Identification-based Diagnosis of Rainfall – Stream Flow Data: the Tinderry Catchment

Karel J. Keesman[1,2] , Peter C. Young[2,3] and Anthony J. Jakeman[2]

[1] Systems and Control Group, Wageningen University, Bornse Weilanden 9, 6708 WG Wageningen, The Netherlands
[2] Fenner School of Environment and Society and National Centre for Groundwater Research and Training, The Australian National University, Canberra ACT 0200
[3] Lancaster Environment Centre, Lancaster University, Lancaster, LA1 5BL

Abstract System identification tools, such as transfer function (TF) model structure identification, recursive estimation, time-varying parameter (TVP) estimation and assessment of data information, are used to evaluate the quality of rainfall-stream flow data from the Tinderry catchment (ACT, Australia) and the time-varying behaviour of the rainfall-stream flow dynamics. For the catchment, given the wide range and the abrupt changes of the single input-single output transfer functions describing different periods or events, we conclude that further investigation of (i) local rainfall effects, (ii) time-varying time delays (travelling time), (iii) time-varying residence times related to the base flow and (iv) occurrence of negative residues is needed. Periods with high and low data information content, for further use in effective parameter estimation procedures, are clearly indicated by the analysis.

Keywords Diagnosis, rainfall-stream flow data, system identification, recursive parameter estimation

INTRODUCTION

Usually relatively long data records of rainfall and stream flow are available: e.g. 20 years of daily measurements from rainfall and stream flow gauges. This even holds for remote catchments due to the wide availability of wireless communication techniques. However, these data are not always of good quality and the question arises: “what can be said about the quality of these long data records?”. Instead of statistical, model-free data analysis, we analyse the data using (dynamic) transfer function (TF) model identification. In particular, we introduce system identification tools, such as structure/order identification (model selection), recursive and time-varying parameter estimation and the assessment of data information content (Young, 1984, Norton, 1986; Keesman, 2011), to diagnose the data quality. These tools allow us not only to investigate the data quality, but also to interpret the results in a hydrological context. The goal of the paper is to show how such diagnostic tools can be used in the case of the Tinderry catchment (near Canberra, ACT, Australia).

BACKGROUND

Dynamic, linear, time-invariant (LTI) systems can be represented by a transfer function (TF) model. In short hand notation, a noise-free TF model, relating the input u_k, at the discrete time instant t_k, to the resulting output y_k , with k the time index, is presented as

$$y_k = \frac{B(q^{-1})}{A(q^{-1})} u_{k-n_k} \tag{1}$$

where y_k is the stream flow, u_k the effective (excess) rainfall, which is usually calculated from a non-linear module that accounts for evapotranspiration and soil moisture, and n_k is a pure time delay of n_k samples, introduced to allow for any delay that occurs between the occurrence of rainfall and its first effect on flow. The polynomials $A(q^{-1})$ and $B(q^{-1})$ are given by

$$\begin{aligned} A(q^{-1}) &= 1 + a_1 q^{-1} + a_2 q^{-2} + \ldots + a_{n_a} q^{n_a} \\ B(q^{-1}) &= b_0 + b_1 q^{-1} + b_2 q^{-2} + \ldots + b_{n_b} q^{n_b} \end{aligned} \tag{2}$$

where $A(q^{-1})$ is chosen to be monic to avoid structural identifiability problems. In these polynomials, q^{-1} is the backward-shift operator (sometimes denoted by z^{-1}). Consequently, the input-output relationship can also be written in terms of a difference equation, that is

$$y_k = -a_1 y_{k-1} - a_2 y_{k-2} - \ldots - a_{n_a} y_{k-n_a} + b_0 u_{k-n_k} + \ldots + b_{n_b} u_{k-n_k-n_b} \qquad (3)$$

Both of these model representations will be used in the analysis of the data. For a further interpretation of the black-box TF model, a so-called partial fraction expansion (see e.g., Jakeman et al, 1990, Goodwin et al., 2001) can be applied. This expansion allows for a decomposition of the TF model in terms of a series of first-order TF models, normally arranged in parallel (see e.g. Young and Wallis, 1985; Young, 1992); plus, in some cases, a constant gain. In addition to this analysis of the TF model, we exploit an eigenvalue decomposition of a square matrix and its interpretation. Consider the $n \times n$ - dimensional matrix Σ. Then, the eigendecomposition of Σ is given by $\Sigma V = V\Lambda$, where V is the eigenmatrix, containing n eigenvectors, and Λ is a diagonal matrix with n eigenvalues λ_i associated with the eigenvectors v_i for i = 1, ..., N. For each eigenvalue, in general a complex number, the following holds: $\Sigma v_i = \lambda v_i$. If Σ is symmetric and positive-definite, then all eigenvalues are positive real numbers and V is an orthogonal matrix (see Horn and Johnson, 1985).

METHODS

Identification of system TF models

It has been recognized (Jakeman et al, 1990) that the input-output relationship between rainfall and stream flow contains nonlinearities due, for instance, to the wetness of the catchment. However, as will be shown later, this non-linearity can also be directly obtained from the data using the Data-Based Mechanistic (DBM) modelling approach (see e.g. Young, 1993, Young and Beven, 1994; Young, 1998, 2003). The intermediate variable u_k (see Eq. 1), is called the effective rainfall. The key question is how to identify the model structure of these TF models, as defined by n_a, n_b, and the time delay n_k (see Eq. 1-2). Several criteria for model identification have been proposed in the past. However, all these criteria depend on the data given and include a term that is related to the residuals or prediction errors. Hence, there is a need to estimate the unknown parameters in the discrete-time transfer function and provide an estimate of the residuals.

Since the data are usually corrupted with noise, we need to introduce an error term into our general TF model (1). In what follows, we consider the process dynamics and effective rainfall data to be free of errors and introduce the (noise free) auxiliary output variable x_k, such that

$$x_k = \frac{B(q^{-1})}{A(q^{-1})} u_{k-n_k} \qquad (4)$$

together with the output equation,

$$y_k = x_k + e_k \qquad (5)$$

which then relates the measured output to the estimated noise-free model output and the output error e_k. For analytical convenience, we presume that $\{e_k\}$ is zero-mean white noise, although this is rarely the case in practice. We have chosen this so-called *output-error model* structure, Eq. 4-5, because we are mostly interested in the mechanisms of the process and not so much in the one-step ahead predictions (see e.g. Young, 1984; Norton, 1986; Ljung, 1999; Keesman, 2011 for details on this). Substituting (5) into (4) gives

$$A(q^{-1})(y_k - e_k) = B(q^{-1}) u_{k-n_k} \qquad (6)$$

and thus

$$A(q^{-1}) y_k = B(q^{-1}) u_{k-n_k} + v_k \qquad (7)$$

with $v_k := A(q^{-1})\, e_k$. Consequently, the error term v_k in (7) is autocorrelated (coloured) and thus direct least-squares estimation does not automatically lead to unbiased estimates, as would be the case if the error term $\{v_k\}$ was zero-mean white noise. Thus, there is a need for pre-whitening, using Instrumental Variable (IV) estimation, or other methods that take into account the coloured error structure. In addition to this, in order to discover irregularities in the data, recursive estimation schemes (see e.g. Young, 1984) can be used for updating the parameter estimates at the current time instant, given estimates from a preceding time instant and current measurements.

Assessment of data information

For illustrative purposes we will neglect, for the moment, the coloured error structure and presume that some effective pre-whitening has taken place, changing (y_k, u_k) into (y'_k, u'_k) and $v_k = e'_k$. Then, Eq. 7 can be written as

$$\begin{aligned} y'_k &= -a_1 y'_{k-1} - a_2 y'_{k-2} - \ldots - a_{n_a} y'_{k-n_a} + b_0 u'_{k-n_k} + \ldots + b_{nb} u'_{k-n_k-nb} + e'_k \\ &= \begin{bmatrix} -y'_{k-1} & \ldots & u'_{k-n_k-n_b} \end{bmatrix} \begin{bmatrix} a_1 \\ \vdots \\ b_{nb} \end{bmatrix} + e'_k \\ &= \phi_k^T \vartheta + e'_k \end{aligned} \tag{8}$$

where e'_k originates from a zero-mean white noise process. In what follows, and for ease of notation, the prime is dropped. The vector $\phi_k^T = \begin{bmatrix} -y_{k-1} & -y_{k-2} & \ldots & -a_{n_a} y_{k-n_a} & u_{k-n_k} & \ldots & u_{k-n_k-n_b} \end{bmatrix}$ is the regression vector and $\vartheta = \begin{bmatrix} a_1 & a_2 & \ldots & a_{n_a} & b_0 & \ldots & b_{n_b} \end{bmatrix}^T$ is the parameter vector. In vector-matrix form, using N data points, this results in

$$y = \Phi \vartheta + e \tag{9}$$

where $y = \begin{bmatrix} y_1 & y_2 & \ldots & y_N \end{bmatrix}^T$, $\Phi = \begin{bmatrix} \phi_1 & \phi_2 & \ldots & \phi_N \end{bmatrix}^T$ and $e = \begin{bmatrix} e_1 & e_2 & \ldots & e_N \end{bmatrix}^T$. Consequently, the ordinary least-squares estimate is optimal with estimation error covariance matrix given by

$$\Sigma_{\hat{\vartheta}} = \hat{\sigma}_{\hat{e}}^2 \left(\Phi^T \Phi \right)^{-1} \tag{10}$$

for Φ deterministic. The estimated residuals are denoted here by $\hat{e}_k$. Hence, the information matrix $\left(\Phi^T \Phi \right)$ plays a key role in the assessment of the estimation errors, since the estimated variance of the residuals ($\hat{\sigma}_{\hat{e}}^2$) is only a scaling factor. Given the information matrix and variance of the residuals, the uncertainty ellipsoid related to the estimated parameters is given by

$$\mathrm{E}(\vartheta) = \left\{ \vartheta : (\vartheta - \hat{\vartheta})^T \Sigma_{\hat{\vartheta}}^{-1} (\vartheta - \hat{\vartheta}) = 1 \right\} \tag{11}$$

with the main axes determined by the eigenvectors of the positive-definite, symmetric matrix $\Sigma_{\hat{\vartheta}}$ and with semi-axis length equal to the square root of the inverse eigenvalues (see e.g. Bard, 1974). Hence, the information content of the data is assessed by these eigenvalues and eigenvectors, which can also be geometrically interpreted in terms of an uncertainty ellipsoid.

Procedure

Given the tools from the previous section, we propose the following procedure for the dynamic diagnosis of the data:

Step 1: Identify (i) the *non-linear* relationship between rainfall and effective rainfall, using e.g. IHACRES (Jakeman et al, 1990; Jakeman and Hornberger, 1993), or the DBM modelling approach (see e.g. Young, 1993, 2003; Young and Beven, 1994); and (ii) the *linear* relationship between effective rainfall and stream flow. Usually, the first, nonlinear relationship can be expressed in terms of temperature-dependent soil moisture dynamics; or, more simply in the DBM case, estimated by non-parametric, recursive state-dependent parameter (SDP) estimation (Young 2000). However, at a final stage in the model estimation, both the non-linear and linear relationships have to be estimated concurrently. Within the DBM modelling approach, the SDP identified nonlinearity is parameterized in a simple manner (usually by a power law in y_k: see later) and then the complete model is estimated using an optimization procedure that incorporates recursive Refined Instrumental Variable (RIV) estimation.

Step 2: Given the estimated TF model from step 1, focus on statistically significant changes in either the gain or the full dynamic behaviour of the linear TF model (i.e. from effective rainfall to stream flow) over the whole observation interval. Effectively, this exploits either a recursive *time-varying* parameter (TVP) estimation scheme, or repeated constant parameter estimation over a fixed or variable data window (rectangular-weighted-past (RWP) estimation: see Young, 1984), in order to estimate changes in the parameter estimates (see, for instance, Young (1984, 2000) for an overview of TVP estimation). This can be as simple as estimating the time varying gain of the TF model using recursive least squares, which provides a very quick indication of the temporal changes required to make the constant parameter model explain the data better, or as an RIV or RWP estimation of all the TF model parameters. For a more complete hydrological interpretation of these results, however, it is necessary to evaluate not only the time-varying parameter estimates, but also the poles and residues of the partial fraction expansion of the TF, as these are related to the residence times and gains of the inferred parallel flow paths (see previously cited references). In addition to this TVP analysis, a ‘movie’ of the changing dynamics can be generated from the impulse response (unit hydrograph) estimated for each window, so providing a clear visual illustration of how the system dynamics may have changed over the whole period.

Step 3: Given the non-linear and linear sub-models, in step 3 analyse the data information content for each data window in an RWP approach. Thus, for each window, calculate ($\Phi^T\Phi$) with Φ the data matrix and subsequently perform an eigendecomposition. Plot the eigenvalues and the elements of the eigenvector, which can be interpreted as weights for each individual parameter. The eigenvalues will reveal how the data content will change for each window. Windows with small eigenvalues indicate intervals on which the parameters may be difficult to identify and which thus may be neglected when estimating the parameters from the data.

RESULTS AND DISCUSSION

System TF modelling

Given the rainfall – stream flow data (Fig. 1), our first step was to estimate the parameters of the simple, first order output-error model without time delay

$$\begin{aligned} x_k &= \frac{b_0}{1+a_1q^{-1}}u_k \\ y_k &= x_k + e_k \end{aligned} \qquad (12)$$

using the recursive SDP estimation algorithm, which is available in the CAPTAIN Toolbox for Matlab[1]. This non-parametric (graphical) estimation identifies that the TF numerator parameter (b_0) is a function of the measured flow, which is acting as a surrogate measure of the changing soil moisture. As in previous DBM modelling, this simple nonlinearity can be parameterized quite well by the following power law in the measured flow, i.e.,

$$\hat{b}_{0,k} \approx b'_0(y_k)^{0.4} \qquad (13)$$

where b'_0 is the new ‘constant’ input parameter, so that $b_0\,u_k$ becomes $b'_0\,u_k\,(y_k)^{0.4}$. Notice, however, that after this substitution the model cannot be used in a predictive sense, which is not a real limitation as our focus is on diagnostics. Hence, in what follows, we will use the "effective rainfall": $u_k\,y_k^{0.4}$, under the constraint that $u_k\,y_k^{0.4} < u_k$, instead of the measured rainfall (u_k) to

[1] A fully functional trial version of the CAPTAIN Toolbox can be downloaded from http://www.es.lancs.ac.uk/cres/captain/

identify the linear (state-dependent) sub-model. Given this newly defined input, the RIV algorithm in CAPTAIN identifies an output-error model with $n_a = 2$, $n_b = 3$ and $n_k = 0$, in short-hand notation OE(2,3,0), which explains the flow data quite well. Hence, the resulting TF model is given by

$$y_k = \frac{b_0 + b_1 q^{-1} + b_2 q^{-2}}{1 + a_1 q^{-1} + a_2 q^{-2}} \tilde{u}_k + v_k \tag{14}$$

where $\tilde{u}_k = u_k y_k^{0.4}$, the "effective rainfall" calculated from the raw rainfall - stream flow data, and $\{v_k\}$ is a (coloured) noise sequence. Most often, after a partial fraction expansion of Eq. 14, the model can be decomposed into two first-order systems in parallel, with time constants τ_q and τ_s, respectively; as well as a direct constant gain parameter. The time constants τ_q and τ_s are related to quick (run-off) flow and to slow (base) flow, while the direct term reflects the instantaneous effect of the effective rainfall on the output (in this example n_k=0).

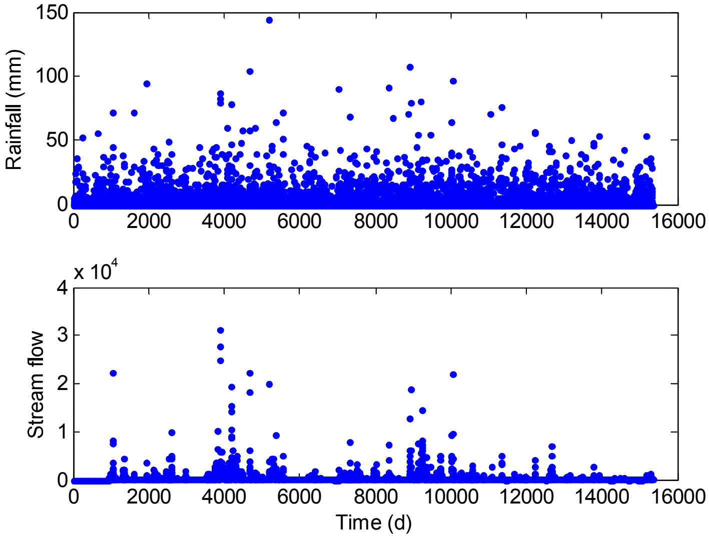

Fig. 1. Rainfall – stream flow data from the Tinderry catchment.

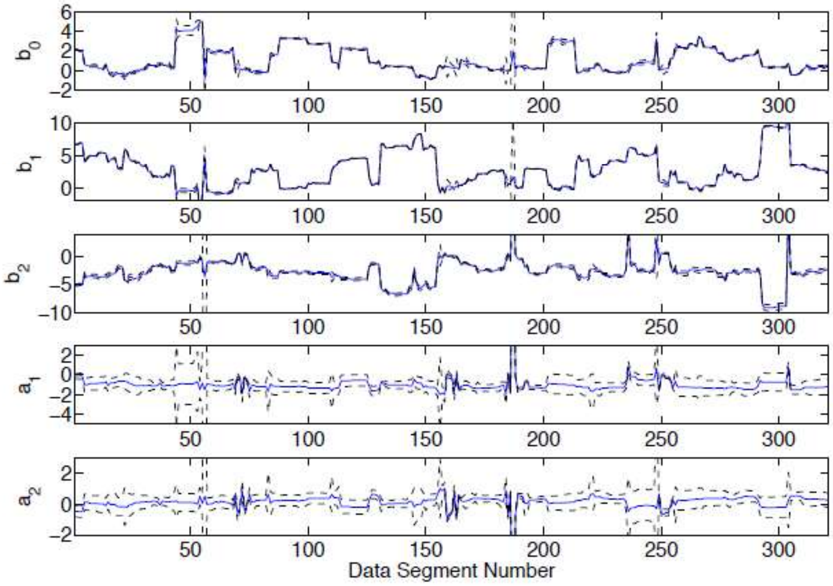

Fig. 2. Recursive estimates and error bounds of an OE(2,3,0) model (Eq. 15) with fixed window of 365 d and a shift of 30 d.

Analysing TF parameter trajectories

The trajectories of the recursive parameter estimates for the first 10,000 days showed considerable jumps in the first 10 years of data. In order to amplify these changes in the RWP estimation case, we either used a fixed window of 365 days that is shifted every 30 days, thus in total approximately 320 windows; or an event-based window. In the latter case, an event starts when $u_k > 20$ mm and the window length is at least 30 days. These results are not shown here, as they do not differ too much from the fixed window case, shown in Fig. 2.

Analysing data content

In the simplest recursive estimation analysis, the gain of the TF model is estimated recursively, as shown in Fig. 3 over a short segment of the data. Here, given the OE(2,3,0) model structure, the narrowing of the standard error bounds when rainfall and resultant high flows occur (e.g. around 3730 days), illustrates when there is high information in the data. In the more detailed RWP analysis, the five parameters in the OE(2,3,0) model structure have to be estimated. Hence, the minimum number of data points needed for a time-varying parameter estimation with fixed window is seven. Shifting this window each time by five days results in an overview of the main directions and associated semi-axis lengths, as presented in Fig. 4. Here, the left panels display the eigenvalues of the data information matrix for each window on a log-scale, where large eigenvalues imply high data information content. The right panels present the absolute values of the eigenvector elements, as these can be interpreted as weights on the individual parameter axes in the parameter space. Fig. 5 presents the rainfall-stream flow data for periods with high and low data information content around 3730 (5x746) days. As expected from the initial recursive estimation, high data information content occurs with high rainfall and stream flows: see the left panels of Fig. 5 and the low data information content in dry periods.

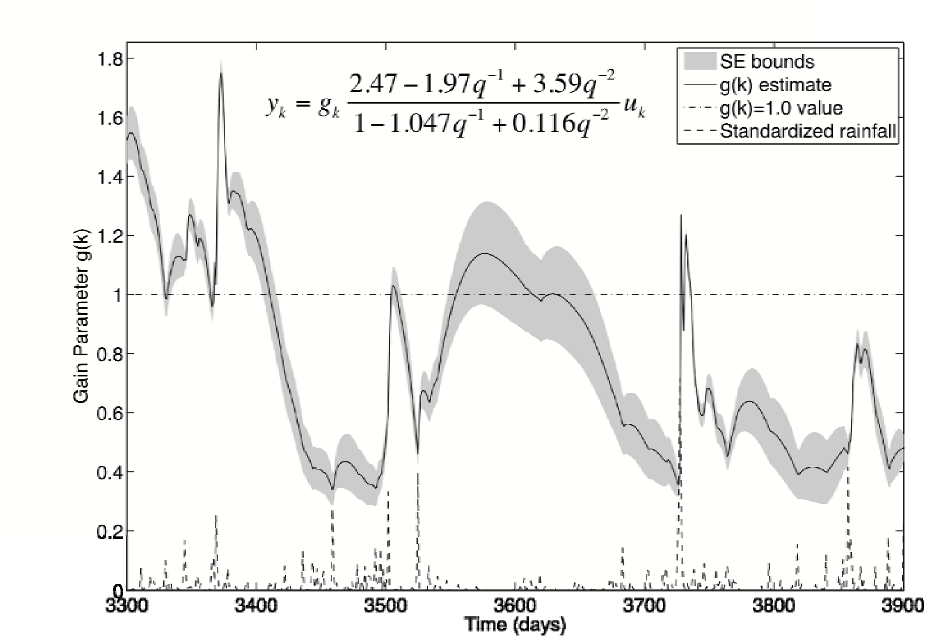

Fig. 3. Recursive estimate of the changing gain parameter showing the changes and associated standard error bounds, based on the effective rainfall series $u_k\, y_k^{0.4}$ (with u_k and y_k shown in Fig. 5).

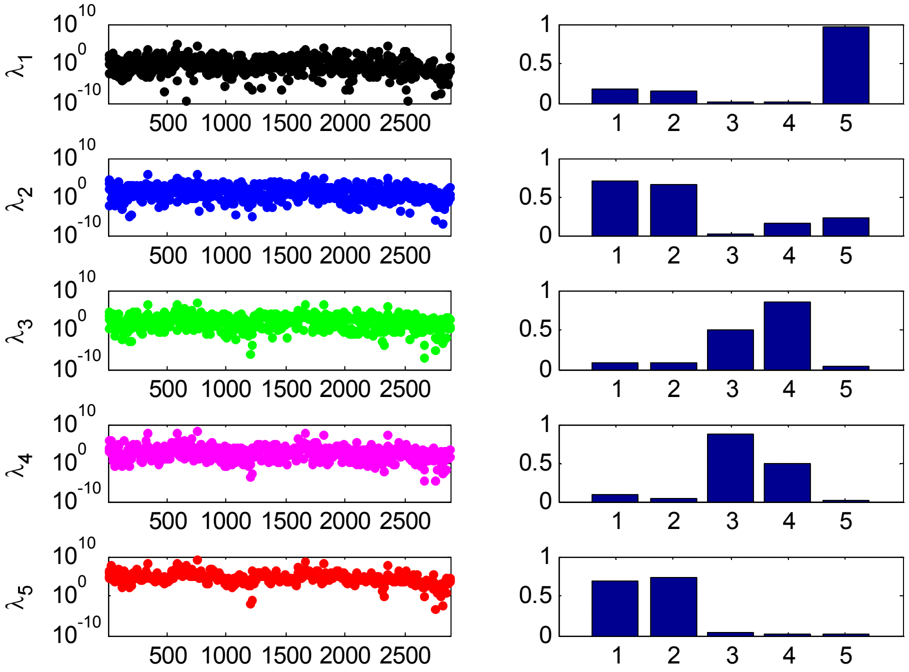

Fig. 4. Eigenvalues and eigenvector (only at the end) elements related to the time-varying data information matrix.

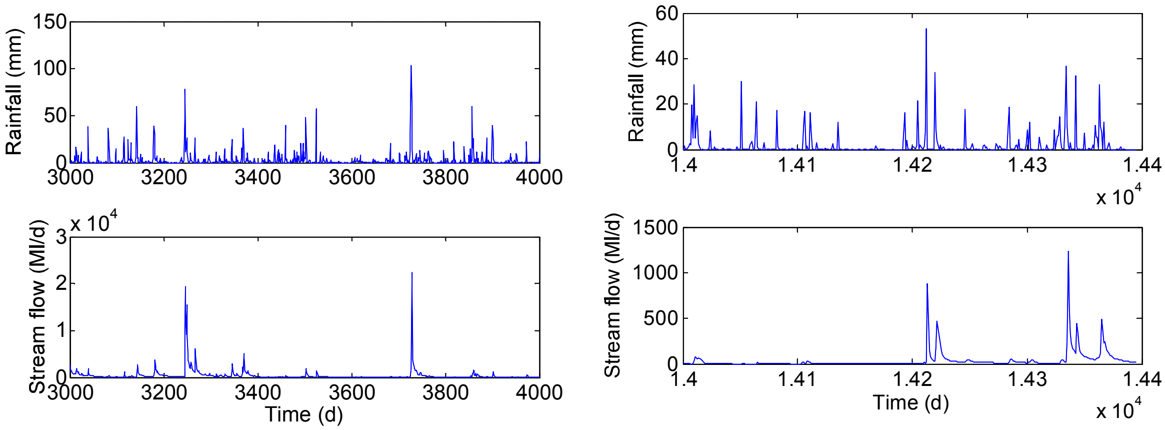

Fig. 5. Rainfall – stream flow data at ranges with maximum (left) and minimum (right) data information.

CONCLUDING REMARKS

1. Dynamic transfer function model-based diagnosis revealed periods/events in which significant parameter changes take place in the linear part of the state-dependent model.
2. These changes can be driven by the rainfall data, which needs a further investigation in terms of local effects; or by the resulting flow dynamics in the catchment.
3. For further understanding of the system, the problems of changing time delays (Fig. 2), i.e. changes in the estimates of b_0, b_1 and b_2, and significant changes in the residence time of the base

flow (not shown) need to be investigated. The problem of negative residues (implying physically implausible negative flow pathways), that sometimes appeared after the partial fraction expansion of the TF, needs further research, in particular with respect to hypotheses of potential cross-flows.
4. In addition to this, the choice of the window length and the effect of heteroscedasticity on the data information content are subjects for further research, too.
5. The data information content significantly changes over a period of 42 years. Hence, for accurate parameter estimation results, the estimation should be focussed on the identified periods with relatively high data information content.

References

Bard, Y. (1974) Nonlinear Parameter Estimation, Academic Press, New York.

Goodwin, G.C. and S.F. Graebe and M.E. Salgado (2001) *Control System Design*, Prentice Hall.

Horn, R.A. and C.R. Johnson (1985) *Matrix analysis*, Cambridge University Press.

Jakeman, A.J., Littlewood, I.G. and Whitehead, P.G. (1990) Computation of the instantaneous unit hydrograph and identifiable component flows with application to two small upland catchments. *Journal of Hydrology* **117** (1-4), 275-300.

Jakeman, A.J. and G.M. Hornberger (1993) How much complexity is warranted in a rainfall-runoff model? *Water Resources Research*, 29(8), 2637-2649.

Keesman, K.J. (2011) *System Identification: an Introduction*. Springer-Verlag, London.

Norton, J.P. (1986) *An Introduction to Identification*, Academic Press.

Young, P.C. (1984) *Recursive Estimation and Time-Series Analysis*. Springer-Verlag, Berlin (new revised and enlarged edition in preparation for publication in 2011).

Young. P.C. (1992) Parallel processes in hydrology and water: a unified time series approach. *Jnl. Inst. of Water and Environmental Management*, 6:598–612.

Young, P.C. (1993) Time variable and state dependent modelling of nonstationary and nonlinear time series. In T. Subba Rao, editor, *Developments in Time Series Analysis*, pages 374–413. Chapman and Hall: London.

Young, P.C. (1998) Data-based mechanistic modeling of environmental, ecological, economic and engineering systems. *Environmental Modelling and Software*, 13:105–122.

Young, P.C. (2000) Stochastic, dynamic modelling and signal processing: time variable and state dependent parameter estimation. In W. J. Fitzgerald, A. Walden, R. Smith, and P. C. Young, editors, *Nonlinear and Nonstationary Signal Processing*, pages 74–114. Cambridge University Press: Cambridge.

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

Young, P.C. and S.G. Wallis (1985) Recursive estimation: a unified approach to the identification, estimation an forecasting of hydrological systems. *Applied Mathematics and Computation* **17**, 299-334.

Young, P.C. and Beven, K.J. (1994) Data-based mechanistic modelling and the rainfall- flow non-linearity. *Environmetrics*, vol. 5, 335-363.