynamical (solid).

$$\sigma^2_{t+1} = a_0 + \sum_{n=1}^{5} a_n \cos\left(n\frac{2\pi}{T}t\right) + b_n \sin\left(n\frac{2\pi}{T}t\right) \qquad (6)$$

where $T = 365$ (days).
In figure 5, the 99% confidence intervals of the flow prediction, based on the three variance models (4), (5) and (6), are compared. At each time t, the extremes of the confidence interval are given by $q_{t|t-1} \pm 3\sigma_t$ where $q_{t|t-1}$ is computed based on model (3) and σ_t is given by one of the three above models. If constant or periodic variance is used, the confidence interval is too large for low water events and too narrow for flood events. The confidence interval based on a dynamical variance, instead, is narrower for low flow values and wider in correspondence to floods, when unpredictability increases.

Another tool that was used to assess the adequacy of the variance model is the probability plot (Laio and Tamea (2007)). The idea is as follows. At each time t of the validation set, a prediction $q_{t|t-1}$ based on model (3) is available, as well as an estimate of the prediction error variance σ_t^2. Since the prediction error is Gaussian and it enters model (3) as an additive term, we can conclude that the conditional probability distribution of the flow is also Gaussian. Unfortunately, we can not evaluate the goodness-of-fit of such distribution with conventional statistical tests, because only one extraction from that distribution is available, i.e. the measure q_t. However, from the probability integral transform it follows that if the estimated cumulative distribution function (cdf) $\hat{P}_{q_t}(\cdot)$ of the flow coincides with the true cdf $P^0_{q_t}(\cdot)$, then the value $z_t = \hat{P}_{q_t}(q_t)$ is an extraction from a uniform distribution over $[0, 1]$, i.e. $z_t \sim U(0, 1)$. Since this is true for any $t = 1, ..., N$, N being the number of flow measures in the validation data set, we can evaluate the good-of-fit of the N cdfs $\hat{P}_{q_t}(\cdot)$ by checking if $Z = \{z_1, ..., z_N\}$ is a sample of mutually independent and identically distributed $U(0, 1)$ observations. Independence of the sample was checked by looking at the autocorrelation function. As for the uniformity hypothesis, we compute the value of the empirical cdf of z_t as $F_{z_t}(z_t) = R_t/N$, where R_t is the number of elements in Z lower than z_t, and compare it with the value

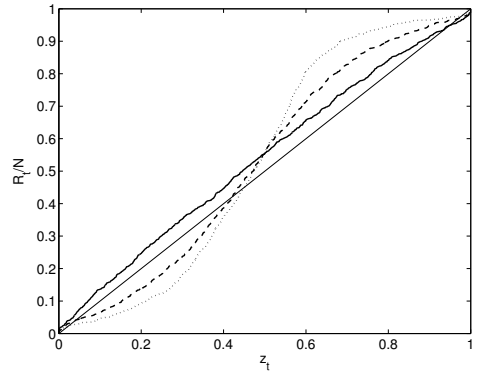

Fig. 6. Sorted $z_t = \hat{P}_{q_t}(q_t)$ versus their empirical cdf values R_t/N; the conditional cdf $\hat{P}_{q_t}(\cdot)$ is Gaussian with parameter σ_t given by the dynamical model 4 (solid line), σ_t constant (dotted) and σ_t periodic (dashed). The bisector (thin line) is the cdf of z_t if z_t is uniform over $[0, 1]$.

of the uniform cdf, $z^0_t = P^0_{z_t}(z_t) = z_t$. Figure 6 shows the probability plot obtained with the three variance models (4), (5) and (6). None of the empirical cdfs lays exactly on the bisector, i.e. none of them coincides with the theoretical cdf $P^0_{z_t}(\cdot)$, however the empirical cdf corresponding to the dynamical model (4) is much closer to the bisector than the other two. In particular, the probability plots based on constant and periodic variance are s-shaped, which means that the z_t points are concentrated towards the centre of the interval $[0, 1]$, i.e. the confidence interval is too wide. The discrepancy from the bisector in the probability plot based on the dynamical model of the variance, instead, is not due to a systematic error in the variance estimate but reveals an overestimation of the flow expected value.

5. FINAL REMARKS AND FUTURE RESEARCH

The paper presents a model for describing the flow formation process from an uncontrolled catchment, to be used for one-step-ahead prediction. It is a nonlinear model that exploits past flow and precipitation measures and past prediction errors to predict the flow over the following 24 hours. The main advantage of this model, with respect to conventional linear models on the flow logarithm, is that it can efficiently handle the rainfall input, avoiding the loss of robustness that is encountered by linear models. However, the drawback is that the flow process does not belong to any particular probability distribution, and thus the model cannot be used for simulation without further improvement.
The paper also presents a model for estimating the variance of the prediction error. It is a dynamical model thus accounting for the heteroscedasticity of the prediction error process. The confidence interval of the flow prediction based on this model seems to be more adequate than those based on a constant or periodic variance. However, assessing the goodness-of-fit of a variance model is still an open issue (see for example Laio and Tamea (2007)) and further research is needed for defining adequate tools to this purpose. Finally it must be noted that the proposed variance model presents a main limit in the fact that, assuming the prediction error as a Gaussian variable, the conditional distribution of the flow turns out to be

Gaussian too, which is not acceptable from the physical standpoint. The difficulty might be overcome by simply neglecting negative predictions, i.e. by forcing the lower bound in Figure 5 to be nonnegative, however further research is advocated to find more satisfactory solutions.

ACKNOWLEDGEMENTS

The data used in this reasearch were kindly provided by Consorzio Ticino (inflow data), ARPA Piemonte and MeteoSwiss (precipitation measures). The research was supported by Fondazione CARIPLO within the TWOLE PROJECT (2004).

REFERENCES

T. Bollerslev. Generalized autoregressive conditional heteroskedasticity. *Journal of Econometrics*, 31(3):307–327, 1986.

R.F. Engle. Autoregressive conditional heteroscedasticity with estimates of the variance of united kingdom inflation. *Econometrica*, 50(4):987–1007, 1982.

K.W. Hipel and A.I. McLeod. *Time Series Modelling of Water Resources and Environmental Systems*. Elsevier, Amsterdam, 1994.

F. Laio and S. Tamea. Verification tools for probabilistic forecast of continuous hydrologycal variables. *Hydrology and Earth System Sciences*, 11:1267–1277, 2007.

R.J. Romanowicz. Data based mechanistic model for low flows: Implications for the effects of climate change. *Journal of Hydrology*, 336:74–83, 2006.

W. Wang, P.H.J.M. Van Gelder, J.K. Vrijling, and J. Ma. Testing and modelling autoregressive conditional heteroskedasticity of stramflow processes. *Nonlinear processes in Geophysics*, 12:55–66, 2005.

Appendix A. EXPECTED VALUE OF THE ABSOLUTE VALUE OF A NORMAL VARIABLE

First it can be proved that if $X \sim N(0, \sigma^2)$ and $Y = |X|$, then

$$f_Y(y) = 2 \cdot f_X(x) \quad \forall x \in \mathbb{R}, y = |x|$$

where $f_Y(\cdot)$ and $f_X(\cdot)$ are the pdfs of X and Y respectively. In fact, from the definition of Y, it follows that $P[Y < y] = P[|X| < y] = P[-y < X < y]$ and thus

$$\int_0^y f_Y(\xi)d\xi = \int_{-y}^y f_X(\xi)d\xi$$

Since $f_X(\cdot)$ is an even function,

$$\int_0^y f_Y(\xi)d\xi = \int_0^y 2 \cdot f_X(\xi)d\xi$$

from which the thesis follows.

Therefore the expected value of Y is given by

$$\begin{aligned} E(Y) &= \int_0^\infty y\, 2 f_X(y)\, dy \\ &= \frac{2}{\sqrt{2\pi}} \int_0^\infty \frac{y}{\sigma} \exp\left(-\frac{y^2}{2\sigma^2}\right) dy = \frac{2}{\sqrt{2\pi}}\sigma \end{aligned}$$