Nordic Hydrology, 33 (5), 2002, 331-346

Rainfall Runoff Modelling Based on Genetic Programming

Vladan Babovic and Maarten Keijzer

Danish Hydraulic Institute, Water and Environment
DK-2970 Hørsholm, Denmark

The runoff formation process is believed to be highly non-linear, time varying, spatially distributed, and not easily described by simple models. Considerable time and effort has been directed to model this process, and many hydrologic models have been built specifically for this purpose. All of them, however, require significant amounts of data for their respective calibration and validation. Using physical models raises issues of collecting the appropriate data with sufficient accuracy. In most cases it is difficult to collect all the data necessary for such a model.

By using data driven models such as genetic programming (GP), one can attempt to model runoff on the basis of available hydrometeorological data. This work addresses use of genetic programming for creating rainfall-runoff models on the basis of data alone, as well as in combination with conceptual models (*i.e* taking advantage of knowledge about the problem domain).

Introduction

The runoff formation process is believed to be highly non-linear, time varying, spatially distributed, and not easily described by simple models. Considerable time and effort has been devoted to model these processes, and many hydrologic models have been built specifically for this purpose. These models are generally referred to as rainfall-runoff (R-R) models. The rainfall-runoff model is a hydrologic model, which basically determines the runoff signal that leaves the watershed basin from the rainfall signal received by this basin. According to the traditional hydrologic

classifications rainfall-runoff models are grouped into three categories, namely: empirical black-box models, lumped conceptual models and distributed physically based modelling systems. The great majority of the rainfall-runoff modelling systems used in practice are from the first two categories.

Empirical black-box models are entirely lacking an explicitly well-defined representation of the physical processes involved in the transformation of rainfall into runoff. A large number of black box models have their origin in the unit hydrograph theory of (Sherman 1932) and are considered to be at the lower end of the scale in terms of inclusion of physical laws into the model structure. These models depend on rainfall and discharge observations for the estimation of their parameters and for further refinement of their structure. It is believed that the black-box models do not work very well outside the conditions used for their development and calibration. However, experience over decades has shown that these models are useful operational tools and indeed they are the only option in cases where there are no any other meteorological data available except rainfall or these data are of poor quality.

Conceptual rainfall-runoff models are designed to approximate (in some physically realistic manner) the general internal subprocesses and physical mechanisms which govern the hydrologic cycle. Conceptual models are usually based on simplified forms of the physical laws and are generally non-linear, time-invariant, and deterministic with parameters that are representative of the watershed characteristics. Such models ignore the spatially distributed, time-varying, and stochastic nature of the rainfall-ruonff process and attempt to incorporate realistic representations of the major non-linearities inherent in the R-R relationships. Again, despite their simplicity, many such models have proven quite successful in representing an already measured hydrograph. However, the implementation and calibration of such a model typically presents various difficulties, requiring sophisticated mathematical tools, significant amounts of calibration data, and some degree of expertise and experience with the model.

While there is a large number of existing black-box and conceptual models, there are only a few distributed physically based hydrologic modelling systems suitable for research purposes and for real world projects. Deterministic models are explicitly based on our current understanding of the physics of the constituent hydrological processes. Perhaps the most widely known such system is the Systeme Hydrologique Europeen (SHE) (Abbott *et al.* 1987) created jointly by the Institute of Hydrology, the Danish Hydraulic Institute and SOGREAH. SHE is a general, physically based, distributed modelling system for constructing and running models of all or any part of the land phase of the hydrological cycle for any geographical area. These types of modelling systems have extensive data demands. They utilize quite a large number of parameters in their operation, which have direct relation to physical catchment characteristics (topography, soil, vegetation, and geology) and operate within a distributed framework to account for the spatial variability of both physical characteristics and meteorological conditions. Even so, deterministic models also

need calibration mainly because the parameters they require could not or are not directly measured everywhere in the modeled basin. The physically based distributed models do not have the applicability shortcomings of the models from the first two groups. In general they are not directed only towards studying the rainfall-runoff processes but also some other processes like erosion, conjunctive use of ground water and surface water, and environmental impacts of land use changes related to the agricultural and forestry practices, which are much more important than rainfall-runoff alone. To model the runoff of a certain river basin using physical models raises issues of collecting the appropriate data with sufficient accuracy. In most cases it is difficult to collect all the data necessary for such a model. Furthermore, this kind of model requires significant amounts of data for their calibration and validation.

An alternative to the outlined approaches may be to use new data driven black- or grey-box type techniques that can model the process using only basic hydrometeorological data. Artificial neural networks (ANNs) have already gained much popularity in hydrologic circles (Minns and Hall 1996) Another such technique is genetic programming (Koza 1992). Genetic programming (GP)is a relatively new domain-independent method for evolving computer programs for solving or approximately solving problems. GP's learning algorithm is inspired by the theory of natural evolution and our current understanding of biology and natural evolution.

The road map for the rest of the paper is as follows. In the continuation, evolutionary algorithms as a method for constructing equations on the basis of data are described. Then, a case study, Orgeval catchment, is described to a greater detail and finally rainfall runoff process in the catchment is modeled using genetic programming. Several approaches are presented and discussed in concluding chapters.

Equation Building

When refining a model of a physical process, a scientist focuses on the agreement of theoretically predicted and experimentally observed behaviour. If these agree in some accepted sense, then the model is 'correct' within that context. Here, we consider the problem inverse to verification of theoretical models: how can we obtain the governing equations directly from measurements? To do this, we will extend the notion of qualitative information contained in a sequence of observations to consider directly the underlying dynamics. We will show that, using this information, one can deduce the effective governing equations. The latter summarize up to an a priori specified level of correctness, or accuracy, the deterministic portion of the observed behaviour. The observed behaviour on short time scales unaccounted for by the reconstructed equations will be considered as extrinsic noise.

Evolutionary Computation

According to Darwinian theory of evolution, all animals and plants inhabiting our planet are actually descendants of few primitive progenitors. Darwin in the illustri-

ous work *The Origin of Species by Means of Natural Selection* (Darwin 1859) claims that all complex and intricate life forms that surround us are actually direct offspring of these original prototypes. However, the offspring differ from the original ancestors. They are not exact copies of their ancestors, but rather variations that possibly provide competitive advantages over other, similar specimen, in the same environment. And so, claims Darwin, through the process of copying (reproduction) with variations (mutation) and competition for resources, the organisms evolve that posses capabilities that are best adapted to the environment they are situated in. Survival of the fittest thus results in a situation in which given environment is populated with best adapted (most fit) organisms.

Evolutionary algorithms (EAs) are processes that are closely inspired by the Darwinian theory of evolution and have one principal objective: to evolve solutions to the problems, rather than to solve problems directly. The fundamental idea is no more original than plagiarism of natural processes, which corresponds to providing 'algorithmic organisms' with hereditary capabilities, allowing them to reproduce and let them, through competition for resources, evolve those traits that maximize their benefits in a given environment. The environment to which entities adapt in the EA context is actually formed by a problem domain for which solutions are being evolved. Thus, EAs attempt to mirror evolutionary processes from nature that allow for adaptation of evolving entities to problem domain, which in turn emerges to a solution of a problem in question.

In the continuation we first outline properties of natural evolution, and then attempt to mirror those in an artificial media, as exemplified through evolutionary algorithms.

Properties of Natural Evolution

Natural evolution has been extremely successful in creating many 'useful' things. Technology can be nothing but jealous about the successes of natural evolution. The success of adaptation achieved by living organisms to their environment can hardly be matched by human creations. For example, the rate of energy consumption for a given speed of any modern submarine, let alone surface vessel, exceeds that of a fish swimming in the water, by several orders of magnitude. What are the processes that enabled natural evolution to construct such an effective creations? According to the prevalent views, there are three main criteria for an evolutionary process to occur (Maynard-Smith 1975):

Criterion of Heredity	Offspring are similar to their parents: the genotype copying process maintains a high fidelity;
Criterion of Variability	Offspring are not exactly the same as their parents: the genotype copying process is not perfect;
Criterion of Fecundity	Variants leave different number of offspring; specific variations have an effect on behaviour and behaviour has an effect on reproductive success.

The three requirements above are necessary and sufficient conditions for an evolutionary process to occur. The criterion of heredity assures that offspring inherits information from parents, assuring their similarity. Variability is ensured through mutations, whereas the criterion of fecundity provides, on the average, more fit individuals with possibilities to reproduce more often thus generating more and better-surviving offspring.

Evolutionary Algorithms

Evolutionary algorithms (EAs) are engines simulating grossly simplified processes occurring in nature and implemented in artificial media – such as a computer. The family of evolutionary algorithms today is divided into four main streams: Evolution Strategies (Schwefel 1981), Evolutionary Programming (Fogel *et al.* 1966), Genetic Algorithms(Holland 1975) and Genetic Programming (Koza 1992). Although different andintended for different purposes, all EAs share a common conceptual base (schematized in Fig. 1). In principle, an initial population of individuals is created in a computer and allowed to evolve using the principles of inheritance (so that offspring resemble parents), variability (the process of offspring creation is not perfect—some mutations occur) and selection (more fit individuals are allowed to reproduce more often whereas less fit less often so that their "genealogical" trees disappear in time). One of the main advantages of EAs is their domain independence. EAs can evolve almost anything, given an appropriate representation of evolving structures. Similarly to processes observed in nature, one should distinguish between an evolving entity's genotype and its phenotype. The genotype is basically a code to be executed (such as a code in a DNA strand), whereas the phenotype represents a result of the execution of this code (such as any living being). Although the information exchange between evolving entities (parents) occurs at the level of genotypes, it is the phenotypes in which one is really interested.

The phenotype is actually an interpretation of a genotype in a problem domain. This interpretation can take the form of any feasible mapping. For example, for optimization and constraint satisfaction purposes, genotypes are typically interpreted as independent variables of a function to be optimised. Along these lines, one can employ mappings in which genotypes are interpreted as roughness coefficients in a

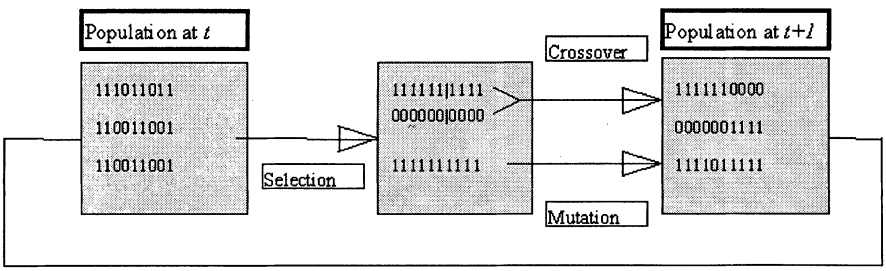

Fig.1. Schematic illustration of an evolutionary algorithm.

free surface pipe flow model with the genetic algorithms (GAs) directed towards the minimization of the discrepancies between model output and measured water level and discharge values. The resulting GA represents an automatic calibration model of hydrodynamic systems (Babovic *et al.* 1994; Madsen 2000). Several other applications of GAs, which make use of various kinds of genotype-phenotype mappings and with a specific emphasis on water resources, are described, for example in (Babovic 1996).

Genetic Programming

Genetic Programming is one instance of the evolutionary algorithms family. In GP the evolutionary force is directed towards the creation of models that take a symbolic form. In this evolutionary paradigm, evolving entities are presented with a collection of data and the evolutionary process is directed towards the creation of closed-form symbolic expression describing the data. In its primitive form, GP lends itself quite naturally to the process of induction of mathematical models based on observations: GP is an efficient search algorithm that need not assume the functional form of the underlying relationship. Given an appropriate set of basic functions, GP discovers a (sometimes very surprising) mathematical model that approximates the data well.

Individual solutions in GP are computer programs represented as parse trees (Fig. 2). The population of the very first generation is usually generated through a random process. However, subsequent generations are evolved through genetic operators of selection reproduction, crossover and mutation. GP thus iteratively applies variation and selection on a population of evolving parse trees representing symbolic expressions. Standard variation operators in genetic programming are subtree mutation (replace a randomly chosen subtree with a randomly generated subtree) and subtree crossover (replace a randomly chosen subtree from a formula with a randomly chosen subtree from another formula—Fig. 3). For a detailed description, see, for example, Babovic and Keijzer (2000).

The types of functions used in this tree structure are user-defined. This means that they can be algebraic operators, such as sin, log, +, -, *etc.*, but they can also take the form of if-the-else rules, making use of logical operators such as OR, AND, *etc.*

The search process in GP is guided by fitness (i:e: a measure of accuracy). Determination of the fitness function to be adopted is an important aspect in GP since its performance largely depends upon how well this fitness function represents the objective or goal of the problem at hand. In the present work we adopt a multi-objective approach in which both Root Mean Squared Error (RMS) and Coefficient of Determination (CoD) are used as fitness functions. The evolutionary process is then guided towards simultaneously minimizing RMS and maximizing CoD towards the value of unity. It has been shown empirically (Babovic and Keijzer 2000) that this combination of objective functions implicitly promotes parsimony and results in simpler expressions.

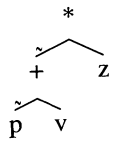

Fig.2. An equation $(p+v)z$ represented as a parse tree.

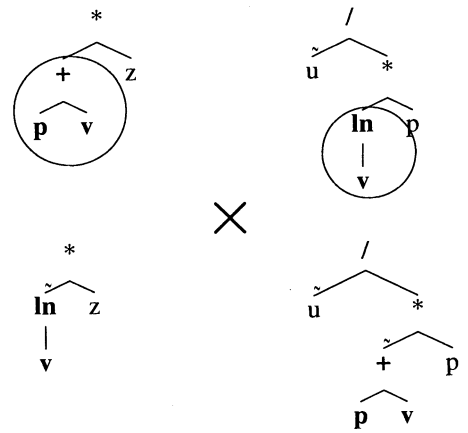

Fig.3. The action of the crossover operator: subtrees of selected parents (above) are swapped in crossover to generate the offspring (below).

A number of applications of GP has been reported, such as studies in which salt intrusion data were analyzed (Babovic and Minns 1994), experimental data for bed concentration of suspended sediment (Babovic and Keijzer 1999) analysis of roughness forces induced by vegetation (Babovic and Keijzer 2000) as well as rainfall runoff modelling (Babovic and Abbott 1997a; Khu *et al.* 2001; Liong *et al.* 2001). In all of the above-mentioned studies, GP-induced relationships provided more accurate descriptions of data than those obtained using more conventional methodologies. An extensive survey of the applications of GP in water resources is provided in (Babovic and Abbott 1997b; Babovic 1996).

Symbolic Regression

Regression — linear or nonlinear — plays a central role in the process of finding empirical equations. In its most general form, regression techniques proceed by selecting a particular model structure and then estimate the accompanied coefficients based on the available data. The model structure can be linear, polynomial, hyperbolic, logarithmic etc. The only requirement in such an approach is that the coefficients in the model can be estimated using an optimization technique. In generalized linear regression for instance, the only requirement is that the model is linear in the coefficients. The model itself can consist of any functional form. Another technique may be a nonlinear regression where the only requirement is that the model is dif-

ferentiable both in the inputs and the coefficients. Supervised Artificial Neural Networks belong to this class of regression techniques.

Genetic programming can also be understood as a regression technique, a so-called symbolic regression. The specific model structure is not chosen in advance, but is part of the search process. In this algorithm, both model structure and coefficients are simultaneously being searched for. The user has to define some basic building blocks (function and variables to be used); the algorithm tries to build a model using only those specified blocks. As a space of model structures is in general not smooth, not differentiable nor linear in any useful sense (it is in fact highly discontinuous), standard optimization techniques fail when trying to find both the model structure and the coefficients.

Case Study

The catchment under consideration is the Orgeval catchment, in France (Fig. 4), which has been studied extensively in the World Meteorological Organization's intercomparison project (WMO 1992). The catchment is located about 80 km east of Paris and the main river that drains the catchment runoff is the Orgeval. The catchment has an area of about 104 km^2. The catchment comprises mainly of rural area, with only 1% of the total being urban areas or roads and 18% of the total being covered by forest.

In this study, a total of 10 storm events from 1972-1974 hourly flow record are selected for training the GP while a total of 6 storm events (denoted as Storms 1-6) between 1979-1980 are selected and used for the verification of the updating procedure.

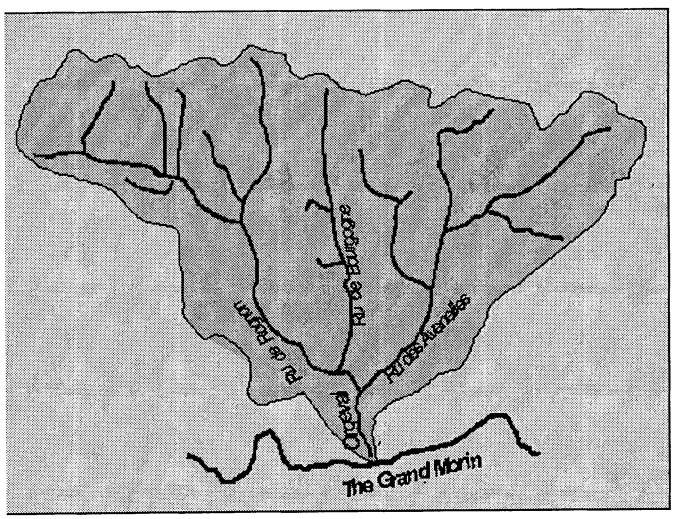

Fig.4. The Orgeval Catchment

Table 1 – Statistical measures of accuracy (Mean absolute error – MAE, Correlation coefficient – r, and Pearson's R^2) for the GP forecast as well as for the naïve forecast 1-12 hours.

Forecast	GP Forecast			Naïve forecast		
Horizon	*MAE*	*r*	*R2*	*MAE*	*r*	*R2*
1 hour	0.0161	0.9995	0.9991	0.0483	0.9973	0.9946
2 hour	0.0245	0.9990	0.9980	0.0954	0.9899	0.9798
3 hour	0.0322	0.9984	0.9969	0.1412	0.9787	0.9578
4 hour	0.0390	0.9979	0.9957	0.1857	0.9646	0.9305
5 hour	0.0445	0.9973	0.9947	0.2285	0.9484	0.8995
6 hour	0.0495	0.9969	0.9938	0.2698	0.9307	0.8662
7 hour	0.0537	0.9966	0.9932	0.3099	0.9118	0.8315
8 hour	0.0571	0.9964	0.9927	0.3482	0.8921	0.7958
9 hour	0.0597	0.9962	0.9924	0.3851	0.8715	0.7594
10 hour	0.0623	0.9960	0.9920	0.4202	0.8500	0.7224
11 hour	0.0647	0.9958	0.9916	0.4533	0.8276	0.6850
12 hour	0.0682	0.9955	0.9911	0.4846	0.8044	0.6471

Forecast Based on Conceptual Model – NAM

In order to establish grounds for intercomparison the widely used rainfall-runoff simulation model NAM (Nielsen and Hansen 1973) is used to simulate the runoff for the entire period of interest. Since the main interest is the investigation of the skill related to the modelling of runoff processes (*i.e.* runoff as a response to forcing by rain) in all cases a so-called ideal rainfall forecast (measured rainfall was used in place of forecasted values) is assumed.

NAM represents a model of a rainfall-runoff process. Given the ideal rainfall forecast, the quality of runoff forecast will not deteriorate with forecast horizon. For the present case the forecast skill is summarized in Table 2.

Naïve Forecast

Another, almost trivial, possibility is to use a so-called naïve forecast: one simply issues the forecast value which is exactly the same as the presently observed discharge. Due to the strong autocorrelation, the forecast skill is expected to be good for very short lead times, but also to quickly deteriorate with forecast horizons. The results for the naive forecast are summarized in Table 1.

Forecast Based on Genetic Programming

A forecasting system is based on information of the past and current states of hydrometeorological and catchment conditions as inputs as well as forecasted values of forcing term (rainfall R in this case) in order to forecast the catchment's response (runoff Q) in the future. Mathematically, this relationship can be expressed as

$$\hat{Q}(t+1) \equiv F(Q_{obs}(t), Q_{obs}(t-1), \ldots, Q_{obs}(t-5) \quad R(t+1), R(t), \ldots, R(t-5)) \tag{1}$$

In the present case the choice of orders for $Q_{obs}(t)$ and $R(t)$ of the immediate past 5 time steps were based on the catchment's concentration time, which varies up a maximum of 5 hours, *i.e.* 5 time steps (WMO 1992).

For forecasts which extend longer in the future (α time steps into the future) a slightly different, so-called iterative approach was utilized

$$\hat{Q}(t+\alpha) = F(\hat{Q}(t+\alpha-1), \hat{Q}(t+\alpha-2), \ldots, \hat{Q}(t+\alpha-5), \quad R(t+\alpha+1), R(t+\alpha), \ldots, R(t+\alpha-5)) \tag{2}$$

To be precise, in the present case, GP was utilized to forecast the temporal difference between the current and the future discharge, $dQ(t+1)$, rather than the absolute value of the discharge $Q(t+1)$. There are two strong reasons for adopting such a setup. Firstly, due to a very strong autocorrelation of discharges, there is a pronounced local optimum for forecasting discharges of the form $Q(t+1)=\beta Q(t)$ with β being a constant, typically smaller than one. Such a local optimum may be statistically very accurate, but the associated phase error discredits its use as a forecasting tool. Once the temporal differences are introduced, the strong autocorrelation is removed, and the GP is forced to approximate change in response $dQ(t+1)$ as a function of forcing terms (rainfall R) and past discharges $Q(t)$. Secondly, temporal differencing of time series of discharges $Q(t)$ removes any of the possible trends which may exist in raw data and consequently yielding a less biased forecaster

$$dQ(t+1) \equiv (Q(t)-Q(t-1)) \sqrt{\frac{R(t)+\sqrt{Q(t-1)+Q(t-2)^2}}{Q(t)/Q(t)+Q(t-1)+Q(t)}} + Q(t) \tag{3}$$

Eq.(3) fundamentally models $dQ(t+1)$ by multiplying $dQ(t)=Q(t)-Q(t-1)$ with a non-linear, time-varying correction factor. The correction factor is based on past discharges $Q(t)$, $Q(t-1)$ and $Q(t-2)$ as well as forecasted rainfall intensity $R(t)$which explains the absence of phase error (see Fig. 5) For longer lead times the quality of iterative forecast deteriorates only due to error introduced through calculated discharge. The ultimate results is that an approach based on GP outperforms NAM even for lead times of 12 hours (see Fig. 7).

Updating

The previous two chapters demonstrated that forecast based on data deteriorates as a function of forecast lead time. At the same time, the forecast based on NAM was not as accurate, however, the quality of the forecast did not deteriorate with the forecasting horizon. A logical idea is to combine the two and provide a hybrid in which the best of the two approaches is combined, yielding a highly accurate forecast which does not deteriorate with forecasting lead times. This corresponds to a form of

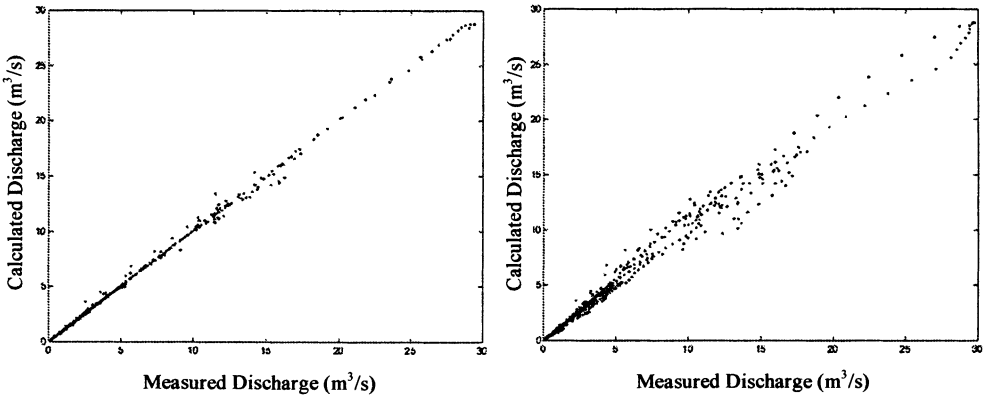

Fig.5. Scatter plots for GP based forecast utilizing Eq.(3) for lead time of (a) 1 hour and (b) 12 hours.

data assimilation, namely the one of error-correction (for more details see (Refsgaard 1997)).

This method is particularly interesting in real-time forecasting, where the originally forecasted values may be updated or modified as measured data become available and, thus prediction errors can be determined and used to improve forecast skill. In real-time runoff forecasting with rainfall runoff simulation models, rainfall time series up to the desired runoff forecast horizon must be available. A similar idea has been utilized before for hydrological problems (see for example (Khu *et al.* 2001; Madsen *et al.*2000)) as well as in marine problems albeit using neural networks and for the forecast of current speed in Danish coastal waters (Øresund) (Babovic *et al.* 2001).

Here, NAM is firstly used to simulate the discharge, Q_{sim} for the entire period of interest based on the rainfall data R. Then the prediction error ε is obtained by comparing the simulated discharge Q_{sim} with the observed discharge, Q_{obs}. The improved discharge $\hat{Q}$ is computed by adjusting Q_{sim} for each forecast lead-time within forecast horizon. Mathematically, the measured discharge $Q_{sim}(t)$ can be expressed as

$$Q_{obs}(t) = Q_{sim}(t) + \varepsilon(t) \tag{4}$$

Obviously,

$$\varepsilon(t) = Q_{obs}(t) - Q_{sim}(t) \tag{5}$$

Genetic programming can then be used to approximate the functional relationship between the prediction error and the simulated discharges, the past simulation errors up to the current time as well as rainfall intensity up to forecast horizon. For lead time of 1 hour, the functional relationship for the prediction error $\hat{\varepsilon}$ may be expressed as follows

$$\hat{\varepsilon}(t+1) \equiv F(Q_{sim}(t+1), Q_{sim}(t), \ldots, Q_{sim}(t-5), \varepsilon(t), \varepsilon(t-1), \ldots, \varepsilon(t-5), R(t+1), R(t), \ldots, R(t-5)) \quad (6)$$

while the forecast for improved discharge $\hat{Q}(t+1)$can be calculated as

$$\hat{Q}(t+1) = Q_{sim}(t+1) + \hat{\varepsilon}(t+1) \quad (7)$$

For longer lead times of 2,3,...,α hours, the recursive form of Eq.(7) can be written as

$$\hat{\varepsilon}(t+\alpha) = F(Q_{sim}(t+\alpha), Q_{sim}(t+\alpha-1), \ldots, Q_{sim}(t+\alpha-5), \hat{\varepsilon}(t+\alpha-1), \hat{\varepsilon}(t+\alpha-2), \ldots, \hat{\varepsilon}(t+\alpha-5), R(t+\alpha), R(t+\alpha-1), \ldots, R(t+\alpha-5)) \quad (8)$$

Note the use of error estimates $\hat{\varepsilon}$ instead of true error ε. The forecast for improved discharge $\hat{Q}(t+\alpha)$can now be calculated as

$$\hat{Q}(t+\alpha) = Q_{sim}(t+\alpha) + \hat{\varepsilon}(t+\alpha) \quad (9)$$

The real-time flood forecasting updating procedure for 1-hour lead-time could be summarized as follows:

1) Surface runoff has been simulated with parameter values of the NAM model calibrated on 1972-1974 period for the validation period of 1979-1980;
2) The prediction errors, ε between the NAM simulated and observed runoff for each time interval are computed;
3) GP is then used to derive the functional relationship between the present prediction error $\hat{\varepsilon}$ the NAM simulated discharge Q_{sim}, the past prediction errors $\hat{\varepsilon}$ and rainfall intensities $R(t)$ as given in Eq.(6);
4) The improved simulated discharge, $\hat{Q}$, is finally calculated, using Eq.(7);
5) For $t>1$ the above procedure is repeated following Eqs.(8) and (9).

Similarly as before and for the same reasons, GP was actually used to approximate a temporal difference in error evolution $dE(t+1)$ rather than $E(t+1)$.

$$d\varepsilon(t+1) = \varepsilon(t)-\varepsilon(t-1)+((\varepsilon(t-3)-\varepsilon(t-1))(0.054-\varepsilon(t-1))- -0.054(0.064(\varepsilon(t)-6.536)+((2\varepsilon(t)-0.0644)R(t-3)Q_{SIM}(t-3))) \quad (10)$$

The first two terms in Eq.(10) are the error at the t^{th} time step $\varepsilon(t)$-$\varepsilon(t-1)$. This error is then corrected by introducing a high-order correction term which utilizes past rainfall $R(t-3)$as well as output of conceptual model $Q_{sim}(t-3)$. Once this equation is used to calculate the error, and this error is in turn appended to NAM output Q_{sim} the

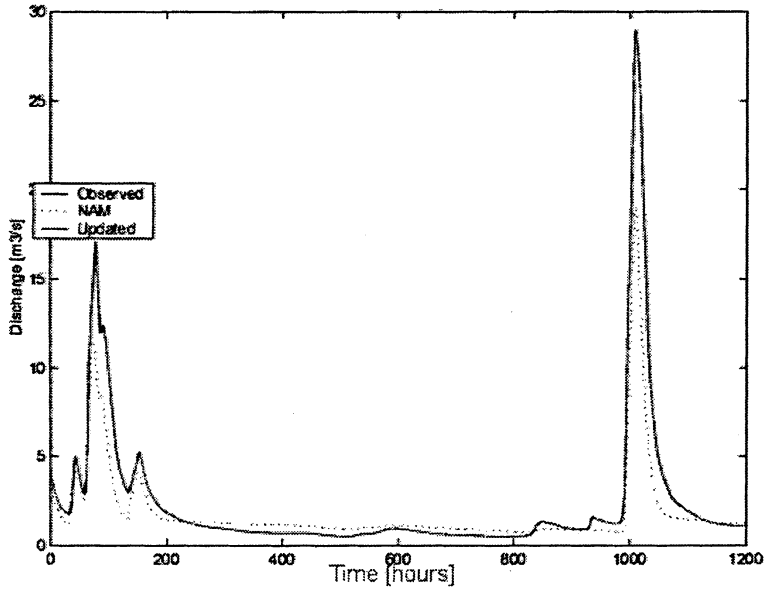

Fig.6. Time series of observed discharge, the one calculated using NAM and the one obtained through updating for two validation events. Lead time is one hour and the difference between observed and the updated cases is so small that it cannot be optically distinguished.

Table 2 – Statistical measures of accuracy (Mean absolute error – MAE, Correlation coefficient – r, and Pearson's R^2) for 'raw' NAM values as well as for updated model. Lead time 1 hour.

Statistic	NAM	After Update
MAE	0.3784	0.0279
r	0.9028	0.9612
R2	0.8150	0.9240

updated results provide a many fold improvement over raw model outputs (see Table 2 and Fig. 6).

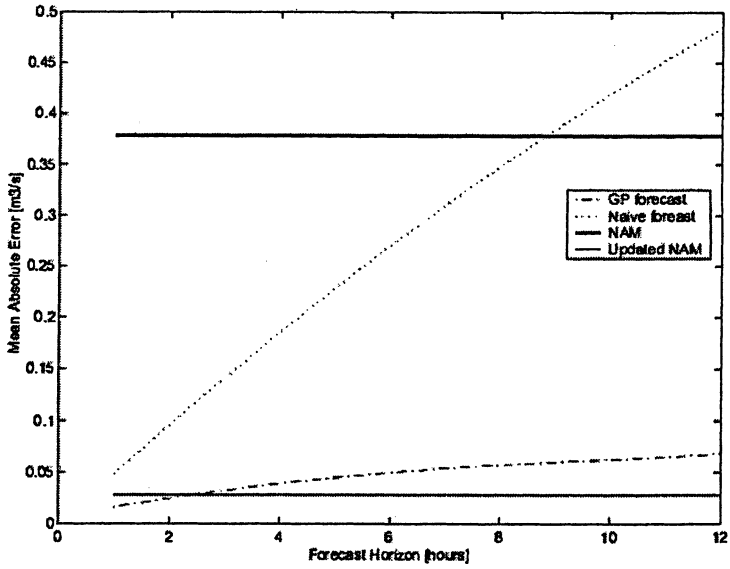

Fig.7. Evolution of mean absolute error as a function of forecasting lead time. The performances is calculated for 6 verification storm events.

Conclusions and Discussion

Several issues have emerged in the preceding chapters:

1) Forecasting on the basis of data is possible and in some cases can do considerably better in short term than forecasting and modelling on the basis of (conceptual) models.
2) The quality of forecasts created on the basis of data alone deteriorates with forecast horizon. This is perfectly reasonable since the initial conditions (in this case observed discharges) are 'washed out' and replaced by calculated discharges. Through the iterative process inaccuracies are introduced which amplify with the forecast horizon.
3) Modelling on the basis of our (albeit conceptualized) insights about the physical processes cannot match short term forecast skills created on the basis of data alone. However, the quality of such forecasts does not deteriorate with time.
4) It appears that the best approach is to combine the best of the two worlds. Use the data to improve the short term forecast and use knowledge (in the form of a conceptual model) to help with extending the forecasting horizon without deterioration of the forecast skill. It is rather interesting to observe that the updated model is more accurate than the NAM model alone for the lead times well beyond catchment's concentration time (in this case around 5 hours). This is due to the fact that GP forecasts errors created by NAM and in principle 'explains' phenomena not resolved by a conceptual model.

5) Genetic programming proved to be a powerful tool in the context of the rainfall forecast. The convenience of a single and simple equation, yet of extreme accuracy defends its use as an approach to short-term forecast.
6) Finally, it is very important to emphasize that it is the updating approach that provides the most accurate results. This clearly demonstrates that an amalgamation of knowledge (in a form of conceptual rainfall-runoff model) with a data driven approach (in the present case in the form of genetic programming) provides the best forecast skill. The authors strongly believe that it is the combination of the two approaches that will enable us to gain new insights which may ultimately lead to better and more accurate rainfall runoff models.

Acknowledgments

This work was in part funded by the Danish Technical Research Council (STVF) under the Talent Project N 9800463 entitled "Data to Knowledge – D2K". More information on the project can be obtained through http://www.d2k.dk

References

Abbott, M. B., Bathurst, J.C., Cunge, J., O'Connel, P.E, and Rasmunsen, J. (1987) An introduction to the european hydrological system – systeme hydrologique europeen (she) 1: History and philosophy of physically-based, distributed modelling system, *J. of Hydrol.*, *Vol. 87*, pp.45–59.

Babovic, V. (1996) *Emergence, Evolution, Intelligence: Hydroinformatics,* Balkema, Rotterdam.

Babovic, V., and Abbott, M.B (1997b) Evolution of equation from hydraulic data: Part i – theory, *J.Hydr.Res.*, *Vol.35*, pp.1–14.

Babovic, V., and Abbot, M.B. (1997a) Evolution of equation from hydraulic data: Part ii – applications, *J.Hydr.Res.*, *Vol.35*, pp.15–34.

Babovic, V., Cãnizares, R., Jensen, H.R., and Klinting, A. (2001) Artificial neural networks as a routine for updating of numerical models, *ASCE J. of Hydr.Eng.*, *Vol.127*, pp.181–193.

Babovic, V., and Keijzer, M. (1999) Data to knowledge – the new scientific paradigm, *Water Industry Systems: Modelling and Optimisation Applications*, D. Savic and G. Walters, eds., Research Studies Press, Exeter, pp.3–14.

Babovic, V., and Keijzer, M. (2000) Genetic programming as a model induction engine, *J. of Hydroinformatics*, *Vol.2*, pp.35–60.

Babovic, V., Larsen, L.C., and Wu, Z. (1994) Calibrating hydrodynamic models by means of simulated evolution, *Proceedings of the First International Conference on Hydroinformatics*, A. Verwey, A. W. Minns, V. Babovic, and C. Maksimovic, eds., Balkema, Rotterdam, pp.193–200.

Babovic, V., and Minns, A. W. (1994) Use of computational adaptive methodologies in hydroinformatics, *Proceedings of the First International Conference on hydroinformatics*, A. Verwey, A. W. Minns, V. Babovic, and C. Maksimovic, eds., Balkema, Rotterdam, pp.201–210.

Darwin, C. (1859) *The Origin of Species by Means of Natural Selection,* John Murray, London, sixth edition.

Fogel, L., Owens, A., and Walsh, M. (1966) *Artificial intelligence through simulated evolution.* Ginn, Needham Height.

Holland, J. (1975) *Adaptation in natural and artificial systems,* University of Michingan, Ann Arbor.

Khu, S. T., Liong, S.-Y., Babovic, V., Madsen, H., and Muttil, N (2001) Genetic programming and its application in real-time runoff forecasting, *J. of Am. Wat. Resour. Assoc.*, *Vol. 37*, pp.439–451.

Koza, J. R. (1992) *Genetic Programming: On the Programming of Computers by Means of Natural Selection*, MIT Press, Cambridge, MA, USA.

Liong, S.-Y., Gautam, T. R. , Khu, S. T. , Babovic, V. , Keijzer, M. , and Muttil N. (2001) Genetic programming: A new paradigm in rainfall runoff modeling, *J. of Am. Water Resour. Assoc.*, to appear.

Madsen, H. (2000) Automatic calibration of a conceptual rainfall-runoff model using multiple objectives, *J. Hydrol., Vol., 235*, pp.276–288.

Madsen, H., Butts, M., Khu, S. T., and Liong S. Y. (2000) Data assimilation in rainfall runoff forecasting, Proceedings of 4th International Conference on Hydroinformatics, Cedar Rapids, Iowa, USA.

Maynard-Smith, J. (1975) *The Theory of Evolution,* Penguin books, Harmondsworth, England, third edition.

Minns, A.W., and Hall, M. J. (1996) Artificial neural networks as rainfall-runoff models, *J. of Hydrol. Sci.*, *Vol. 41*, pp.399–417.

Nielsen, S., and Hansen, E. (1973) Numerical simulation of rainfall runoff process on a daily basis, *Nord. Hydrol.*, *Vol. 4*, pp.171–190.

Refsgaard, J. C. (1997) Validation and intercomparison of different updating procedures for real-time forecasting, *Nord. Hydrol.*, *Vol. 28*, pp.65–84.

Schwefel, H-P. (1981) *Numerical Optimization of Computer Models,* Wiley, Chichester.

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

WMO (1992) Simulated real-time intercomparison of hydrological models, WMO Operational Hydrology Report 38 – WMO No. 779, World Meteorological Organisation, Geneva.

Received: 13 July, 2001
Revised: 2 July, 2002
Accepted: 2 September, 2002

Address:
Danish Hydraulic Institute,
Water and Environment,
Agern Allé 11,
DK2970 Hørsholm,
Denmark,
Email: vmb@dhi.dk