Identification of a catchment model via errors-in-variables approaches - a preliminary study

J. G. Linden * **P. C. Young** ** **T. Larkowski** * **K. J. Burnham** *

* *Control Theory and Applications Centre, Coventry University, UK, (email: j.linden@coventry.ac.uk)*
** *Department of Environmental Science, Lancaster University; Fenner School of Environment and Society, Australian National University, Canberra; and School of Electrical Engineering & Telecommunications, University of NSW, Sydney, Australia*

Abstract: Rainfall flow-data from the Leaf River in Mississippi, USA, is used to identify dynamic catchment models within an errors-in-variables framework. Extended bias compensating least squares and extended bias compensating instrumental variable methods are utilised in order to estimate the parameters of a linear model, as well as the variance of the input measurement disturbance. These techniques are also extended to deal with multiplicative noise on the input, rather than assuming an input disturbance of an additive character, which seems to be a more appropriate assumption for the data considered in the paper. In the multiplicative noise case, the resulting model residuals are shown to be smaller than in the additive noise case and so the noise level on the input is estimated to be rather low, indicating that it is sufficient to consider conventional maximum likelihood, non-errors-in-variables techniques for these particular rainfall-flow data.

Keywords: Catchment model; Errors-in-variables identification; Recursive identification; System identification.

1. INTRODUCTION

In order to predict the flood levels of rivers, a catchment model, which describes the dynamic behaviour between the measured rainfall and the measured river flow, is usually required. Once a model has been identified, this can be utilised as the basis of a forecasting engine, such as the Kalman filter, in order to predict future flood levels (see e.g. Young [2002, 2009]).

When a suitable model is obtained via dynamic system identification applied to the measured input and output data, it is convenient to represent it as a discrete-time transfer function, which is characterised by a finite number of parameters. Whilst the parameters of the transfer function do not necessarily exhibit any direct physical meaning, the model obtained in this manner can be transformed into a continuous-time state space form that can be interpreted in physically meaningful terms. In fact, the state space model parameters will normally contain information on the various 'quick' and 'slow' pathways involved in the transfer of the rainfall into river flow (see above references). Such information can provide valuable insight into the dynamic properties of the catchment area.

A common assumption within system identification is that the system input (here the rainfall) is exactly known. However, this assumption is clearly violated in the present application, since both the flow and the rainfall data are measured. It seems reasonable, therefore, to assume that both data series are corrupted (at least to some degree) by measurement noise. When the noise on the input signal is ignored during the parameter estimation task, the result can lead to a systematic, asymptotic bias error on the parameter estimates. Whilst such a bias might have a minor effect on the predictive performance of the model (see Young [1984], Söderström [2007]), it would not be desirable in cases where the parameters are being estimated in order to infer physical meaning. Fortunately, in order to avoid such bias on the parameter estimates (or simply to evaluate whether such bias may exist), the input measurement noise can be taken into account during the estimation process by using *errors-in-variables* identification techniques. In order to emphasize the practical utility of the approach, this case study is based on real data: namely, daily 'effective' rainfall (see later) and flow data from the Leaf River in Mississippi, USA.

Section 2 presents the rainfall-flow data, provides some details about catchment modelling approaches and provides a (non-errors-in-variables) reference model, which will be used to compare and contrast the subsequently identified errors-in-variables models. Section 3 gives a short introduction to the errors-in-variables framework by presenting some required notation and assumptions. Two well known errors-in-variables identification algorithms are reviewed in Section 4, whilst an extension to the multiplicative input noise case is discussed in Section 5. These errors-in-variables identification algorithms are then applied to the

rainfall-flow data in Section 6, where the results obtained are assessed and critically appraised. Some conclusions about the results are reported in Section 7.

Note that, throughout this paper, general time dependent quantities have a sub- or superscript k and estimated quantities are marked by a hat: e.g. $y_k; \hat{\vartheta}_k$.

2. CATCHMENT MODELLING

2.1 Overall catchment model

The '*Data-Based Mechanistic*' (DBM) catchment model [Young, 2002] is normally identified in a Hammerstein form, in which an '*effective rainfall nonlinearity*' is cascaded with a linear dynamic model for flow in the river. Within this study, attention is focused on the identification of this second, linear river flow model, where the input data is the effective rainfall, as estimated from the measured data during previous exercises (Young [2009]). In particular, the identification task is concerned with estimating a linear transfer function model that relates the effective rainfall in the the Leaf River catchment to the river flow data at a given location in the catchment, over the complete year of daily data shown in Figure 1.

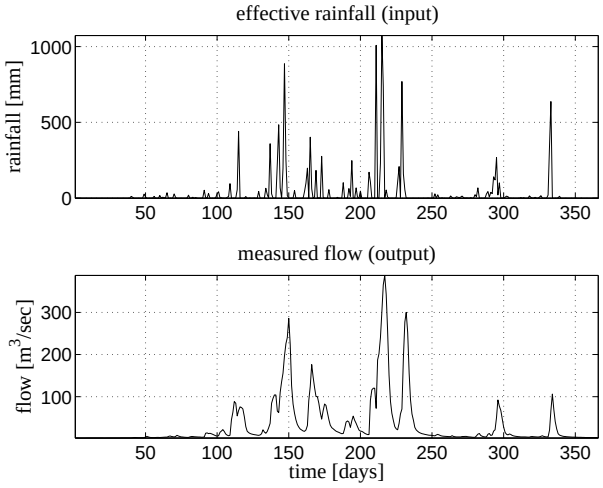

Fig. 1. Rainfall-flow data from the Leaf River.

2.2 Box-Jenkins effective rainfall-flow model

During previous identification tasks, the data has been successfully modelled using a discrete-time Box-Jenkins model structure (see e.g. Young [1984], Ljung [1999]) of the form:

$$y_k = \frac{B(q^{-1})}{A(q^{-1})}u_k + \frac{D(q^{-1})}{C(q^{-1})}e_k, \tag{1}$$

where y_k denotes the system output (flow), u_k is the input (effective rainfall), e_k denotes a zero mean, white noise term with variance [1] σ_e and k represents the discrete-time index. The quantities $A(q^{-1}), B(q^{-1}), C(q^{-1})$ and $D(q^{-1})$ are polynomials in the backward shift operator q^{-1}: i.e. $q^{-1}u_k = u_{k-1}$. The polynomials are defined as follows:

$$A(q^{-1}) = 1 + a_1 q^{-1} + \cdots + a_{n_a} q^{-n_a}, \tag{2a}$$

$$B(q^{-1}) = b_0 + b_1 q^{-1} + \cdots + b_{n_b} q^{-n_b}, \tag{2b}$$

$$C(q^{-1}) = 1 + c_1 q^{-1} + \cdots + c_{n_c} q^{-n_c}, \tag{2c}$$

$$D(q^{-1}) = 1 + d_1 q^{-1} + \cdots + d_{n_d} q^{-n_d}. \tag{2d}$$

Note that whilst $A(q^{-1})$ and $B(q^{-1})$ describe the dynamics of the system, $C(q^{-1})$ and $D(q^{-1})$ aim to describe the disturbance acting on the output, which is driven by the assumed random white noise sequence e_k. Using the optimal *Refined Instrumental Variable* (RIVBJ) algorithm in the CAPTAIN Toolbox for Matlab [2], the model structural parameters are identified as: $n_a = 3$, $n_b = 3$, $n_c = 1$ and $n_d = 3$) and the associated maximum likelihood parameter estimates obtained by RIVBJ are given in Table 1, together with their estimated standard errors (SE). The estimated variance of the white noise e_k is $\hat{\sigma}_e = 100.39$ and its autocorrelation function shows no significant correlation, as required.

$\hat{a}_1$	$-1.1858(0.079)$	$\hat{c}_1$	$-2.9826{\cdot}10^{-1}(0.086)$
$\hat{a}_2$	$7.0533{\cdot}10^{-1}(0.118)$		
$\hat{a}_3$	$-2.5497{\cdot}10^{-1}(0.061)$		
$\hat{b}_0$	$3.5848{\cdot}10^{-2}(0.0043)$		
$\hat{b}_1$	$1.0455{\cdot}10^{-1}(0.0043)$	$\hat{d}_1$	$9.9310{\cdot}10^{-1}(0.078)$
$\hat{b}_2$	$1.3810{\cdot}10^{-2}(0.0079)$	$\hat{d}_2$	$6.9380{\cdot}10^{-1}(0.091)$
$\hat{b}_3$	$8.1269{\cdot}10^{-2}(0.0063)$	$\hat{d}_3$	$4.7677{\cdot}10^{-1}(0.055)$

Table 1. RIVBJ parameter estimates (SE in parentheses).

This model is statistically well defined: the recursive RIVBJ parameter estimates exhibit little significant variation after about 240 samples, with a small standard error band; and the model explains 97.34% of the variance in y_k. Note, however, that it has been assumed in this exercise that the effective rainfall contains no measurement errors and this is the assumption that is critically evaluated in subsequent sections of the paper.

Remark 1. (Comparison with PEM Estimation). Not surprisingly, the PEM algorithm in the Matlab SID Toolbox yields very similar parameter estimates and standard errors to those shown in Table 1. This is because the RIVBJ and PEM algorithms both produce maximum likelihood estimates if the statistical assumptions apply, although they employ quite different algorithmic procedures: see the comparisons in Young [2008], where two rainfall-flow inspired examples demonstrate the distinct advantages of the RIVBJ algorithm when the system has real poles and 'stiff' dynamics. Also, in the present example, while the RIVBJ recursive estimates converge on the same final values as the *en bloc* estimates in Table 1, those produced with similar user-specifications by the RPEM algorithm in the SID Toolbox do not do this and are more volatile.

Remark 2. (Constraints on poles). Note that in order to build a state space system with a physically meaningful parametrisation from the Box-Jenkins transfer function model (1), it is necessary to constrain the roots of the $A(q^{-1})$ polynomial, i.e. the poles of the dynamical system, to be real [Young, 2009]. For simplicity, however, such a constraint is not imposed on the rainfall-flow model in this

[1] Note that for the ease of notation variances are denoted here with σ rather than σ^2, the latter being the preferred notation within the statistical literature.

[2] Available from http://www.es.lancs.ac.uk/cres/captain/

study but it will be considered as an immediate step in further research.

3. ERRORS-IN-VARIABLES MODELLING

In order to identify a rainfall-flow model in which the effect of any input measurement disturbance is taken into account, the presence and variance of such input noise has to be identified concurrently with the other model parameters. If this estimated variance is found to be large, then the associated bias on the estimates is likely to be significant, so justifying the use of the more complex errors-in-variables approach to identification, rather than the simpler maximum likelihood RIVBJ approach used in the previous section.

Within the present study, 'bias compensating least squares' approaches for coloured output noise (see Söderström [2007]) are considered initially. An errors-in-variables transfer function model is considered in the form

$$A(q^{-1})y_{0_k} = B(q^{-1})u_{0_k}, \tag{3}$$

where the polynomials $A(q^{-1})$ and $B(q^{-1})$ now describe the dynamic relationship between the noise-free (but unknown) input and noise-free output, which are denoted u_{0_k} and y_{0_k}, respectively. The noisy input and output signals, u_k and y_k, are then defined by the measurement equations:

$$u_k = u_{0_k} + \tilde{u}_k, \tag{4a}$$
$$y_k = y_{0_k} + \tilde{y}_k, \tag{4b}$$

where $\tilde{u}_k$ denotes zero mean, white measurement noise at the input, whilst $\tilde{y}_k$ denotes the noise on the output. Since this output noise not only accounts for measurement errors, but is also required to accommodate for model mismatch and the effect of unobserved inputs, it is assumed to be coloured noise, as in the standard model (1). Such an errors-in-variables system setup is depicted in Figure 2.

3.1 Notation

In order to proceed, it is necessary to introduce some commonly used notation. Within this paper, cross- and auto-correlation functions are defined by

$$r_{cd}(\tau) \triangleq E\left[c_k d_{k-\tau}\right], \tag{5a}$$
$$r_c(\tau) \triangleq E\left[c_k c_{k-\tau}\right], \tag{5b}$$

where c_k and d_k denote two arbitrary ergodic zero mean stochastic processes and where $E[\cdot]$ denotes the expected value operator. The matrices and vectors comprising these covariance elements are denoted as Σ and ξ, respectively. Consequently, the cross/auto-covariance matrices of two arbitrary random vectors v_k and w_k are denoted

$$\Sigma_{vw} \triangleq E\left[v_k w_k^T\right], \tag{6a}$$
$$\Sigma_v \triangleq E\left[v_k v_k^T\right], \tag{6b}$$

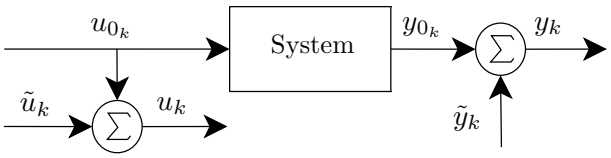

Fig. 2. Errors-in-variables setup.

whilst the cross-covariance vector between an arbitrary random vector v_k and a scalar stochastic process c_k is denoted

$$\xi_{vc} \triangleq E\left[v_k c_k\right] \qquad \text{(column vector)}, \tag{7a}$$
$$\xi_{cv} \triangleq E\left[c_k v_k^T\right] \qquad \text{(row vector)}. \tag{7b}$$

Finally, the variance of a scalar process is denoted by

$$\sigma_c \triangleq E\left[c_k^2\right]. \tag{8}$$

3.2 Assumptions

The following assumptions are required for the development of the bias-compensating least squares techniques:

A1 The dynamic system (3) is asymptotically stable, i.e. $A(q^{-1})$ has all zeros inside the unit circle.
A2 All system modes are observable and controllable, i.e. $A(q^{-1})$ and $B(q^{-1})$ have no common factors.
A3 The polynomial degrees n_a and n_b are known *a priori* with $n_b \leq n_a$.
A4 The true input u_{0_k} is a zero-mean ergodic process and is persistently exciting of sufficiently high order.
A5 The sequence $\tilde{u}_k$ is an additive, stationary, zero-mean, ergodic, white noise process with unknown variance $\sigma_{\tilde{u}}$.
A6 The sequence $\tilde{y}_k$ is an additive, stationary, zero-mean, ergodic noise process with unknown auto-correlation sequence $\{r_{\tilde{y}}(0), r_{\tilde{y}}(1), ...\}$.
A7 The sequences $\tilde{u}_k$ and $\tilde{y}_k$ are mutually uncorrelated and also uncorrelated with u_{0_k} and y_{0_k}.

3.3 Errors-in-variables model in regression form

Introducing the parameter vectors

$$\theta \triangleq \left[a^T\ b^T\right]^T = \left[a_1\ ...\ a_{n_a}\ b_0\ ...\ b_{n_b}\right]^T, \tag{9}$$

an alternative description of (3)-(4) is given by

$$y_k = \varphi_{0_k}^T \theta, \tag{10a}$$
$$\varphi_k = \varphi_{0_k} + \tilde{\varphi}_k, \tag{10b}$$

where

$$\varphi_{0_k} \triangleq [-y_{0_{k-1}}\ ...\ -y_{0_{k-n_a}}\ u_{0_k}\ u_{0_{k-1}}\ ...\ u_{0_{k-n_b}}]^T, \tag{11a}$$
$$= \left[\varphi_{y_{0_k}}^T\ \varphi_{u_{0_k}}^T\right]^T,$$
$$\varphi_k \triangleq [-y_{k-1}\ ...\ -y_{k-n_a}\ u_k\ u_{k-1}\ ...\ u_{k-n_b}]^T, \tag{11b}$$
$$= \left[\varphi_{y_k}^T\ \varphi_{u_k}^T\right]^T,$$
$$\tilde{\varphi}_k \triangleq [-\tilde{y}_{k-1}\ ...\ -\tilde{y}_{k-n_a}\ \tilde{u}_k\ u_{k-1}\ ...\ \tilde{u}_{k-n_b}]^T, \tag{11c}$$
$$= \left[\tilde{\varphi}_{y_k}^T\ \tilde{\varphi}_{u_k}^T\right]^T.$$
(11d)

The errors-in-variables identification problem can be stated as follows:

Problem 3. Given an incrementally increasing number k of measured input-output samples

$$Z^k \triangleq \{u_1, y_1, ..., u_i, y_i, ..., u_k, y_k\}, \tag{12}$$

determine an estimate of the augmented parameter vector

$$\vartheta \triangleq \left[a_1\ ...\ a_{n_a}\ b_0\ ...\ b_{n_b}\ \sigma_{\tilde{u}}\right]^T. \tag{13}$$

Note that it is also possible to estimate the auto-correlation sequence of the output noise

$$\rho_{\tilde{y}} \triangleq \left[r_{\tilde{y}}(0)\ r_{\tilde{y}}(1)\ ...\ r_{\tilde{y}}(n_a)\right]^T. \tag{14}$$

as outlined in Söderström [2008].

4. BIAS-COMPENSATING LEAST SQUARES FOR COLOURED OUTPUT NOISE

In principle, it is possible to apply the least squares estimator to obtain an estimate $\hat{\theta}$ of the parameter vector θ of the system (10). Using the notation of covariance matrices and covariance vectors given in Section 3.1, the least squares estimator is given by

$$\hat{\theta}^{\mathrm{LS}} = \Sigma_{\varphi}^{-1}\xi_{\varphi y} = \begin{bmatrix} \Sigma_{\varphi_y} & \Sigma_{\varphi_y\varphi_u} \\ \Sigma_{\varphi_u\varphi_y} & \Sigma_{\varphi_u} \end{bmatrix}^{-1} \begin{bmatrix} \xi_{\varphi_y y} \\ \xi_{\varphi_u y} \end{bmatrix}. \tag{15}$$

However, the estimate obtained will be asymptotically biased, i.e. no matter how many measurements are taken, $\hat{\theta}^{\mathrm{LS}}$ will always contain a systematic error, which depends on the input measurement noise variance $\sigma_{\tilde{u}}$ and the auto-correlation sequence of the output noise given by $\rho_{\tilde{y}}$. If these quantities are known, it is possible to compensate for the bias in $\hat{\theta}^{\mathrm{LS}}$ by solving the compensated normal equations (see e.g. Söderström [2008] for details):

$$\left(\begin{bmatrix} \Sigma_{\varphi_y} & \Sigma_{\varphi_y\varphi_u} \\ \Sigma_{\varphi_u\varphi_y} & \Sigma_{\varphi_u} \end{bmatrix} - \begin{bmatrix} \Sigma_{\tilde{\varphi}_y} & 0 \\ 0 & \sigma_{\tilde{u}} I_{n_b+1} \end{bmatrix} \right) \theta = \begin{bmatrix} \xi_{\varphi_y y} \\ \xi_{\varphi_u y} \end{bmatrix} - \begin{bmatrix} \xi_{\tilde{\varphi}_y \tilde{y}} \\ 0 \end{bmatrix}, \tag{16}$$

where I_{n_b+1} denotes the identity matrix of size n_b+1 and

$$\Sigma_{\tilde{\varphi}_y} = \begin{bmatrix} r_{\tilde{y}}(0) & \cdots & r_{\tilde{y}}(n_a-1) \\ \vdots & \ddots & \vdots \\ r_{\tilde{y}}(n_a-1) & \cdots & r_{\tilde{y}}(0) \end{bmatrix} \tag{17}$$

is a symmetric Toeplitz matrix, whilst

$$\xi_{\tilde{\varphi}_y \tilde{y}} = \left[r_{\tilde{y}}(1) \ \cdots \ r_{\tilde{y}}(n_a) \right]^T. \tag{18}$$

Hence, the principle of the bias compensating least squares approaches is to estimate the the noise parameters (here $\sigma_{\tilde{u}}$ and $\rho_{\tilde{y}}$) and then to compute θ using the compensated normal equations (16).

4.1 Extended bias compensating least squares

For the estimation of the parameter vector $\theta \in \mathbb{R}^{n_a+n_b+1}$, the n_a+n_b+1 normal equations, as specified in (16), can be utilised. However, in order to compensate for the asymptotic bias, not only θ, but also $\sigma_{\tilde{u}}$ and $\rho_{\tilde{y}}$ are required to be estimated from the data. Hence more equations are needed, which can be obtained by appending instruments to the regression vector. Such an approach is known as *extended bias compensating least squares* and has been addressed in Ekman [2005].

In order to develop this approach, we introduce the instrument vector

$$\zeta_k \triangleq \left[u_{k-n_b-1} \ \ldots \ u_{k-n_b-n_\zeta} \right]^T, \tag{19}$$

with $n_\zeta > n_a+1$. This leads to an overdetermined system of normal equations (cf. (16)) given by

$$M(\sigma_{\tilde{u}}, \rho_{\tilde{y}})\theta = m(\rho_{\tilde{y}}), \tag{20}$$

where

$$M(\sigma_{\tilde{u}}, \rho_{\tilde{y}}) \triangleq \begin{bmatrix} \Sigma_{\varphi_y} & \Sigma_{\varphi_y\varphi_u} \\ \Sigma_{\varphi_u\varphi_y} & \Sigma_{\varphi_u} \\ \Sigma_{\zeta\varphi_y} & \Sigma_{\zeta\varphi_u} \end{bmatrix} - \begin{bmatrix} \Sigma_{\tilde{\varphi}_y} & 0 \\ 0 & \sigma_{\tilde{u}} I_{n_b} \\ 0 & 0 \end{bmatrix}, \tag{21a}$$

$$m(\rho_{\tilde{y}}) \triangleq \begin{bmatrix} \xi_{\varphi_y y} \\ \xi_{\varphi_u y} \\ 0 \end{bmatrix} - \begin{bmatrix} \xi_{\tilde{\varphi}_y \tilde{y}} \\ 0 \\ 0 \end{bmatrix}. \tag{21b}$$

Equation (20) consists of a sufficient number of equations to estimate θ, $\sigma_{\tilde{u}}$ and $\rho_{\tilde{y}}$. Note that this estimation problem is nonlinear due to the multiplication of θ with $\sigma_{\tilde{u}}$ and $\rho_{\tilde{y}}$. A natural way to estimate the parameters would be to search for those quantities, which produce the smallest residuals, i.e.

$$\{\hat{\theta}^{\mathrm{EBCLS}}, \hat{\sigma}_{\tilde{u}}, \hat{\rho}_{\tilde{y}}\} = \arg\min_{\theta, \sigma_{\tilde{u}}, \rho_{\tilde{y}}} \ \left\| M(\sigma_{\tilde{u}}, \rho_{\tilde{y}})\theta - m(\rho_{\tilde{y}}) \right\|_2^2. \tag{22}$$

However, the problem is seperable in θ and $\{\sigma_{\tilde{u}}, \rho_{\tilde{y}}\}$, so it is beneficial to solve the nonlinear least squares problem via the technique of variable projection (see Golub and Pereyra [1973, 2002]), which is also known as separable nonlinear least squares. The idea is to compute the least squares residuals for different values of $\{\sigma_{\tilde{u}}, \rho_{\tilde{y}}\}$ and choose that estimate $\{\hat{\sigma}_{\tilde{u}}, \hat{\rho}_{\tilde{y}}\}$, which produces the residual vector of smallest magnitude: i.e.,

$$\{\hat{\sigma}_{\tilde{u}}, \hat{\rho}_{\tilde{y}}\} = \arg\min_{\sigma_{\tilde{u}}, \rho_{\tilde{y}}} \ V_1(\sigma_{\tilde{u}}) \tag{23}$$

with

$$V_1(\sigma_{\tilde{u}}, \rho_{\tilde{y}}) = \left|\left| M(\sigma_{\tilde{u}}, \rho_{\tilde{y}}) M^\dagger(\sigma_{\tilde{u}}, \rho_{\tilde{y}}) m(\rho_{\tilde{y}}) - m(\rho_{\tilde{y}}) \right|\right|_2^2, \tag{24}$$

where $M^\dagger(\sigma_{\tilde{u}}, \rho_{\tilde{y}}) \triangleq [M^T(\sigma_{\tilde{u}}, \rho_{\tilde{y}}) M(\sigma_{\tilde{u}}, \rho_{\tilde{y}})]^{-1} M^T(\sigma_{\tilde{u}}, \rho_{\tilde{y}})$ denotes the Moore-Penrose pseudo inverse of $M(\sigma_{\tilde{u}}, \rho_{\tilde{y}})$. Once the estimate $\hat{\sigma}_{\tilde{u}}$ has been obtained, the compensated parameter vector is found by solving (20) as

$$\hat{\theta}^{\mathrm{EBCLS}} = M^\dagger(\hat{\sigma}_{\tilde{u}}, \hat{\rho}_{\tilde{y}}) m(\hat{\rho}_{\tilde{y}}). \tag{25}$$

4.2 Extended bias compensating instrumental variables

Whilst for the coloured output noise case, the compensation of the least squares estimate requires the estimation of $\rho_{\tilde{y}}$, it might be more practical to consider an instrumental variable approach, which avoids the need for computing the auto-correlation sequence $\rho_{\tilde{y}}$. Therefore, we define the instrument vector comprised solely of delayed inputs as

$$\delta_k \triangleq \left[u_k \ u_{k-1} \ \ldots \ u_{k-n_\delta} \right]^T, \tag{26}$$

where $n_\delta \geq n_a+n_b+2$ is the number of instruments, which is a user-chosen quantity. The compensated least squares normal equations then become

$$G(\sigma_{\tilde{u}})\theta = \xi_{\delta y} \tag{27}$$

with

$$G(\sigma_{\tilde{u}}) \triangleq \Sigma_{\delta\varphi} - \sigma_{\tilde{u}} \begin{bmatrix} 0 & I_{n_b+1} \\ 0 & 0 \end{bmatrix}. \tag{28}$$

Note that (27) constitutes n_δ equations in n_a+n_b+2 unknowns (relating to θ and $\sigma_{\tilde{u}}$) and that, in order to compensate for the bias in θ, only $\sigma_{\tilde{u}}$ is required to be estimated. However, the estimation problem is nonlinear due to the multiplication of θ and $\sigma_{\tilde{u}}$ so, by applying the technique of variable projection, $\hat{\sigma}_{\tilde{u}}$ can be estimated as

$$\hat{\sigma}_{\tilde{u}} = \arg\min_{\sigma_{\tilde{u}}} \ V_2(\sigma_{\tilde{u}}) \tag{29}$$

with

$$\begin{aligned} V_2(\sigma_{\tilde{u}}) &= \left|\left| G(\sigma_{\tilde{u}}) G^\dagger(\sigma_{\tilde{u}}) \xi_{\delta y} - \xi_{\delta y} \right|\right|_2^2 \\ &= \xi_{\delta y}^T \xi_{\delta y} - \xi_{\delta y}^T G(\sigma_{\tilde{u}}) \left[G^T(\sigma_{\tilde{u}}) G(\sigma_{\tilde{u}}) \right]^{-1} G^T(\sigma_{\tilde{u}}) \xi_{\delta y}. \end{aligned} \tag{30}$$

Once the estimate $\hat{\sigma}_{\tilde{u}}$ has been obtained, the compensated parameter vector is obtained by solving (27) as

$$\hat{\theta}^{\mathrm{EBCIV}} = G^\dagger(\hat{\sigma}_{\tilde{u}}) \xi_{\delta y}. \tag{31}$$

5. BIAS-COMPENSATING LEAST SQUARES FOR MULTIPLICATIVE INPUT NOISE

For the input rainfall data at hand, the assumptions introduced in Section 3.2 may not all be justifiable. In particular, Assumption **A5a** might not be realistic when inspecting the effective rainfall data in Figure 1. For example, during periods of long dryness, which are observed during the first 40 days of the data, the effective rainfall is zero and it does seem unreasonable to assume an additive measurement error of constant variance during these periods. In addition, one might postulate that, during intervals of minor rainfall, the measurement disturbance is rather small, whereas it may increase during periods of heavy rainfall. One way to model this phenomena could be to assume an additive heteroscedastic noise with time variable variance, i.e. to assume a non-stationary input measurement noise process. Whilst this is likely to lead to some difficulties during the estimation task, a more appealing alternative would be to assume a multiplicative noise on the input having a unity mean value and fixed variance. Thus, Assumption **A5a** is replaced by:

A5b The sequence $\tilde{u}_k$ is a multiplicative, stationary, ergodic, white noise process with unity mean and unknown variance $\sigma_{\tilde{u}}$.

Then the accordingly modified errors-in-variables model is given by

$$A(q^{-1})y_{0_k} = B(q^{-1})u_{0_k}, \tag{32a}$$
$$u_k = u_{0_k}\tilde{u}_k, \tag{32b}$$
$$y_k = y_{0_k} + \tilde{y}_k, \tag{32c}$$

where the characteristics of the output noise sequence remain unchanged.

In order to illustrate the modification of the bias-compensating techniques to accommodate the case of multiplicative input noise, it is derived here for the extended bias compensating instrumental variable approach, which has been reviewed for the additive noise case in Section 4.2. The bias compensating least squares approach presented in Section 4.1 can also be adjusted to accommodate for the multiplicative input noise case in a similar manner.

Consider the covariance matrix $\Sigma_{\delta\varphi}$, as well as the vector $\xi_{\delta y}$. These quantities essentially consist of three different forms of entries, namely $r_{uu}(0)$, $r_{uu}(\tau)$ and $r_{uy}(\tau)$, which can be computed in a straightforward manner, as outlined in Appendix A. By introducing the partition of the instrument vector δ_k as

$$\delta_k = \begin{bmatrix}\varphi_{u_k}^T & \zeta_k^T\end{bmatrix} \tag{33}$$

with ζ_k as defined in (19), the uncompensated normal equations for the multiplicative input noise case are given by

$$\begin{bmatrix}\Sigma_{\varphi_u\varphi_y} & \Sigma_{\varphi_u} \\ \Sigma_{\zeta\varphi_y} & \Sigma_{\zeta\varphi_u}\end{bmatrix}\theta = \begin{bmatrix}\xi_{\varphi_u y} \\ \xi_{\zeta y}\end{bmatrix}, \tag{34}$$

and since the input measurement noise variance given by

$$\sigma_{\tilde{u}} = r_{\tilde{u}}(0) - 1 \tag{35}$$

occurs exclusively in entries of the form $r_u(0)$ (cf. (A.1)), only the diagonal elements of Σ_{φ_u} are effected by the input noise. Using (A.1)-(A.2), the (uncompensated) matrix Σ_{φ_u} is given by

$$\Sigma_{\varphi_u} = \begin{bmatrix} r_{\tilde{u}}(0)r_{u_0}(0) & r_{u_0}(1) & \cdots & r_{u_0}(n_b) \\ r_{u_0}(1) & r_{\tilde{u}}(0)r_{u_0}(0) & \cdots & \vdots \\ \vdots & & \ddots & \vdots \\ r_{u_0}(n_b) & & \cdots & r_{\tilde{u}}(0)r_{u_0}(0)\end{bmatrix}. \tag{36}$$

By dividing the diagonal elements of Σ_{φ_u} by $r_{\tilde{u}}(0)$, and by making use of (A.1)-(A.2), the compensated covariance matrix is obtained as

$$\Sigma_{\varphi_{u_0}}(\sigma_{\tilde{u}}) = \begin{bmatrix} \dfrac{r_u(0)}{\sigma_{\tilde{u}}+1} & r_u(1) & \cdots & r_u(n_b) \\ r_u(1) & \dfrac{r_u(0)}{\sigma_{\tilde{u}}+1} & \cdots & \vdots \\ \vdots & & \ddots & \vdots \\ r_u(n_b) & & \cdots & \dfrac{r_u(0)}{\sigma_{\tilde{u}}+1}\end{bmatrix}, \tag{37}$$

which yields an unbiased estimate for θ given by

$$\hat{\theta}^{\mathrm{EBCIV}} = H^{\dagger}(\sigma_{\tilde{u}})\xi_{\delta y}, \tag{38}$$

where

$$H(\sigma_{\tilde{u}}) = \begin{bmatrix}\Sigma_{\varphi_u\varphi_y} & \Sigma_{\varphi_{u_0}}(\sigma_{\tilde{u}}) \\ \Sigma_{\zeta\varphi_y} & \Sigma_{\zeta\varphi_u}\end{bmatrix}. \tag{39}$$

Note that the overall estimation problem is again nonlinear, due to the product between $\sigma_{\tilde{u}}$ and θ. As in the additive noise case, it is possible to use the variable projection approach to obtain an estimate of $\sigma_{\tilde{u}}$ via

$$\hat{\sigma}_{\tilde{u}} = \arg\min_{\sigma_{\tilde{u}}} V_3(\sigma_{\tilde{u}}) \tag{40}$$

with

$$V_3(\sigma_{\tilde{u}}) = \left|\left|H(\sigma_{\tilde{u}})H^{\dagger}(\sigma_{\tilde{u}})\xi_{\delta y} - \xi_{\delta y}\right|\right|_2^2. \tag{41}$$

6. APPLICATION TO RAINFALL-FLOW DATA

Four different algorithms are utilised in order to obtain an errors-in-variables rainfall-flow model based on the rainfall-flow data given in Figure 1. In all cases, a fixed model structure with $n_a = 3$, $n_b = 3$, $n_c = 1$ and $n_d = 3$ is assumed. The first two algorithms are the extended bias compensating least squares (EBCLS) approach (cf. Section 4.1) and the extended bias compensating instrumental variable (EBCIV) approach (cf. Section 4.2). Both algorithms assume an additive measurement noise of constant variance on the input signal. The third and fourth algorithms are the modified versions, which assume an input disturbance of multiplicative character, as discussed in Section 5. All four algorithms are summarised in Table 2. For the EBCLS approaches the number of instruments

Algorithm	Approach	Input noise
EBCLS_a	EBCLS	additive
EBCIV_a	EBCIV	additive
EBCLS_m	EBCLS	multiplicative
EBCIV_m	EBCIV	multiplicative

Table 2. Algorithms used to identify errors-in-variables rainfall-flow model.

is chosen to be $n_\zeta = 12$, whereas in the EBCIV case, $n_\delta = 25$ has been selected. Both of these user-selected numbers for the instruments have been determined via simulation experiments based on the initially determined system given in Table 1. All estimates are obtained in a recursive manner, since it is considered that this might

give some insight into the performance of the identification algorithms, as was done in the RIVBJ estimation of the standard non-errors-in-variables model.

6.1 Bias compensation using additive input noise

The results of EBCLS$_a$ and EBCIV$_a$ for the additive input noise case are presented in Figure 3 and Figure 4, respectively. In the case of EBCLS$_a$, the input measurement noise variance after 366 days has been estimated to be around 150. The model parameters vary only a little after 240 days and are relatively close to the model parameters obtained by the RIVBJ algorithm in the CAPTAIN toolbox (cf. Table 1). This observation might be explained by investigating the signal-to-noise ratio on the input, which is defined as

$$\mathrm{SNR}_u = 10\log\left(\frac{E\left[u_{0_k}^2\right]}{\sigma_{\tilde{u}}}\right) = 10\log\left(\frac{E\left[u_k^2\right] - \sigma_{\tilde{u}}}{\sigma_{\tilde{u}}}\right). \tag{42}$$

Computing SNR$_u$ for this case gives a rather low value, of SNR$_u$ = 46dB, which means that the contribution of the input measurement noise is estimated to be marginal. The estimated model can, therefore, be considered to be of a standard (non-errors-in-variables) type with correlated output noise, which might explain why the model parameter estimates are in reasonable agreement with the RIVBJ estimates of the Box-Jenkins model parameters given in Table 1. It is interesting to observe, therefore, that both approaches yield rather similar estimates of the parameter vector θ even though the coloured output disturbance in the EBCLS approach is not explicitly parametrised as an ARMA process. In this connection, note that the the RIVBJ estimates in Table 1 are optimal in a maximum likelihood sense if there is no noise on the input variable.

Considering the estimates obtained via EBCIV$_a$, the estimate of $\sigma_{\tilde{u}}$ after 366 days is around 1,300, which corresponds to a SNR$_u$ of 24dB. Consequently, in this case, the influence of the input disturbance is estimated to be more significant. However, by inspecting the time trajectory of $\hat{\sigma}_{\tilde{u}}$ it is observed that the estimate is rather volatile, which might be an indication of identifiability issues or ill-conditioning of the estimation problem. One potential source of problems for this particular algorithm could be that the chosen instruments do not exhibit a sufficient amount of correlation with the regression vector to form a meaningful estimate. The confidence in these results is, therefore, considered to be rather low.

Since it has been argued in Section 5, that the assumption of the input measurement noise being of a multiplicative character might be more feasible for the data at hand, it is interesting to investigate the results obtained via the EBCLS$_m$ and EBCIV$_m$ algorithms, as presented in Figures 5 and 6, respectively.

6.2 Bias compensation using multiplicative input noise

Considering the results obtained via EBCLS$_m$, it is observed that the input measurement noise variance after 366 days is estimated to be around $\sigma_{\tilde{u}} = 7.8{\cdot}10^{-3}$. This is again a very low estimate for the noise level, since this implies that, most of the time, the measured signal is corrupted by

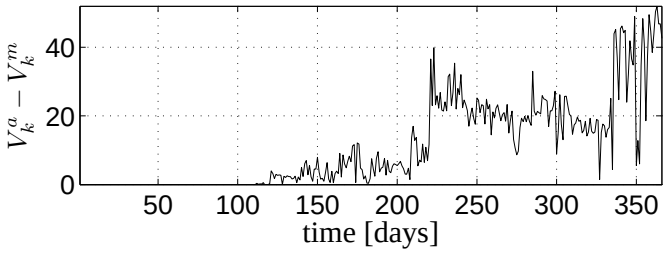

Fig. 8. Difference between V_k in the case of EBCLS$_a$ (denoted V_k^a) and V_k in the case of EBCLS$_m$ (denoted V_k^m).

less than 1% of measurement disturbances. The estimates obtained are barely changing after 250 data samples and they are very close to the estimates obtained by the RIVBJ algorithm.

In the case of EBCIV$_m$, the input measurement noise variance is estimated to be $\sigma_{\tilde{u}} = 8.2 \cdot 10^{-2}$, hence the noise level is estimated to be around 10 times larger than in the case of EBCLS$_m$. However, this is still quite small and would engender only small bias.

6.3 Model comparison

In order to assess the quality of the identified models, it is possible to compare the magnitude of the resulting residuals. This can be achieved by considering the cost function

$$V_k \triangleq \frac{1}{n_z}\left\|\Sigma_{z\varphi}^k \theta_k - \xi_{z\varphi}^k\right\|_2^2 \tag{43}$$

for each of the four algorithms, where the general vector z_k is either equal to $\left[\varphi_k^T \;\; \zeta_k^T\right]^T$, in the case of the EBCLS approaches, or equal to δ_k when the EBCIV techniques are utilised. Note that the cost function is normalised by the length of the vector z_k, which is denoted n_z. This allows for a 'fair' comparison, even when the number of instruments used for the techniques differ. The values of the cost functions evolving with time are shown in Figure 7 for the four algorithms and are compared to the least squares (LS) cost function, i.e. when $z_k = \varphi_k$.

From these results, it is observed that the EBCLS approaches seem to be superior with respect to the EBCIV approaches, since the cost function values of the latter are larger. This seems to be in accordance with the inspection of the parameter values, where the EBCIV approaches appear to be more erratic, hence less trustworthy. After 366 days, the cost functions of the EBCIV approaches are very close to the LS cost function, whereas the EBCLS approaches are distinguishably smaller. When comparing the approaches assuming additive input noise with the approaches assuming multiplicative noise on the input, the quality with respect to the cost V_k appear to be very similar. However, when comparing the V_k for the EBCLS$_a$ and EBCLS$_m$, it is observed that the multiplicative input noise does explain the data more exactly than in the additive noise case, as illustrated in Figure 8, where the difference of both cost functions is given. This justifies the assumptions of using an input noise of a multiplicative character for the Leaf River rainfall-flow data.

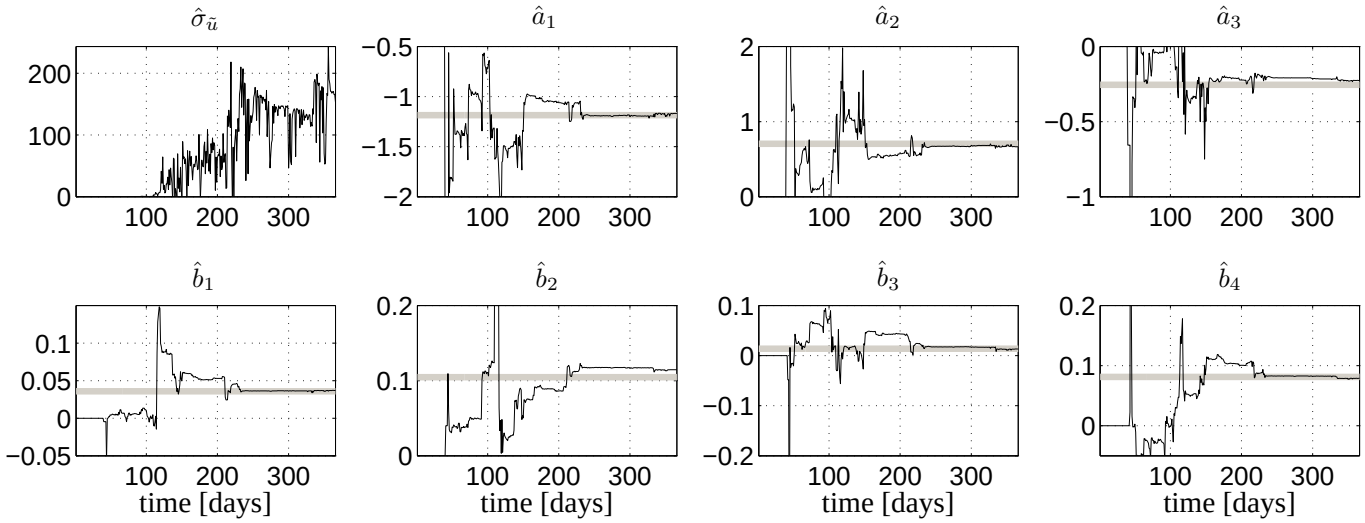

Fig. 3. Estimates using EBCLS$_a$ (black) in comparison to the estimates given in Table 1 (grey).

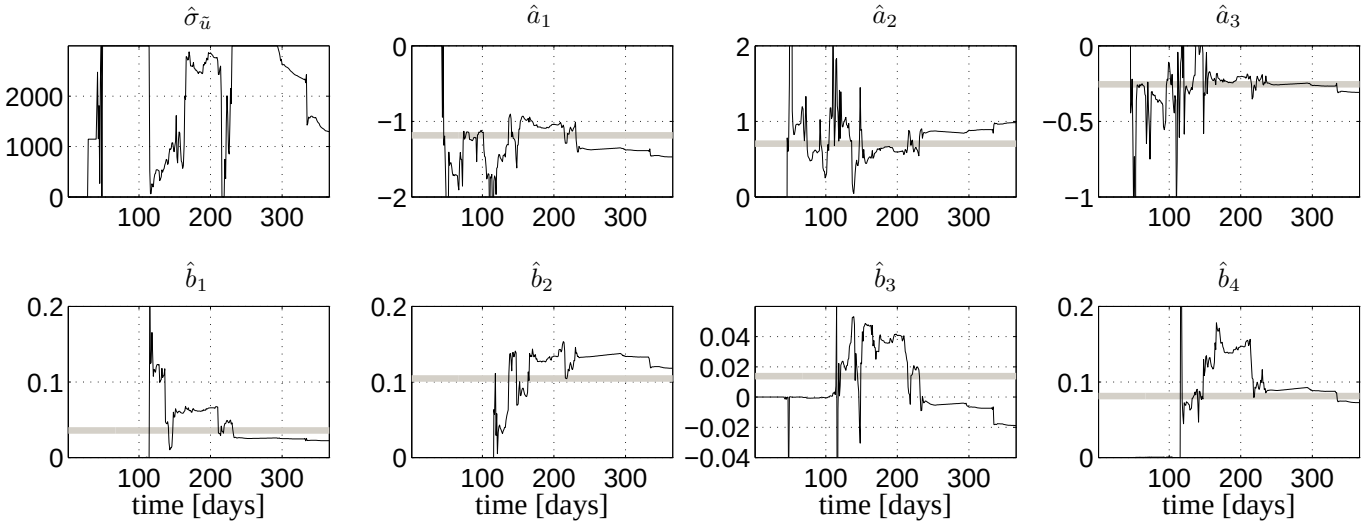

Fig. 4. Estimates using EBCIV$_a$ (black) in comparison to the estimates given in Table 1 (grey).

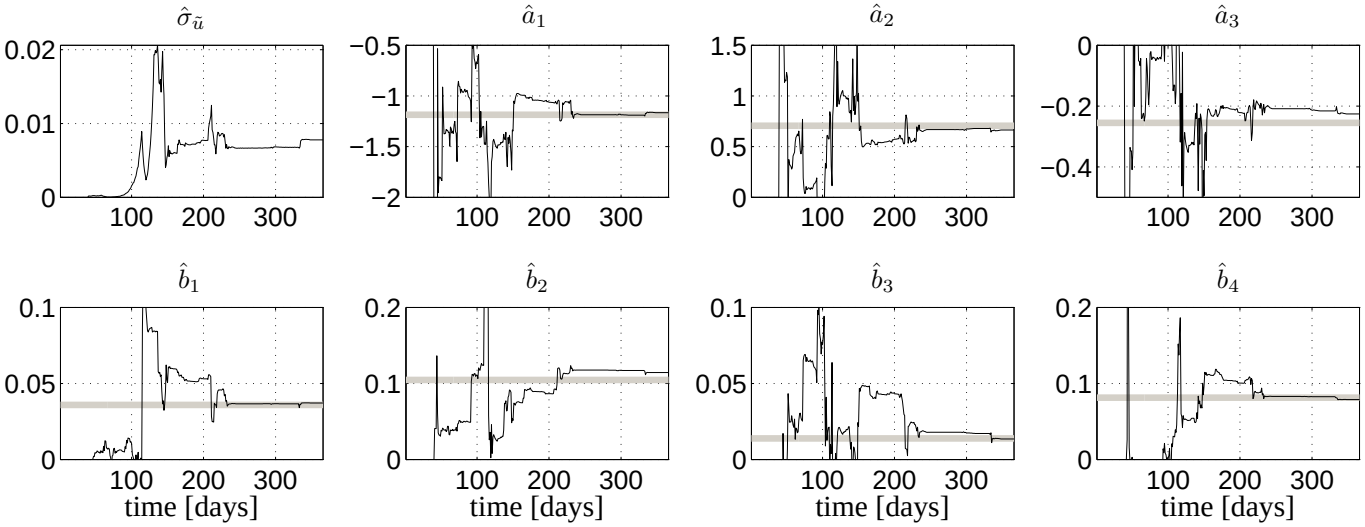

Fig. 5. Estimates using EBCLS$_m$ (black) in comparison to the estimates given in Table 1 (grey).

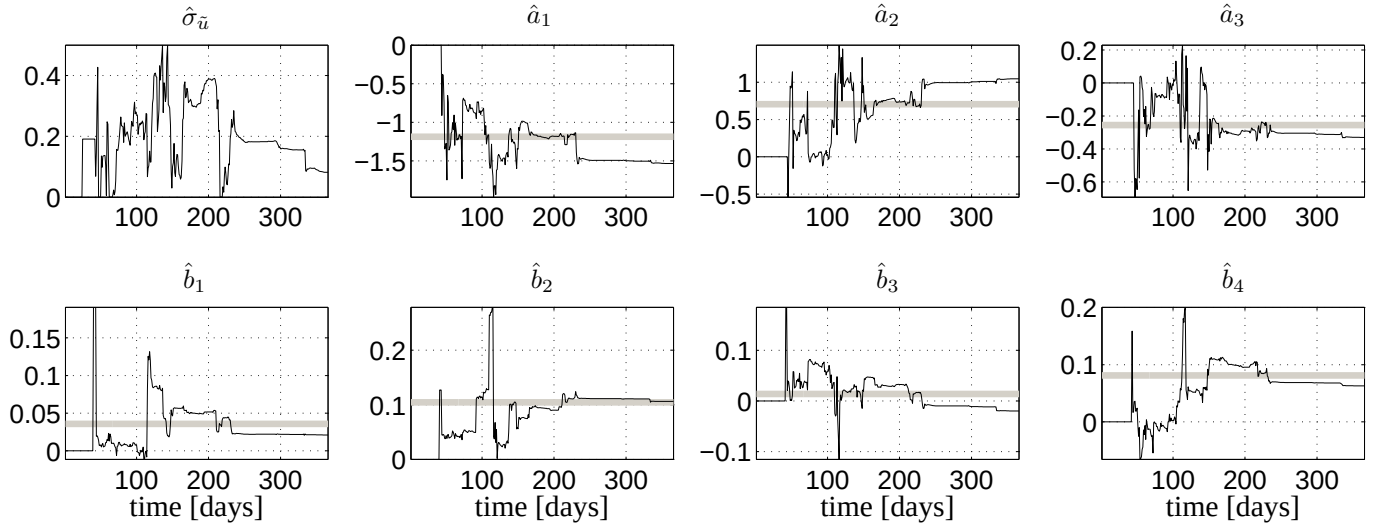

Fig. 6. Estimates using EBCIV$_m$ (black) in comparison to the estimates given in Table 1 (grey).

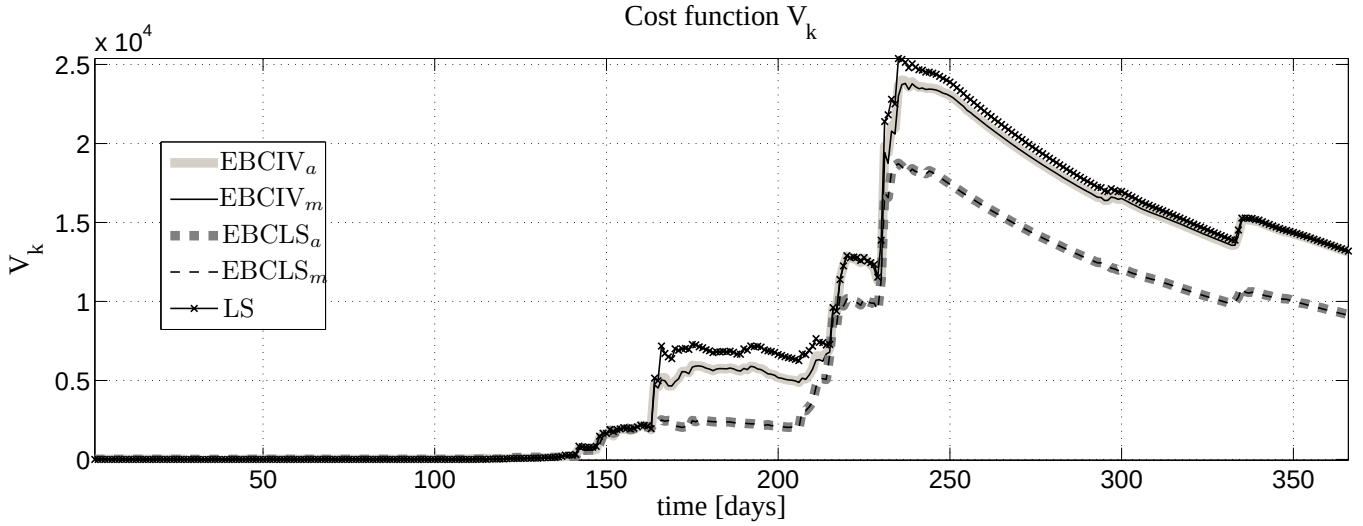

Fig. 7. Cost function values V_k for the different algorithms in comparison to the least squares (LS) cost function.

7. CONCLUSIONS

Rainfall-flow data from the Leaf River has been investigated in order to determine the potential need for modelling the input measurement noise explicitly by the means of errors-in-variables system identification techniques. Bias compensating techniques using least squares and instrumental variable approaches have been used to estimate the model parameters, as well as the additive input measurement noise variance, which provides a measure for the 'magnitude' of the input noise contamination. Both of these well-known errors-in-variables identification algorithms have been adjusted in order to deal with the case of multiplicative input measurement noise. The results indicate that, in this particular case, it seems to be beneficial to model the input uncertainty as a multiplicative disturbance, rather than additive noise. In general, it appears that the bias compensating least squares approaches are superior to their instrumental variables counterparts. One potential reason for this observation is that, in the errors-in-variables case, the latter approach uses only delayed inputs to form an instrument vector (in contrast to the optimal instruments used by the RIVBJ algorithm which are generated by an 'auxiliary model' of the system). As a result, they might not be sufficiently correlated with the regression vector, so reducing the statistical efficiency of the estimates.

The noise level on the input for both bias compensating least squares variants (additive and multiplicative noise) is estimated to be relatively small. This suggests that, for this particular rainfall-flow model identification problem, it seems unnecessary to adopt an errors-in-variables approach in order to estimate the model parameters, since the bias introduced on the parameter estimates is likely to be rather small. This is confirmed by the fact that estimates obtained by the bias compensating least squares approach are very similar to the estimates obtained by the optimal RIVBJ algorithm in the CAPTAIN toolbox, which does not allow for noise on the input variable. Note, however, that whilst the RIVBJ algorithm explicitly models the output disturbance by assuming a Box-

Jenkins model structure with an ARMA noise process, the estimates of the extended bias compensating least squares assume an arbitrary correlation of the output noise. In addition, the RIVBJ algorithm includes explicit, optimal pre-filtering of the input and output data which attenuates noise outside the bandwidth of the system model and pre-whitens noise within the bandwidth [Young, 2008]. It is interesting, therefore, that the estimates are very similar in both cases, and this increases confidence in the errors-in-variables estimation procedures and the RIVBJ estimated Box-Jenkins model parameters in Table 1.

REFERENCES

M. Ekman. *Modeling and Control of Bilinear Systems*. PhD thesis, Uppsala University, 2005.

G. H. Golub and V. Pereyra. The differentiation of pseudo-inverses and nonlinear least squares problems whose variables are separate. *Siam J. Numer. Anal.*, 10(2): 413–432, 1973.

G. H. Golub and V. Pereyra. Separable Nonlinear Least Squares: the Variable Projection Method and its Applications. Technical Report SCCM-02-07, Stanford University, Stanford, USA, 2002.

L. Ljung. *System Identification - Theory for the user*. PTR Prentice Hall Information and System Sciences Series. Prentice Hall, New Jersey, 2nd edition, 1999.

T. Söderström. Errors-in-variables methods in system identification. *Automatica*, 43(6):939–958, 2007.

T. Söderström. Extending the Frisch scheme for errors-in-variables identification to correlated output noise. *Int. J. of Adaptive Control and Signal Proc.*, 22(1):55–73, 2008.

P. C. Young. *Recursive Estimation and Time Series Analysis*. Springer-Verlag, Berlin, 1984.

P. C. Young. Advances in real-time flood forecasting. *Phil. Trans. R. Soc. Lond.*, 360:1433–1450, 2002.

P. C. Young. The refined instrumental variable method: unified estimation of discrete and continuous-time transfer function models. *Journal Européen des Systèmes Automatisés*, 42:149–179, 2008.

P. C. Young. *The Flood Management Handbook (in press)*, chapter Real-Time Updating in Flood Forecasting and Warning. Wiley-Blackwell, 2009.

Appendix A. COMPUTATION OF CORRELATION ENTRIES

Diagonal auto-correlation terms of u_k:

$$r_u(0) = E\left[u_{0_k}^2 \tilde{u}_k^2\right] = r_{u_0}(0) r_{\tilde{u}}(0) \tag{A.1}$$

Off-diagonal auto-correlation terms of u_k:

$$r_u(\tau) = E\left[u_{0_k} \tilde{u}_k u_{0_{k-\tau}} \tilde{u}_{k-\tau}\right] = r_{u_0}(\tau) \tag{A.2}$$

Cross-correlation terms of u_k and y_k:

$$\begin{aligned} r_{uy}(\tau) &= E\left[u_{0_k} \tilde{u}_k \left(y_{0_{k-\tau}} + \tilde{y}_{k-\tau}\right)\right] \\ &= E\left[u_{0_k} \tilde{u}_k y_{0_{k-\tau}}\right] + E\left[u_{0_k} \tilde{u}_k \tilde{y}_{k-\tau}\right] \\ &= r_{u_0 y_0}(\tau) \end{aligned} \tag{A.3}$$