Stochastic formulation of a conceptual hydrological model

Stochastic formulation of a conceptual hydrological model

Thorsten Wagener[1], Hoshin V. Gupta[1] and Soroosh Sorooshian[2]

[1]The University of Arizona; University of California, USA

INTRODUCTION

Computer-based hydrological models are indispensable tools in today's research and operational practice, and are used for a variety of purposes, ranging from the prediction of floods and droughts to water quality and water resource studies. Such models attempt to represent the spatially distributed complex interactions of water, energy and vegetation processes using relatively simple mathematical equations.

Two characteristics of such models are important in the context of this research. First, all hydrological models are at some level lumped, in the sense that their equations (and therefore their parameters) are aggregate descriptions (in space and time) of real world processes. As a consequence of this, at least some of the model parameters lose their direct physical interpretation and measurability, and are therefore referred to as conceptual parameters. The scale at which conceptual parameters are defined in the model is typically different from the scale at which they can be measured. Therefore, they are often estimated through an indirect process in which the value of the parameter is adjusted to bring the model simulated input–output behaviour as close as possible to the system input–output behaviour observed in the field.

A second characteristic of such models is that it is common to specify/select the model structural equations *prior* to any modelling being undertaken (Wheater *et al.*, 1993). There appear to be no well-defined pathways or objective procedures leading to an unambiguous selection of an appropriate model structure. Rather, this process is influenced by a combination of factors including observations about the characteristics of the catchment, available data, modelling objective and personal preference. Hydrological modelling is therefore often referred to as being as much an 'art' as a 'science'.

Uncertainty is an unavoidable element of any modelling exercise in hydrology. Uncertainty is introduced, for example, through the inability to find a single best model structure suitable for a particular application, and/or by the inability to uniquely identify values for the model parameters. There has recently been a surge in attention given to methods for the treatment of model uncertainty as (a) decision makers push for better quantification of the *accuracy and precision* of hydrological model predictions, (b) interest is growing in methods for properly merging data with models and for reducing predictive uncertainty, and (c) scientists push to better represent what is, and is not, well understood about the hydrological systems we study. Modellers realise that it is advantageous to consider uncertainty as an intrinsic part of any hydrological modelling study, because (among other reasons) *"giving a best estimate of the range of possible predictions is good protection against being wrong"* (Beven, in Wagener *et al.*, 2004).

This paper discusses briefly the current state of uncertainty estimation, including current limitations and bottlenecks. A new approach to the stochastic formulation of hydrological models is then presented and its applicability is discussed. The methodology is introduced using a simple illustrative case study involving a linear reservoir model.

SOURCES OF UNCERTAINTY IN HYDROLOGICAL MODELLING

Uncertainties stem from a variety of sources (Melching, 1995; Gupta *et al.*, 2004), mainly:

- *Data uncertainty*, i.e. uncertainty introduced by errors in the measurement of input (including forcing) and output data itself, or by data pre-processing. Additional uncertainty is introduced if long-term predictions are made which include climate change scenarios.
- *Parameter estimation uncertainty*, i.e. the inability to converge to a single 'best' parameter set (model) using the information provided by the available data. The lack of correlation often found between conceptual model parameters and physical catchment characteristics commonly results in significant uncertainty if land use change scenarios are to be investigated.
- *Model structural uncertainty* introduced through

Hydrology: Science & Practice for the 21st Century. Volume I

simplifications and/or inadequacies in the description of real world processes.

- *State uncertainty*, which can refer to the uncertainty in the states at the beginning of the modelling exercise or to the uncertainty during predictions. The inclusion of a warm-up period can help to reduce the former, whereas the latter can be treated via data assimilation (i.e. state-updating) techniques.

Additionally, one has to keep in mind that even if the uncertainties mentioned above could be removed, there would remain some (immeasurable) randomness in the natural processes themselves (Melching *et al.*, 1990), which introduces uncertainty that cannot be reduced.

Research in the past has mainly focused on the treatment of parameter and data uncertainty. However, it has become apparent in recent years that the *"effects of model structural error are often even more severe than that of uncertain parameters"* (Carrera and Neuman, 1986; see also James and Oldenburg, 1997).

CURRENT TREATMENT OF UNCERTAINTY AND RECENT DEVELOPMENTS

Most approaches currently available for the treatment of uncertainty consider only some of these uncertainties or combine them in a way that does not allow for the analysis of the influence of individual elements. It is most common to map the overall uncertainty into the parameter space (Figure 1), using a sampling strategy to approximate the shape of the response surface (or rather hypercube) and to propagate the predictions of the behavioural region (parameter sets that give an acceptable reproduction of the system response) into the output space. A wide range of methods is available, differing in underlying assumptions and in sampling schemes used. However, they typically contain similar elements and a simple classification can be made as follows,

- *Stochastic optimisation methods* that attempt to estimate the full posterior distribution using methods for sampling the parameter space such as the Monte Carlo Markov Chain approach (e.g. Vrugt *et al*., 2003).
- *Bayesian methods* using Bayes law to recursively update the parameter population when new information (*i.e.* data) becomes available (e.g. Thiemann *et al*., 2001).
- *Set-theoretic approaches* that partition the model/ parameter space into behavioural and non-behavioural models, usually based on uniform random sampling. They can use batch (e.g. Beven and Binley, 1992) or recursive (Wagener *et al*., 2003) data processing schemes.
- *Multiple-objective approaches* that can be used if the classification of a behavioural model is based on more than one criterion (e.g. Gupta *et al*., 1998).
- *First-order approximation methods* that make simplifying assumptions about the shape of the response surface (e.g. quadratic) so that only the second order statistics (mean and covariance matrix) of the distributions are propagated. These are often referred to as engineering methods (e.g. Melching, 1995; Melching *et al*., 1990).
- *State-Estimation methods* (such as Kalman filtering) which propagate the uncertainty in the state variables through time, for real-time forecasting purposes, and which recursively assimilate observational information back into model to update the state estimate (e.g. Kitanidis and Bras, 1980).

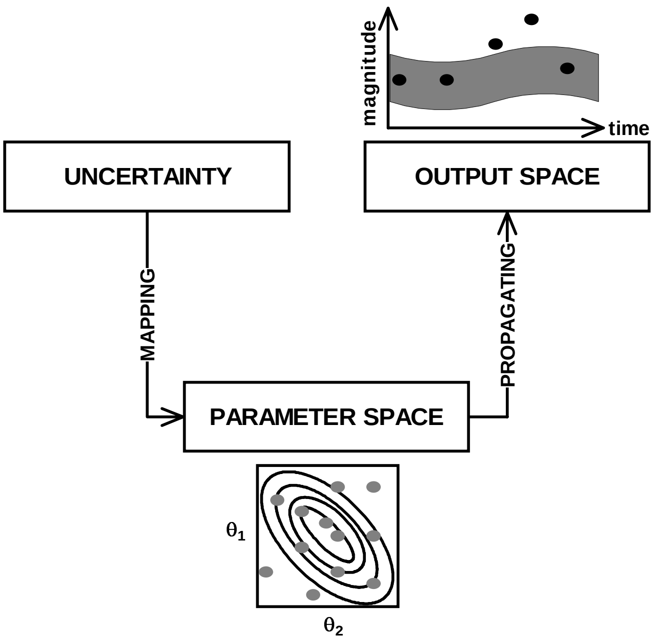


Fig. 1 Current procedure for uncertainty estimation and propagation.

Recent developments in adaptive parameter estimation techniques, Bayesian analysis and Monte Carlo Markov Chain algorithms offer new possibilities to address the problem of simultaneous consideration of all the uncertainties. One current trend is to use the following components: Bayesian based ideas (e.g. Thiemann *et al*., 2001), new sampling methods (usually Monte Carlo Markov Chain) (e.g. Vrugt *et al*., 2003), recursive techniques (e.g. Wagener *et al.*, 2003) and a flexible objective function that allows for the analysis of underlying assumptions (e.g. Thiemann *et al*., 2001).

A major problem is the appropriate treatment of model structural uncertainty in the modelling process. The process of simplifying the natural hydrological system into mathematical form introduces ambiguity (due to the lack of an objective procedure) and the resulting model structures exhibit shortcomings (e.g. Gupta *et al.*, 2004) that indicate that we currently do not possess a system representation capable of consistently reproducing observed catchment behaviour. Even worse, there is also little indication about the structure of this uncertainty.

To examine the problem of structural uncertainty, it is instructive to examine the two extremes of how model structures are built. On the one hand are procedures based on *prior knowledge*, and on the other are approaches based on *data* (i.e. observations of input-output behaviour). In the first approach a model structure is built based on our perceptual and conceptual understanding of the hydrological system (e.g. Freeze and Harlan, 1969). In the second approach, an empirical model is constructed to represent the observed input–output behaviour of the system (e.g. Young, 2001). These two approaches are used here as boundaries to discuss how model structural uncertainty could be considered. Figure 2 shows a schematic of this idea. On the left hand side it is assumed that considerable prior knowledge is available to build the model structure. The uncertainty in the resulting model structure can be explored using a perturbation approach where the model equations are perturbed by a chosen error model (see next section for details on this approach). If several candidate model structures provide equally likely system representations, ensembles of all applicable model structures can be used to yield a combined prediction (Chatfield, 1995). Bayesian Model Averaging (BMA), for example, combines different model structures based on their probability of being good representations of the system under investigation (e.g. Hoeting *et al.*, 1999; Neuman, 2003). Similar effects can be achieved in a less statistically rigorous framework using Monte Carlo methods (e.g. Beven, 2004).

On the right hand side of the continuum lie data-based fully stochastic approaches. For example, empirical approaches based on Artificial Neural Networks (ANN) are often used to derive model structures when no prior knowledge is available. An interesting modification on this type of method is the data-based mechanistic (DBM) approach by Young (2001), in which several transfer function type model structures are initially derived from the data alone, and then scrutinised with respect to their physical realism. Only structures that are physically plausible are retained as possible system representations.

STOCHASTIC MODEL FORMULATION

As mentioned earlier, a major problem with respect to the assessment of uncertainty is that we know little about their structure. While some of them may often be reasonably approximated using Gaussian distributions (e.g. output data uncertainty), others such as the form of model structural uncertainty usually can not. This limits our ability to consider these uncertainties in a realistic manner. Additionally, even if the structure of the uncertainties were to be known, there is still no consensus on how to implement them in a fully stochastic model. The formulation presented here is an attempt to address these current limitations.
The characteristics of the uncertainty estimation method under development here can be summarised as follows,

- it allows for the description of *a priori* uncertainty,
- it explicitly incorporates multiple sources of information,
- it permits recursive processing of data,
- it provides probabilistic estimates of model outputs.

The general notation used is shown in Figure 3. Every dynamic mathematical model can be written in a general form using an output, an input and a state equation (though slight modifications might be necessary). An additional parameter equation can be formulated if the parameters are time-varying. These equations can be written as follows,

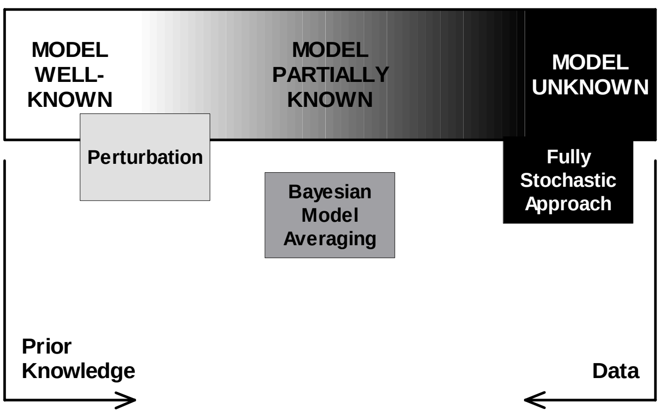


Fig. 2 Approaches to consider model structural uncertainty.

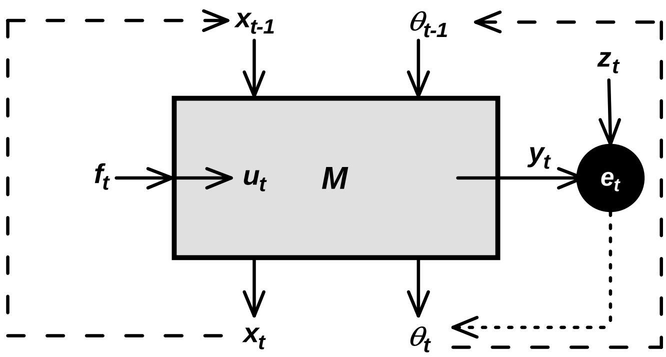


Fig. 3 General stochastic model formulation.

Output equation

$$y_t = m_y\left(x_{t-1}, \theta_{t-1}\right) + \varepsilon_{y_t} \tag{1}$$

Input equation

$$u_t = m_u\left(x_{t-1}, \theta_{t-1}, f_t\right) + \varepsilon_{u_t} \tag{2}$$

State equation

$$x_t = m_x\left(x_{t-1}, \theta_{t-1}, f_t\right) + \varepsilon_{x_t} \tag{3}$$

Parameter equation

$$\theta_t = \theta_{t-1} + \varepsilon_{\theta_t} \tag{4}$$

where f is the forcing (e.g. precipitation), u is the input that enters the model component (e.g. part of the precipitation that infiltrates the soil), θ is the parameter vector, x is the state, y is the model output, z is the observation, and M is the model structure (consisting of the constituent equations m_y, m_u and m_x). The index t refers to the current time step. In this format, the component is drained in the first step, and then the state is updated with the input at the current time step. Uncertainties are present in the individual variables (f, u, x, y, z) and parameters (θ), as well as in the formulation of the model itself (M, *i.e.* m_y, m_u and m_x). The latter being represented by an additive error term ε_i, which is also a function of time. We therefore treat the uncertain model elements as probabilistic variables represented by their probability density functions,

$$\varepsilon_{y_t} \sim p\left(\varepsilon_{y_t}\right) \tag{5a}$$

$$\varepsilon_{u_t} \sim p\left(\varepsilon_{u_t}\right) \tag{5b}$$

$$\varepsilon_{x_t} \sim p\left(\varepsilon_{x_t}\right) \tag{5c}$$

$$\varepsilon_{\theta_t} \sim p\left(\varepsilon_{\theta_t}\right) \tag{5d}$$

$$x_{t-1} \sim p\left(x_{t-1}\right) \tag{5e}$$

$$\theta_{t-1} \sim p\left(\theta_{t-1}\right) \tag{5f}$$

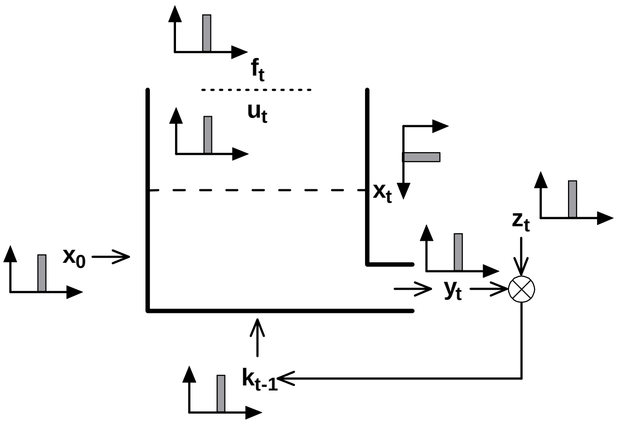
$$f_{t-1} \sim p\left(f_{t-1}\right) \tag{5g}$$

We explore ways to consider all these uncertainties in an adaptive method to update and keep track of the probability distributions of model states and parameters. The idea underlying this approach is to make assumptions about the uncertainties involved (i.e. the probability distributions of the random variables), test their impact on the prediction uncertainty and reduce the number of assumptions as knowledge and understanding are gained. This approach enables us to investigate the individual and combined influences of the different uncertainties. In principle, this generic approach can be used to formulate any hydrological model in a stochastic way. The model equations applied are, of course, despite the perturbation, still deterministic equations. A different approach must be found if our knowledge about the physical system is not sufficient to reasonably estimate these equations.

APPLICATION EXAMPLE – LINEAR CONCEPTUAL RESERVOIR

A linear conceptual reservoir model component, commonly used in hydrological modelling, is utilised here as an example. Figures 4a and 4b show the conventional deterministic formulation and the new stochastic formulation in line with the notation introduced earlier. The example calculations are solely based on synthetic, error-free data.

(a) Deterministic formulation



(b) Stochastic formulation

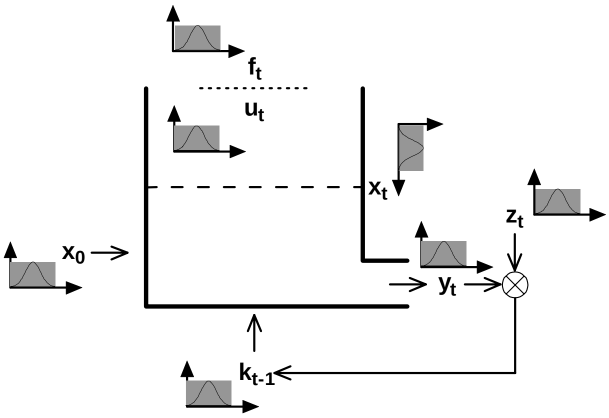


Fig. 4 Deterministic and stochastic formulation of a linear conceptual reservoir.

Stochastic formulation

The equations of a standard deterministic conceptual linear reservoir can be written as follows in the above-introduced form (Fig. 4a),

Output equation

$$y_t = k_{t-1} \cdot x_{t-1} \tag{6}$$

Input equation

$$u_t = f_t \tag{7}$$

State equation

$$x_t = x_{t-1} - y_t + u_t \tag{8}$$

Parameter equation

$$k_t = k_{t-1} \tag{9}$$

where k is the reservoir coefficient in the dimension of 1/time in this formulation. Considering the above-mentioned uncertainties can lead to the following form of a stochastic conceptual linear reservoir (Fig. 4b),

Output equation

$$y_t = k_{t-1} \cdot x_{t-1} + \varepsilon_{y_t} \tag{10}$$

Input equation

$$u_t = f_t + \varepsilon_{ut} \tag{11}$$

State equation

$$x_t = x_{t-1} - y_t + u_t + \varepsilon_{xt} \tag{12}$$

Parameter equation

$$k_t = k_{t-1} + \varepsilon_{kt} \tag{13}$$

where all the variables (x, y, etc.) are not deterministic, but probabilistic variables with a particular probability density function $p(.)$ (Eq. 5e–g). The perturbations of the model equations, ε_i (Eq. 5a–d), are also considered random variables as a first stage.

Results

A synthetic time-series of forcing data is used to drive the model and perform some simple numerical experiments to visualise the general strategy introduced above.

(a) Propagation of uncertainty through the model without updating of probabilities (Figure 5a, b). All probabilities sampled are taken from normal distributions around the true values, *i.e*. a heteroscedastic error is introduced. The ‘true’ values are shown as grey dots in Figure 5. The values show in the graphs have the index 1, meaning that they refer to the particular variables at the end of a time-step.

- As a first experiment, x_0 (initial state) and k_0 are sampled from normal distributions with standard deviations of 15% and 10% respectively at the first time-step. The input f_1 is sampled at every time-step from a normal distribution with 15% standard deviation. Taking the uncertainties forward in time without updating results in a reduction of the predictive uncertainty, down to a certain range, because of the fact that the model reaches a steady-state situation. The output uncertainty y_1 reduces, despite the fact that the variation in the state variable x_1 keeps increasing. Figure 5a shows the result for the case where the model equation uncertainty is not considered.
- The perturbation of the model equations is added in Figure 5b. Perturbations of the output (ε_y), input (ε_u) and state (ε_x) equations are considered. They are assumed to follow normal distributions with a 10% standard deviation. One can see that the uncertainty in the output increases considerably, roughly by a factor two, when this uncertainty is added. However, the general tendency to assume a steady-state situation remains.

(b) Two-dimensional plots can be used to show the relationship between parameters, state and output if different sources of uncertainty are considered individually (Figure 6). All uncertainties are again sampled from normal distributions with the same standard deviations as used under [a]. The indices, 0 and 1, indicate whether the value shown relates to the beginning or the end of a time step.

- Figure 6a shows these relationships (after the model has been run for one time-step) for the cases in which the uncertainties in input, initial state and in the parameter k are considered. The uncertainty in the initial state only impacts the model behaviour in the first few time-steps and is therefore of no consequence for these plots. Not introducing uncertainty in the initial state means that all the x_0 values are identical, but the x_1 values differ due to the variation in forcing etc. One can note that the uncertainty introduced through the parameter has very little impact on the predictive uncertainty in this case (stars on x_1 *v*. y_1 plot). The uncertainty introduced through the model forcing is much larger, i.e. the range of y_1 is much wider.

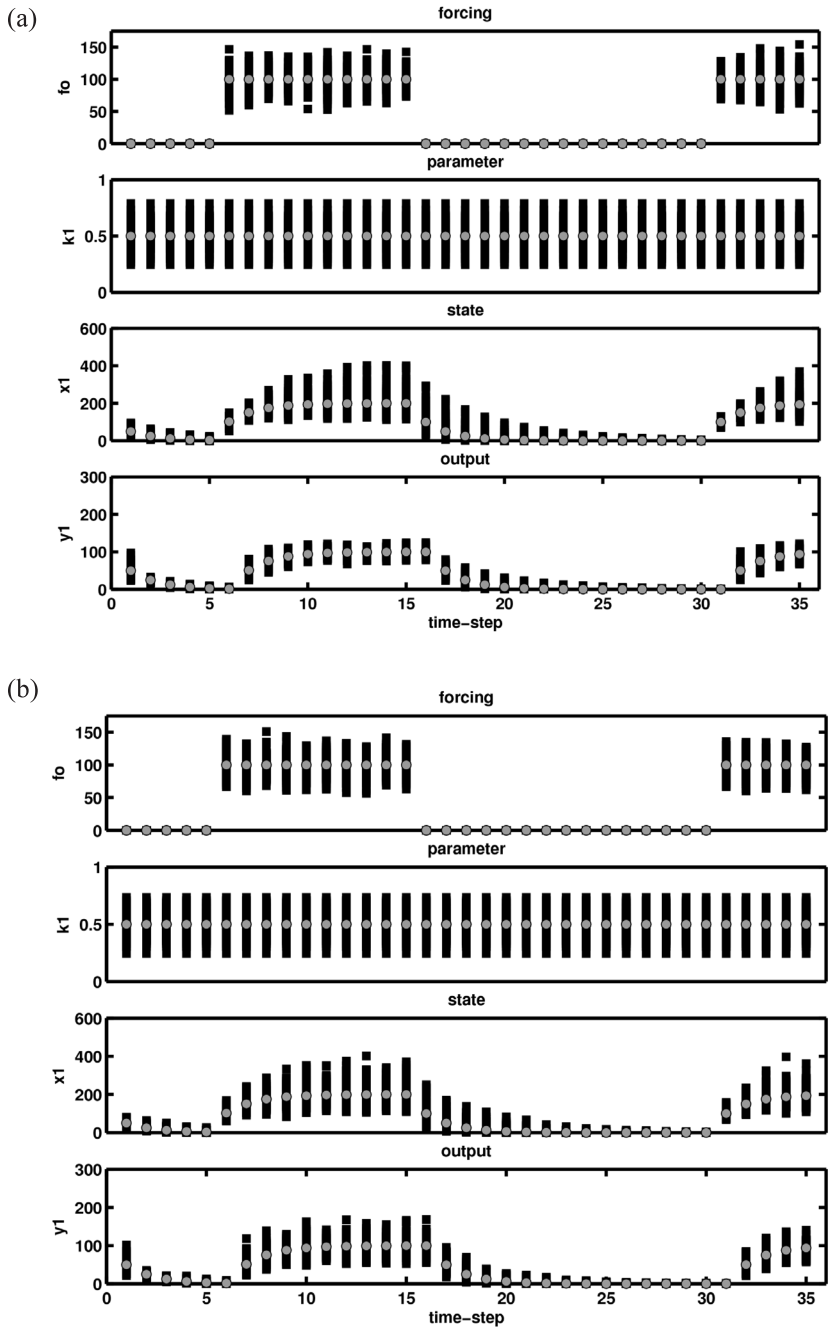


Figure 5. Forward propagation based on normal uncertainty distributions. The gray dots indicate the 'true' values. (a) Without perturbation of the model equations. (b) Including perturbation of the model equations (state, output and input).

- Figure 6b shows that the impacts of the uncertainties in the model equations (output, state and input) are of similar magnitude for the conceptual linear reservoir. The range of y_1 produced is very close for all three sources, with the uncertainty on the input equation being seemingly slightly less influential.
- Comparing Figures 6a and 6b, the impact of the uncertainty in the variables and parameter to that of the model equations shows that the uncertainty introduced by perturbing the model structure seems to have a greater impact on the output uncertainty. Particularly since the standard deviation used for the equation uncertainties was smaller (10%) than that for input (15%) and state (15%), and equal to the parameter uncertainty.

(a) Uncertainty in input (o), initial state (+), and parameter (*).

(b) Uncertainty in output (□), state (◆) and input (∇) equations.

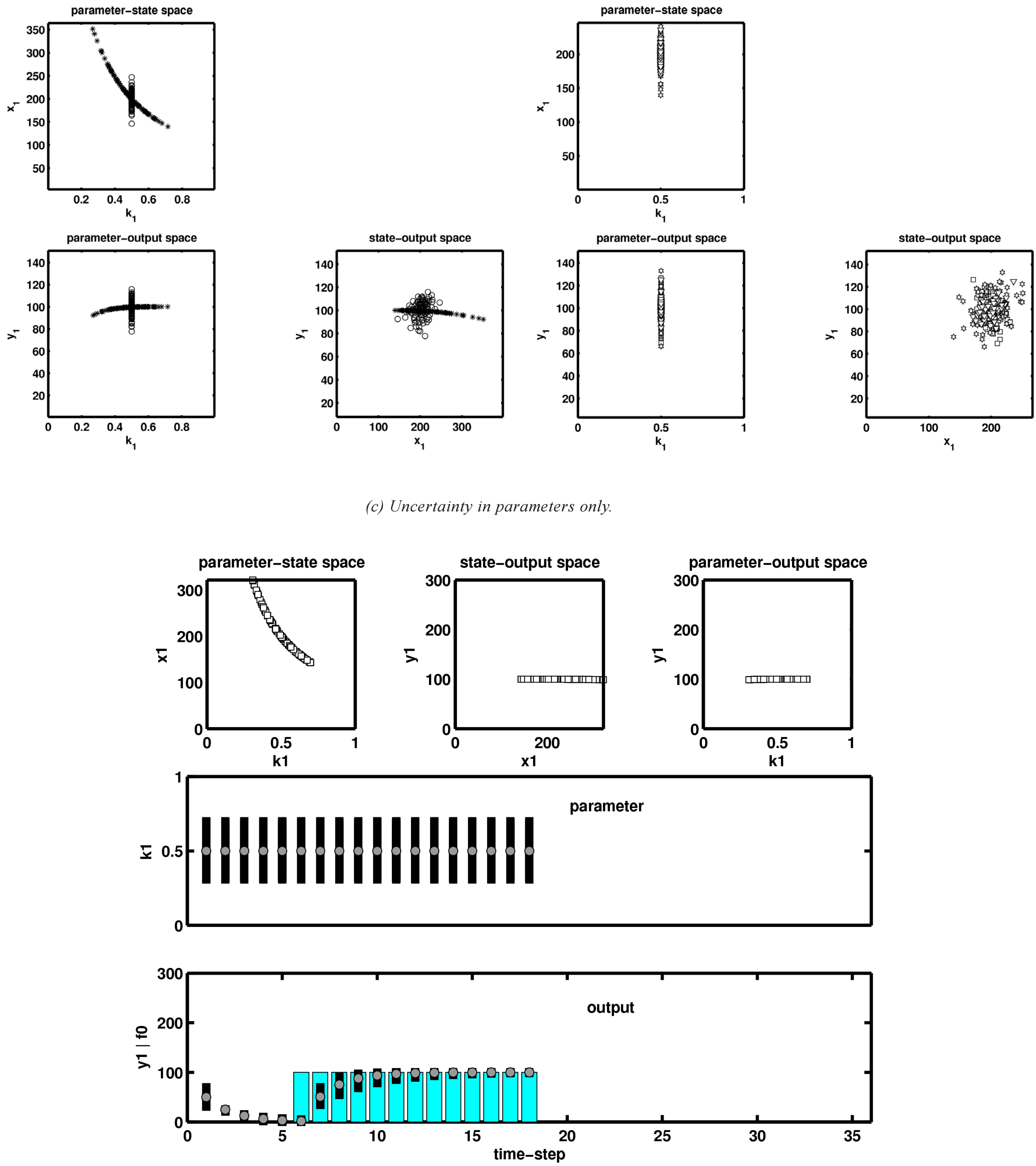


Fig. 6 Relationship between parameters, state and output if different sources of uncertainty are considered in Figures 6a and b. The model was run up to time-step 14 of the input sequence shown in Fig. 5 Figure 6c shows the state-parameter-output spaces for the case if only parameter uncertainty (based on a uniform distribution) is considered. The forcing f0 *is shown as bars at every time-step (either 0 or 100) in the bottom plot. The model was run up to time-step 18 of the input sequence shown in Fig.5. The plots in the top row of Figure 6c only show the individual spaces after the last time-step.*

Figure 6c shows a detailed analysis of the case when parameter uncertainty is considered only. One can see that x_1 and k_1 show an almost linear relationship (after 17 time-steps) that results in almost no uncertainty in the predictions, i.e. very similar y_1 value for every x_1–k_1 combination. One can see that the prediction uncertainty is much larger when the output y_1 is decreasing (time-step 1) or increasing (time-steps 7–11). The system reaches, for this particular case, a steady state for a certain range of x_1–k_1 combinations. A higher residence time (i.e. a smaller k_1 value) results in a larger state, yielding the same output as a smaller residence time with a smaller state (close to the size of the forcing).

CONCLUSIONS AND FUTURE WORK

The approach presented here is limited to the case for which we can assume that the chosen model structure is a reasonable description of the hydrological or environmental system under investigation. In this case, the remaining uncertainty might be captured using the perturbation of the model equations as suggested here. The scheme is not yet implemented to function recursively and improved sampling schemes must be tested. In addition, other approaches, for example those including time varying perturbation, have to be tested. This simple case study has been used to implement the technique and establish the notation.

The basic idea underlying the approach presented here is to make assumptions about the uncertainties involved and test the impact of these assumptions on the resulting predictive uncertainty. Our aim is to progressively reduce these assumptions, leading to a new approach to stochastic modelling.

Two limitations will remain even if the proposed approach is fully implemented is that (1) the model equations are still deterministic in character and pre-defined by the modeller; and (2) a range of assumptions still has to be made (e.g. underlying distributions). Work is ongoing to understand and treat these limitations.

ACKNOWLEDGEMENTS

Support of the National Science Foundation Science and Technology Center on “ Sustainability of semi-Arid Hydrology and Riparian Areas” (SAHRA) (EAR-9876800) is greatly appreciated. Thorsten Wagener also acknowledges support from the German Academic Exchange Service (DAAD) through its post-doctorate fellowship program.

REFERENCES

Beven, K. 2004. A manifesto for the equifinality thesis. *Water Resour. Res*., in review.

Beven, K.J. and Binley, A.M. 1992. The future of distributed models: Model calibration and uncertainty in prediction. *Hydrol. Process*., **6**, 279–298.

Carrera, J. and Neuman, S.P. 1986. Estimation of aquifer parameters under transient and steady state conditions: 1. Maximum likelihood method incorporating prior information. *Water Resour. Res*. **22**, 199–210.

Chatfield, C. 1995. Model uncertainty, data mining and statistical reference. *J. Royal Statist. Soc. Series A*, **158**, 419–466.

Freeze, R.A. and Harlan, R.L. 1969. Blueprint for a physically-based, digitally-simulated hydrologic response model. *J. Hydrol.*, **9**, 237–258.

Gupta, H.V., Beven, K. and Wagener, T. 2004. Model calibration and uncertainty analysis. In: *Encyclopaedia of Hydrological Sciences,* M.G. Anderson (ed.) John Wiley and Sons, Chichester, UK, in press.

Hoeting, J. A., Madigan, D., Raftery, A.E. and Volinsky, C.T. 1999. Bayesian model averaging: A tutorial. *Statist. Sci.*, **14**, 382–417.

James, A.L. and Oldenburg, C.M. 1997. Linear and Monte Carlo uncertainty analysis for subsurface contaminant transport simulation. *Water Resour. Res.*, **33**, 2495–2508.

Kennedy M.C. and O'Hagan 2001. A Bayesian calibration of computer models (including discussion). *J. Roy. Statist. Soc., Series B*-Statistical Methodology, **63**, 425–464.

Kitanidis, P.K. and Bras, R.L. 1980. Real-time forecasting with a conceptual hydrologic model 1. Analysis of uncertainty. *Water Resour. Res*.*,* **16**, 1025–1033.

Melching, C.S. 1995. Reliability estimation. In: *Computer models of watershed hydrology,* V.P. Singh (ed.). Water Resources Publishers, USA, 69–118.

Melching, C.S., Yen, B.C. and Wenzel, H.G., Jr. 1990. A reliability estimation in modeling watershed runoff with uncertainties. *Water Resour. Res*.*,* **26**, 2275–2286.

Neuman, S.P. 2003. Maximum likelihood Bayesian averaging of uncertain model predictions. *Stochastic Environ. Res. & Risk Assess*., **17**(5), 291–305.

Thiemann, M., Trosset, M.W., Gupta, H.V. and Sorooshian, S. 2001. Recursive Bayesian inference for hydrologic models. Water Resour. Res.*,* **37**, 2521–2535.

Vrugt, J.A., Gupta, H.V., Bouten, W. and Sorooshian, S. 2003. A shuffled complex evolution Metropolis algorithm for optimization and uncertainty assessment of hydrologic model parameters. *Water Resour. Res*., **39**, 1201, 10.1029/2002WR001642.

Wagener, T., McIntyre, N., Lees, M.J., Wheater, H.S. and Gupta, H.V. 2003. Towards reduced uncertainty in conceptual rainfall-runoff modelling: Dynamic identifiability analysis. *Hydrol. Process.*, **17**, 455–476.

Wagener, T., Wheater, H.S. and Gupta, H.V. 2004. *Rainfall-runoff modelling in gauged and ungauged catchments*. Imperial College Press, London, UK, in press.

Wheater, H. S., Jakeman, A. J. and Beven, K. J. 1993. Progress and directions in rainfall-runoff modelling. In: *Modelling change in environmental systems,* A.J. Jakeman, M.B. Beck and M.J. McAleer (eds.). John Wiley & Sons, Chichester, UK, 101–132.

Young, P.C. 2001. Data-based mechanistic modelling and validation of rainfall-runoff processes. In: *Model validation,* M.G. Anderson and P.D. Bates (eds.). John Wiley & Sons, Chichester, UK, 117–161.