Hydrology and Earth System Sciences, 9, 209–224, 2005
www.copernicus.org/EGU/hess/hess/9/209/
SRef-ID: 1607-7938/hess/2005-9-209
European Geosciences Union

Hydrology and Earth System Sciences

Using a spatio-temporal dynamic state-space model with the EM algorithm to patch gaps in daily riverflow series

B. A. Amisigo[1,2] **and N. C. van de Giesen**[2]

[1]Center for Development Research, Bonn University, Bonn, Germany
[2]Delft University of Technology, Delft, Netherlands

Received: 2 March 2005 – Published in Hydrology and Earth System Sciences Discussions: 6 April 2005
Revised: 29 June 2005 – Accepted: 30 August 2005 – Published: 20 September 2005

Abstract. A spatio-temporal linear dynamic model has been developed for patching short gaps in daily river runoff series. The model was cast in a state-space form in which the state variable was estimated using the Kalman smoother (RTS smoother). The EM algorithm was used to concurrently estimate both parameter and missing runoff values. Application of the model to daily runoff series in the Volta Basin of West Africa showed that the model was capable of providing good estimates of missing runoff values at a gauging station from the remaining time series at the station and at spatially correlated stations in the same sub-basin.

1 Introduction

A major requirement for the assessment, development and sustainable use of the water resources of any river basin is the availability of good quality runoff series of sufficiently long duration. In the Volta Basin (Fig. 1) both daily and monthly river discharge series exist for a good number of gauging stations. However, many of these records are of poor quality and contain gaps, from several days to several years.

In an assessment of monthly flow series of river discharge from the main river gauging stations in the basin, Taylor (2003) observed that in general 20% of monthly discharge data over a 20-year period are missing from the available series in the basin with some gauges having as many as 50% gaps in their series. By regressing rainfall with the various series, the above study determined that only half of the gauging stations examined had reliable flow series, though these also had gaps of varying lengths. Filling gaps in existing river flow series is necessary for the design of water management plans that depend on complete water balances. In addition, a full series greatly facilitates data-driven model development.

Correspondence to: N. C. van de Giesen
(n.c.vandegiesen@citg.tudelft.nl)

Several methods are available for filling gaps in data series in general and for hydrological data series in particular. Data imputation methods (Dempster et al., 1977; Schafer, 1997; Little and Rubin, 2003) are generally difficult to apply to hydrological data such as monthly river runoff series because of autocorrelations at high lags and seasonal effects. A few reviews are available (Kottegoda and Elgy, 1979; Gyau-Boakye and Schultz, 1994) on methods that have been used successfully for hydrological data infilling. Gyau-Boakye and Schultz (1994) provided a framework for filling in gaps of various lengths in daily runoff series in West Africa including the Volta Basin. Among methods they recommended for such data in-filling are autoregression with or without rainfall, simple and multiple regressions with neighbouring gauges, interpolation, recession methods and linear storage model formulations – the method used depending, among others, on the length of the data gaps to be filled and the season in which these gaps occur. Papadakis et al. (1993) have also demonstrated the strength of satellite imagery with non-linear modelling in stream flow generation. Taylor (2003) used the Thornthwaite-Mather (TM) method to model river runoff in the Volta Basin and concluded that the method could be used to fill gaps in runoff series if properly formulated.

Autoregressions without rainfall, simple and multiple linear regressions with neighbouring gauges, interpolation and recession methods of data in-filling, where they work, have the advantages of simplicity and not requiring any other input data such as rainfall, evaportranspiration, soil moisture status etc, that other models need. However, they still need separate complete and extensive data sets for their calibration and verification, a requirement that may be a luxury in data-poor areas such as the Volta Basin. Recession and autoregressive methods are unsuitable for periods of flow with rainfall as they ignore the effect of the driving rainfall input. More suitable are multiple regressions that account for rainfall input and catchment moisture status by using runoff from

© 2005 Author(s). This work is licensed under a Creative Commons License.

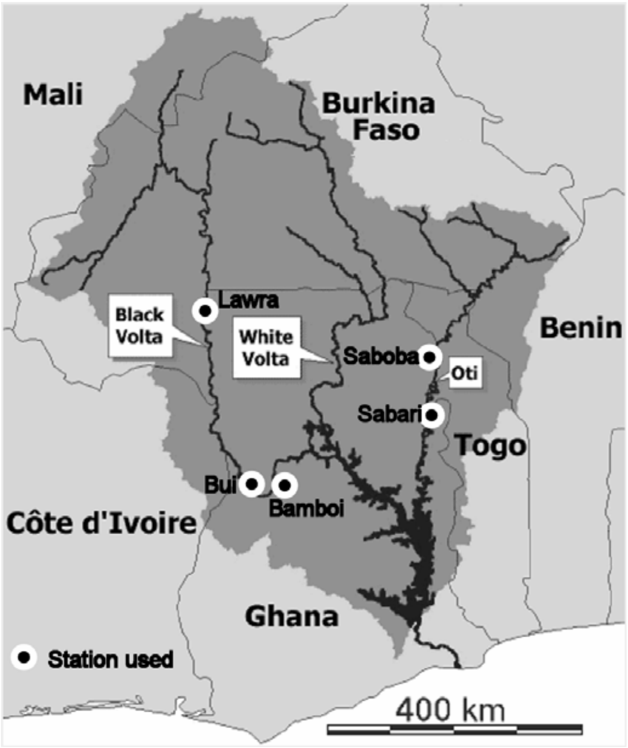

Fig. 1. Map of the Volta Basin showing the Gauging Stations used in the Study.

neighbouring gauges. Multiple regressions are, therefore, more suitable for data in-filling when there is rainfall. However, by being fitted to "global" data, consisting of runoff series for several years together, they may not provide very good estimates for short "local" gaps. Multiple regressions also ignore available observed runoff at the station for the period considered. Because not all available information is used, the quality of the estimates will be sub-optimal.

The gap-filling method proposed here makes good use of all available spatial and temporal information simultaneously. Spatio-temporal dynamic models have been applied successfully to environmental systems (Shumway and Stoffer, 1982; Hasket, 1989; Rouhani and Myers, 1990; Goodall and Mardia, 1994; Guttorp and Sampson, 1994; Mardia et al., 1998). When cast in state-space form, they can be used with the Kalman smoother and the Expectation-Maximisation (EM) algorithm to estimate missing values in environmental data including runoff data. A full Bayesian approach would be the preferred approach but this requires knowledge of prior distributions, and subsequent estimation of all probability density functions involved. Here, we choose to make some reasonable simplifications. Specifically, we consider the dynamic system to be linear with Gaussian errors and initial system state. The important advantage is that with these assumptions, the full Bayesian approach reduces to the Kalman filter and smoother. The Kalman filter/smoother simplifies the analysis to the estimation of the mean and covariance of the system states instead of the estimation of the conditional probabilities of the states at each time step. We have chosen to use this simplification due to the difficulty of specifying an appropriate prior distribution for the system states (or surrogate states, such as catchment daily runoff) in our analysis. The combined use of Kalman smoother and EM algorithm has the advantage of not needing separate calibration data. The use of the EM algorithm ensures that model parameters, missing observations and state vectors can be adequately estimated with the same data set and concurrently. Another advantage is that available measured streamflow data within the modelling period are also used in the estimation process thereby ensuring that important information contained in these data is fully used.

In this study, the spatio-temporal state space dynamic model with time invariant parameters was formulated using the Kalman smoother and the EM algorithm and applied to filling short gaps in riverflow data in the Volta Basin of West Africa using series of up to one year. The gaps in these series span a few days to a month. A typical daily time series with gaps is given in Fig. 2. The reason to use the spatio-temporal modeling framework with the Kalman smoother and EM algorithm is twofold. The first reason is to introduce the framework as a powerful tool for data assimilation in hydrology. The second reason is to ascertain the applicability and effectiveness of such a model to runoff patching in the Volta basin, where the authors currently conduct part of their hydrological research. A similar combination of EM and Kalman filter was originally developed for general ecological and environmental models (Xu and Wikle, 2004). Besides providing a novel and efficient procedure to fill gaps in hydrological time series, we believe that the proposed modelling framework will have many hydrological applications whenever both spatial and temporal information is available to estimate states and parameters. In this light, the present study can be seen as an introduction of this technique to hydrological problems. Therefore, some extra room is given to the mathematical development to facilitate further applications.

The presentation of the methodology is divided over three parts. First, we present the "Discrete spatial-temporal dynamic modelling framework", which is based on the Expectation-Maximisation (EM) algorithm, the core algorithm used here. Second, the framework is adapted to address the issue of missing values. Thirdly, the EM-algorithm is generalized to allow for the inclusion of constraints and prior knowledge of model structure and parameters (Xu and Wikle, 2004). This complete framework is subsequently applied to fill gaps in runoff series. The success of this application is shown and discussed. The article ends with the conclusions drawn from the general framework and its application.

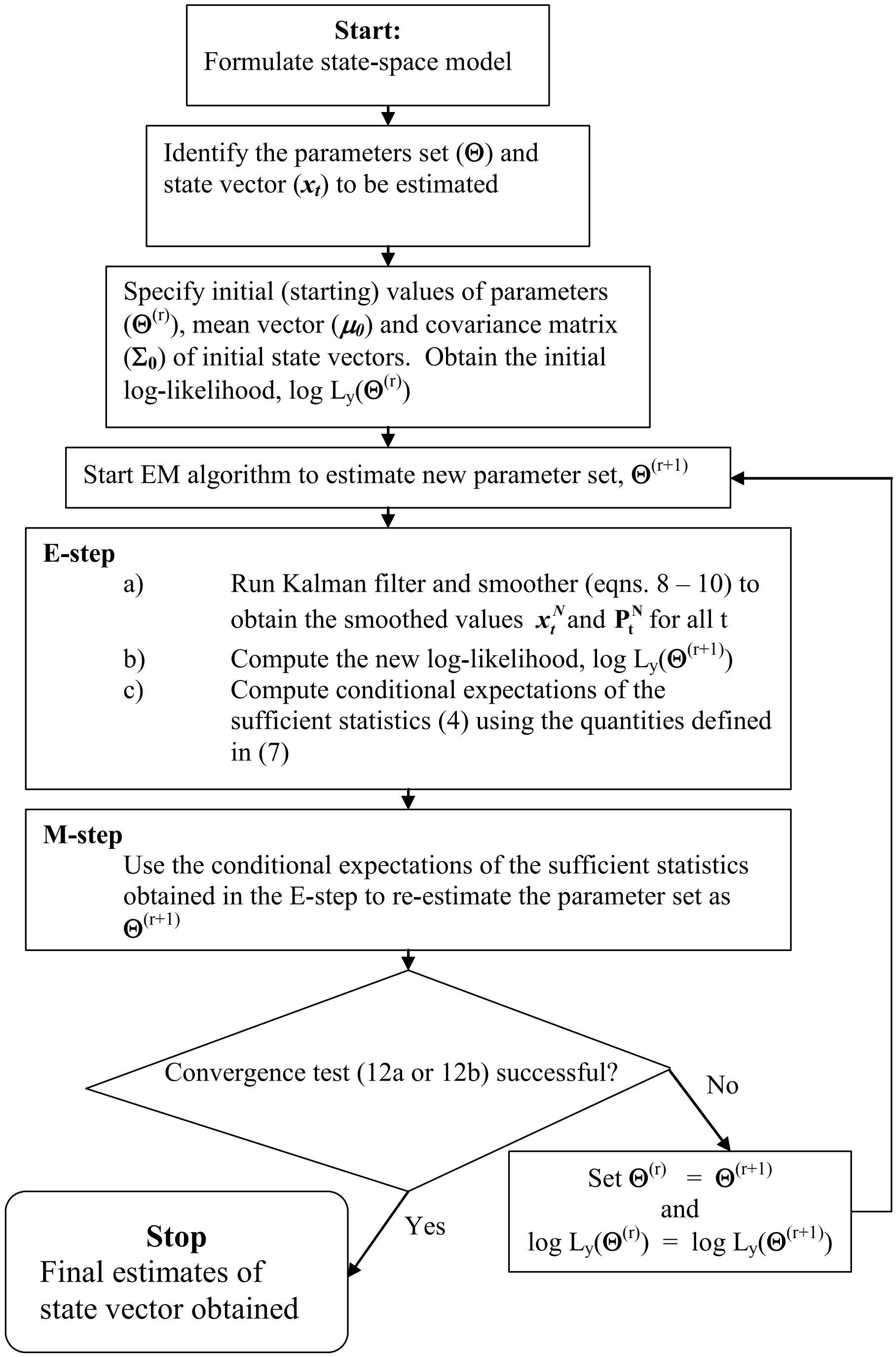

Fig. 2a. Flow diagram spatio-temporal framework.

2 The discrete spatio-temporal dynamic modelling framework

The first step in the development of the framework is to represent the system at hand as a set of discrete updating equations that represent the evolution of the system state and the associated observations. Subsequently, the EM algorithm is presented as an iteration scheme with each iteration consisting of an Expectation and a Maximization step. As an end result, EM produces the set of system parameters with the highest likelihood, given the complete set of observations. For implementation of EM, one needs to calculate for each iteration the so-called "sufficient statistics", which will be defined below. Because the problem was formulated as a

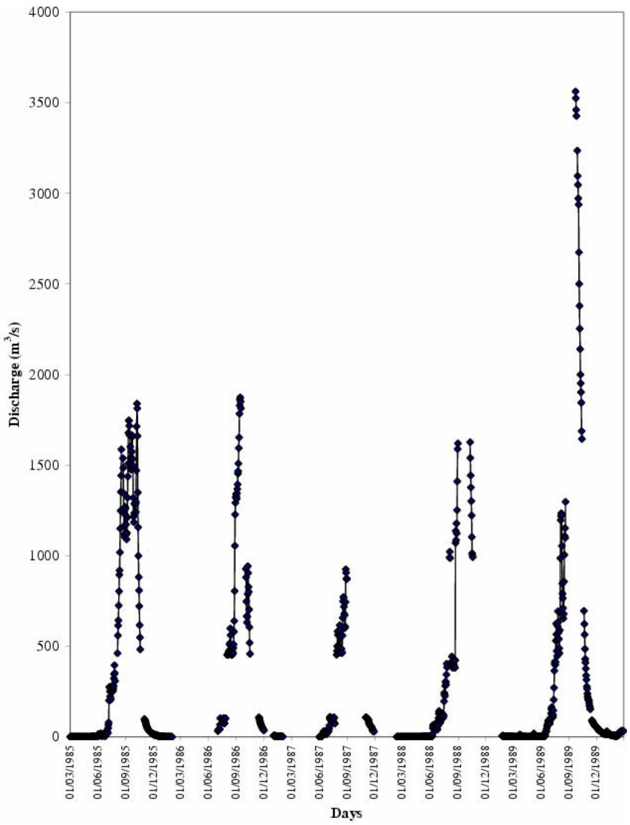

Fig. 2b. Daily Hydrograph at Sabari on the Oti River (1985–1989). Typical pattern of Gaps considered in the Spatio-temporal Modelling.

discrete state-space model, these "sufficient statistics" can be calculated with a Kalman smoother. The presentation of the Kalman smoother, consisting of the normal Kalman filter, first run in forward mode and then in backward or smoothing mode, concludes this section. The steps involved in the modeling framework are detailed below and also summarised in the chart in Fig. 2a.

2.1 State-space representation of the dynamics of the system's state

The discrete spatio-temporal dynamic model is formulated to predict a $n\times 1$ state vector $\boldsymbol{x_t}=(x_{\alpha 1}(t), x_{\alpha 2}(t), \ldots, x_{\alpha n}(t))'$ of an unobserved spatio-temporal state process at a fixed network of n locations. In addition, there is the $m\times 1$ vector $\boldsymbol{y}_t=(y_{\beta 1}(\mathrm{t}), y_{\beta 2}(\mathrm{t}), \ldots, y_{\beta m}(\mathrm{t}))'$ of observed or measured values at m locations at time t, where the two sets of spatial locations α and β need not be the same (Xu and Wikle, 2004). Here, and throughout, the prime denotes the transpose vector or matrix. Bold small letters indicate vectors and bold capitals are matrices. The state-space representation for the prediction of the unobserved process without external input and for the linear dynamic case with time invariant parameters, consists of the following process/state and measurement equations:

$$\boldsymbol{x_t} = \boldsymbol{F}\boldsymbol{x}_{t-1}+\boldsymbol{\omega}_t, \quad \boldsymbol{x_0} \sim N(\boldsymbol{\mu_0}, \boldsymbol{\Sigma_0}), \quad \boldsymbol{\omega} \sim N(\boldsymbol{0}, \boldsymbol{Q}) \tag{1a}$$

$$\boldsymbol{y_t} = \boldsymbol{H}\boldsymbol{x}_t + \boldsymbol{\upsilon}_t, \quad \boldsymbol{\upsilon} \sim N(\boldsymbol{0}, \boldsymbol{R}) \tag{1b}$$

F is the $n\times n$ transition or state propagation matrix that describes the dynamics of the system – a first-order Markov process in which the current state (x_t) depends on only the immediate past state (x_{t-1}). **H** is the $m\times n$ measurement matrix that relates the estimated state vectors to the vector of actual observations. The additive $n\times 1$ state-estimation errors, $\boldsymbol{\omega_t}$, and the $m\times 1$ measurement errors $\boldsymbol{\upsilon_t}$ are uncorrelated Gaussian white noises with zero mean and covariance **Q** $(n\times n)$ and **R** $(m\times m)$. The initial $n\times 1$ state vector, $\boldsymbol{x}_0$, is considered normally distributed with mean $\boldsymbol{\mu_0}$ and covariance $\boldsymbol{\Sigma_0}$. Equation (1) is the finite dimensional linear dynamical system from which the vector time series of observations $\mathbf{Y}=(\boldsymbol{y_0}, \boldsymbol{y_1}, \boldsymbol{y_2}, \ldots, \boldsymbol{y_N})$ is assumed to be generated. As noted in Ribeiro (2004), when $\boldsymbol{x_0}$ is a Gaussian vector, $\boldsymbol{\omega_t}$ and $\boldsymbol{\upsilon_t}$ Gaussian white noises and when the state and observation dynamics are linear, the conditional probability density function $p(\boldsymbol{x_t}|\boldsymbol{Y})$, is normally distributed with mean $\boldsymbol{x_t^N}$ and covariance $\boldsymbol{P_t^N}$, i.e., $p(\boldsymbol{x_t}|\boldsymbol{Y}) \sim N\left(\boldsymbol{x_t^N}, \boldsymbol{P_t^N}\right)$. The conditional mean of this Gaussian pdf is equivalent to the estimate $\boldsymbol{x}_t^N$ of the state $\boldsymbol{x}_t$ given the N observations at each of the m sites. The covariance (also called "dispersion") matrix $\boldsymbol{P_t^N}$ quantifies the uncertainty of the state estimate given the same N observations. In our specific case there is no need to estimate the uncertainty associated with the state-space model parameters because they are not of hydrological interest. If, however, one would need to know more about the uncertainty of state-space model parameters one could consider them as part of the state vector. Alternatively, one could use Markov Chain Monte Carlo simulations (Wang, 2001). Note that $\boldsymbol{x_t}$ without any superscript refers to actual, unobserved values while superscripted variables such as $\boldsymbol{x_t^N}$ refer to simulated or estimated values; $\boldsymbol{y_t}$ always refers to actual observations.

2.2 Parameter estimation

In general, some or all of the system parameters $\boldsymbol{\Theta}$={**F, H, Q, R,** $\boldsymbol{\mu_0}$, $\boldsymbol{\Sigma_0}$} are not known and would have to be estimated from the observations. This is a system identification problem and, in the Gaussian framework under consideration, the parameter estimation can be undertaken by the method of maximum likelihood. The maximum likelihood estimate of $\boldsymbol{\Theta}$ given $\mathbf{X}=(\boldsymbol{x_0}, \boldsymbol{x_1}, \boldsymbol{x_2}, \ldots, \boldsymbol{x_N})$ and $\mathbf{Y}=(\boldsymbol{y_0}, \boldsymbol{y_1}, \boldsymbol{y_2}, \ldots, \boldsymbol{y_N})$ is obtained by maximizing the joint log-likelihood of **X**, **Y**, and $\boldsymbol{\Theta}$ with respect to $\boldsymbol{\Theta}$. This log-likelihood function is given as (Shumway and Stoffer, 1982):

$$\log L_Y(\boldsymbol{\Theta}) = \log L(\boldsymbol{X}, \boldsymbol{Y}, \boldsymbol{\Theta}) = -\frac{1}{2}\left\{\begin{array}{l} \log|\boldsymbol{\Sigma_0}| + (\boldsymbol{x_0}-\boldsymbol{\mu_0})' \boldsymbol{\Sigma_0^{-1}} (\boldsymbol{x_0}-\boldsymbol{\mu_0}) \\ +N\log|\boldsymbol{Q}| + \sum_{t=1}^{N}(\boldsymbol{x}_t-\boldsymbol{F}\boldsymbol{x}_{t-1})' \boldsymbol{Q}^{-1}(\boldsymbol{x}_t-\boldsymbol{F}\boldsymbol{x}_{t-1}) \\ +N\log|\boldsymbol{R}| + \sum_{t=1}^{N}(\boldsymbol{y}_t-\boldsymbol{H}\boldsymbol{x}_t)' \boldsymbol{R}^{-1}(\boldsymbol{y}_t-\boldsymbol{H}\boldsymbol{x}_t) \end{array}\right\} \quad (2)$$

When there are no constraints placed on the structure of the system matrices **F**, **H**, **Q**, and **R**, the estimates of the components of $\boldsymbol{\Theta}$ are (Digalakis et al., 1993; Xu and Wikle, 2004):

$$\hat{\boldsymbol{F}} = \boldsymbol{A_4}\boldsymbol{A_3^{-1}} \quad (3a)$$

$$\hat{\boldsymbol{H}} = \boldsymbol{A_6}\boldsymbol{A_1^{-1}} \quad (3b)$$

$$\hat{\boldsymbol{Q}} = \boldsymbol{A_2}-\boldsymbol{A_4}\boldsymbol{A_3^{-1}}\boldsymbol{A_4'} \quad (3c)$$

$$\hat{\boldsymbol{R}} = \boldsymbol{A_5} - \boldsymbol{A_6}\boldsymbol{A_1^{-1}}\boldsymbol{A_6'} \quad (3d)$$

$$\boldsymbol{\mu_0} = \boldsymbol{x_0} \quad (3e)$$

$$\boldsymbol{\Sigma_0} = P_0 \quad (3f)$$

Where:

$$\boldsymbol{A_1} = \frac{1}{N+1}\sum_{t=0}^{N}\boldsymbol{x}_t\boldsymbol{x}_t' \quad (4a)$$

$$\boldsymbol{A_2} = \frac{1}{N}\sum_{t=1}^{N}\boldsymbol{x}_t\boldsymbol{x}_t' \quad (4b)$$

$$\boldsymbol{A_3} = \frac{1}{N}\sum_{t=1}^{N}\boldsymbol{x}_{t-1}\boldsymbol{x}_{t-1}' \quad (4c)$$

$$\boldsymbol{A_4} = \frac{1}{N}\sum_{t=1}^{N}\boldsymbol{x}_t\boldsymbol{x}_{t-1}' \quad (4d)$$

$$\boldsymbol{A_5} = \frac{1}{N+1}\sum_{t=0}^{N}\boldsymbol{y}_t\boldsymbol{y}_t' \quad (4e)$$

$$\boldsymbol{A_6} = \frac{1}{N+1}\sum_{t=0}^{N}\boldsymbol{y}_t\boldsymbol{x}_t' \quad (4f)$$

are known as the sufficient statistics.

2.3 The E-M algorithm for parameter estimation

In the state-space formulation of interest, the set of state vectors $\mathbf{X}$ $(N\times n)$ are not observed directly and not available a priori for the computation of the sufficient statistics and hence the parameter estimates. In addition some of the observations in the set $\mathbf{Y}$ $(N\times m)$ may be missing. In these circumstances, the Expectation-Maximisation or EM algorithm (Dempster et al., 1977) has been found to be a powerful tool for the maximum likelihood estimation of the system parameters (Shumway and Stoffer, 1982; Digalkis et al., 1993; Ghahramani and Hinton, 1996; Bilmes, 1998; Xu and Wikle, 2004).

The EM algorithm is designed for parameter estimation of incomplete or missing data problems by the method of maximum likelihood (Dempster et al., 1997). By treating the state vector as missing observations, the problem is now the same as a problem with incomplete data, which justifies the use of the EM algorithm. Thus, both the recursions for the computations of the state vector, such as by the Kalman filter used here, and the estimation of the state-space model parameters can be undertaken concurrently – offline computations of the parameters is not necessary. These state-space model parameters are not hydrologically meaningful; they describe the evolution of the state vector. The EM procedure involves computing the time-invariant state-space model parameter set $\boldsymbol{\Theta}$ and then the state vector, for all time steps (batch mode), over and over again in a series of iterations until a set of convergence conditions is met. The computations for each iteration are carried out in two main steps, the E-step and the M-step.

Consider the $(r+1)^{th}$ iteration when $\boldsymbol{\Theta}^{(r)}$, the parameter set at the r^{th} iteration, is known and it is required to find $\boldsymbol{\Theta}^{(r+1)}$, the parameter set at the $(r+1)^{th}$ iteration. In the E-step of the EM algorithm, the expected value of the complete-data log-likelihood log p($\mathbf{X},\mathbf{Y}|\boldsymbol{\Theta}$) with respect to the unobserved and missing data $\mathbf{X}$, given the observations $\mathbf{Y}$ and the current parameter estimates $\boldsymbol{\Theta}^{(r)}$, is evaluated. This expectation is defined as:

$$\begin{aligned} G\left(\boldsymbol{\Theta}^{(r+1)}\right) &= G\left(\boldsymbol{\Theta}^{(r+1)}, \boldsymbol{\Theta}^{(r)}\right) \\ &= E\left[\log p(\boldsymbol{X}, \boldsymbol{Y}|\boldsymbol{\Theta}^{(r+1)})|\boldsymbol{Y}, \boldsymbol{\Theta}^{(r)}\right] \\ &= E[L(\mathbf{X}, \mathbf{Y}, \boldsymbol{\Theta}^{(r+1)}|\boldsymbol{Y}, \boldsymbol{\Theta}^{(r)}] \end{aligned} \quad (5)$$

The expectation is evaluated with the known parameter set $\boldsymbol{\Theta}^{(r)}$ (the new parameter set $\boldsymbol{\Theta}^{(r+1)}$ is obtained by optimising G in the M step). The set $\mathbf{Y}$, excluding any missing observations, constitutes the incomplete data set, while the full data set ($\mathbf{X}$, $\mathbf{Y}$) contains both observed and missing data.

In the M-step, $\boldsymbol{\Theta}^{(r+1)}$, the new estimate of $\boldsymbol{\Theta}$, is computed by maximising the conditional expectation evaluated in the E-step. That is,

$$\boldsymbol{\Theta}^{(r+1)} = \underset{\Theta}{\operatorname{argmax}}\, G\left(\boldsymbol{\Theta}^{(r+1)}\right) \quad (6)$$

The log-likelihood is guaranteed to increase with each iteration and the algorithm is guaranteed to converge to at least a local minimum (Dempster, 1977; Bilmes, 1998).

For the regular exponential distributions (e.g. the normal, binomial, poisson, gamma distributions), the E-step of the EM algorithm consists of the computation of the conditional expectations of the complete data sufficient statistics, as given under Eq. (4) (Dempster, 1977). In the M-step, these conditional expectations of the complete-data sufficient statistics are then used instead of the (unknown) actual complete-data sufficient statistics. Thus, the following quantities are computed in the E-step and used to evaluate the sufficient statistics in Eq. (4), (Digalakis, 1993):

$$E\left\{\boldsymbol{x}_t|\boldsymbol{Y},\boldsymbol{\Theta}^{(r)}\right\}=\boldsymbol{x}_t^N \tag{7a}$$

$$E\left\{\boldsymbol{x}_t\boldsymbol{x}_t'|\boldsymbol{Y},\boldsymbol{\Theta}^{(r)}\right\}=\boldsymbol{P}_t{}^N+\boldsymbol{x}_t{}^N\left(\boldsymbol{x}_t{}^N\right)' \tag{7b}$$

$$E\left\{\boldsymbol{x}_t\boldsymbol{x}_{t-1}'|\boldsymbol{Y},\boldsymbol{\Theta}^{(r)}\right\}=E\left\{\left(\boldsymbol{x}_t-\boldsymbol{x}_t{}^N\right)\left(\boldsymbol{x}_{t-1}-\boldsymbol{x}_{t-1}{}^N\right)'|\boldsymbol{Y}\right\}$$
$$+\boldsymbol{x}_t{}^N\left(\boldsymbol{x}_{t-1}{}^N\right)'=\boldsymbol{P}_{t,t-1}^N+\boldsymbol{x}_t{}^N\left(\boldsymbol{x}_{t-1}{}^N\right)' \tag{7c}$$

$$E\left\{\boldsymbol{y}_t\boldsymbol{x}_t{}'|\boldsymbol{Y},\boldsymbol{\Theta}^{(r)}\right\}=\boldsymbol{y}_t E\left\{\boldsymbol{x}_t{}'|\boldsymbol{Y},\boldsymbol{\Theta}^{(r)}\right\}=\boldsymbol{y}_t\left(\boldsymbol{x}_t{}^N\right)' \tag{7d}$$

$$E\left\{\boldsymbol{y}_t\boldsymbol{y}_t{}'|\boldsymbol{Y},\boldsymbol{\Theta}^{(r)}\right\}=|\boldsymbol{y}_t\boldsymbol{y}_t{}',\ \text{(no missing values in } \boldsymbol{Y}) \tag{7e}$$

2.4 Kalman smoother for the computation of the conditional expectations of the complete-data sufficient statistics

As it turns out, all expectations in Eqs. (7a–d) above at iteration r+1 can be computed for all t=1, 2, … , N from the fixed interval Kalman smoother, also known as Rauch-Tung-Striebel or RTS smoother (Haykin, 2001), using the parameter estimates obtained at iteration r. The smoother consists of a forward and a backward pass, given as (Haykin, 2001; Xu and Wikle, 2004):

Filter equation – Forward pass (t=1, 2, …, N)

a. Prediction:

$$\boldsymbol{x}_t^{t-1}=\boldsymbol{F}\boldsymbol{x}_{t-1}^{t-1} \tag{8a}$$
$$\boldsymbol{P}_t^{t-1}=\boldsymbol{F}\boldsymbol{P}_{t-1}^{t-1}\boldsymbol{F}'+\boldsymbol{Q} \tag{8b}$$

with

$$\boldsymbol{x}_0^0=\boldsymbol{\mu}_0 \text{ and } \boldsymbol{P}_0^0=\boldsymbol{\Sigma}_0 \tag{8c}$$

b. Update (Filter):

$$\boldsymbol{e}_t=\boldsymbol{y}_t-\boldsymbol{H}\boldsymbol{x}_t^{t-1} \tag{9a}$$
$$\boldsymbol{\Sigma}_{et}=\boldsymbol{H}\boldsymbol{P}_t^{t-1}\boldsymbol{H}'+\boldsymbol{R} \tag{9b}$$
$$\boldsymbol{K}_t=\boldsymbol{P}_t^{t-1}\boldsymbol{H}'\boldsymbol{\Sigma}_{et}^{-1} \tag{9c}$$
$$\boldsymbol{x}_t^t=\boldsymbol{x}_t^{t-1}+\boldsymbol{K}_t\boldsymbol{e}_t \tag{9d}$$
$$\boldsymbol{P}_t^t=\boldsymbol{P}_t^{t-1}-\boldsymbol{K}_t\boldsymbol{H}\boldsymbol{P}_t^{t-1} \tag{9e}$$

Smoothing – Backward pass (t=N, N-1, …,1)

$$\boldsymbol{x}_N^N=x_t^t,\quad \boldsymbol{P}_N^N=\boldsymbol{P}_t^t,\quad t=N \text{ only} \tag{10a}$$
$$\boldsymbol{J}_{t-1}=\boldsymbol{P}_{t-1}^{t-1}\boldsymbol{F}'\left(\boldsymbol{P}_t^{t-1}\right)^{-1} \tag{10b}$$
$$\boldsymbol{x}_{t-1}{}^N=\boldsymbol{x}_{t-1}^{t-1}+\boldsymbol{J}_{t-1}\left(\boldsymbol{x}_t{}^N-\boldsymbol{x}_t{}^{t-1}\right) \tag{10c}$$
$$\boldsymbol{P}_{t-1}^N=\boldsymbol{P}_{t-1}^{t-1}+\boldsymbol{J}_{t-1}\left(\boldsymbol{P}_t^N-\boldsymbol{P}_t^{t-1}\right)\boldsymbol{J}_{t-1}' \tag{10d}$$

Smoothed lag-one covariance (t=N, N-1, …, 2):

$$\boldsymbol{P}_{N,N-1}^N=(\boldsymbol{I}-\boldsymbol{K}_N\boldsymbol{H})\boldsymbol{F}\boldsymbol{P}_{N-1}^{N-1},\ t=N \text{ only} \tag{10e}$$
$$\boldsymbol{P}_{t-1,t-1}^N=\boldsymbol{P}_{t-1}^{t-1}\boldsymbol{J}_{t-2}+\boldsymbol{J}_{t-1}\left(\boldsymbol{P}_{t,t-1}^N-\boldsymbol{F}\boldsymbol{P}_{t-1}^{t-1}\right)\boldsymbol{J}_{t-2}' \tag{10f}$$

In the above equations, $\boldsymbol{x}_t^{t-1}, \boldsymbol{x}_t^t, \boldsymbol{x}_t^N$ and $\boldsymbol{P}_t^{t-1}, \boldsymbol{P}_t^t, \boldsymbol{P}_t^N$ are the predicted, filtered (updated) and smoothed values respectively, of the state vector $\boldsymbol{x}_t$ and its covariance $\boldsymbol{P}_t$. The values of interest are the smoothed values,$\boldsymbol{x}_t^N$ and $\boldsymbol{P}_t^N$, which are inserted in the right-hand side of Eqs. (7a–e) to calculate the expected values needed for the M-step. The log-likelihood function can also be conveniently computed as a by-product of the Kalman filter as follows:

$$\log L_Y(\boldsymbol{\Theta})=-\frac{1}{2}\left\{\sum_{t=1}^{N}\log|\boldsymbol{\Sigma}_{et}|+\sum_{t=1}^{N}\boldsymbol{e}_t'\boldsymbol{\Sigma}_{et}^{-1}\boldsymbol{e}_t\right\} \tag{11}$$

where $\boldsymbol{e}_t$ and $\boldsymbol{\Sigma}_{et}$, the innovation vector and innovation covariance matrix, respectively, are computed for t=1, 2, …, N as in Eqs. (9a) and (9b).

The EM algorithm for the maximum likelihood estimation of the linear time invariant dynamic system represented in Eq. (1) can now be summarised as follows:

2.4.1 E-step at iteration r+1

(i) Use the parameter set $\boldsymbol{\Theta}^{(r)}$={$\mathbf{F}^{(r)}$, $\mathbf{H}^{(r)}$, $\mathbf{Q}^{(r)}$, $\mathbf{R}^{(r)}$, $\boldsymbol{\mu}_0{}^{(r)}$, $\boldsymbol{\Sigma}_0{}^{(r)}$} obtained from the previous iteration, r, and the Kalman smoother presented in Eqs. (8–10) to compute the statistics in Eq. (7). Also compute the log-likelihood, $L_Y(\boldsymbol{\Theta}^{(r+1)})$ using Eq. (11).
(ii) Use the statistics computed in (i) to compute the conditional expectations of the sufficient statistics in Eq. (4).

2.4.2 M-step at iteration r+1

Re-estimate (update) the parameter set as $\boldsymbol{\Theta}^{(r+1)}$={$\mathbf{F}^{(r+1)}$, $\mathrm{H}^{(r+1)}$, $\mathbf{Q}^{(r+1)}$, $\mathbf{R}^{(r+1)}$, $\boldsymbol{\mu}_0{}^{(r+1)}$, $\boldsymbol{\Sigma}_0{}^{(r+1)}$} using the relationships in Eq. (3), with the conditional expectations of the sufficient statistics calculated under E-ii in Eqs. (3a–d), $\boldsymbol{x}_0$ in (3e), and $\boldsymbol{P}_0$ in (3e).

2.4.3 Convergence

Test for the convergence of either the parameters or the log-likelihood, i.e., perform one of the following tests:

$$\left\|\boldsymbol{\Theta}^{(r+1)}-\boldsymbol{\Theta}^{(r)}\right\|<\varepsilon_\theta \tag{12a}$$

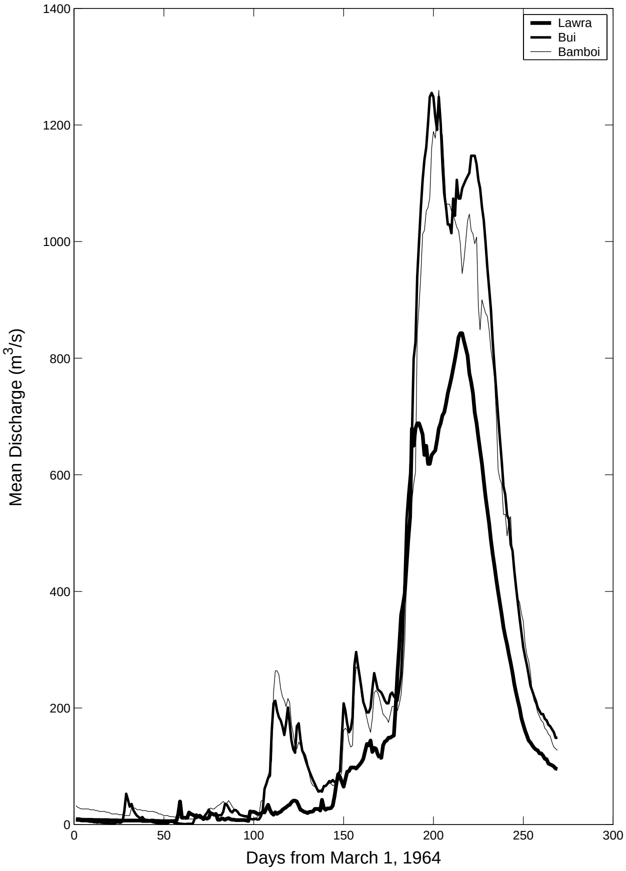

Fig. 3a. 1964 Daily Hydrographs for Lawra, Bui and Bamboi on the Black Volta River.

$$\left\| \log L_Y \left(\Theta^{(r+1)} \right) - \log L_Y \left(\Theta^{(r)} \right) \right\| < \varepsilon_L \tag{12b}$$

where we use the Euclidean norm and ε_θ and ε_L are sufficiently small positive numbers.

If the test succeeds, the iterations are stopped and $\Theta^{(r+1)}$ is retained as the final set of estimates of the system parameters, otherwise the iterations continue. Xu and Wikle (2004) prefer the use of the parameter values as the test criterion (Eq. 12a) to the use of the log-likelihood (Eq. 12b) as the log-likelihood can be unstable for spatio-temporal problems due to the high spatial correlations in the innovations as a result of processes at adjacent spatial locations being often very similar.

The "discrete spatio-temporal dynamic modelling framework" in its general form now stands. In the following, it is shown how the estimation of missing observations can readily be included in this framework. Before the EM algorithm is applied, it will be generalized to allow for inclusion of parameter constraints and knowledge of the structure of the physical processes (Xu and Wikle, 2004).

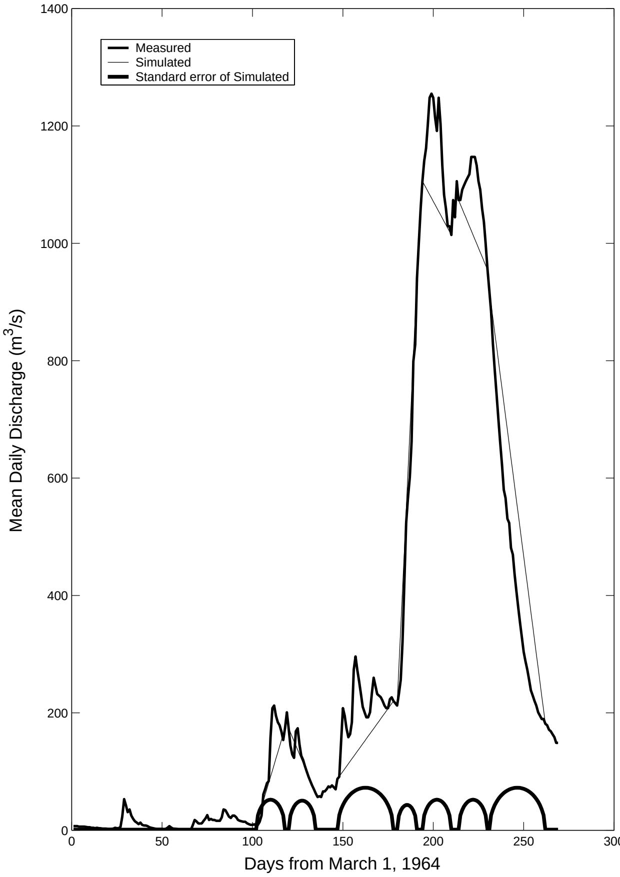

Fig. 3b. Measured and Simulated Daily Discharge for Bui on the Black Volta River (Black Volta Basin) – State Space Model on only Bui Discharge (1964).

3 Missing observations

If there are missing observations, then $\boldsymbol{y_t}$ and $\mathbf{H}$ in the update equations of the forward pass of the Kalman filter would have to be modified as follows:
(i) Replace all missing values in $\boldsymbol{y_t}$ in Eq. (9a) with zeroes.
(ii) Replace entries in the corresponding rows in $\mathbf{H}$ in Eq. (9a) with zeroes.

In addition the conditional expectations given in Eq. (7) would have to be modified as follows (Digalakis et al., 1993):

$$E\left\{\boldsymbol{y_t}|Y,\boldsymbol{\Theta}^{(r)}\right\} = \left\{\begin{array}{ll}\boldsymbol{y_t} & \text{, if observed} \\ \boldsymbol{H}^{(m)}E\left\{\boldsymbol{x_t}|Y,\boldsymbol{\Theta}^{(r)}\right\} & \text{, if missing}\end{array}\right\} \tag{13a}$$

$$E\left\{\boldsymbol{y_t y_t'}|Y,\boldsymbol{\Theta}^{(r)}\right\} = \left\{\begin{array}{ll}\boldsymbol{y_t y_t'}, & \text{if observed} \\ \boldsymbol{R}^{(m)}+\boldsymbol{H}^{(m)}E\left\{\boldsymbol{x_t x_t'}|Y,\boldsymbol{\Theta}^{(r)}\right\}\left(\boldsymbol{H}^{(m)}\right)', & \text{if missing}\end{array}\right\} \tag{13b}$$

$$E\left\{\boldsymbol{y_t x_t'}|Y,\boldsymbol{\Theta}^{(r)}\right\} = \left\{\begin{array}{ll}\boldsymbol{y_t}E\left\{\boldsymbol{x_t}|Y,\boldsymbol{\Theta}^{(r)}\right\}, & \text{if observed} \\ \boldsymbol{H}^{(m)}E\left\{\boldsymbol{x_t x_t'}|Y,\boldsymbol{\Theta}^{(r)}\right\}, & \text{if missing}\end{array}\right\} \tag{13c}$$

where $\mathbf{H^{(m)}}$ and $\mathbf{R^{(m)}}$ are the $\mathbf{H}$ and $\mathbf{R}$ matrices, respectively, corresponding to the missing values.

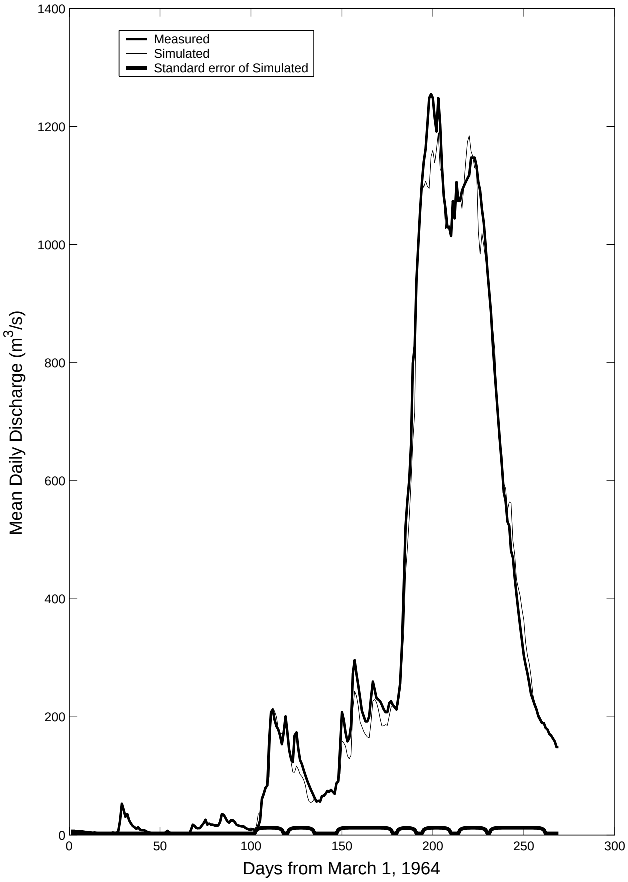

Fig. 3c. Measured and Simulated Daily Discharge for Bui on the Black Volta River (Black Volta Basin) – State Space Model on 1964 Bui, Lawra and Bamboi Discharges.

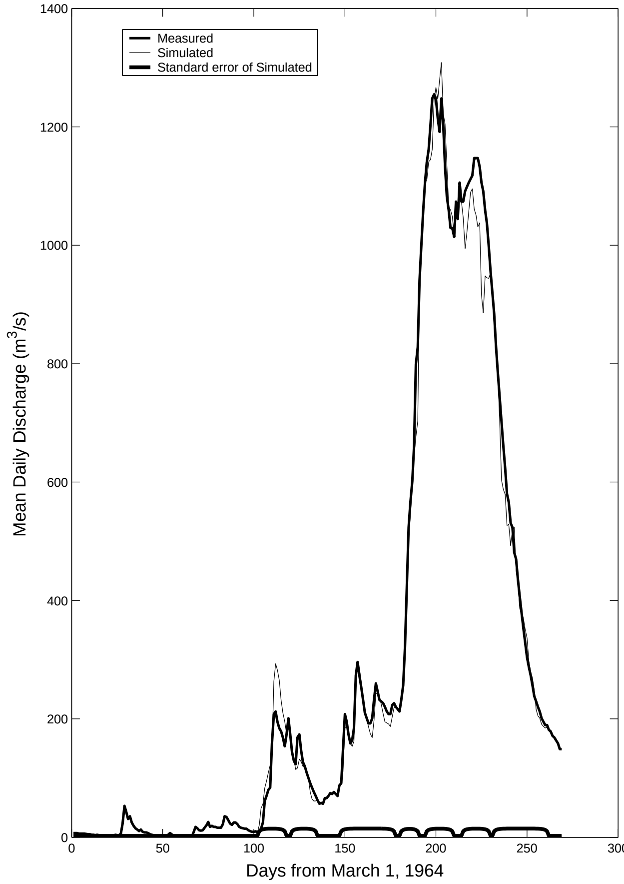

Fig. 3d. Measured and Simulated Daily Discharge for Bui on the Black Volta River (Black Volta Basin) – State Space Model on 1964 Bui and Bamboi Discharges.

4 System matrices parameterisation

Equation (3) is used in the M-step to update the parameter values when these parameters are not constrained or parameterised in any way. The parameters thus obtained are the maximum likelihood estimates. To avoid identifiability problems, some or all of the system matrices may be constrained or parameterised directly. In such cases, the parameters are no longer maximum likelihood estimates. However, the parameterisation can be undertaken in such a way as to still result in an increase in the log likelihood at each iteration and lead to a convergence of the parameters. The algorithm is then called the General EM (GEM) (Xu and Wikle, 2004). Most often, **Q** is constrained to a diagonal matrix while **R** is modelled as $\mathbf{R}=\sigma^2\mathbf{I_m}$, where $\mathbf{I_m}$ is an $m\times m$ identity matrix. The process matrix **F** can also be parameterised if its form and structure are known while **H** may be specified a priori as a design matrix and would no longer be updated in the M-step.

In GEM, the $(r+1)^{th}$ update formula for the general unconstrained **Q**, whether **F** is parameterised or not, is given as (Xu and Wikle, 2004):

$$\boldsymbol{Q}^{(r+1)} = \frac{1}{N}\boldsymbol{A}, \text{ where}$$
$$\boldsymbol{A} = \left(\boldsymbol{A_2} - \boldsymbol{A_4}\boldsymbol{F}' - \boldsymbol{F}\boldsymbol{A_4} + \boldsymbol{F}\boldsymbol{A_3}\boldsymbol{F}'\right) \tag{14a}$$

For the case when **F** is not parameterised and is estimated as in relation (3a), Eq. (14a) reduces to Eq. (3c). For a diagonal **Q** matrix, its M-step update is:

$$\text{diag}\left(\boldsymbol{Q}^{(\boldsymbol{r+1})}\right) = \frac{1}{N}\text{diag}\,(\boldsymbol{A}) \tag{14b}$$

where diag(**A**) is the diagonal vector of **A**.

When **R** is parameterised as $\mathbf{R}=\boldsymbol{\sigma}^2\mathbf{I_m}$, $\boldsymbol{\sigma}^2$ is updated in the M-step as follows:

$$\sigma^{2(r+1)} = \frac{1}{Nm}\sum_{t=1}^{N}\boldsymbol{tr}\left\{\left(y_t - \boldsymbol{H}\boldsymbol{x}_t^N\right)\left(y_t - \boldsymbol{H}\boldsymbol{x}_t^N\right)' + \boldsymbol{H}\boldsymbol{P}_t^N\boldsymbol{H}'\right\} \tag{14c}$$

The relations in Eq. (14) ensure the log likelihood increases monotonically even though the final parameter estimates would not be maximum likelihood estimates.

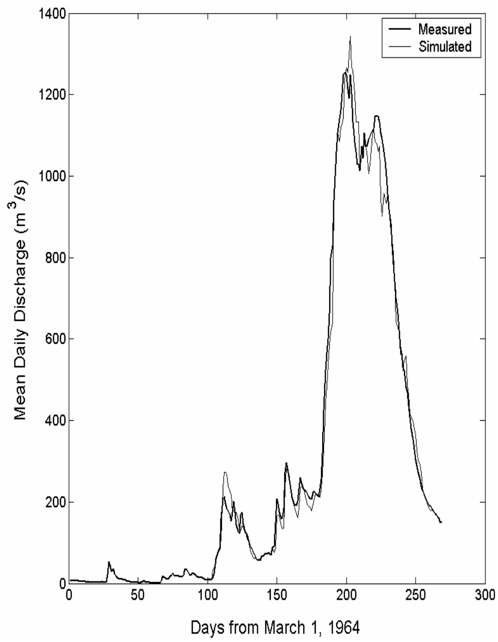

Fig. 3e. Measured and Simulated Daily Hydrographs for Bui on the Black Volta River (Black Volta Basin) – Linear Regression on 1964 Bamboi and Non-missing Bui Daily Discharges.

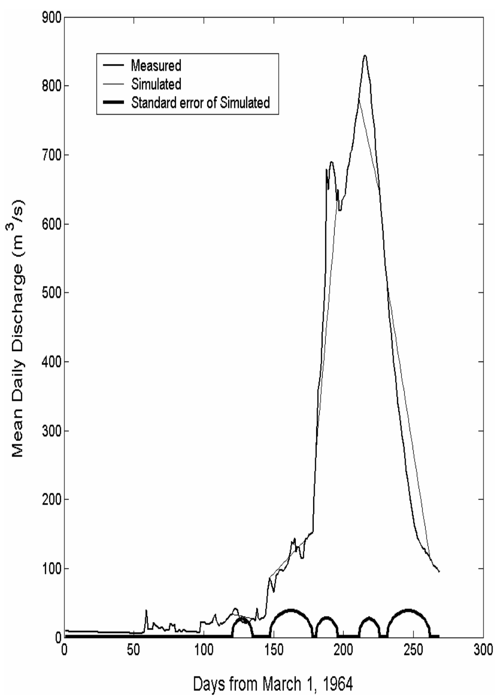

Fig. 3f. Measured and Simulated Daily Hydrographs for Lawra on the Black Volta River (Black Volta Basin) – State Space Model on only Lawra Daily Discharge – (Missing Data Pattern 1).

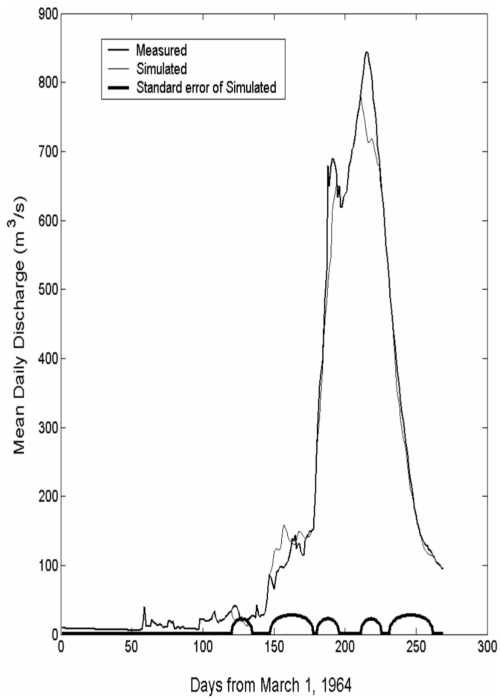

Fig. 3g. Measured and Simulated Daily Hydrographs for Lawra on the Black Volta River (Black Volta Basin) – State Space Model on Lawra and Bamboi Daily Discharges (Missing Data Pattern 1).

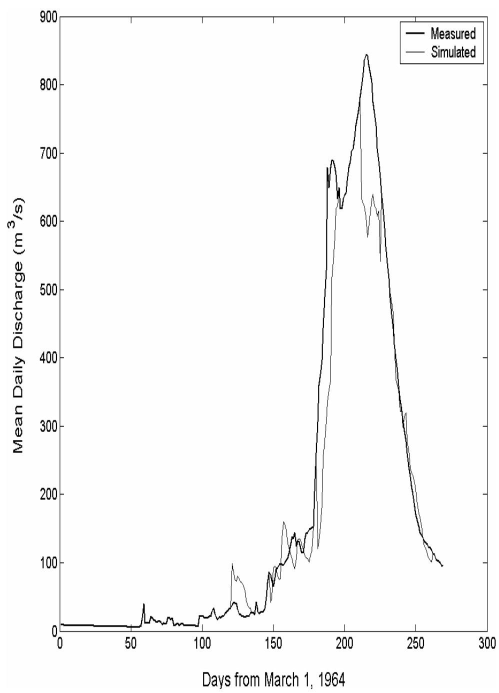

Fig. 3h. Measured and Simulated Daily Hydrographs for Lawra on the Black Volta River (Black Volta Basin) – Linear Regression on 1964 Bamboi and Non-missing Lawra Daily Discharges – (Missing Data Pattern 1).

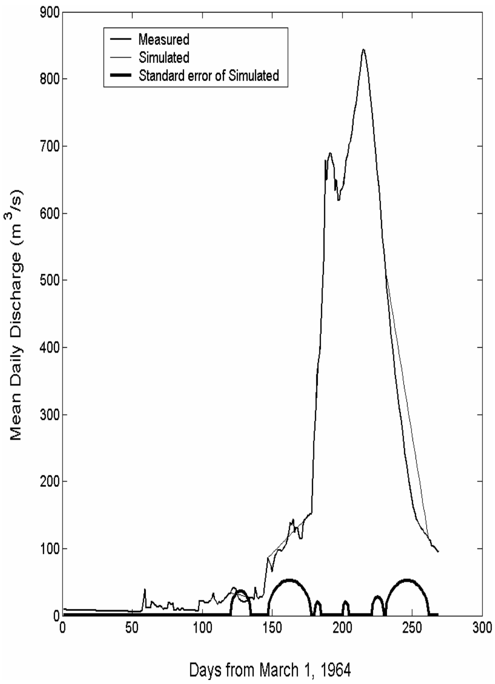

Fig. 3i. Measured and Simulated Daily Hydrographs for Lawra on the Black Volta River (Black Volta Basin) - State Space Model on only Lawra Daily Discharge – (Missing Data Pattern 2).

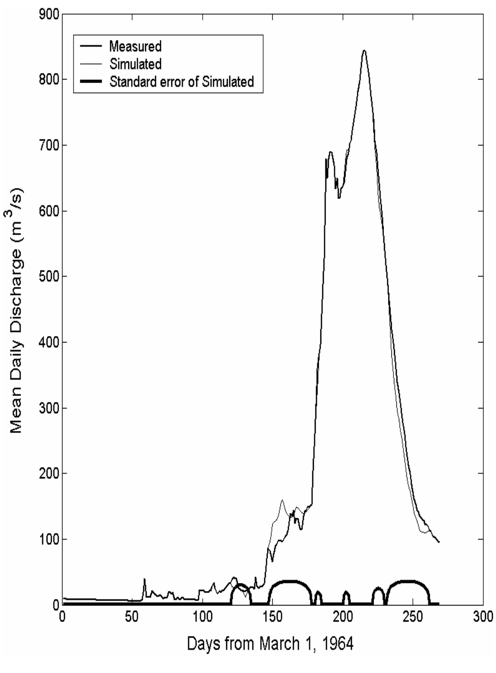

Fig. 3j. Measured and Simulated Daily Hydrographs for Lawra on the Black Volta River (Black Volta Basin) – State Space Model on Lawra and Bamboi Daily Discharges (Missing Data Pattern 2).

5 Application of the modeling framework

The methodology developed here was designed to enable the estimation of short gaps of a few days to one month in the

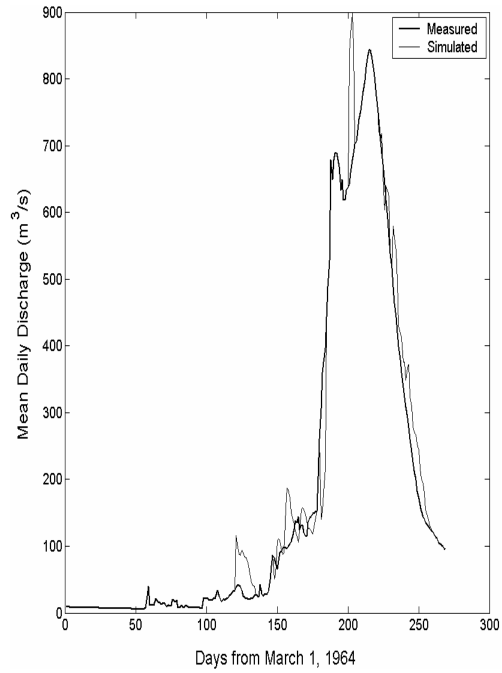

Fig. 3k. Measured and Simulated Daily Hydrographs for Lawra on the Black Volta River (Black Volta Basin) – Linear Regression on 1964 Bamboi and Non-missing Lawra Daily Discharges – (Missing Data Pattern 2).

annual daily runoff series (temporal) at a given gauging station using the available observed runoff series of the same period measured at the station and at one or more other stations (spatial) in the same main sub-basin. Thus the missing runoff data at a station in the Black Volta basin would be estimated using its available runoff data and those from other stations in the same main sub-basin for the year. The lengths of missing data considered in the study would be comparable to the typical real gaps shown in the hydrograph in Fig. 2.

Thus Eq. (1) was used to model the runoff process both spatially, with discrete locations at the gauging stations, and temporally, with annual daily time series of runoff at the stations. Catchment wetness is an appropriate state of the hydrological system but cannot be handled directly in this study. However, catchment runoff has been found to be a very good surrogate for catchment wetness (Young and Beven, 1994; Young, 2001) and so in this study, the state vector, $\boldsymbol{x}_t$, represents the unobserved actual catchment runoff (acting as surrogate for catchment wetness) and $\boldsymbol{y}_t$ is the vector with measured runoff at the m gauging stations at sampling time t. Both $\boldsymbol{x}_t$ and $\boldsymbol{y}_t$ are taken as $m\times 1$ vectors of actual catchment and measured runoff series of length N$\leq$366, i.e., $\boldsymbol{x}_t$ and $\boldsymbol{y}_t$ occur at the same locations. The dimensionality of the system state used here is low so that parameterisation of the system matrices should be unnecessary. Nonetheless, parameterisation of selected matrices may enhance the stability of the simulations and speed up model convergence.

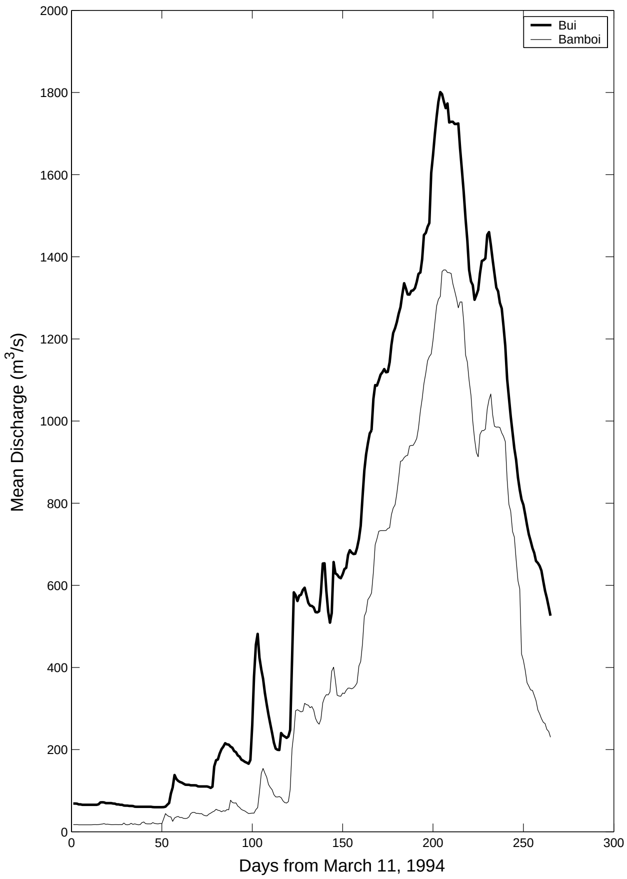

Fig. 4a. 1994 Daily Hydrographs for Bui and Bamboi on the Black Volta River.

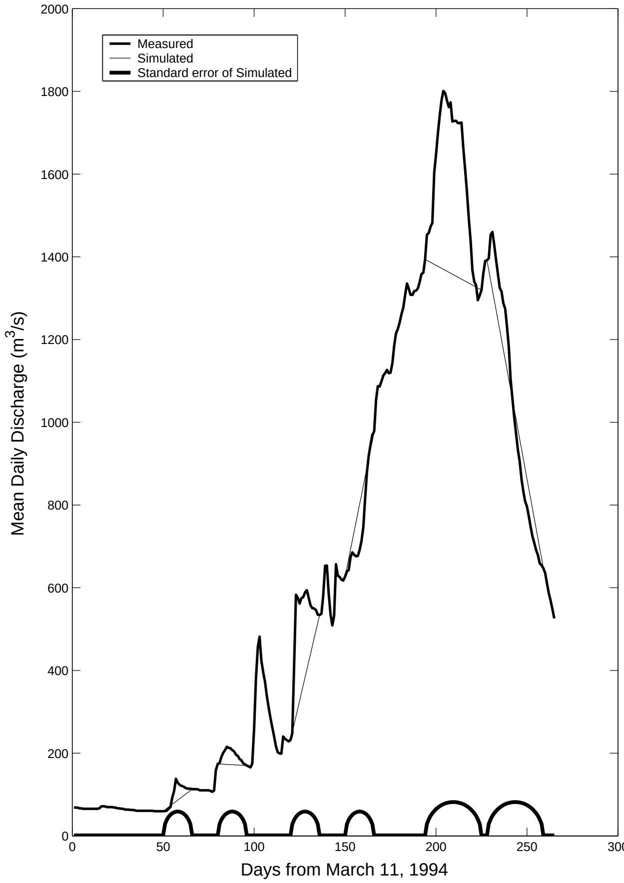

Fig. 4b. Measured and Simulated Daily Discharge for Bui on the Black Volta River (Black Volta Basin) – State Space Model on only Bui Discharge (1994).

In this study, the design matrix $\mathbf{H} = \mathbf{I_m}$, the process matrix $\mathbf{F}$ and error covariance matrix $\mathbf{Q}$ are unconstrained. However, $\mathbf{R}$ had to be constrained to $\mathbf{R}=\sigma^2\mathbf{I_m}$, because the unconstrained $\mathbf{R}$ caused stability problems and convergence was not achieved in the simulations with unconstrained $\mathbf{R}$. It was noted that when $\mathbf{Q}$ was constrained as well, $\mathbf{Q}$=diag($\mathbf{A}$), the iterations oscillated much before settling convergence than when $\mathbf{Q}$ was unconstrained. At least one of the stations would have missing riverflow observations, the estimation of which is the main task at hand. Only gaps of a few days to a maximum of one month were considered. As covergence criterion, $\left\|\boldsymbol{\Theta}^{(r+1)}-\boldsymbol{\Theta}^{(r)}\right\| < 0.001$ was adopted. It should be mentioned that the results are relatively sensitive to the initial values used.

The model was applied to daily time series of riverflows of about one year measured at the stations Lawra (Black Volta, 2°55′ W; 10°38′ N), Bui (Black Volta, 2°14′ W; 8°17′ N), Bamboi (Black Volta, 1°54′ W; 8°15′ N), Saboba (Oti, 0°24′ E; 9°45′ N), and Sabari (Oti, 0°12′ E; 9°17′ N), indicated in Fig. 1. By blacking out some of the observed data at a target gauging station, a maximum of 30 consecutive days of missing data in selected periods of the year were artificially created in the flow series of the target station. Three predictions of the missing data were made and compared by running the model with (i) the remaining samples of the target station as observed series (ii) the target station's series and the series from the rest of the stations in the same main sub-basin and (iii) the target station's series and that from only one other station in the same main sub-basin. Reliable daily flows for at least two stations for a full year could be obtained only for the main sub-basins of the Black Volta and Oti. The EM algorithm as described earlier not only provides estimates of the parameters but also those of the missing observations and state variables. It was used here to obtain estimates of the artificially created missing riverflow observations of the target stations.

In addition, linear regressions for the non-missing flows at each target station on the other stations were established. The results of the regression were subsequently used as simple models to predict the missing flows. We used these regression-based models in order to evaluate the efficacy of the more elaborate spatio-temporal framework.

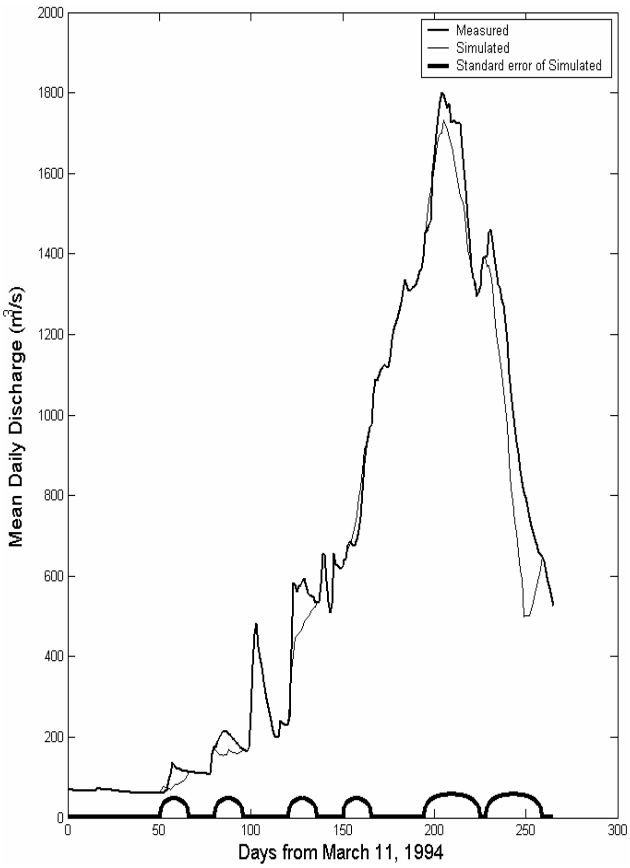

Fig. 4c. Measured and Simulated Daily Discharge for Bui on the Black Volta River (Black Volta Basin) – State Space Model on 1994 Bui and Bamboi Discharges.

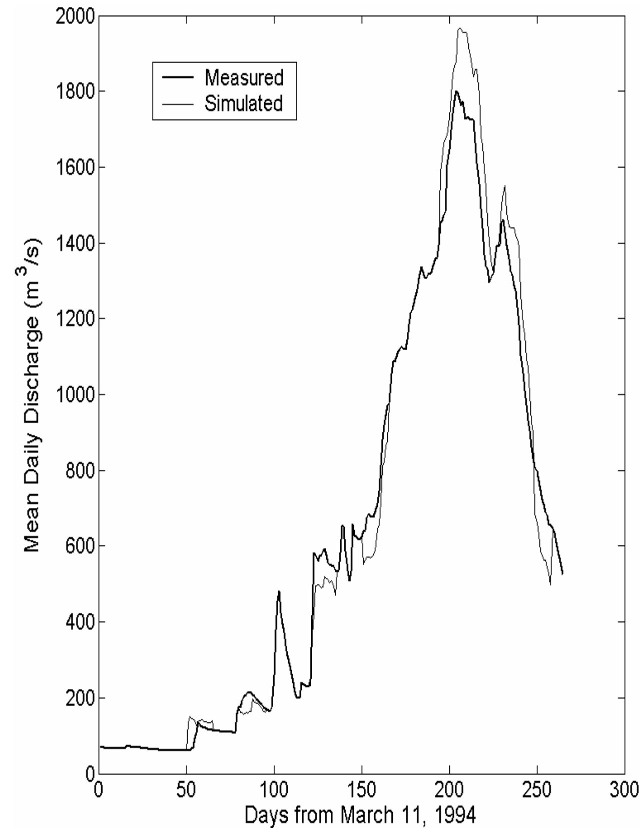

Fig. 4d. Measured and Simulated Daily Hydrographs for Bui on the Black Volta River (Black Volta Basin) – Linear Regression on 1994 Bamboi and Non-missing Bui Daily Discharges.

The performance of the models was evaluated in each case by the following Nash-Sutcliffe Efficiency (Nash and Sutcliffe, 1970) criterion:

$$\mathrm{NSE} = 100\left(1 - \frac{\sigma_e^2}{\sigma_y^2}\right) \tag{15}$$

where σ_e^2 is the variance of the residuals and σ_y^2 the variance of the measured runoff at the target station.

6 Results and discussion

Figures 3a, 4a, and 5a are the observed runoff series used in the modelling exercise. They show varying degrees of spatial correlation between the individual series in a plot. Figure 3a shows, for example, that there is a high correlation between the Bui and Bamboi flows but not much correlation between the Lawra series and the others. The spatial correlation between the two series in Fig. 5a is not very good but not too poor either. As in Fig. 3a, Fig. 4b shows very good spatial correlation between the two series plotted. The high correlation or lack of it in any set of runoff series could result in good or poor predictions of missing values in one series with the others as predictors. However, since in the framework used here the temporal correlation of the discharge data at the target station is also made use of (as opposed to the case of simple spatial regressions when this temporal information is discarded after fitting the regression), this framework could still produce good predictions where spatial correlations are poor.

Predicted missing values for target station Bui in the Black Volta Basin are presented in Figs. 3b, c and d for the three cases (i), (ii) and (iii) above, for the year 1964. The standard errors of the estimates for each case are also shown in the respective plots (the bottom plots). These errors are small where observations are available and larger where they are missing, as expected. The “hills” and “valleys” in the error plots indicate clearly the ranges of the missing and observed flows. Figures 3b and c and the relevant NSE values in Table 1 show that the use of the series from both Lawra and Bamboi, together with the remaining observations from Bui, provides better predictions of the missing values of Bui and with less uncertainty than using the observed series of Bui alone. The spatial interpolation obtained from the model is therefore adequate in this case. Figure 3d and Table 1 show

Table 1. NSEs for the various tests and R^2 values for the linear regressions.

Test Series		NSE (%)
1964	Flows in the Black Volta	
	State-space on only target station, Bui	95.0
	State-space on Bui, Lawra and Bamboi	99.0
	State-space on Bui and Bamboi only	98.6
	Linear Regression on Bamboi and non-missing Bui discharges (R^2=0.996)	98.1
Missing Data Pattern 1 for Lawra		
	State-space on only, Lawra	93.9
	State-space on Lawra and Bamboi	97.5
	Linear Regression on Bamboi and non-missing Lawra discharges (R^2=0.96)	86.3
Missing Data Pattern 2 for Lawra		
	State-space on only Lawra	97.0
	State-space on Lawra and Bamboi	98.3
	Linear Regression on Bamboi and non-missing Lawra discharges (R^2=0.94)	89.0
1994	Flows in the Black Volta	
	Only target station, Bui	93.5
	Bui and Bamboi	97.6
	Linear Regression on Bamboi and non-missing Bui discharges (R^2=0.97)	96.0
1976	Flows in the Oti	
	Only target station, Sabari	61.6
	Sabari and Saboba	91.4
	Linear Regression on Saboba and non-missing Sabari discharges (R^2=0.96)	89.0

that the use of Bamboi's series without that of Lawra has not reduced the quality of the predictions significantly, caused by the fact that the spatial correlation between the flows at Bui and at Lawra is low, as can be seen from the hydrographs in Fig. 3a. Figure 3e is a plot of the observed and predicted hydrographs at Bui from a linear regression of the non-missing discharges at Bui on the corresponding values at Bamboi. The figure and the R^2 from the regression shown in Table 1 show the very good spatial correlation between the Bui and Bamboi discharges. The NSE values in the Table also show that the quality of predictions from the linear regression and the state-space formulations are similar and suggest, it would appear, that the use of the spatio-temporal framework is unnecessary.

However, Figs. 3f–k, portray a different picture and illustrate the main point of the analysis undertaken in this study. The figures show observed and predicted hydrographs at Lawra using predictions from non-missing discharges from Lawra only (Figs. 3f and i), spatio-temporal predictions with both Bamboi and non-missing Lawra discharges (Figs. 3g and j). Figures 3h and k show the prediction results for missing Lawra discharges for two different missing data patterns based on corresponding Bamboi discharges using the linear regression between Lawra and Bamboi. As shown by the relevant NSE values in Table 1, in both cases of missing data patterns, predictions from the Lawra only non-missing discharges and those from the spatio-temporal model are much better than predictions from the spatial regressions of Lawra on Bamboi discharges. In these cases, the spatio-temporal model has made use of the remaining discharges for Lawra, values that still have high enough information content as to produce very good predictions with the Lawra-only non-missing discharges. The Lawra-only predictions in the two missing-data patterns show also that the information content of the remaining flow series for Lawra in pattern 2 is better than for pattern 1 (compare Figs. 3f and i and the NSE values in Table 1) and this also shows in the spatio-temporal model predictions.

Thus, the spatio-temporal model behaves just as its name implies – as both a spatial and temporal interpolator – shifting more to spatial interpolation when the correlation in this dimension is very good and to temporal interpolation when correlation shifts substantially to this dimension. The critical point to note is that in real situations, the full series for the target stations would not be available so whether good spatial correlations between the flows at the target and other stations exist or not would be difficult to ascertain visually. Values of R^2 from linear regression analysis of the non-missing

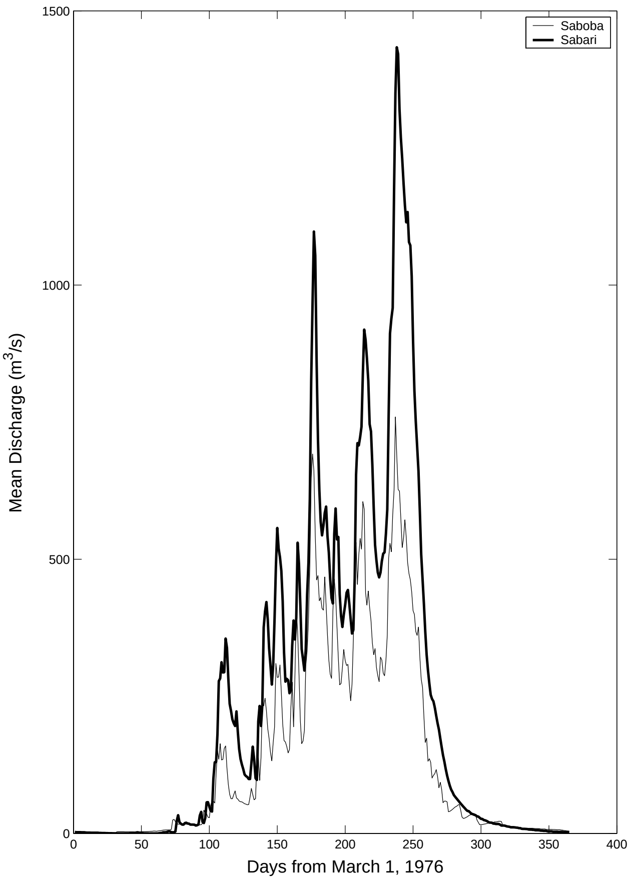

Fig. 5a. 1976 Daily Hydrographs for Saboba and Sabari on the Oti River.

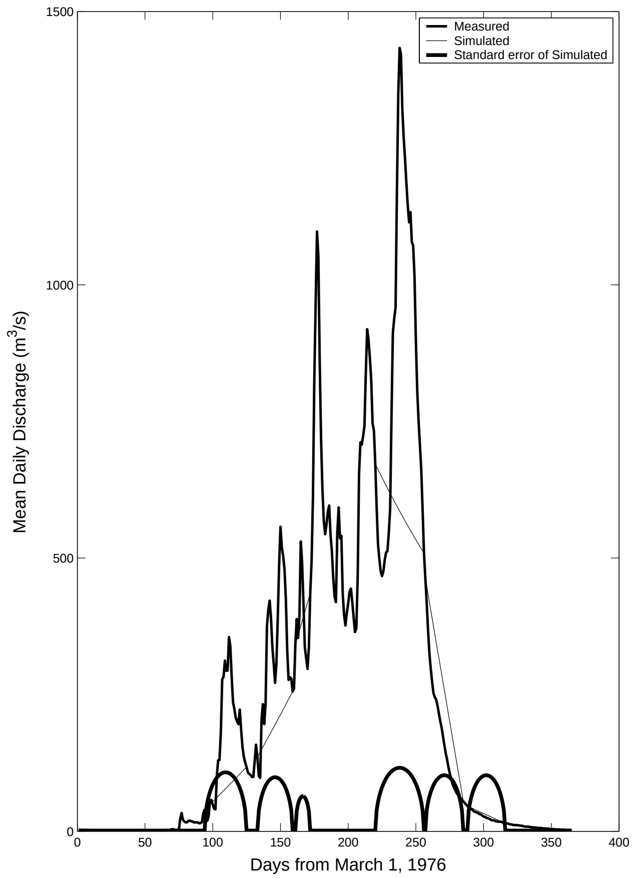

Fig. 5b. Measured and Simulated Daily Discharge for Sabari on the Oti River (Oti Basin) – State Space Model on only Sabari Discharge (1976).

discharges at the target stations and the corresponding discharges at the other stations, for example, would most likely be relied upon in judging the existence of good spatial correlations. In that case, the R^2 values in Table 1 for the case of Lawra could be very misleading. These values, of more than 0.9, suggest strong spatial correlation between the Lawra and Bamboi discharges and so would discourage the use of the spatio-temporal model – a model that provides much better predictions than the regressions!

Figures 4b and d show plots of predicted and observed flows at Bui for the year 1994 for cases (i) and (iii), with Fig. 4d showing plots from the linear regression of 1994 Bamboi and non-missing Bui discharges. These plots and relevant NSE and R^2 values in Table 1 show the good spatial correlation between the Bui and Bamboi flows, the good information content of the non-missing Bui discharges and the ability of the spatio-temporal model to exploit both information types.

Predicted and observed 1976 flows for Sabari on the Oti River are shown in the plots in Figs. 5b and c for cases (i) and (iii) with the respective standard errors of the predictions at the bottom of the plots. Figure 5d is the plot of observed and predicted Sabari hydrographs from the linear regression of the non-missing Sabari discharges on the corresponding Saboba discharges. Here too, the R^2 of 0.96 of the linear regression gives the impression of extremely good spatial correlation between these discharge data – an impression shown clearly to be misleading from the plots and the relevant NSE values in Table 1. Therefore, an important strength of the spatio-temporal model is its good use of the information contained in both the spatial correlation between discharges at different gauging stations and the temporal distribution of the non-missing discharges of the target station.

The second potentially equal important advantage of the used method over the use of simpler methods is the calculation of standard errors over the periods with missing data. As shown above, goodness of fit of simple models is in our case not a good measure for the error associated with the prediction of missing values. Also in other applications of the presented methodology in hydrological data assimilation, this provided insight in errors and error structure would be very valuable.

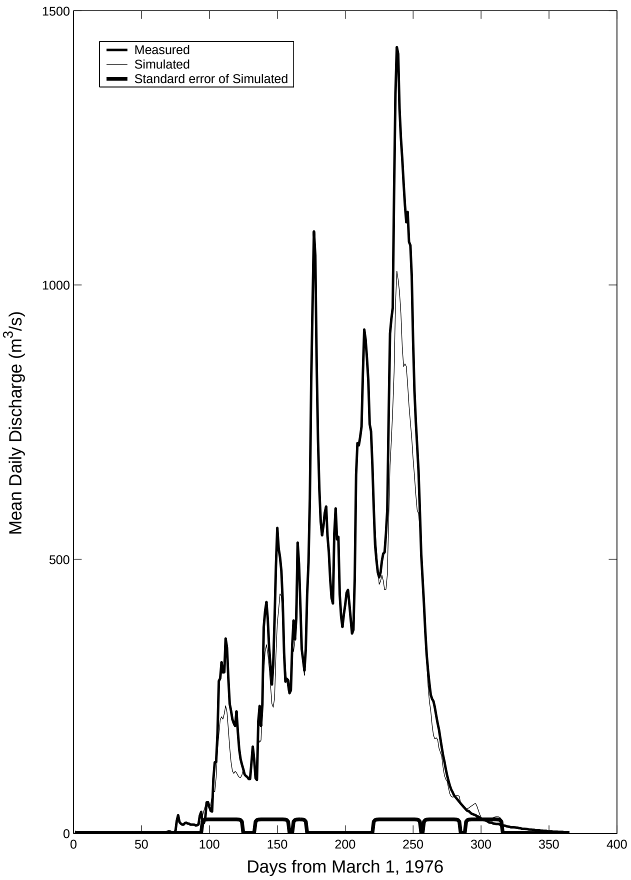

Fig. 5c. Measured and Simulated Daily Hydrograph for Sabari on the Oti River (Oti Basin) – State Space Model on 1976 Sabari and Saboba Discharges.

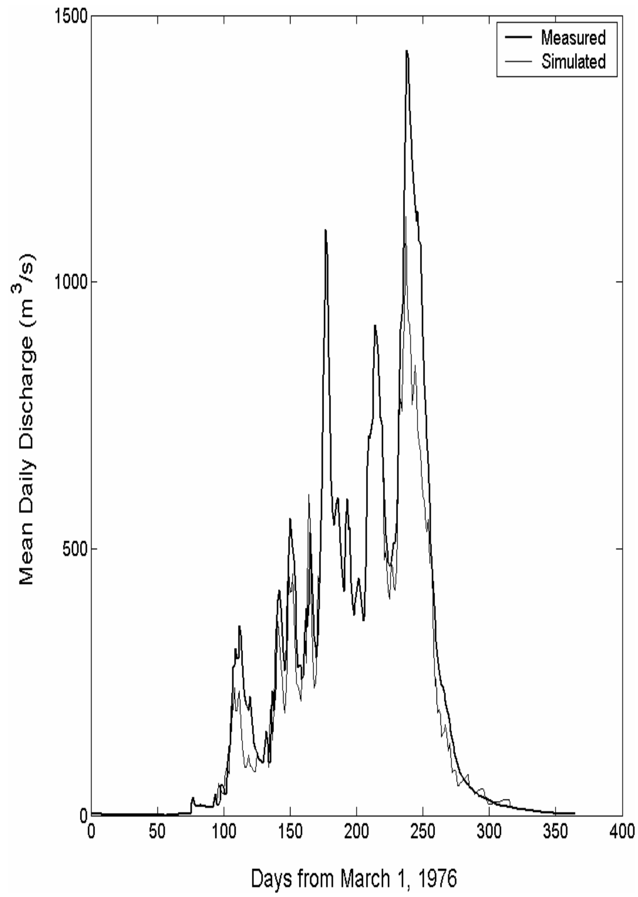

Fig. 5d. Measured and Simulated Daily Hydrographs for Sabari on the Oti River (Oti Basin) – Linear Regression on 1976 Saboba and Non-missing Sabari Daily Discharges.

7 Conclusions

A spatio-temporal state space linear dynamic model was developed to fill short gaps in daily runoff series using other, spatially correlated, daily series. Parameter estimation was done using the EM algorithm. Application of the model in the Volta Basin of West Africa shows that it is capable of providing good estimates of short gaps in river flows. The model has two main strengths – its ability to make good use of both spatial and temporal information in its predictions and also to provide estimates of both the parameter and the missing values concurrently and without the need for separate calibration or training series. It is, therefore, very suitable for short gaps in-filling in river basins such as the Volta where missing flows in runoff series at many gauging stations abound.

The critical assumption in the spatio-temporal model is that the catchment runoffs for all stations considered in a model run are generated by the same process. The model thus works well when one or both of the following conditions are satisfied:

– There is good spatial correlation between the runoff series involved.
– There is good temporal information in the remaining, non-missing, values at the target station.

The results obtained in this study show this very clearly. The presented method to fill data gaps should be seen as a relatively simple application of the powerful combination of EM and Kalman smoother. This combination allows us to estimate states and parameters concurrently, using spatial and temporal data. Further research will need to address some of the simplifying assumptions made, specifically those concerning linear updates and Gaussian error structures of the underlying processes.

Acknowledgements. Support of the GLOWA Volta Project, under Grant 07GWK01 provided by the German Federal Ministry of Education and Research and the State of North-Rhine Westphalia is gratefully acknowledged. Very constructive suggestions by the two anonymous reviewers are also gratefully acknowledged.

Edited by: T. Wagener

References

Bilmes, J. A.: A Gentle Tutorial of the EM Algorithm and its Application to Parameter Estimation for Gaussian Mixture and Hidden Markov Models, ftp://ftp.icsi.berkeley.edu/pub/techreports/1997/tr-97-021.pdf, 1998.

Dempster, A. P., Laird, N. M., and Rubin, D. B.: Maximum Likelihood from Incomplete Data via the EM Algorithm, J. Royal Statistical Soc., Series B (Methodological), 39(1), 1–38, 1977.

Digalakis, V., Rohlicek, J. R., and Ostendorf, M.: ML Estimation of a Stochastic Linear System with the EM Algorithm and its Application to Speech Recognition, IEEE Transactions on Speech and Audio Processing, 1(4), 431–442, 1993.

Ghahramani, Z. and Hinton, G. E.: Parameter Estimation for Linear Dynamical Systems, Technical Report CRG-TR-96-2, Department of Computer Science, University of Toronto, Canada, 1996.

Goodall, C. and Mardia, K. V.: Challenges in Multivariate Spatial modelling, Proceedings of the XVIIth International Biometric Conference, Hamilton, Ontario, Canada, 8–12, 1994.

Guttorp, P. and Sampson P. D.: Methods for estimating heterogeneous spatial covariance functions with environmental applications, Environmental Statistics, edited by: Patil, G. P. and Rao, C. R., Handbook of Statistics, North Holland, Amsterdam, New York, 12, 661–689, 1994.

Gyau-Boakye, P. and Schultz, G. A.: Filling gaps in runoff time series in West Africa. Hydro.Sci.J 39(6): 621–636, 1994.

Haslett, J.: Space time modelling in meteorology – a review, Bulletin of the International Statistical Institute, 51, 229–246, 1989.

Haykin, S.: Kalman Filters, in: Haykin, S: Kalman Filtering and Neural Networks, John Wiley & Sons, Inc., 2001.

Kottegoda, N. T. and Elgy, J.: Infilling missing flow data, in: Modeling Hydrologic Processes, edited by: Morel-Seytoux, H. J., Salas, J. D., Sanders, T. G., and Smith, R. E., Proc. Fort Collins 3rd Int. Hydrol. Symp., On Theoretical and Applied Hydrology, Colorado State University, Fort Collins, Colorado, USA, 27–29 July, 1977, 60–73, 1979.

Little, R. J. A. and Rubin D. B.: Statistical Analysis with Missing Data, 2nd Editon, Wiley, New York, 2003.

Mardia, K., Goodall, C., Redfern, E., and Alonso, F.: The kriged kalman filter (with discussion), Test, 7, 217–285, 1998.

Nash, J. E. and Sutcliffe, J. V.: River Flow Forecasting through conceptual models, J. Hydrol., 10, 282–290, 1970.

Papadakis, I., Napiorkowski, J. and Schultz, G. A.: Monthly runoff generation by non-linear model using multispectral and multi-temporal satellite imagery, Adv. Space Res., 13(5), 181–186, 1993.

Rouhani, S. and Myers D. E.: Problems in space-time Kriging of geohydrological data, Math. Geology, 22, 611–623, 1990.

Riberiro, M. I.: Kalman and Extended Kalman Filters: Concept, Derivation and Properties, http://omni.isr.ist.utl.pt/~mir/pub/kalman.pdf, 2004.

Schafer, J. L.: Analysis of Incomplete Multivariate Data, New York, CRC Press, 1997.

Shumway, R. H. and Stoffer, D. S.: An Approach to Time series Smoothing and Forecasting Using The EM Algortihm, J. Time Series Analysis, 3(4), 253–264, 1982.

Taylor, J.: Watershed management in the Volta River Basin, West Africa: Providing information through hydrological modelling, M.Sc. Thesis, Cornell University, 2003.

Wang, Q. J.: A Bayesian joint probability approach for flood record augmentation, Water Resour. Res., 37(6), doi:10.1029/2000WR900401, 2001

Xu, K. and Wikle, C. K.: Estimation of Parameterized Spatio-Temporal Dynamic models, http://www.stat.missouri.edu/~wikle/xu_wikle_072104.pdf, 2004.

Young, P.: Data-based mechanistic modelling and validation of rainfall-flow processes, in: Model Validation – Perspectives in hydrological science, edited by: Anderson, M. G. and Bates, P. D., Chichester, John Wiley & Sons Ltd., 117–161, 2001.

Young, P. C. and Beven, K. J.: Data-based mechanistic modelling and the rainfall-flow nonlinearity, Environmetrics, 5, 335–363, 1994.