Medium-range flow prediction for the Nile: a comparison of stochastic and deterministic methods

Demetris Koutsoyiannis

Department of Water Resources and Environmental Engineering, Faculty of Civil Engineering, National Technical University of Athens, Heroon Polytechneiou 5, GR-157 80 Zographou, Greece (corresponding author; dk@itia.ntua.gr)

Huaming Yao & Aris Georgakakos

Georgia Water Resources Institute, School of Civil and Environmental Engineering, Georgia Institute of Technology, 790 Atlantic Drive, Atlanta, Georgia, 30332-0335 USA

Abstract Due to its great importance, the availability of long flow records, contemporary as well as older, and the additional historical information of its behaviour, Nile is an ideal test case for identifying and understanding hydrological behaviours, and for model development. Such behaviours include the long term persistence, which historically has motivated the discovery of the Hurst phenomenon and has put into question classical statistical results and typical stochastic models. Based on the empirical evidence from the exploration of the Nile flows and on the theoretical insights provided by the principle of maximum entropy, a concept newly employed in hydrological stochastic modelling, an advanced yet simple stochastic methodology is developed. The approach is focused on the prediction of the Nile flow a month ahead but it is fairly general. The stochastic methodology is also compared with deterministic approaches, specifically an analogue (local nonlinear chaotic) model and a connectionist (artificial neural network) model based on the same flow record. All models have good performance with the stochastic model outperforming in prediction skills and the analogue model in simplicity. In addition, the stochastic model has other elements of superiority such as ability to provide long-term simulations and to improve understanding of natural behaviours.

Keywords stochastic models, deterministic models, artificial neural networks, linearity, nonlinearity, maximum entropy, prediction, simulation, Hurst phenomenon, Nile.

First draft, August 2006; second draft May 2007

1. Introduction

We predict, God laughs (Paraphrase of an old proverb)

The use of stochastic models in hydrological tasks such as simulation and prediction has a history of half a century, since the pioneering works of Barnes (1954) and Thomas & Fiering (1964). The classical book on time series analysis, forecasting and control by Box and Jenkins (1970) has greatly influenced stochastic hydrology. The model classes it proposed (autoregressive – AR; moving average – MA and combinations of the two – ARMA) have become classical and are still very popular.

However, despite making a large family, these model types are not ideal for hydrological processes for several reasons. First, all Box-Jenkins models are essentially of short-range dependence (SRD), that is, their autocorrelation structure decays exponentially with lag time; in contrast, there is evidence that hydrological processes exhibit long-range dependence (LRD), i.e., power-type decay of autocorrelation also known as the Hurst phenomenon (e.g., Koutsoyiannis, 2005b). Second, these models rely largely on a normality assumption, whereas it is known that hydrological processes (mostly on sub-annual scales) depart from normality and perhaps have distribution tails of power-type (e.g., Koutsoyiannis, 2005a). Third, the seasonal behaviour exhibited by hydrological processes at sub-annual scales is complex (with distribution type and dependence structure that change within the year) and cannot be handled by "deseasonalization" techniques typically used in companion to Box-Jenkins models (Koutsoyiannis & Georgakakos, 2006). Fourth, except for simplified processes, such as AR(1) or ARMA(1,1), the models are not parsimonious as they involve many parameters to be estimated from the data. In fact, the structure of these models is tightly linked to the number of parameters and one cannot change the structure (e.g., increase the tail of the dependence) without using additional parameters. However, typical statistical samples do not allow a reliable estimation of many parameters. This is particularly the case for processes with LRD, as this behaviour entails parameter uncertainty dramatically higher than in SRD (Koutsoyiannis, 2003; Cohn & Lins, 2005, Koutsoyiannis & Montanari, 2007). Fifth, again except for low order processes (i.e., AR(1) and ARMA(1,1)), other processes of this family do

not have a physical meaning and thus are used in a rather black-box setting.

Several of the above drawbacks have been remedied by adaptations or even introduction of different model types. Thus, LRD can be reproduced by fractional Gaussian noise processes (Mandelbrot, 1965), fractionally differenced ARMA processes (Hosking, 1984) or by generalized generation schemes applied on generalized autocorrelation structures (Koutsoyiannis, 2000). The latter technique also tackles some of the other problems listed above as it is parametrically parsimonious (i.e., the generation scheme is not tied to the autocorrelation structure), it can handle non-normal distributions and is also multivariate. The seasonality problem has been tackled by cyclostationary processes. However, these are necessarily SRD because only low-order processes (such as periodic AR(1) or periodic ARMA(1,1); Bras & Rodriguez-Iturbe, 1985; Koutsoyiannis, 1999) are computationally feasible in a cyclostationary setting. Yet, however, in simulation mode (as opposed to prediction mode) the stochastic generation problem has been tackled using a disaggregation logic. Thus, the LRD properties are handled by an appropriate stationary model such as those stated above, the seasonality is handled by a cyclostationary model with SRD, and the two models are coupled so that the latter becomes operationally consistent with the former (Koutsoyiannis, 2001). A different approach that can generate cyclostationary time series with LRD without disaggregation was proposed recently (Langousis & Koutsoyiannis, 2006) but this is again for simulation.

Thus, while for stochastic simulation there exist advanced techniques (generally departing from ARMA model types) that are consistent with the peculiarities of hydrological processes, this does not happen in stochastic prediction. The modelling techniques for simulation may not be directly adaptable for prediction in a cyclostationary setting. For instance, a disaggregation framework is not appropriate for prediction.

In the last decade, this gap has been covered by techniques structurally different from stochastic techniques. These are based on recent advances on nonlinear dynamical systems (or chaotic systems) and have a deterministic basis. Most popular among these deterministic model categories are the ‘analogue’ models also called ‘local nonlinear chaotic’ models and the ‘connectionist’ models or metaphorically “artificial neural network” models. The

application of such models in hydrological prediction are numerous in the last years (a sample from *Hydrological Sciences Journal* in the last few years includes Tomasino, 2004; Hu et al., 2005; Shouyu & Honglan, 2005; Kisi, 2005; Giustolisi & Laucelli, 2005; Giustolisi & Simeone, 2006; Jayawardena et al., 2006; Abrahart et al., 2007; Corzo & Solomatine, 2007; de Vos & Rientjes, 2007; Muluye, & Coulibaly, 2007; see also See et al., 2007) and their performances are impressive in most cases. A major weak point of these model types is that, despite being deterministic in conception, in most cases they are data driven and black box, thus providing no process insight and involving no structured reasoning in their formulation. Another weak point is the fact that they do not provide tools for Monte Carlo simulation.

The nonlinear and data-driven character of these models have inspired many to devise stochastic models with such features. Such models (pioneered by Lall & Sharma, 1996) can implement a nonlinear data-driven (as opposed to linear parametric) dependence structure of the process and can reproduce the historical histogram in lieu of a theoretical distribution function. In our opinion, however, these may be weak rather than strong points of these model types. As we will discuss later, linearity in a deterministic approach may not be justifiable, but it makes sense in a stochastic approach. In other words, linearity is a concept with a totally different meaning in deterministic and stochastic approaches. Also the use of theoretical models justified by reasoning (e.g. the normal distribution justified by the central limit theorem or by the maximum entropy principle) is a powerful and insightful feature in stochastic modelling and should not be replaced by high uncertainty estimates of merely empirical basis, which after all are not appropriate to perform extrapolations that are largely needed.

In this paper we propose a general stochastic framework which is in full agreement with the features of hydrological process and the requirements for prediction, i.e., (a) it admits and utilizes LRD; (b) it can perform with distributions of either exponential or hyper-exponential tails; (c) it is cyclostationary; (d) it is parameter parsimonious; (e) it is insightful as it has a strong theoretical basis (principle of maximum entropy); (f) it can perform in both prediction and simulation; and (g) it is simple and easily applicable. In addition, we compare this stochastic approach with two data-driven models following a deterministic logic, namely an

analogue model and a connectionist model. The comparison includes both theoretical issues as well as the prediction skills as derived from a test case study.

The test case pertains to the Nile river and has significant interest both from a research as well as an operational standpoint. The Nile River is spread over 10 East African countries with numerous water uses including water supply for domestic, industrial, and agricultural use, energy generation, flood protection, and environmental management, among others (Georgakakos, 2007). Medium and long range Nile flow prediction is critical for the operation of the existing water control projects from Lake Victoria (in Kenya, Tanzania, and Uganda) to the High Aswan Dam in Egypt, and several planned facilities in the middle reaches (Ethiopia and Sudan). The forecast lead time considered in the case study and model intercomparison is one month; however the framework is general enough and can perform in longer lead times albeit with lower skills. The Nile is the world's longest river (6 670 km) with water travel times that, depending on the season, vary from 20 days (Blue Nile tributary) to more than 45 days (White Nile tributary). This induces strong dependence on a monthly time scale and, along with other storage mechanisms in the catchment, makes monthly forecast feasible.

The modern flow record at Aswan is one of the longest worldwide (131 years) and makes analysis and modelling reliable. In addition, there exist older instrumental records of annual maximum and minimum water level at the Roda Nilometer for more than 800 years. All flow records as well as additional historical and archaeological data (Said, 1993) affirm the LRD behaviour of the Nile flows and raise the demand that this dependence should be incorporated in a stochastic model, either for simulation or prediction.

Prediction models in this study are based on the available time series of Nile flows and make use of no additional explanatory variables. For the stochastic model, the 131 year available record is divided into two periods, an 78-year period for model fitting (60% of the total length; from hydrologic year 1870-71 to 1947-48) and a 53-year period for model validation (from 1948-49 to 2000-01). For the data-driven models, the fitting period is further subdivided into two sub-periods, a 52 year calibration period (2/3 of the total, from 1870-71 up to 1921-22) and a 26-year verification period (from 1922-23 to 1947-48). In this way, all

model fitting procedures are done exclusively within the 78-year period in all cases, thus enabling a fair comparison of all models in the validation period, the data of which were not used in model calibration.

2. Modelling approaches and underlying concepts

Let us consider a simple stochastic model which attempts to issue a prediction W of the flow at a specified month, say December, based on a single explanatory variable Z, say the flow at the same year in November. In a stochastic approach, W and Z are thought of as random variables (for this reason we have used an upper case convention, whereas for values, e.g., observations of the process, we use lower case letters). W is assumed stochastically dependent on Z. The dependence manifests the stochastic dynamics of the process, which can be represented by

$$W = g_s(Z, V) \tag{1}$$

where $g_s(\,)$ is an appropriate function and V is an additional random variable (assumed independent of Z) whose involvement manifests the fact that the dependence between the variables W and Z is not fully deterministic. If we know a realization of Z which is an observation z of the natural process, then we can calculate a point prediction of W by $E[W|Z = z] = E[g(z, V)]$, where $E[\,]$ denotes expectation. Using (1), we can also calculate prediction intervals of W for any desired confidence, either by analytical means or by Monte Carlo simulation.

To obtain any type of prediction we need to know the function $g_s(\,)$. We can have an idea of the type of this function by constructing a plot of historical observations of W and Z, such as that of Figure 1 (left panel). This plot reveals a practically (macroscopically) linear arrangement of points and suggests that $W = g_s(Z, V) = a\,Z + V$ (where a is a parameter). The fact that in this case the stochastic relationship appears to be so simple (linear) should not be regarded as a fortunate coincidence. Rather it seems to be the rule in hydrological and other processes. In addition, it has been observed that a heavily nonlinear system may become approximately linear again (Penland, 2006). An explanation for this could be traced on the

principle of maximum entropy (ME), which (as demonstrated in Koutsoyiannis, 2005a, b) can explain several behaviours of hydrological processes. Particularly, it is known (Papoulis, 1991) that maximization of the standard joint entropy (see definition on section 3.1) results in multivariate normal distribution. This entails a linear dependence of the lagged flows W and Z, as depicted in Figure 1 (left panel) and marked as "stochastic linear model 2". However, as discussed in Koutsoyiannis (2005a) and will be explained further in section 3.1, standard entropy may not be an appropriate representation of hydrological processes and the generalized entropy definition by Tsallis (2004) should be used instead. In this case, maximization of entropy will result in linear dependence of nonlinearly transformed flows using a normalizing transformation that will be discussed in section 3.1. The result in this case is also depicted in Figure 1 (left panel) marked as "stochastic linear model 1", which is only slightly different from model 2.

According to the deterministic approach, the lagged flows are not random variables and the system dynamics is a deterministic relationship of them, i.e.,

$$w = g_{\mathrm{d}}(z) \tag{2}$$

where $g_{\mathrm{d}}(\ )$ is a nonlinear function, such as the hypothetical (caricature) function shown in Figure 1 (right panel). This is a non-intersecting curve passing from all 78 points of the fitting period. If this caricature function represented the true dynamics of the process, then the additional 53 points of the validation period would lie on the curve. Since in fact they lie outside the curve, we should either change the function $g_{\mathrm{d}}(\ )$ or add explanatory variables, e.g., additional lagged flows, and replace the single variable z in (2) with a vector of variables $\mathbf{z} = [z_{(1)}, z_{(2)}, \ldots, z_{(m)}]^T$. (Clearly, additional variables could be also used in the stochastic case (1)). For a sufficient number of variables (or embedding dimension) m and an appropriate function $g_{\mathrm{d}}(\ )$ we may anticipate to reconstruct the dynamics of the system if the system is in reality deterministic.

According to the analogue (local linear) approach, the prediction is done without explicitly determining the function $g_{\mathrm{d}}(\ )$. Instead, the method tries to locate within the historical record a vector $\mathbf{z}^1$ that is nearest (in state) to the "current" vector $\mathbf{z}$ (i.e., an analogue past state of the

system). The value w^1 next (in time) to $\mathbf{z}^1$, which is known from the record, could be used as a prediction for the future of the current system state. In fact, the algorithm uses more than one nearest past states as will be explained later (section 4.1).

In contrast, in the connectionist (artificial neural network) approach the function $g_d(\ )$ is determined although it is usually so complicated that we do not even write its mathematical expression. An understanding of the complex relationships between inputs (function arguments) $\mathbf{z}$ and outputs w within a connectionist model is offered by the so called Kolmogorov's (1957) superposition theorem, according to which any continuous real function $g_d(\mathbf{z})$ of a vector variable $\mathbf{z}$ (defined on the m-dimensional hypercube $[0, 1]^m$) can be represented as a superposition and composition of continuous functions of only one variable. Formally, the theorem says that there exist continuous real functions $h_{ij}(z)$ and $g_i(z)$ such that

$$g_d(\mathbf{z}) = \sum_{i=1}^{2m+1} g_i\left(\sum_{j=1}^{m} h_{ij}(z_j)\right) \tag{3}$$

As shown by Kurkova (1992; based on earlier results by Hecht-Nielsen, 1987) and discussed by Beiu & Zawadzki (2005), it is possible to adapt Kolmogorov's theorem to an artificial neural network and to approximate the functions h and g by staircase-like functions.

Connectionist models typically use sigmoidal elementary functions ($\sigma(z) = 1/(1 + e^{b\,z - c})$) and perform weighted sums and compositions of many of them, according to some rules determined by a geometric analogue of nodes and arcs forming a network. The network topology includes an 'input layer' with m nodes, an output layer with one node and one or more "hidden" layers (in the case that (3) applies, there is only one hidden layer with $2m + 1$ nodes). Kurkova (1992) showed that connectionist models with standard sigmoidal functions and only two hidden layers could approximate any continuous function with arbitrary precision, but the number of units needed for a good approximation is exponential on m.

The three models, one stochastic and two deterministic with the above general features, are applied here in several configurations that will be detailed in the following sections. A synopsis of models and configurations is given in Table 1.

3. Stochastic model

Before we can construct a stochastic model it is necessary to study the marginal and dependence properties of the process of interest, here the Nile flow. Its summary statistics on a monthly and annual basis are given in Table 2. The convention of a hydrologic year is used, which for the Nile flows at Aswan, Egypt, it is assumed to start on 1 August. As observed in Table 2, two different regimes are typically observed. The flood period from August to October, when most of the flow comes from the Blue Nile, and the November to July base flow period when the flow is sustained by the White Nile. The two periods will be referred to as high-flow and low-flow periods, respectively.

3.1 Marginal distribution

Generally, the selection of a distribution function for use in a stochastic hydrological model is done empirically, based on comparisons of empirical statistics with theoretical ones of a repertoire of common distribution functions. Koutsoyiannis (2005a) proposed that theoretical reasoning could also assist this selection, and pointed to ME as the physical and mathematical principle that can be the basis for such reasoning. For a continuous random variable X with density $f(x)$ the standard entropy, also known as Boltzmann-Gibbs-Shannon entropy is by definition (e.g., Papoulis, 1991)

$$\varphi := E[-\ln f(X)] = -\int_{-\infty}^{\infty} f(x) \ln f(x)\, dx \tag{4}$$

A generalization of this definition, fruitfully used in several scientific fields including physics, chemistry, biology, economics, medicine, computer sciences and social sciences, and also useful in hydrology has been offered by Tsallis (1998, 2004):

$$\varphi_q = \frac{1 - \int_{-\infty}^{\infty} [f(x)]^q\, dx}{q - 1} \tag{5}$$

It can be easily checked that the limit for $q \to 1$ precisely reproduces the Shannon entropy, i.e., $\varphi_1 \equiv \varphi$. The above definitions can be generalized for vectors of random variables, where $f(\mathbf{x})$ is the joint density and φ_q or φ are joint entropies.

Maximization of standard entropy (4) (i.e., application of the ME principle) with simple constraints of known mean μ and variance σ^2 results in (e.g., Papoulis, 1991)

$$f(x) = \exp(-\lambda_0 - \lambda_1 x - \lambda_2 x^2) \tag{6}$$

where λ_0, λ_1 and λ_2 are parameters depending on μ and σ. Inspection of (6) shows that it is the normal density function. If the variable under study X is by definition non negative, as is the case for hydrological and most geophysical variables, maximization of entropy should incorporate the additional inequality constraint $x \geq 0$. In this case the resulting ME distribution is given by (6) again, but defined on $x \geq 0$, and it is the truncated normal distribution.

As discussed in Koutsoyiannis (2005a), the truncated normal distribution fails to describe cases in which the variation $\sigma/\mu > 1$. To find a ME solution for such cases, one should use Tsallis entropy (also known as nonextensive or nonadditive entropy) in lieu of standard entropy. Maximization of Tsallis entropy φ_q in (5) with known μ and σ^2 yields a hyper-exponential (power-type distribution), i.e.,

$$f(x) = [1 + \kappa (\lambda_0 + \lambda_1 x + \lambda_2 x^2)]^{-1 - 1/\kappa} \tag{7}$$

where $\kappa := (1-q)/q$ (Koutsoyiannis, 2005). For $\kappa \to 0$ ($q \to 1$), (7) switches to (6). Equation (7) is mathematically equivalent to the so called Tsallis distribution (Tsallis et al., 1995; Prato and Tsallis, 1999), as can be verified replacing $1-q$ with $q-1$. Note that the latter distribution has been obtained by constraining optimization not with the typical first and second moments as above but with generalized ones, known as q-expectations.

The fact that high variation σ/μ is common in hydrological variables at fine time scales is a strong indication of the applicability of Tsallis ME principle in hydrology. The most essential difference of (7) with respect to (6) is the implied hyper-exponential tail of distribution, which is quantified by parameter κ (for $\kappa > 0$). When the process of interest is aggregated from fine to coarser scales, σ/μ becomes smaller and smaller and the hyper-exponential behaviour of the tail becomes less and less visible from data due to the central limit theorem. However, it can be easily shown that theoretically the tail of the distribution is still hyper-exponential with the same κ, although the mathematical form of the distribution (7) is not preserved exactly (in contrast to (6) which is preserved in aggregation). Nonetheless, (7) can be used as an

approximation over a range of scales.

The above are investigated for the Nile flow as shown on Table 2 and Figure 2. The latter depicts normal probability plots of the distribution functions of annual and monthly Nile flows for August and December, which are representative of the high-flow and low-flow periods, respectively. The variation σ/μ is 0.20 for the annual flows and ranges from 0.23 to 0.43 for the monthly flows; these values can support the appropriateness of the (truncated) normal distribution. It can be seen that the normal distribution is a satisfactory approximation for the annual flows as well as for the monthly flows in August. This is also the case for September and October. Clearly, however, the flows in December exhibit a heavier tail, and this is observed for all months in the low flow period. In all these months, the empirical coefficients of skewness are positive (see Table 2) and those of kurtosis (and L-skewness and L-kurtosis) are higher than those of the normal distribution. Therefore, for the nine months of low flows, the departures of the empirical distributions from the normal could be attributed to over-exponential tails and could justify the use of (7) instead of (6).

As the normal distribution is very convenient in building a stochastic model either for simulation or prediction, one can think of applying a normalizing transformation $Z = g(X)$ to the variable of interest X, instead of using the non-normal distribution (7). In this case, Z will have a distribution that is approximated by (6), whereas X has a distribution that is approximated by (7). For appropriate selection of a translation parameter c, based on (7) and (6), we can write

$$\exp[-\lambda_2 (z-c)^2] \sim [1 + \kappa \lambda'_2 (x-c)^2]^{-1-1/\kappa} \tag{8}$$

from which (after algebraic manipulation and change of parameters) we obtain the following normalizing transformation:

$$Z = g(X) - g(0); \quad g(x) = c + \mathrm{sgn}(x-c)\, \lambda \sqrt{\left(1 + \frac{1}{\kappa}\right) \ln\left[1 + \kappa \left(\frac{x-c}{\lambda}\right)^2\right]} \tag{9}$$

This, in addition to the tail-determining dimensionless parameter κ, contains a scale parameter λ with the same units as x, which enables physical consistency of the transformation, and a translation parameter c, again with the same units as x. It is easily seen that: (a) z has the same units as x; (b) for x/λ ranging in $[0, \infty)$, z/λ also ranges in $[0, \infty)$; and (c) for $\kappa = 0$, z is identical

to x.

To apply the normalizing transformation to the monthly Nile flows in the low flow period, one may think of using different parameters κ, λ and c for each month. However, due to the already mentioned high estimation uncertainty, accurate estimation of $9 \times 3 = 27$ parameters is hardly attainable. Therefore, we prefer to assume $c = 0$ and a single pair of parameters κ, λ for all nine months, which we estimate by minimizing the departures of empirical skewness, kurtosis, L-skewness, L-kurtosis of z, aggregated over all months, from those of the normal distribution. The resulting parameters, estimated for the fitting period (78 years) were $\kappa = 2.76$ and $\lambda = 0.47$ km^3. The last panel of Figure 2, which depicts a normal probability plot of the transformed monthly flows z in December, indicates that transformation (9) performed a satisfactory normalization of the distribution. The statistical characteristics of the transformed monthly flows are shown in Table 3.

3.2 Dependence

Table 2 shows that the autocorrelations of the monthly Nile flows ($\rho_j = \mathrm{Corr}[X_i, X_{i+j}]$ for month $i = 1$ (August) to 12 (July) and lag $j = 1, 2, 12$) are very high but differ from month to month. Table 3 shows that the autocorrelations of the normalized flows (here $\rho_j = \mathrm{Corr}[Z_i, Z_{i+j}]$) have essentially the same values as in the natural flows. Figure 3 depicts the autocorrelograms for two months, August and December, representative for the high- and low-flow periods respectively, for lags up to 60 (corresponding to five years). Here we can observe that monthly autocorrelations differ significantly from month to month for small lags (periodicity) but become very similar for large lags, for which they keep high values thus suggesting LRD. These autocorrelograms are constructed from the 78-year fitting period; had the complete 131-year record been used, the peaks of autocorrelograms would be higher, indicating enhancement of both the periodicity and LRD. LRD is better seen on Figure 4, which depicts higher lag autocorrelations of monthly flows for lags that are multiples of 12 ($\mathrm{Corr}[X_i, X_{i+12j}]$), so as to eliminate periodicity, as well as autocorrelations of annual flows ($\mathrm{Corr}[Y_i, Y_{i+j}]$, where $Y_i := X_{12(i-1)+1} + \ldots + X_{12i}$, the annual flow at year i). Note that the presence of LRD implies much higher uncertainty than in classical statistics, as well as bias in

classical statistical estimators (Koutsoyiannis, 2003). Therefore, in Figure 4, two series of estimates of ρ_j have been plotted, the classical statistical ones (marked as "empirical classical") and the adapted ones (marked as "empirical SSS" where SSS stands here for simple scaling statistics; see discussion later) that recover from bias (Koutsoyiannis, 2003). In addition to empirical estimates, some model curves are also plotted, which will be discussed later. All panels in Figure 4 verify the presence of LRD and interestingly, indicate that this behaviour is virtually the same in all months as well as annually.

All above observations support a modelling approach of a mixed type, with a cyclostationary description of dependence at small lags (different for each month) and a stationary description for large lags (same for all months). The former can be done easily, using a small number of empirical autocorrelation coefficients (as those shown in Tables 2 and 3 estimated from the fitting period). To specify this number, we use the notion of explained variance. We observed that the portion of total variance explained by two autocorrelations (ρ_1 and ρ_2) is increased considerably from the case of using only one (ρ_1) whereas the addition of additional autocorrelation coefficients (ρ_3, ρ_4, …) essentially does not increase the explained variance. Thus, the values of ρ_1 and ρ_2 given in Table 3 suffice to describe the dependence for small lags. This should not be confused with the adoption of a periodic AR(2) or ARMA(1,1) model as happens typically in stochastic modelling using two autocorrelations. These models would imply SRD, while here we introduce LRD.

A simple stationary structure with LRD is the simple scaling stochastic process (SSS process) with autocorrelation (e.g., Koutsoyiannis, 2002)

$$\rho_j = (1/2)\,[|j+1|^{2H} + |j-1|^{2H}\,] - |j|^{2H} \approx H(2H-1)\,|j|^{2H-2} \tag{10}$$

where H is the so-called Hurst coefficient with values in the interval (0.5, 1) for positively autocorrelated processes. Here j is meant as the lag for the annual scale or the lag divided by 12 for the monthly scale. However, Figure 4 suggests that autocorrelation in the Nile decays with lower rates than implied by (10). Therefore, we need to investigate it further, again using the ME principle.

To determine the dependence structure of a stochastic process, Koutsoyiannis (2005b)

maximized average entropy on a range of timescales with appropriate constraints. That ME framework was revisited and advanced in light of the statistical behaviours observed in the Nile (Koutsoyiannis and Georgakakos, 2006). The results of the latter work are also used here. To summarize them, the entropy maximization is done on a time scale d tending either to zero (a local setting) or to infinity (a global setting). Maximization of entropy is done numerically using a parametric form of the autocorrelation function, initially formulated in continuous time, as shown in Appendix 1. This parametric autocorrelation includes three components: a white noise term, a SRD term and an LRD term. This parametric representation is deliberately rich (it includes three parameters) in order to provide appropriate degrees of freedom for the entropy maximization.

Entropy maximization either at the local or the global setting results in virtually the same solution if two autocorrelation constraints are used. This solution is more complicated than SSS but tends to SSS as scale increases. For this reason, we call it an asymptotic scaling stochastic process (ASS process). The exact mathematical form is given in Appendix 1. Both SSS and ASS structures are plotted in terms of the resulting autocorrelation functions in Figure 4 also in comparison with empirical autocorrelations. Practically, the autocorrelation of either SSS and ASS is a power function of the lag; the difference is that in SSS both the slope and the intercept are dependent on each other (they are functions of the single parameter H, namely $2H-2$ and $\rho_1 = 2^{2H-1} - 1$, respectively) whereas in ASS the intercept ρ_1 does not determine the slope of the autocorrelation decay. Here the SSS and ASS models were determined for the annual series and applied also in the monthly series. Thus, the model curves in all panels of Figure 4 are exactly the same. It seems that ASS is in all cases in better agreement with the empirical points and therefore we use this in the development of the model. Even though, due to the adopted rich parameterization, the ASS autocorrelation is nominally dependent on six parameters, in fact two parameters suffice to describe it (i.e., the intercept and slope in the plots of Figure 4, which correspond to the two autocorrelations that were used as constraints).

The link of short-term dependence, described by the two small lag autocorrelations, with the long-term dependence, described with the ASS model, will be done in the next subsection

to construct an operational stochastic model for prediction as well as for simulation.

3.3 Model formulation

In accordance to (1) and the linearity assumption justified in section 2, our stochastic model will be a general linear one, i.e.,

$$W = \mathbf{a}^T \mathbf{Z} + V \tag{11}$$

where **a** is a vector of weights and V is a random term assumed independent of **Z**. The variable to be predicted, W, and all items of the vector of explanatory variables, **Z**, are normalized monthly flows. Specifically, assuming that the current time is $i - 1$, denoting Z_i the normalized flow at time i and $Z_{(i)}$ the ith item of **Z**, we will have

$$W \equiv Z_i,\ Z_{(1)} \equiv Z_{i-1},\ Z_{(2)} \equiv Z_{i-2},\ Z_{(3)} \equiv Z_{i-12},\ Z_{(4)} \equiv Z_{i-24},\ \ldots,\ Z_{(m)} \equiv Z_{i-12(m-2)} \tag{12}$$

Thus, the first two items of **Z** are the nearest in time normalized monthly past flows whereas all other items are normalized past flows of the same month of the year as the month in time i. With this composition of **Z**, the model takes account of both long-term and short-term dependence. To account for LRD as much as possible, we should make m as large as possible. In fact, the size m of **Z** is determined by the available data record that conditions prediction. In our case, $m - 2 = 78$ (the length of the fitting period), so $m = 80$. It should be emphasized that the size m is not at all related to the number of parameters to be estimated from the data, and there is no reason to seek parsimony in this case.

To specify the model we need to determine the weights **a** and the statistical characteristics of V. Assuming (for convenience) that W and **Z** have zero mean and unit variance, V will have zero mean, variance $\mathrm{Var}[V] < 1$, and normal distribution. Multiplying both sides of (11) by $\mathbf{Z}^T$ and taking expected values we obtain

$$\mathbf{a}^T = \boldsymbol{\eta}^T \mathbf{h}^{-1} \tag{13}$$

where $\boldsymbol{\eta} := \mathrm{Cov}[W, \mathbf{Z}]$ and $\mathbf{h} := \mathrm{Cov}[\mathbf{Z}, \mathbf{Z}]$. Squaring both sides of (11) and taking expected values we obtain

$$\mathrm{Var}[V] = 1 - \boldsymbol{\eta}^T \mathbf{h}^{-1} \boldsymbol{\eta} = 1 - \mathbf{a}^T \boldsymbol{\eta} \tag{14}$$

We observe that the vector **η** contains 80 monthly autocorrelation items, the lag one and two, which are model parameters estimated from the data, and the lag 12, 24, and so on, which, as described above, are equal to the lag 1, 2, … autocorrelations of the annual flows. The latter are determined from the ASS model described in section 3.2 in terms of a couple of free parameters (annual autocorrelations for two lags). Thus the first two items of **η** change from month to month whereas all others are the same for all months. The matrix **h** contains numerous items (80 × 80 = 6400 for each month). However most of them (the lower 78 × 78 part of the matrix) are determined from the ASS model and few more (the upper 2 × 2 part of the matrix) contains lag one and two autocorrelations already appearing in **η**. The remaining part of the matrix (two 2 × 78 areas, symmetric to each other because **h** is symmetric) contains unknown autocorrelations (Cov[W, Z_i] for several Z_i).

According to prevailing practices in stochastic modelling, these unknown autocorrelations, whose number is very large (12 × 2 × 78 = 1872) would be estimated from the data. Even though this is technically feasible (and done in some cases such as in most disaggregation models) it makes no sense, given that the available data values are 12 × 78 = 936, i.e., half the number of these unknown autocorrelations.

Here we propose that these parameters should be left 'unestimated' in the statistical sense and should be calculated by applying the ME principle. In this way, no additional parameter is introduced in the stochastic model. As shown in Appendix 2, the entropy maximization in this case has an easy analytical solution that can be formulated as a generalized Cholesky decomposition of the matrix **h** (assuming that $\mathbf{h} = \mathbf{b}\,\mathbf{b}^T$, where **b** is a lower triangular matrix to be calculated). In this case the total number of autocorrelation parameters to be estimated statistically does not exceed (12 +1) × 2 = 26 (two autocorrelations per month plus two annual autocorrelations) for the entire model; thus, the proposed model is indeed parametrically parsimonious.

After the calculation of the matrix **h**, all other computational effort is trivial (typical matrix operations). Model (11) can perform either in forecast mode or in simulation mode. To apply model (11) to obtain a point prediction W for the observed values $\mathbf{Z} = \mathbf{z}$, it suffices to set V = 0; the resulting value of W from (11) will be the expected value conditional on $\mathbf{Z} = \mathbf{z}$. Interval

predictions can be easily derived analytically based on the distribution of V which is normal with zero mean and variance Var[V]. Furthermore, stochastic simulation is also easily performed using the same model with V generated from the normal distribution.

3.4 Results

Figure 5 provides a graphical depiction of the vector of weights **a** for two months. Due to the higher value of lag one autocorrelation in December (0.94 against 0.71 in August) the relative weight of distant lagged past flows is much smaller in December rather than in August. The entire picture of weights changes significantly from month to month (cyclostationarity) even though most parts of matrices **η** and **h** represent stationary components. Generally, the weights decrease with the increase of lag but this is reversed for very high lags. This seems counterintuitive but it is totally justified: the non availability of information for lags higher than 78 years results in relatively higher weights near the 78-year lags.

Based on the twelve vectors of weights, the model was applied to predict the flows of the 53-year validation period. Each time, the most recent 78-year historical information was used to condition the prediction. The results are shown graphically in Figure 6, both in terms of natural and standardized (by month) values, and indicate a very satisfactory proximity with actual values. A numerical index of performance index is the attained coefficient of efficiency

$$C_E = 1 - E[(W - X)^2] / \mathrm{Var}[X] \tag{15}$$

where X is the actual variable that is predicted by W. As shown in Table 6, the performance index is very high.

In addition to this full configuration of the stochastic model, abbreviated as S1, two additional ones were examined (see Table 1), whose performances are also shown in Table 6. In configuration S2, which is similar to S1 but without normalizing transformation, the performance is only slightly lower than in S1. Even stochastic model S3, which is a typical PAR(2) process without a normalizing transformation, the performance is good, but inferior to those of S1 and S2.

4. Deterministic modes

4.1 Analogue model

As described in section 2, the logic and the algorithm of the analogue model are very easy (Kantz & Schreiber, 1997; Georgakakos & Yao, 1995, Yao & Georgakakos, 2001). In operational mode, the only difference from the general description of in section 2 is that a number of neighbours $\mathbf{z}^1$, …, $\mathbf{z}^n$, instead of a single vector $\mathbf{z}^1$, is used and the prediction w is extracted as the average of w^1, …, w^n, the states next (in time) to the latest coordinate of $\mathbf{z}^1$, …, $\mathbf{z}^n$, respectively. The number n can be fixed or varying determined in a manner that the vectors $\mathbf{z}^i$ have distance from $\mathbf{z}$ smaller than a threshold; here the first option has been used. The vector $\mathbf{z}$ is formed from the current state and some earlier ones whereas the vectors $\mathbf{z}^i$ are sought in the calibration period exclusively.

Thus, the analogue model involves no parameters except the size of the vector $\mathbf{z}$ (embedding dimension, m) and the number of neighbours n. These adjustable quantities are determined by a trial-and-error procedure aiming at finding the optimal m and n that make the prediction error minimum at the verification period. Application of the method with the Nile flows resulted in the variation of the coefficient of efficiency in the verification period with m and n that is shown in Figure 7. It is generally observed that if we exclude too low values (i.e. m = 1 and n = 1-2) the efficiency is very good. Two local optima were found corresponding to (m = 12, n = 11) and (m = 13, n = 24). The efficiencies of these two model configurations, abbreviated as A1 and A2, respectively (see Table 1) are shown in Table 4.

We also studied an additional model configuration, inspired by the stochastic model. In this to construct the vector $\mathbf{z}$ we assumed (similar to (12)) a variable (rather than constant) time delay, i.e., $\mathbf{z} = [z_{i-1}, z_{i-2}, z_{i-12}, z_{i-24}, \ldots, z_{i-12(m-2)}]^T$. In this case, however, we cannot use a high m (like 78 in the stochastic model) because we would run out of a pool of neighbours. Thus, we only tested the cases m = 3 and 4; with these values we cannot anticipate to capture the long-term dependence properties of the process, but only to simplify the model (using 4 instead of 24 terms for the same effective total lag). The resulting efficiencies for these two cases are also plotted in Figure 7 and are comparable to that of the

constant time lag cases. Among the several configurations of this type shown in Figure 7, the optimal was that with $m = 4$ and $n = 7$; this has also been included in Table 1 and Table 4 abbreviated as A3.

4.2 Connectionist model

The connectionist model used in this study follows the logic described in section 2; its details are described in Georgakakos & Yao (1995). In our case study, structures with one or two hidden layers have been examined. The model fitting, metaphorically known as 'training' or 'learning', is a nonlinear optimization procedure than minimizes fitting errors. In this case it was executed by the "error backpropagation" method which is a version of a gradient descent method.

As opposed to the analogue model case, in which the natural flows were used, here the flows were standardized by the mean and standard deviation of each month. To avoid over-fitting (i.e., the use of too many components of elementary functions, a common propensity of connectionist models) an early stopping method was used combined with two fitting measures: the calibration error (in the calibration period) and the verification error (in the verification period; Georgakakos and Yao, 1995). Typically, the calibration error decreases steadily while the verification error initially decreases and eventually increases, exhibiting a minimum. Model calibration is typically terminated when the verification error achieves a minimum value.

Several model configurations were tested, which make two groups. In the first group a constant time delay (1 month) was assumed, the number of inputs varied from 1 to 15, the hidden layers from 1 to 2, and the hidden nodes in each layer from 1 to 15. In the second group the time delay was variable (as in the analogue model), the number of inputs was fixed to 4, the hidden layers were 1 or 2, and the hidden nodes in each layer varied from 1 to 10. The tradeoff of the two fitting measures for all examined configurations and the Pareto front formed are shown in Figure 8. From the solutions lying in the Pareto front three were chosen as model configurations C1, C2 and C3, whose characteristics are shown in Figure 8 and Table 1. Further, in Table 5, which depicts several fitting criteria, we observe that the

performance of all configurations is very good (having in mind that the values given in the table are for the standardized variables) with C3 slightly outperforming the other two.

5. Model intercomparison

The intercomparison of models in terms of their prediction skill is made for the 53-year validation period, which was not used in any fitting procedure and in any model. Three performance indices which are the coefficients of efficiency of untransformed values, logarithmically transformed values, and monthly standardized untransformed value are shown in Table 6. By all indices, the stochastic models S1 and S2 have the best performance and are followed by the connectionist model C1 (which has almost equal performance with S3) and the analogue model A3.

In terms of simplicity and easiness of application, the analogue model is best. A spreadsheet environment suffices to develop, calibrate, and run it, and its development can take place in very short order. Particularly, the configuration A3, which gave the best, among the three analogue models, performance in the validation period, is also the simplest (as it involves only four variables) and fastest in calibration and running. The next model group in terms of simplicity is the stochastic. The configuration S3 is very simple but even the full proposed model S1 is simple enough to be implemented on a spreadsheet environment. Here we deliberately discussed the model in depth without simplifications. However, in a practical application simplifications are possible. For instance, the SSS model could be adopted by default, without performing entropy maximization to determine LRD. Another option is to use a generalized parametric autocorrelation as in Koutsoyiannis (2000) and determine its parameters by fitting it (e.g., by least squares) to the empirical autocorrelogram, again without entropy maximization. In contrast the connectionist approach is not simple and cannot be implemented on a spreadsheet.

In terms of model ability to perform in simulation mode, in addition to forecast mode, only the stochastic model provides this option, whose procedure was discussed in section 3.3. To illustrate this, a synthetic record of length equal to that of the historical was generated by model S1. Comparisons of the statistics of the synthetic and historical monthly records are

given in Figure 9, which indicates a satisfactory performance. Some discrepancies of the skewness and lag 12 autocorrelations during low flows are usual due to the small sample size; to match such statistics a longer sample size by several orders of magnitude is need, particularly because of the increased sampling variability due to LRD.

The analogue model cannot operate in simulation mode because soon it converges to an "attracting" periodic trajectory, same for all years. The connectionist model, when the number of nodes is small, behaves similarly to the analogue model resulting in an "attracting" periodic trajectory. For more than 15-20 hidden nodes, it produces irregular trajectories, which do not resemble and are statistically dissimilar to the historical flows.

In terms of potential insights of the process, we can argue that the stochastic approach offers some, especially when combined with the maximum entropy principle. The latter, as discussed above provides an explanation for the observed linearity on a stochastic setting, of the marginal distribution and particularly its tail, and of the dependence structure, particularly the long-term one. The parametric setting of the stochastic approach, along with these insights, offers the ability of controlling the entire procedure. In contrast, the deterministic approaches, which are data driven, do not offer any control and are black box rather than insightful.

It has been argued that such or similar deterministic approaches (e.g., time delay embedding, see Koutsoyiannis, 2006) offer insights because they reconstruct the dynamics of the process based on the observed time series and uncover its deterministic attractor. This is true for simple low dimensional experimentation systems (e.g., with one positive Lyapunov exponent) but it is unlikely to be the case for complex natural processes, such as the flow of the Nile. The fact that such deterministic models can cast good predictions should not necessarily be given the interpretation that the process is governed by deterministic dynamics. To illustrate this we used the aforementioned synthetic flows generated by the stochastic model S1, to which we applied the analogue model. As shown in Figure 10, despite the a priori known stochastic character of the inflows, the analogue model gave good predictions and its performance is comparable to that with the historical data. Thus, a good prediction does not necessarily imply deterministic dynamics.

6. Conclusion and discussion

A general conclusion of this work is that it is always worth to construct a good stochastic model of a hydrological process. Such a model can operate in simulation mode as well in forecast mode and thus can support strategic planning as well as real time management of a hydrosystem. In addition, the development of a good stochastic model is closely linked with understanding of natural behaviours in a bidirectional manner: a good model presupposes understanding and also supports understanding by providing insights into natural behaviours. These behaviours include extreme phenomena (distribution tails) and temporal dependences, particularly, LRD. It becomes obvious then that good modelling practices should depart from the typical ARMA formalism.

The principle of maximum entropy can largely support the development of stochastic models, providing both logical foundation and computational tools. Here the principle was used four times, i.e., (a) to infer the marginal distribution of the process (b) to explain and model the long-term dependence of the process, (c) to justify the (macroscopical) linearity in lagged flows of Nile, and (d) to determine unknown covariances in the stochastic model structure. This study offers advances in (a) and (b) whereas to our knowledge propositions (c) and (d) are original. Particularly, proposition (d) offers a powerful yet very simple (because of the closed analytical solution) computational tool for the construction of a generalized stochastic model.

Deterministic modelling alternatives, which recently have been given great attention, are good practical tools, too. Particularly the analogue model is very attractive due to simplicity, non-parametric character and easiness to construct and apply. However, care is needed in interpretation. Good predictions by deterministic models do not necessarily mean consistency of the natural process with determinism. In this case study, all configurations of deterministic models gave performance inferior to the advanced stochastic model. Perhaps however, in another application or with the use of more advanced deterministic techniques, they may perform better. Still, however, deterministic models are inferior on other grounds, such as in supporting Monte Carlo simulation of hydrosystems or in deriving interval predictions, in describing and exploiting the long term persistence, and in offering insights of the process.

It may seem counterintuitive that the particular stochastic model developed in this study, which largely relies on the principle of maximum entropy, in other words on the postulation that uncertainty in nature is as large as possible, yields better forecasts than the deterministic models negating uncertainty. Perhaps stochasticity and the notion of maximum entropy explain natural behaviour better than determinism.

Acknowledgements We wish to thank Dr. Bayoumi Attia of the Egyptian Ministry of Water and Irrigation for providing the Nile flow data as part of the Lake Nasser Flood and Drought Control Project funded by the Government of Netherlands. We also thank Constantino Tsallis for the discussion and suggestions on the generalized entropy and the normalizing transformation. This work was partially supported by the Georgia Water Resources Institute.

.

References

Abrahart, R.J., Heppenstall, A.J., & See, L. M. (2007), Timing error correction procedure applied to neural network rainfall-runoff modelling, *Hydrol. Sci. J.*, 52(3), 414-431.

Barnes, F. B. (1954) Storage required for a city water supply, *J. Inst. Eng. Australia*, 26(9) 198-203.

Beiu, V. & Zawadzki, A. (2005) On Kolmogorov's Superpositions: Novel Gates and Circuits for Nanoelectronics?, *Proceedings of International Joint Conference on Neural Networks*, Montreal, Canada, Jul. 31 – Aug. 4, 2005.

Box, G. E. P., and Jenkins, G. M. (1970) *Time series analysis; Forecasting and control*, Holden Day.

Bras, R. L., & Rodriguez-Iturbe, I., (1985) *Random functions in hydrology*, Addison-Wesley.

Cohn, T. A., & Lins, H. F. (2005) Nature's style: Naturally trendy, *Geophys. Res. Lett.*, 32(23), L23402, doi:10.1029/2005GL024476.

Corzo, G., & Solomatine, D. (2007), Baseflow separation techniques for modular artificial neural network modelling in flow forecasting, *Hydrol. Sci. J.*, 52(3), 491-507.

de Vos, N.J., & Rientjes, T.H.M. (2007), Multi-objective performance comparison of an artificial neural network and a conceptual rainfall-runoff model, *Hydrol. Sci. J.*, 52(3), 397-413.

Georgakakos, A. (2007), Decision Support Systems for Integrated Water Resources Management with an Application to the Nile Basin, Chapter 5 in Topics on System Analysis and Integrated Water Resources Management, eds. A. Castelletti & R. Soncini-Sessa, Elsevier, 2007.

Georgakakos, A. & Yao, H. (1995), Inflow forecasting models for the High Aswan Dam, *Technical Project Report to the Egyptian Ministry of Public Works and Water Resources*, School of Civil and Environmental Engineering, Georgia Tech, Atlanta.

Giustolisi, O., & Laucelli, D. (2005) Improving generalization of artificial neural networks in rainfall–runoff modelling, *Hydrol. Sci. J.*, 50(3), 439-457.

Giustolisi, O. & Simeone, V. (2006) Optimal design of artificial neural networks by a multi-objective strategy: groundwater level predictions, *Hydrol. Sci. J.*, 51(3), 502-523.

Hecht-Nielsen, R. (1987) Kolmogorov's mapping neural network existence theorem, *Proc. IEEE International Conference on Neural Networks*, 3, 11-14.

Hosking, J. (1984), Modeling persistence in hydrological time series using fractional differencing, *Water Resour. Res.*, 20(12), 1898– 1908.

Hu, T. S., Lam, K. C., & Thomas Ng, S. (2005) A Modified Neural Network for Improving River Flow Prediction, *Hydrol. Sci. J.*, 50(2), 299-318.

Jayawardena, A. W., Xu, P. C., Tsang, F. L., & Li W. K. (2006) Determining the structure of a radial basis function network for prediction of nonlinear hydrological time series, *Hydrol. Sci. J.*, 51(1), 21-44.

Kantz, H., and T. Schreiber (1997), *Nonlinear Time Series Analysis*, Cambridge University Press, Cambridge.

Kolmogorov A.N. (1957) On the representation of continuous functions of many variables by superposition of continuous functions of one variable and addition, *Dokl. Akad. Nauk* USSR, 114:953–956, 1957.

Kisi, O. (2005) Suspended sediment estimation using neuro-fuzzy and neural network approaches, *Hydrol. Sci. J.*, 50(4), 683-696.

Koutsoyiannis, D. (1999) Optimal decomposition of covariance matrices for multivariate stochastic models in hydrology, *Water Resour. Res.*, 35(4), 1219-1229.

Koutsoyiannis, D. (2000), A generalized mathematical framework for stochastic simulation and forecast of hydrologic time series, *Water Resour. Res.*, 36(6), 1519-1533.

Koutsoyiannis, D. (2001) Coupling stochastic models of different time scales, *Water Resour. Res.*, 37(2), 379-392.

Koutsoyiannis, D., The Hurst phenomenon and fractional Gaussian noise made easy, *Hydrol. Sci. J.*, 47(4), 573-595, 2002.

Koutsoyiannis, D. (2003), Climate change, the Hurst phenomenon, and hydrological statistics, *Hydrological Sciences Journal*, 48(1), 3-24.

Koutsoyiannis, D. (2005a), Uncertainty, entropy, scaling and hydrological stochastics, 1, Marginal distributional properties of hydrological processes and state scaling, *Hydrol. Sci. J.*, 50(3), 381-404.

Koutsoyiannis, D. (2005b), Uncertainty, entropy, scaling and hydrological stochastics, 2,

Time dependence of hydrological processes and time scaling, *Hydrol. Sci. J.*, 50(3), 405-426.

Koutsoyiannis, D. (2006) On the quest for chaotic attractors in hydrological processes, *Hydrol. Sci. J.*, 51(6), 1065-1091.

Koutsoyiannis, D., & Georgakakos, A. (2006) Lessons from the long flow records of the Nile: Determinism vs. indeterminism and maximum entropy, *20 Years of Nonlinear Dynamics in Geosciences*, Rhodes, Greece, 11-16 June 2006.

Koutsoyiannis, D., & Montanari, A. (2007) Statistical analysis of hydroclimatic time series: Uncertainty and insights, *Water Resources Research*, 43(5), W05429.1-9.

Kurkova, V. (1992) Kolmogorov's theorem and multilayer neural networks, *Neural Networks*, 5(3), 501-506.

Lall, U., & Sharma, A. (1996) A nearest neighbor bootstrap for resampling hydrologic time series. *Water Resour. Res.* 32(3), 679-693.

Langousis, A., & Koutsoyiannis, D. (2006) A stochastic methodology for generation of seasonal time series reproducing overyear scaling, *J. Hydrol.*, 322, 138-154.

Mandelbrot, B. B. (1965) Une class de processus stochastiques homothetiques a soi: Application a la loi climatologique de H. E. Hurst, *Compte Rendus Academie Science*, 260, 3284-3277.

Muluye, G., & Coulibaly, P. (2007), Seasonal reservoir inflow forecasting with low-frequency climatic indices: a comparison of data-driven methods, *Hydrol. Sci. J.*, 52(3), 508-522.

Papoulis, A. (1991), *Probability, Random Variables, and Stochastic Processes* (third edn.), McGraw-Hill, New York.

Penland, C. (2006), A stochastic view of nonlinear geosystems, *20 Years of Nonlinear Dynamics in Geosciences*, Rhodes, Greece, 11-16 June 2006.

Prato, D. & Tsalis, C. (1999) Nonextensive foundation of Lévy distributions, Physical Review E, 60(2), 2398-2401.

Said, R. (1993), *The River Nile: Geology, Hydrology and Utilization*, Pergamon Press, Oxford.

See, L., Solomatine, D., Abrahart, R. & Toth, E. (2007), Hydroinformatics: computational intelligence and technological developments in water science applications – Editorial,

Hydrol. Sci. J., 52(3), 391-396.

Shouyu, C., & Honglan, J. (2005), Fuzzy Optimization Neural Network Approach for Ice Forecast in the Inner Mongolia Reach of the Yellow River, *Hydrol. Sci. J.*, 50 (2), 319-330.

Thomas, H. A., & Fiering, M. B. (1962) Mathematical synthesis of streamflow sequences for the analysis of river basins by simulation, in *Design of Water Resource Systems*, by A. Maass, et al., Harvard Univerity Press, Cambridge, Mass.

Tomasino, M., Zanchettin, D., & Traverso, P. (2004) Long-range forecasts of River Po discharges based on predictable solar activity and a fuzzy neural network model, *Hydrol. Sci. J.*, 49(4) 673-684.

Tsallis, C. (1988) Possible generalization of Boltzmann-Gibbs statistics, *J. Statist. Phys.*, 52, 479–487.

Tsallis, C. (2004) Nonextensive statistical mechanics: construction and physical interpretation, in *Nonextensive Entropy, Interdisciplinary Applications* (edited by M. Gell-Mann & C. Tsallis), Oxford University Press, New York, NY.

Tsallis, C., Levy, S. V. F., Souza, A. M. C., & Maynard, R. (1995), Statistical-mechanical foundation of the ubiquity of Lévy distributions in nature, *Phys. Rev. Lett.*, 75, 3589–3593.

Yao, H, & Georgakakos, A. (2001), Assessment of Folsom Lake response to historical and potential future climate scenarios, *J. Hydrol.*, 249, 176-196.

Appendix 1: Entropy maximization to determine LRD

We assume a continuous time stationary representation of the process of interest with the following three-component parametric autocovariance function:

$$\gamma(\tau) := \mathrm{Cov}[X(t), X(t+\tau)] = \lambda_0\,\delta(\tau) + \lambda_1 \exp(-\kappa_1|\tau|) + \lambda_2\,(1 + \kappa_2\,\beta\,|\tau|)^{-1/\beta} \tag{16}$$

where $\delta(\tau)$ is the Dirac delta function and λ_0, λ_1, λ_2, κ_1, κ_2 and β are parameters, all nonnegative and with the following dimensions: β dimensionless, $[\kappa_1] = [\kappa_2] = [\mathrm{T}^{-1}]$, $[\lambda_1] = [\lambda_2] = [X^2]$, and $[\lambda_0] = [X^2]\,[\mathrm{T}]$. The three components of $\gamma(\tau)$ are respectively, a white noise term (Dirac delta), a SRD term (Markovian or exponential function of time lag) and an LRD term (generalized power function of time lag with exponent β manifesting LRD unless $\beta = 0$). For $\beta = 0$ the third term becomes a second exponential term and represents SRD at a second characteristic time scale. For $\beta > 1$ this term approaches the SSS autocovariance with Hurst exponent $H = 1 - 1/(2\beta)$. When $\lambda_0 = \lambda_1 = 0$, $\lambda_2 \neq 0$, $\beta > 1$ and $\kappa \to \infty$, the process becomes precisely SSS.

Let X_i denote the process at discrete time with scale d, i.e.,

$$X_i := \frac{1}{d} \int_{(i-1)d}^{id} X(t)\,dt \tag{17}$$

Then, the following equations can be obtained from typical calculus of stochastic processes:

$$\gamma_0 := \mathrm{Var}[X_i] = \lambda_0 / d + 2\,\varepsilon_1\,(\xi_{1,1} + \kappa_1 d - 1) + 2\,\varepsilon_2\,[\xi_{2,1} - (2\beta - 1)\,\kappa_2 d - 1] \tag{18}$$

$$\gamma_1 := \mathrm{Cov}[X_i, X_{i+j}] = \varepsilon_1\,(\xi_{1,j-1} + \xi_{1,j+1} - 2\,\xi_{1,j}) + \varepsilon_2\,(\xi_{2,j-1} + \xi_{2,j+1} - 2\,\xi_{2,j}), \quad j > 0 \tag{19}$$

where

$$\varepsilon_1 := \lambda_1 / (\kappa_1^2 d^2),\ \ \varepsilon_2 := \lambda_2 / [(\beta - 1)(2\beta - 1)\,\kappa_2^2 d^2] \tag{20}$$

$$\xi_{1,j} := \exp(-\kappa_1 d j),\ \xi_{2,j} := (1 + \kappa_2\,\beta\,d\,j)^{2 - 1/\beta} \tag{21}$$

It can be seen that the Markovian second term in continuous time becomes ARMA(1, 1) in discrete time (as $\rho_j = \gamma_j / \gamma_0 \sim \exp(-\kappa_1 d j)$) (similar is the behaviour of the third term for $\beta = 0$).

The above process representations are stationary. In can be assumed that the stationary representation in discrete time is satisfactory for time scales d that are multiples of the annual scale. In fact, in the above framework these are the scales of interest for determining the LRD properties; also in these scales the normal distribution is a good approximation and thus the standard entropy definition can be used. In time scales smaller than the annual (and in the continuous time setting) the process can be interpreted as representing an average behaviour within the seasons of the year.

The entropic quantities involved in the application of the ME principle to determine the long-term dependence of the process are the standard (unconditional) entropy φ defined in (4), the conditional entropy φ_c and the information gain $\psi := \varphi - \varphi_c$. The conditional entropy is the entropy of a random variable representing a future step when the present and the entire past are observed and is given by

$$\varphi_c := \lim_{m \to \infty} (\varphi_m - \varphi_{m-1}) \tag{22}$$

where φ_m is the joint entropy of m consecutive variables in the stochastic process of interest, which is given by (4) (replacing the single variable with the vector of variables). In a Gaussian process φ_m simplifies to

$$\varphi_m = \ln\sqrt{(2\,\pi\,e)^m\,|\mathbf{c}_m|} \tag{23}$$

where $\mathbf{c}_m$ is the covariance matrix of the vector of the m consecutive variables and $|\mathbf{c}_m|$ is its determinant (Papoulis, 1991).

The technical details of the entropy maximization procedure, according to which the six parameters of the autocovariance function are determined, are given elsewhere (Koutsoyiannis and Georgakakos, 2006). A summary of the method is as follows: (a) The maximization of entropy (unconditional φ or conditional φ_c) is done numerically based on the above parametric representation of the dependence structure for a time scale tending to 0 or ∞; (b) Numerically, the limiting timescales are approximated by $d = 2^{\pm s}$ where it was chosen $s = 7$; (c) The constraints used in the maximization deal with known variance (the mean is not necessary if the distribution is normal), and known annual autocorrelation at lag one and at a greater lag (the use of a second autocorrelation is useful because it makes the solution

practically independent of the choice of the time scale $d = 2^{-s}$ or $d = 2^{+s}$); (d) In addition to these equality constraints, two inequality constraints (restrictions), related to the information gain $\psi^{(d)}$ at scale d are used to prevent solutions without a physical meaning: (d_1) $\psi^{(0)} \geq \psi^{(d)}$ for any $d > 0$ (meaning that predictability at any timescale d should be lower than that instantaneously after the measurement) and (d_2) $\psi^{(\infty)} < \infty$ (prohibiting an illimitable predictability at very large scales).

Application of the procedure with the 78-year record of standardized annual flows with constraints Var[Y] = 1, $\rho_1 = 0.45$ and $\rho_4 = 0.37$ (which are the SSS sample estimates and differ from the values in Table 2) resulted in $\lambda_0 = 0.00484$, $\lambda_1 = 24.87$; $\lambda_2 = 1$, $\kappa_1 = 46.80$, $\kappa_2 = 0.179$, and $\beta = 5.25$. This solution has the properties of an ASS process as discussed in the text.

Appendix 2 Entropy maximization to determine unknown covariances in the stochastic model structure

Let $\mathbf{Z}^+ = [W, \mathbf{Z}^T]^T$ the vector that contains all variables of the problem, assumed to be standardized. Its covariance matrix is

$$\mathbf{c} := \text{Cov}\,[\mathbf{Z}^+, \mathbf{Z}^+] = \begin{bmatrix} 1 & \boldsymbol{\eta}^T \\ \boldsymbol{\eta} & \mathbf{h} \end{bmatrix} \tag{24}$$

From (23), the joint entropy of $\mathbf{Z}^+$ is $\varphi^+ = \ln(\sqrt{(2\pi e)^{s+1}\,|\mathbf{c}|})$ where $|\mathbf{c}|$ is the determinant of $\mathbf{c}$. It is reasonable to assume that the unknown items of $\mathbf{h}$ are those that maximize φ^+, or equivalently $|\mathbf{c}|$. Any different consideration would imply that the information we have about $\mathbf{Z}^+$ is more than contained in the known elements of $\mathbf{c}$.

The maximization of φ^+ can be performed very easily. Since $\mathbf{c}$ is a symmetric matrix, it can be written as $\mathbf{c} = \mathbf{b}\,\mathbf{b}^T$ where $\mathbf{b}$ is a lower triangular matrix, known as the square root of $\mathbf{c}$. Then,

$$|\mathbf{c}| = \prod_{i=1}^{s+1} b_{ii}^2 \tag{25}$$

so that maximization of $|\mathbf{c}|$ is equivalent to maximization of the product of the diagonal elements of $\mathbf{b}$. It is reminded that the elements of $\mathbf{b}$ are calculated from $\mathbf{c}$ with a step-by-step algorithm (Cholesky decomposition, e.g., Bras and Rodriguez-Iturbe, 1985, p. 96) for

growing i and then j, so that,

$$b_{ij} = \frac{c_{ij} - \sum_{l=1}^{j-1} b_{il}\, b_{jl}}{b_{jj}} \text{ if } j < i;\ \ b_{ii} = \sqrt{c_{ii} - \sum_{j=1}^{i-1} b_{ij}^2} \text{ if } j = i;\ \ b_{ij} = 0 \text{ if } j > i \tag{26}$$

Clearly then, maximization of the diagonal element b_{ii} demands that all "free" b_{ij} (which correspond to an unknown c_{ij}) should be zero. Thus, only a small modification of the algorithm is needed for the non-diagonal elements: if c_{ij} is unknown then $b_{ij} = 0$, else b_{ij} is calculated from (26). Furthermore, one may observe that since Var[V] is related to the entropy of $\mathbf{Z}^+$ conditional on known $\mathbf{Z}$, it can be (alternatively to (14)) estimated from (Papoulis, 1991, pp. 500, 568):

$$\text{Var}[V] = |\mathbf{c}|/|\mathbf{h}| \tag{27}$$

Tables

Table 1 List of all models of the study.

Model type	Model specifications	Model abbreviation
Stochastic	Cyclostationary with short- and long-range dependence, using normalizing transformation of time series	S1
	As S1 but without normalizing transformation	S2
	PAR(2) without normalizing transformation	S3
Analogue (local linear)	Constant delay, 12 consecutive delay items, 11 neighbours	A1
	Constant delay, 13 consecutive delay items, 24 neighbours	A2
	Variable delay, 4 delay items, 7 neighbours	A3
Connectionist (artificial neural network)	Constant delay, 5 inputs, 2 hidden layers, 2+2 hidden nodes	C1
	Constant delay, 14 inputs, 2 hidden layers, 11+11 hidden nodes	C2
	Variable delay, 4 inputs, 2 hidden layers, 4+2 hidden nodes	C3

Table 2 Main marginal and dependence statistics of the untransformed 78-year record on monthly and annual basis.

Month	μ (km^3)	σ (km^3)	C_s	C_k	τ_3	τ_4	H	ρ_1	ρ_2	ρ_{12}
Aug	19.37	4.62	-0.09	-0.14	0.00	0.12	0.76	0.71	0.26	0.16
Sep	22.98	4.29	-0.12	-0.57	-0.02	0.07	0.74	0.80	0.51	0.17
Oct	16.33	3.65	0.41	0.31	0.08	0.14	0.76	0.88	0.70	0.24
Nov	8.79	2.34	0.42	-0.27	0.09	0.11	0.80	0.90	0.77	0.26
Dec	5.92	1.60	0.86	0.60	0.19	0.13	0.89	0.94	0.85	0.42
Jan	4.37	1.20	0.64	0.31	0.15	0.15	0.88	0.98	0.91	0.44
Feb	3.02	1.00	0.85	0.27	0.20	0.12	0.82	0.96	0.92	0.35
Mar	2.51	0.96	1.25	1.34	0.26	0.17	0.78	0.91	0.84	0.31
Apr	1.89	0.75	1.75	3.56	0.33	0.19	0.78	0.94	0.78	0.33
May	1.68	0.63	2.13	6.30	0.33	0.23	0.72	0.93	0.85	0.30
Jun	1.91	0.68	1.89	6.00	0.27	0.20	0.63	0.70	0.59	0.11
Jul	5.06	1.84	0.75	0.24	0.16	0.12	0.89	0.65	0.44	0.47
Average			0.90	1.50	0.17	0.14	0.79	0.86	0.70	0.30
Annual	93.85	20.16	0.35	-0.08	0.09	0.09	0.85	0.35	0.35	

Notation. μ: mean; σ: standard deviation; C_s: standard coefficient of skewness; C_k: standard coefficient of kurtosis, τ_3: L-coefficient of skewness; τ_4: L-coefficient of kurtosis, H: Hurst coefficient; ρ_1 and ρ_2: autocorrelation coefficients for lags 1 and 2 (for the monthly series they are autocorrelations of the current month with one or two months before; for the annual series they are autocorrelations of the current year with one or two years before); ρ_{12} (for the monthly series): autocorrelation of current month to the same month one year before.

Table 3 Main marginal and dependence statistics of the transformed 78-year record on monthly basis.

Month	μ (km^3)	σ (km^3)	C_s	C_k	τ_3	τ_4	H	ρ_1	ρ_2	ρ_{12}
Aug	19.37	4.623	-0.09	-0.14	0.00	0.12	0.76	0.71	0.25	0.16
Sep	22.98	4.292	-0.12	-0.57	-0.02	0.07	0.74	0.80	0.52	0.17
Oct	16.33	3.649	0.41	0.31	0.08	0.14	0.76	0.88	0.70	0.24
Nov	1.42	0.057	-0.30	-0.16	-0.05	0.11	0.76	0.89	0.77	0.23
Dec	1.34	0.058	0.05	0.07	0.05	0.11	0.88	0.95	0.87	0.40
Jan	1.27	0.065	-0.48	1.26	-0.02	0.17	0.83	0.98	0.92	0.36
Feb	1.18	0.081	-0.13	0.18	0.02	0.11	0.77	0.96	0.93	0.30
Mar	1.13	0.090	0.25	-0.19	0.07	0.12	0.80	0.89	0.81	0.36
Apr	1.05	0.093	0.60	0.08	0.13	0.12	0.84	0.95	0.77	0.44
May	1.02	0.087	0.67	0.72	0.13	0.14	0.77	0.93	0.86	0.36
Jun	1.05	0.086	0.32	0.49	0.07	0.15	0.64	0.69	0.57	0.11
Jul	1.30	0.084	-0.31	0.13	-0.04	0.11	0.90	0.64	0.45	0.50
Average			0.07	0.18	0.03	0.12	0.79	0.86	0.70	0.30

Notation as in Table 2.

Table 4. Coefficient of efficiency of the optimal configurations of the analogue model in the verification period; all coefficients refer to natural (not standardized) series.

Model	Untransformed values	Logarithmically transformed values
A1	0.959	0.945
A2	0.945	0.942
A3	0.955	0.954

Table 5 Performance indices of the optimal configurations of the connectionist model for the calibration and verification periods; all coefficients refer to standardized series.

Model	Mean square error		Coefficient of efficiency	
	Calibration	Verification	Calibration	Verification
C1	0.289	0.241	0.749	0.435
C2	0.183	0.309	0.842	0.277
C3	0.241	0.240	0.794	0.438

Table 6 Coefficients of efficiency of the different prediction models for the validation period (53 years).

Model	Untransformed values	Logarithmically transformed values	Monthly standardized untransformed values
S1	0.911	0.904	0.673
S2	0.907	0.899	0.675
S3	0.884	0.884	0.624
A1	0.840	0.613	-0.145
A2	0.847	0.623	-0.126
A3	0.879	0.851	0.490
C1	0.888	0.878	0.583
C2	0.775	0.791	0.280
C3	0.859	0.849	0.472

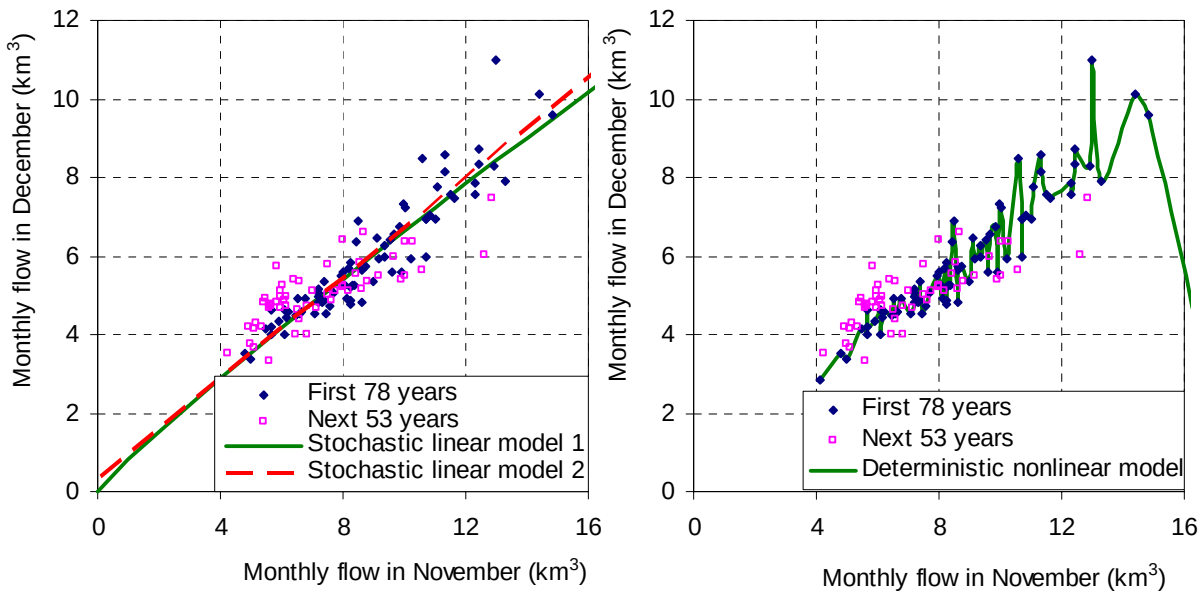

Figure 1 Graphical depiction of the basic concept of (left) a stochastic and (right) a deterministic approach in the hypothetical case of a model with a single time delay component. Stochastic linear models 1 and 2 are models assuming linear dependence of the normalized and natural flows, respectively. Deterministic linear model is an arbitrary hypothetical non-intersecting curve passing from all 78 points.

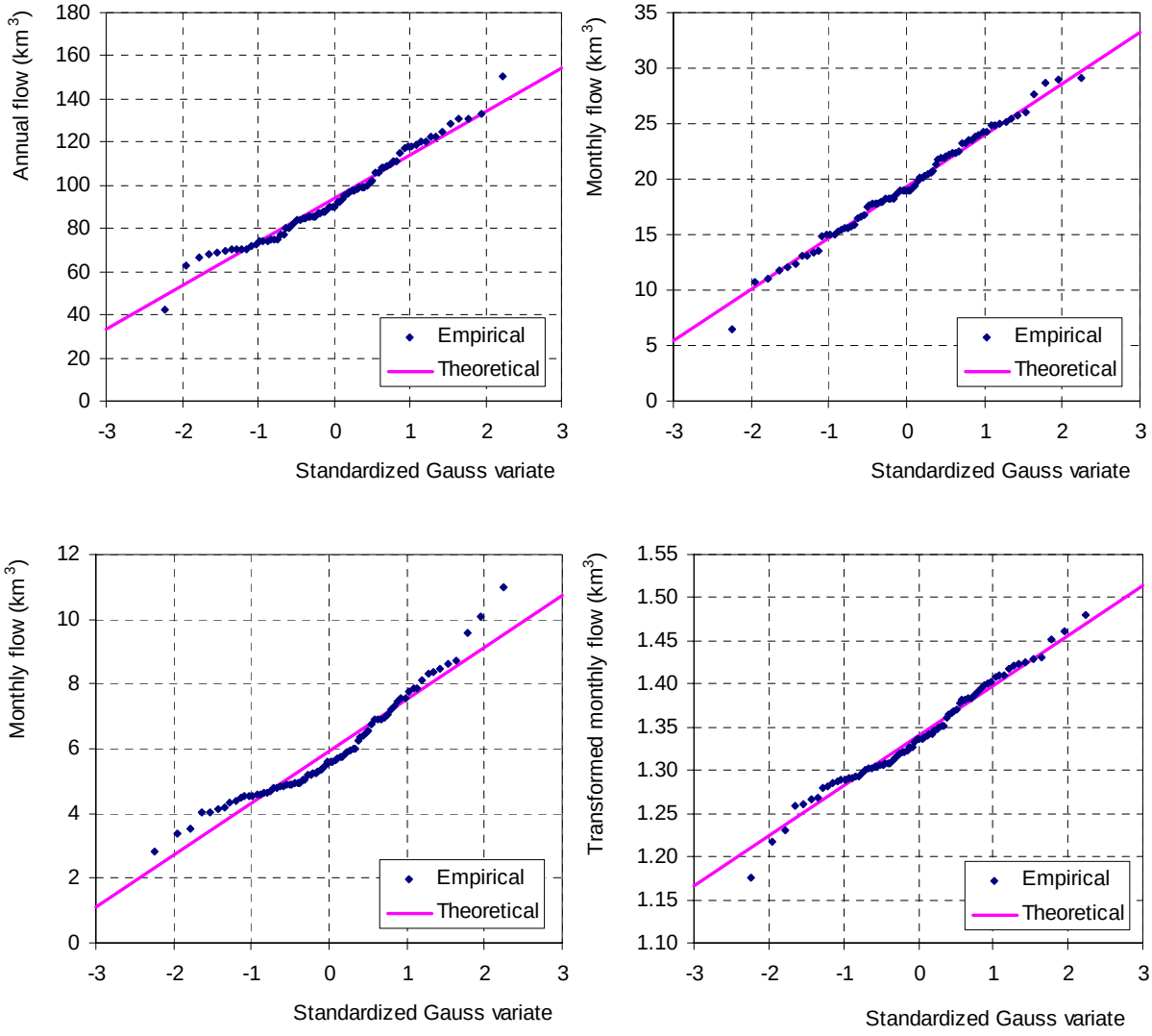

Figure 2. Normal probability plots of the distribution function (empirical and normal) of annual and monthly Nile flows: (up left) annual; (up right) August; (down left); December; (down right) December but after applying normalizing transformation (9).

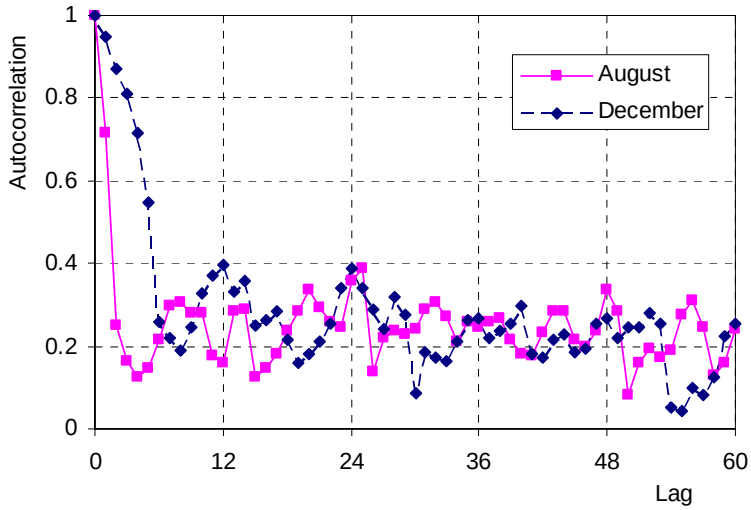

Figure 3. Correlation coefficients of the transformed (by (9)) monthly flows, i.e. Corr[Z_i, Z_{i+j}] for i = 1 (August) and 5 (December) and lag j up to 60.

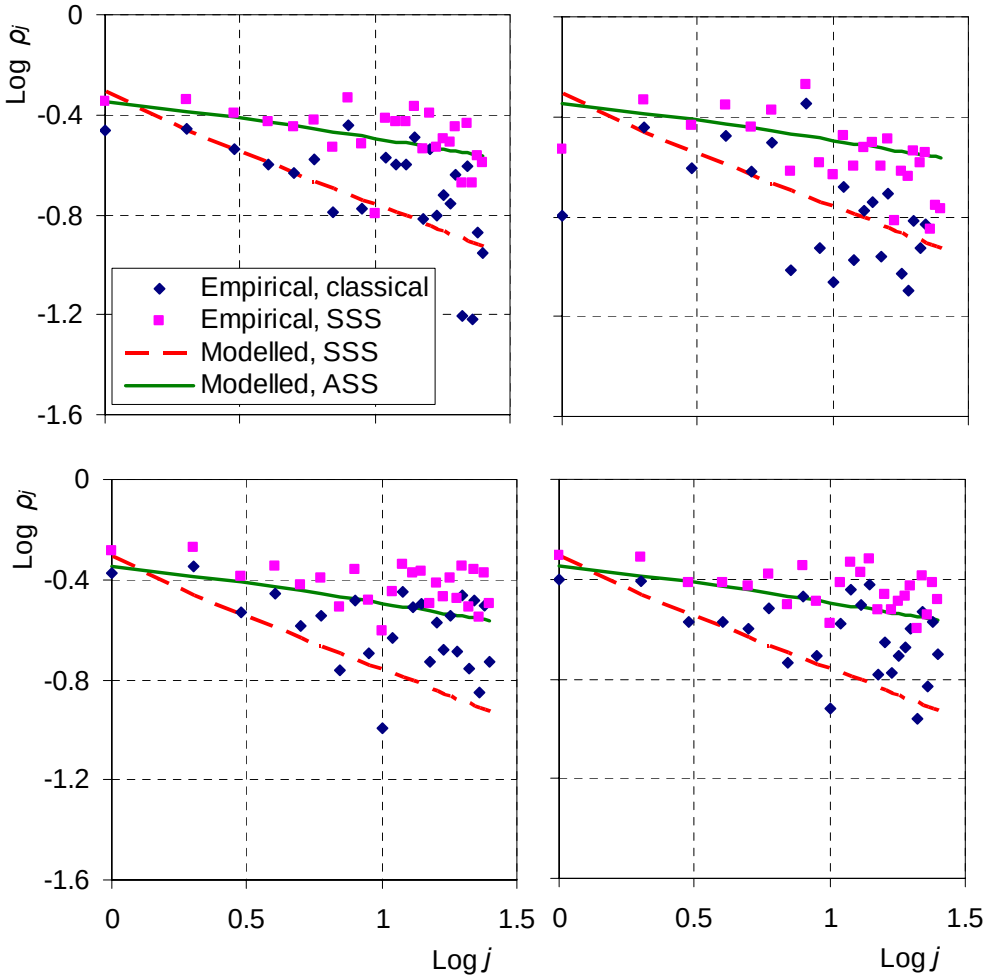

Figure 4 Logarithmic plots of autocorrelations of annual flows (Corr[Y_i, Y_{i+j}]) and monthly flows (Corr[X_i, X_{i+12j}]) versus j: (up left) annual; (up right) August; (down left) December, untransformed time series; (down right) December, transformed time series.

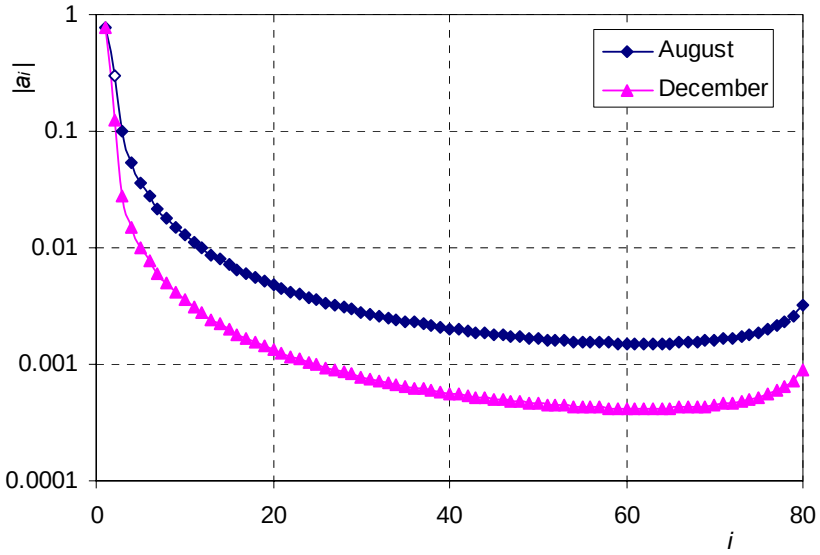

Figure 5 Graphical depiction of the vector of weights **a** for the indicated months; full points indicate positive values and the single empty point indicates a negative value.

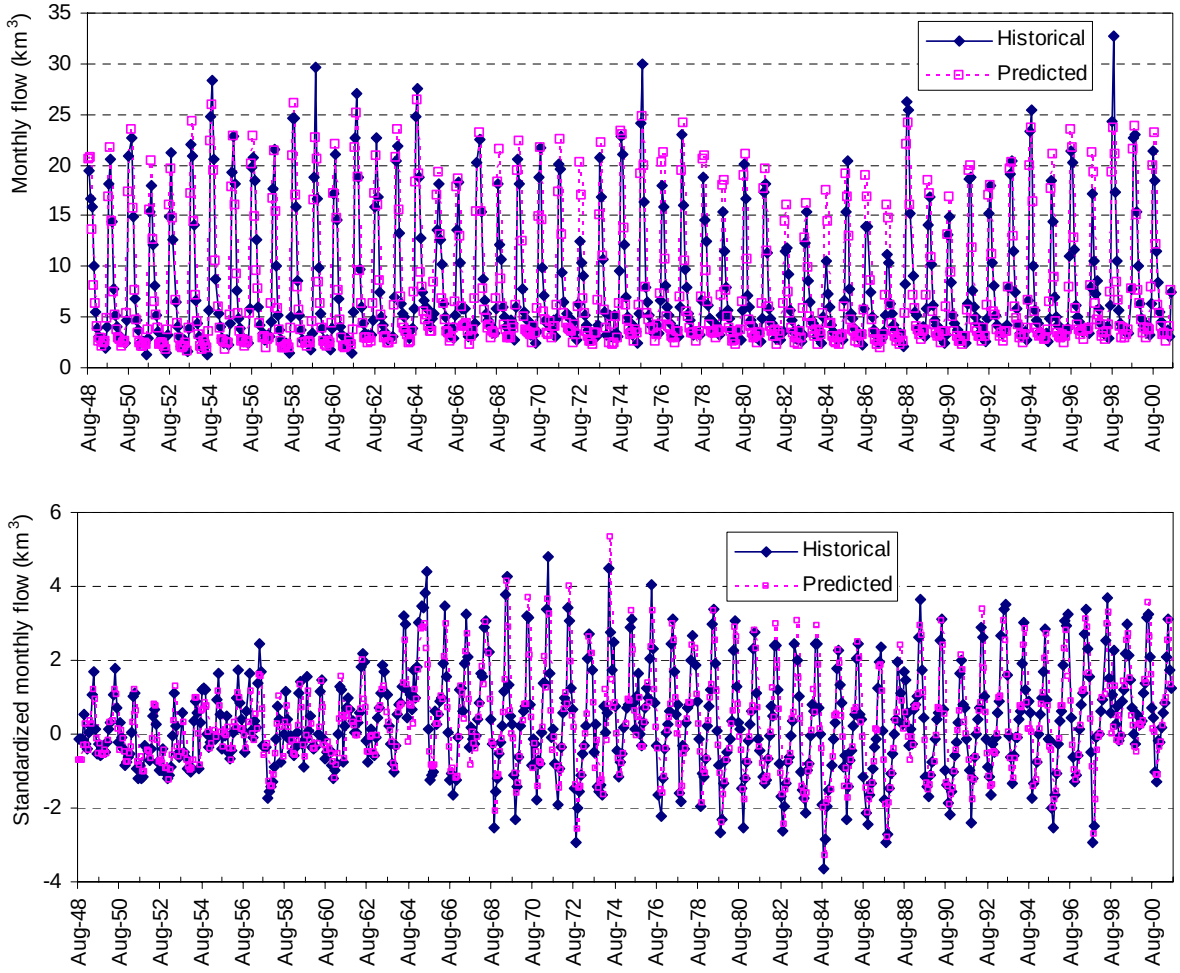

Figure 6 Graphical depiction of the proximity of monthly predictions of model S1 to historical values; (up) natural values; (down) monthly standardized values.

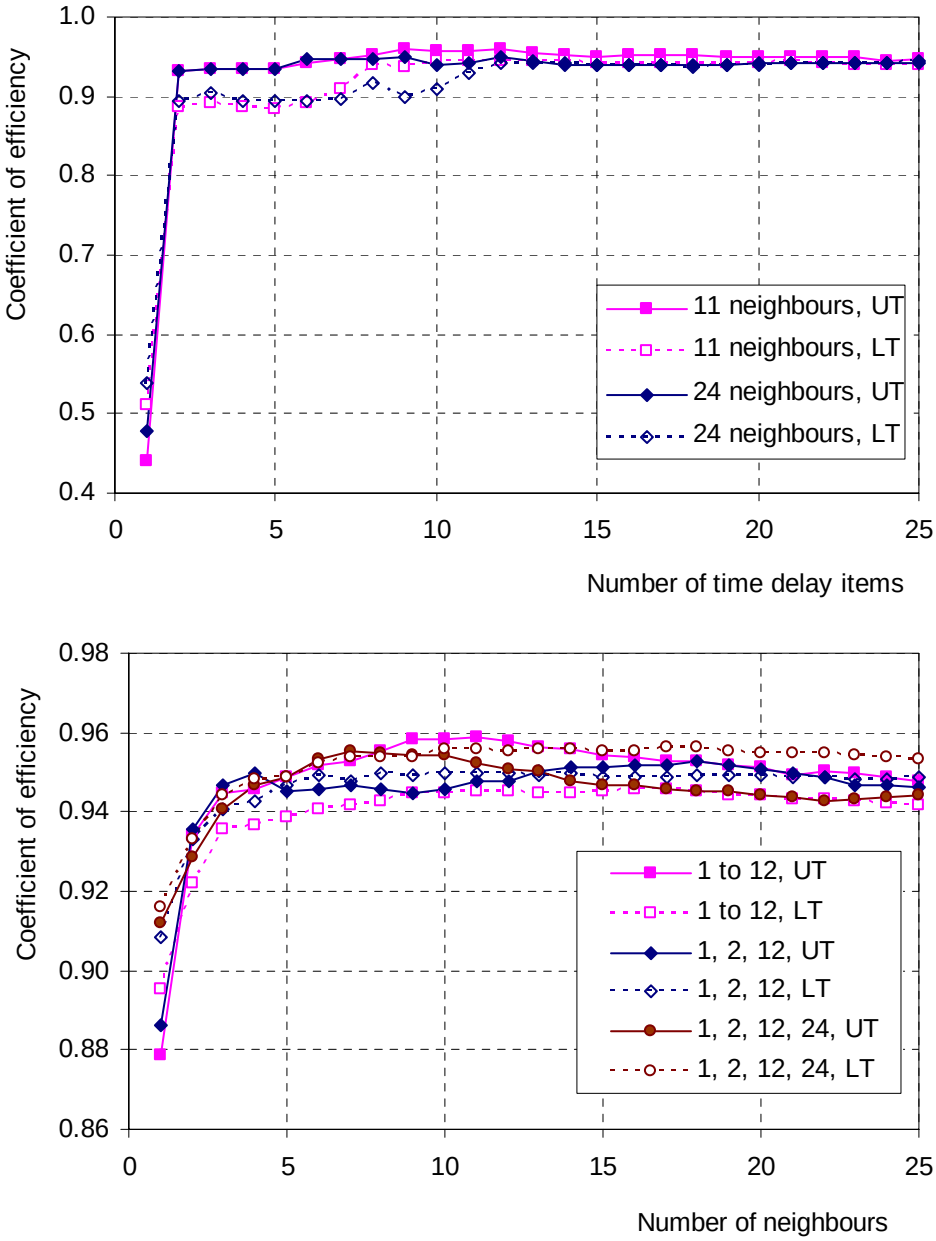

Figure 7 Coefficient of efficiency attained by the analogue model for the verification period, as a function of (up) number of delay items assuming fixed (11 or 24) number of neighbours, and (down) the number of neighbours assuming the indicated delay items; UT untransformed (natural) values; LT logarithmically transformed values.

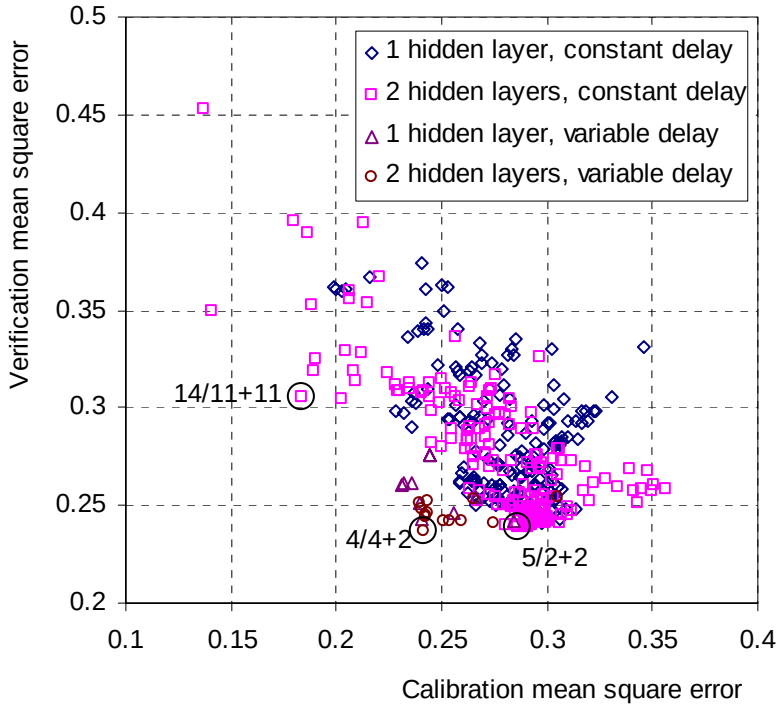

Figure 8. Plot of the attained verification error vs. the attained calibration error of a series configurations of the connectionist model with 1 to 15 input nodes, 1 or 2 hidden layers, and 1 to 15, hidden nodes in each layer. The circled points in the Pareto front, for which the number of input and hidden nodes are marked, depict the solutions further explored (from left to right: C2, C3, C1).

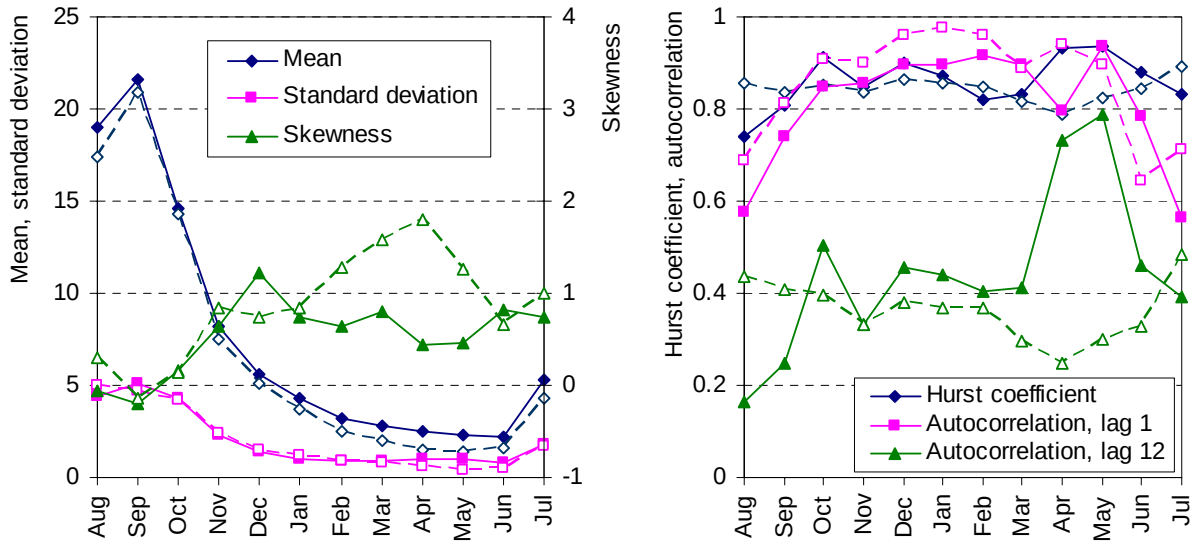

Figure 9. Comparison of statistics of the historic monthly record and a synthetic record of equal length generated by model S1.

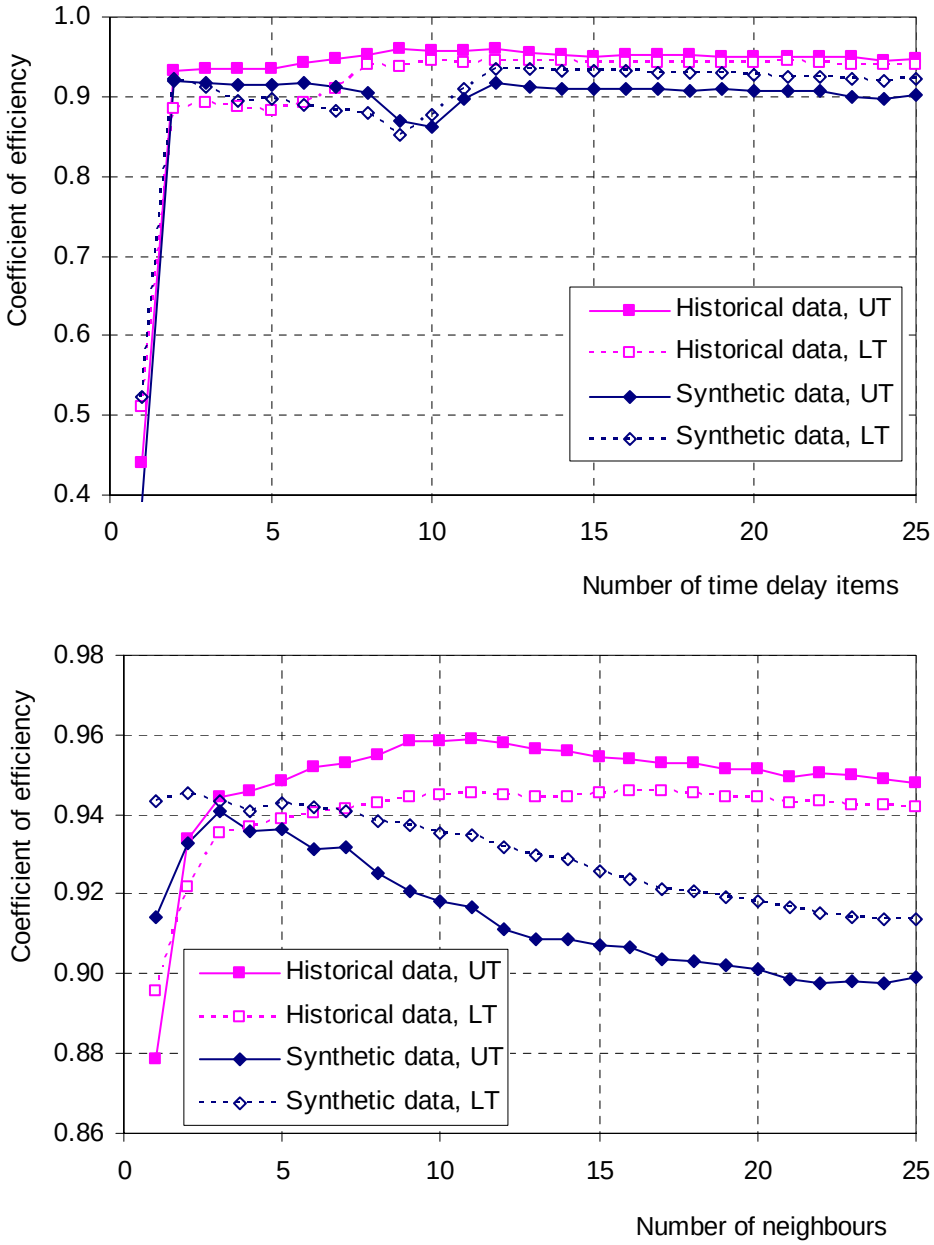

Figure 10 Coefficient of efficiency attained by the analogue model for the verification period for synthetic data generated by model S1 in comparison to the respective values for the historic data, as a function of (up) number of delay items assuming fixed (11) number of neighbours, and (down) the number of neighbours assuming fixed (12) delay items; UT untransformed (natural) values; LT logarithmically transformed values.