DATA-BASED MECHANISTIC MODELLING OF ENVIRONMENTAL SYSTEMS

Peter C. Young

Appears in Proc. IFAC Workshop on Environmental Systems, Yokohama, Japan, 2001

CRES, Institute of Environmental and Natural Sciences, Lancaster University, Lancaster LA1 4YQ, U.K., and
CRES, Australian National University, Canberra, Australia

Abstract: The paper discusses the problems associated with environmental modelling and the need to consider uncertainty in the formulation, identification, estimation and validation of environmental models. It introduces the concept of Data-Based Mechanistic (DBM) modelling and contrasts its inductive approach with the hypothetico-deductive approaches that dominate most environmental modelling research at the present time. The major methodological procedures utilized in DBM modelling are outlined and two practical examples illustrate how it has been applied in a hydrological and water quality context.

Keynote paper presented at the International Federation on Automatic Control (IFAC) Workshop on Environmental Systems, Tokyo, Japan, August 21st, 2001.

1. INTRODUCTION

The environment is a complex assemblage of interacting physical, chemical, and biological processes, many of which are inherently nonlinear, with considerable uncertainty about both their nature and their interconnections. It is surprising, therefore, that stochastic dynamic models are the exception rather than the rule in environmental science research. One reason for this anomaly lies in the very successful history of physical science over the last century. Modelling in deterministic terms has permeated scientific endeavour over this period and has led to a pattern of scientific investigation which is heavily reductionist in nature. Such deterministic reductionism appears to be guided by a belief that physical systems can be described very well, if not exactly, by deterministic mathematical equations based on well known scientific laws, provided only that sufficient detail can be included to describe all the physical processes that are *perceived* to be important by the scientists involved. This leads inexorably to large, nonlinear models reflecting the scientist's perception of the environment as an exceedingly complex dynamic system.

Although deterministic reductionism still dominates environmental modelling, there are some signs that attitudes may be changing. There is a growing realization that, despite their superficially rigorous scientific appearance, simulation models of the environment based on deterministic concepts are more extensions of our mental models and perceptions of the real world than necessarily accurate representations of the real world itself. The recent revived interest in the 'top-down' approach to modelling in the hydrological literature (e.g. Jothityangkoon *et al.*, 2000 and the references therein), for instance, is a response to the relative failure of the alternative reductionist ('bottom-up') philosophy in this area

[1] The author is very grateful to Chris Martinez and Dr. William R. Wise of the Environmental Engineering Sciences Department, University of Florida for giving permission the use their tracer experiment data; to Prof. Tony Jakeman of CRES, Australian National University, for supplying the Canning River data; and to Paul McKenna for carefully checking the manuscript. The author is, of course, responsible for any errors or omissions.

of study. But such scepticism is not new. It has its parallels in the environmental (e.g. Young, 1978, 1983; Beck, 1983) and ecosystems (e.g. see prior references cited in Silvert, 1993) literature of the 1970s and early 1980s, where the present author's contributions were set within the context of 'badly defined' environmental systems. To quote from Young (1983), which echoes earlier ideas (Young, 1978), for instance:

> "Although such reductionist analysis is perfectly respectable, it must be used very carefully; the dangers inherent in its application are manifold, but they are not, unfortunately, always acknowledged by its proponents. It is well known that a large and complex simulation model, of the kind that abounds in current ecological and environmental system analysis, has enormous explanatory potential and can usually be fitted easily to the meagre time-series data often used as the basis for such analysis. Yet even deterministic sensitivity analysis will reveal the limitation of the resulting model: many of the 'estimated' parameters are found to be ill-defined and only a comparatively small subset is important in explaining the observed system behavior.
>
> Of course, over-parameterization is quite often acknowledged, albeit implicitly, by the reductionist simulation model-builder. Realizing the excessive degrees of freedom available for fitting the model to the data, he will often fix the values of certain 'better known' parameters and then seek to fit the model by optimizing the chosen cost function (usually the sum of the squares of the difference between the model outputs and the observations) in relation to the remaining parameters, which are normally few in number. In this manner, the analyst ensures that the cost function-parameter hypersurface is dominated by a clearly defined optimum (a minimum in the least-squares case), so that estimation of the parameters which define the optimum becomes more straightforward.
>
> But what is the value of this optimization exercise in relation to the specification of the overall model? Clearly a lower-dimensional parameter space has been located which allows for the estimation of a unique set of parameter values. However, this has been obtained only at the cost of constraining the other model parameters to fixed values that are assumed to be known perfectly and are defined in relation to the analyst's prior knowledge of the system. As a result, the model has a degree of 'surplus content' not estimated from the available data, but based on a somewhat *ad hoc* evaluation of all available prior knowledge of the system and coloured by the analyst's preconceived notions of its behavioral mechanisms.
>
> On the surface, this conventional simulation modeling approach seems quite sensible: for example, the statistician with a Bayesian turn of mind might welcome its tendency to make use of all *a priori* information available about the system in order to derive the *a posteriori* model structure and parameters. On the other hand, he would probably be concerned that the chosen procedures could so easily be misused: whereas the constrained parameter optimization represents a quantitative and relatively objective approach, it is submerged rather arbitrarily within a more qualitative and subjective framework based on a mixture of academic judgment and intuition. Such a statistician would enquire, therefore, whether it is not possible to modify this framework so that the analyst cannot, unwittingly, put too much confidence in *a priori* perceptions of the system and so generate overconfidence in the resulting model."

These early papers then went on to present initial thoughts on such an objective, statistical approach to modelling poorly defined systems that tried to avoid the dangers of placing too much confidence in prior perceptions about the nature of the model. They also adumbrate very similar anti-reductionist arguments that have appeared recently in the hydrological literature and express some of these same views within a hydrological context (Jakeman and Hornberger, 1993, Beven, 2000). In the subsequent period since the earlier papers were published, however, the author has sought to develop this statistical approach within a more rigorous systems setting that he has termed *Data-Based Mechanistic* (DBM) modelling. Prior to discussing the DBM approach, the present paper will first outline the major concepts of statistical modelling that are important in any modelling process. Subsequently, a number of examples will be presented that illustrate the utility of DBM modelling in practical environmental science and systems analysis.

2. STATISTICAL IDENTIFICATION, ESTIMATION AND VALIDATION

The statistical approach to modelling assumes that the model is stochastic: in other words, no matter how good the model and how low the noise on the observational data happens to be, a certain level of uncertainty will remain after modelling has been completed. Consequently, full stochastic modelling requires that this uncertainty, which is associated with both the model parameters and the stochastic inputs, should be quantified in some manner as an inherent part of the modelling analysis.

Within the control and systems literature such a stochastic modelling procedure is usually termed 'Identification'. In the statistical, time series literature, however, it is normally considered in two main stages: *identification* of an appropriate, identifiable model structure; and *estimation* (optimization, calibration) of the parameters that characterize this structure, using some form of estimation or optimization. Sometimes, a further stage of *validation* (or *conditional validation*: see later) is defined, in which the ability of the model to explain the observed data is evaluated on data sets different to those used in the model identification and estimation stages. In this section, we outline these three stages in order to set the scene for the later analysis. This discussion is intentionally brief, however, since the topic is so large that a comprehensive review is not possible in the present context.

2.1 *Structure and Order Identification*

In the DBM approach to modelling, the identification stage is considered as a most important and essential prelude to the later stages of model building. It usually involves the identification of the most appropriate model order, as defined in dynamic system terms. However, the model structure itself can be the subject of the analysis if this is also considered to be ill-defined. In the DBM approach, for instance, the nature of linearity and nonlinearity in the model is not assumed *a priori* (unless there are good reasons for such assumptions based on previous data-based modelling studies). Rather it is identified from the data using nonparametric and parametric statistical estimation methods based on a suitable, generic model class. Once a suitable model structure has been defined within this class, there are a variety of statistical methods for identifying model order, some of which are mentioned later. In general, however, they exploit some order identification statistics, such as the correlation-based statistics popularized by Box and Jenkins (1970), the well known Akaike Information Criterion (AIC: Akaike, 1974), and the more heuristic YIC statistic (see e.g. Young *et al.*, 1996) which provides an alternative to the AIC in the case of transfer functions (where the AIC tends to identify over-parameterized models).

2.2 *Estimation (Optimization)*

Once the model structure and order have been identified, the parameters that characterize this structure need to be estimated in some manner. There are many automatic methods of estimation or optimization available in this age of the digital computer. These range from the simplest, deterministic procedures, usually based on the minimization of least squares cost functions, to more complex numerical optimization methods based on statistical concepts, such as Maximum Likelihood (ML). In general, the latter are more restricted, because of their underlying statistical assumptions, but they provide a more thoughtful and reliable approach to statistical inference; an approach which, when used correctly, includes the associated statistical diagnostic tests that are considered so important in statistical inference. In the present DBM modelling context, the estimation methods are based on optimal, linear Instrumental Variable (IV) methods for transfer function models (e.g. Young, 1984 and the references therein) and nonlinear modifications of these methods (see later).

2.3 *Conditional Validation*

Validation is a complex process and even its definition is controversial. Some academics (e.g. Konikow and Brederhoeft (1992), within a groundwater context; Oreskes *et al.* (1994), in relation to the whole of the earth sciences) question even the possibility of validating models. To some degree, however, these latter arguments are rather philosophical and linked, in part, to questions of semantics: what is the 'truth'; what is meant by terms such as validation, verification and confirmation? etc. Nevertheless, one specific, quantitative aspect of validation is widely accepted; namely 'predictive validation' (often referred to as just 'validation'), in which the predictive potential of the model is evaluated on data other than that used in the identification and estimation stages of the analysis. While Oreskes *et al.* (1994) dismiss this approach, which they term 'calibration and verification', their criticisms are rather weak and appear to be based on a perception that "models almost invariably need additional tuning during the verification stage". While some modellers may be unable to resist the temptation to carry out such additional tuning, so negating the objectivity of the validation exercise, it is a rather odd reason for calling the whole methodology into question. On the contrary, provided it proves practically feasible, there seems no doubt that validation, in the predictive sense it is used here, is an essential pre-requisite for any definition of model efficacy, if not validity in a wider sense.

It appears normal these days to follow the Popperian view of validation (Popper, 1959) and consider it as a continuing process of falsification. Here, it is assumed that scientific theories (models in the present context) can never be proven universally true; rather, they are not yet proven to be false. This yields a model that can be considered as 'conditionally valid', in the sense that it can be assumed to represent the best theory of behaviour currently available that has not yet been falsified. Thus, conditional validation means that the model has proven valid in this more narrow predictive sense. In the rainfall-flow context considered later, for example, it implies that, on the basis of the new measurements of the model input (rainfall) from the validation data set, the model produces flow predictions that are acceptable within the uncertainty bounds associated with the model.

Note this stress on the question of the inherent uncertainty in the estimated model: one advantage of statistical estimation, of the kind considered in this chapter, is that the level of uncertainty associated with the model parameters and the stochastic inputs is quantified in the time series analysis. Consequently, the modeller should not be looking for perfect predictability (which no-one expects anyway) but predictability which is consistent with the quantified uncertainty associated with the model.

It must be emphasized, of course, that conditional validation is simply a useful statistical diagnostic which ensures that the model has certain desirable properties. It is not a panacea and it certainly does not prove the complete validity of the model if, by this term, we mean the establishment of the 'truth' (Oreskes *et al.*, 1994). Models are, at best, approximations of reality designed for some specific objective; and conditional validation merely shows that this approximation is satisfactory in this limited predictive sense. In many environmental applications, however, such validation is sufficient to establish the credibility of the model and to justify its use in operational control, management and planning studies.

3. DATA-BASED MECHANISTIC (DBM) MODELLING

The term 'data-based mechanistic modelling' was first used in Young and Lees (1993) but the basic concepts of this approach to modelling dynamic systems have developed over many years. It was first applied within a hydrological context in the early 1970s, with application to modelling water quality in rivers (Beck and Young, 1975) and rainfall-flow processes (Young, 1974; Whitehead and Young, 1975). Indeed, the DBM water quality and rainfall-flow models discussed later in the present paper are a direct development of these early models.

In DBM modelling, the most parametrically efficient (parsimonious) model structure is first inferred statistically from the available time series data in an *inductive* manner, based on a generic class of black-box models (normally linear or nonlinear differential equations or their difference equation equivalents). *After this initial black-box modelling stage is complete*, the model is interpreted in a physically meaningful, mechanistic manner based on the nature of the system under study and the physical, chemical, biological or socio-economic laws that are most likely to control its behaviour. By delaying the mechanistic interpretation of the model in this manner, the DBM modeller avoids the temptation to attach too much importance to prior, subjective judgement when formulating the model equations. This inductive approach can be contrasted with the alternative *hypothetico-deductive* 'Grey-Box' modelling, approach, where the physically meaningful but simple model structure is based on prior, physically-based and possibly subjective assumptions, with the parameters that characterize this simplified structure estimated from data only *after* this structure has been specified by the modeller.

Other previous publications, as cited in Young (1998), map the evolution of the DBM philosophy and its methodological underpinning in considerable detail, and so it will suffice here to merely outline the main aspects of the approach:

(1) The important first step is to define the objectives of the modelling exercise and to consider the type of model that is most appropriate to meeting these objectives. Since DBM modelling requires adequate data if it is to be completely successful, this stage also includes considerations of scale and the data availability at this scale, particularly as they relate to the defined modelling objectives. However, the prior assumptions about the form and structure of this model are kept at a minimum in order to avoid the prejudicial imposition of untested perceptions about the nature and complexity of the model needed to meet the defined objectives.

(2) Appropriate model structures are identified by a process of objective statistical inference applied directly to the time-series data and based initially on a given generic class of linear Transfer Function (TF) models whose parameters are allowed to vary over time, if this seems necessary to satisfactorily explain the data.

(3) If the model is identified as predominantly linear or piece-wise linear, then the constant parameters that characterize the identified model structure in step 2. are estimated using advanced methods of statistical estimation for dynamic systems. The methods used in the present paper are based on optimal Instrumental Variable (IV) estimation algorithms (see Young, 1984) that provide a robust approach to model identification and estimation and have been well tested in practical applications over many years. Here the important identification stage means the application of objective statistical methods to determine the dynamic model order and structure. Full details of these time series methods are provided in the above references and they are outlined more briefly in both Young and Beven (1994) and Young *et al.* (1996).

(4) If significant parameter variation is detected over the observation interval, then the model parameters are estimated by the application of an approach to time dependent parameter estimation based on the application of recursive Fixed Interval Smoothing (FIS) algorithms (e.g. Bryson and Ho, 1969; Young, 1984; Norton, 1986). Such parameter variation will tend to reflect nonstationary and nonlinear aspects of the observed system behaviour. In effect, the FIS algorithm provides a method of nonparametric estimation, with the *Time Variable Parameter*

(TVP) estimates defining the nonparametric relationship, which can often be interpreted in *State-Dependent Parameter* (SDP) terms (see later).

(5) If nonlinear phenomena have been detected and identified in stage 4, the nonparametric state dependent relationships are normally parameterized in a finite form and the resulting nonlinear model is estimated using some form of numerical optimization, such as nonlinear least squares or Maximum Likelihood (ML) optimization.

(6) Regardless of whether the model is identified and estimated in linear or nonlinear form, *it is only accepted as a credible representation of the system if, in addition to explaining the data well, it also provides a description that has direct relevance to the physical reality of the system under study.* This is a most important aspect of DBM modelling and differentiates it from more classical 'black-box' modelling methodologies, such as those associated with standard TF, nonlinear autoregressive-moving average-exogenous variables (NARMAX), neural network and neuro-fuzzy models.

(7) Finally, the estimated model is tested in various ways to ensure that it is conditionally valid (see Young, 2001a,b). This can involve standard statistical diagnostic tests for stochastic, dynamic models, including analysis which ensures that the nonlinear effects have been modelled adequately (e.g. Billings and Voon, 1986). It also involves validation exercises, as well as exercises in stochastic uncertainty and sensitivity analysis.

Of course, while step 6. should ensure that the model equations have an acceptable physical interpretation, it does not guarantee that this interpretation will necessarily conform exactly with the current scientific paradigms. Indeed, one of the most exciting, albeit controversial, aspects of DBM models is that they can tend to question such paradigms. For example, DBM methods have been applied very successfully to the characterization of imperfect mixing in fluid flow processes and, in the case of pollutant transport in rivers, have led to the development of the *Aggregated Dead Zone* (ADZ) model (Beer and Young, 1983; Wallis *et al.*, 1989). Despite its initially unusual physical interpretation, the acceptance of this ADZ model (e.g. Davis and Atkinson, 2000 and the prior references therein) and its formulation in terms of physically meaningful parameters, seriously questions certain aspects of the ubiquitous Advection Dispersion Model (ADE) which preceded it as the most credible theory of pollutant transport in stream channels (see the comparative discussion in Young and Wallis, 1994).

One aspect of the above DBM approach which differentiates it from alternative deterministic 'top-down' approaches (e.g. Jothityangkoon *et al.*, 2000) is its inherently stochastic nature. This means that the uncertainty in the estimated model is always quantified and this information can then be utilized in various ways. For instance, it allows for the application of Monte Carlo-based uncertainty and sensitivity analysis, as well as the use of the model in statistical forecasting and data assimilation algorithms, such as the Kalman filter. The uncertainty analysis is particularly useful because it is able to evaluate how the covariance properties of the parameter estimates affect the probability distributions of physically meaningful, derived parameters, such as residence times and partition percentages in parallel hydrological pathways (see e.g. Young, 1992, 1999a and the examples below).

The DBM approach to modelling is widely applicable: It has been applied successfully to the characterization of numerous environmental systems including: the development of the ADZ model for pollution transport and dispersion in rivers (e.g. Wallis *et al.*, 1989; Young, 1992); rainfall-flow modelling and forecasting (Young, 2001b and the prior references therein); adaptive flood warning (Lees *et al.*, 1994; Young and Tomlin, 2000); and the modelling of ecological and biological systems (Jarvis *et al.*, 1999); Other applications, in which the DBM models are subsequently utilized for control system design, include: the modelling and control of climate in glasshouses (e.g. Lees *et al.*, 1996), forced ventilation in agricultural buildings (e.g. Price *et al.*, 1999), and inter-urban road traffic systems (Taylor *et al.*, 1998). They have also been applied in the context of macro-economic modelling (e.g. Young and Pedregal, 1999).

4. THE STATISTICAL TOOLS OF DBM MODELLING

The statistical and other tools that underpin DBM modelling are dominated by recursive methods of time series analysis, filtering and smoothing. These include: optimal *Instrumental Variable* (IV) methods of identifying and estimating discrete and continuous-time transfer function models (e.g. Young, 1984); *Time Variable Parameter* (TVP) estimation and its use in the modelling and forecasting of nonstationary time series (e.g. Young, 1999b and the prior references therein); and *State Dependent Parameter* (SDP) parameter estimation methods for modelling nonlinear stochastic systems (see Young, 1978, 1984, 1993, 1998, 2000, 2001a; Young and Beven, 1994). Here, the TVP and SDP estimation is based on optimized *Fixed Interval Smoothing* (FIS) algorithms.

These recursive statistical methods can be utilized also for other environmental purposes. For example, they can provide a rigorous approach to the evaluation and exploitation of large simulation models (e.g. Young *et al.*, 1996), where the analysis provides a means of simplifying the models. Such *reduced order* representations can then provide a better understanding of the most important mechanisms within the model; or they can provide *dominant mode* models that can be used for control and operational management system design, adaptive forecasting, or data assimilation purposes (see e.g. Young *et al.*, 1996; Young, 1999a). The DBM data analysis tools can also be used for the statistical analysis of nonstationary data, of the kind encountered in many areas of environmental science (see Young, 1999b). For example, they have been applied to the analysis and forecasting of trends and seasonality in climate data (e.g. Young *et al.*, 1991); and the analysis of palaeoclimatic data (Young and Pedregal, 1998).

5. PRACTICAL EXAMPLES

Two practical examples will be considered here, both concerned with hydrological systems. The first will show how even purely linear DBM modelling can provide a powerful approach to analyzing experimental data. However, many environmental systems are nonlinear and so the second example will show SDP modelling procedures can be exploited to handle such nonlinearity.

5.1 *A Linear Example: Modelling Solute Transport*

The first model to be considered seriously in DBM terms was the ADZ model for the transport and dispersion of solutes in river systems, as mentioned earlier. This model has also led to related models that describe the imperfect mixing processes that characterize mass and energy flow processes in the wider environment (see e.g. Young and Lees, 1993; Price *et al.*, 1999)

This example is concerned with the DBM/ADZ modelling of input-output data shown in figure 1. These were obtained from a bromide tracer experiment carried out in a Florida wetland area receiving treated domestic wastewater for further nutrient removal. The experiment was part of a study carried out by Chris Martinez and Dr. William R. Wise of the Environmental Engineering Sciences Department, University of Florida for the City of Orlando. The study objective was to determine residence times for each wetland cell in the system and to assess whether the same degree of treatment could be maintained should the wastewater loading be raised from 16 to 20 million gallons per day. The bromide tracer was injected 765 metres upstream of a weir, at which samples were taken with a sampling interval Δt of 2 hours.

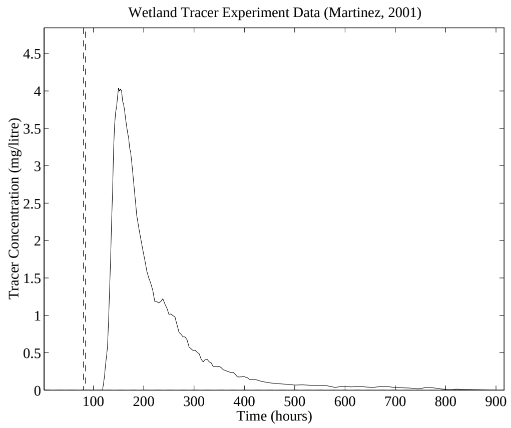

Fig. 1. Wetland tracer experiment data: the input u_t is an impulsive or 'gulp' application of bromide tracer (dashed line) and the output y_t is the concentration of bromide measured every two hours at a downstream weir (full line).

The first step in DBM modelling is to identify a suitable model from a generic model class that is both capable of explaining the data in a parametrically efficient manner and producing a model that can be interpreted in physical terms. Based on the previous research described in the above references, a reasonable model class is the linear TF model in continuous or discrete time form. As we shall see, such TF models are not only able to explain the tracer data well, they can also be interpreted in multi-reach ADZ model terms that have physical meaning. Here, we will consider the discrete-time TF model and utilize the SRIV algorithm (a simplified version of the optimal IV algorithm mentioned earlier) to identify the model order and estimate the parameters [2].

The impulsive input is not persistently exciting but the SRIV algorithm has no difficulty identifying and estimating a low order model. The best identified TF, based on the YIC criterion, is either 3^{rd} or 4^{th} order but subsequent analysis, described below, suggests that the latter is superior from a physical standpoint. The estimated [4 2 22] (4^{th} order denominator, 2^{nd} order numerator and a 22 sampling interval pure time delay) takes the form:

$$y_t = \frac{\hat{B}(z^{-1})}{\hat{A}(z^{-1})} u_{t-22} + \xi_t \qquad (1)$$

where,

$$\hat{A}(z^{-1}) = 1 - 3.67z^{-1} + 5.06z^{-2} - 3.11z^{-3} + 0.72z^{-4}$$

[2] Continuous time TF estimation using the continuous-time SRIV algorithm yields similar (albeit not identical) results but the discrete-time analysis is more convenient in terms of the subsequent analysis.

$\hat{B}(z^{-1}) = 0.00103 - 0.00101z^{-1}$

Here the 'hat' denotes the estimated value; z^{-i} is the backward shift operator (i.e. $z^{-i}y_t = y_{t-i}$); y_t is the observed tracer concentration at the weir and u_t is the impulsive input of tracer (186.33 mg/l), both measured at the t^{th} sampling instant. Note that the large 'advective' time delay of 22 sampling intervals (44 hrs.) is the time taken for the solute to first reach the weir. The noise ξ_t, which represents the quantification of all stochastic influences including measurement noise, is small and the model explains the data very well with a *Coefficient of Determination* (or *Nash Efficiency* in the hydrological literature) based on the response error of $R_T^2 = 0.997$ (i.e. 99.7% of the output variance is explained by the model).

Unfortunately, despite its ability to describe the data very well, the model (1) is not immediately acceptable from a DBM standpoint, primarily because the eigenvalues are $\{0.988, 0.964, 0.860 \pm 0.132j\}$ and the pair of complex roots is difficult to justify in ADZ modelling terms. In particular, the elemental, single reach ADZ model is a first order differential equation and so, other than in exceptional circumstances, multiple reach ADZ models must be characterized by real eigenvalues when considered in TF terms.

In the present circumstances, the most obvious approach is to re-estimate the model in a form where the eigenvalues are constrained to be real. This was carried out by means of constrained nonlinear least squares optimization using the *leastsq* optimization procedure in Matlab$^{\text{TM}}$. To ensure that the most parametrically efficient model was obtained, both [3 2 22] and [4 2 22] models were considered in this analysis but the latter yielded much the best constrained model, which has the following form:

$$y_t = \frac{\hat{B}(z^{-1})}{\hat{A}(z^{-1})} u_{t-22} + \xi_t \qquad (2)$$

where,

$\hat{A}(z^{-1}) = (1 - 0.980z^{-1})(1 - 0.855z^{-1})^3$

$\hat{B}(z^{-1}) = 0.00127 - 0.00121z^{-1}$

This model is well defined statistically and it explains 99.7% of the experimental data ($R_T^2 = 0.997$), the same as the unconstrained model (1). Figure 2 compares the model output (full line) with the measured output y_t (circular points).

Unlike the TF model (1), the model (2) not only has four real eigenvalues, as required, but three of these are repeated, so defining three identical ADZ reaches. These eigenvalues define ADZ residence times (time constants) of 99 hours and 12.8 hours (x3), giving a total estimated residence time for the wetland cell is 137.4 hours (99+3x12.8). One particular physically meaningful decomposition and interpretation of the model defined in this manner is obtained by partial fraction expansion of the TF in (2). This consists of two parallel pathways, each consisting of three ADZ reaches, as shown in the top block diagram of figure 3.

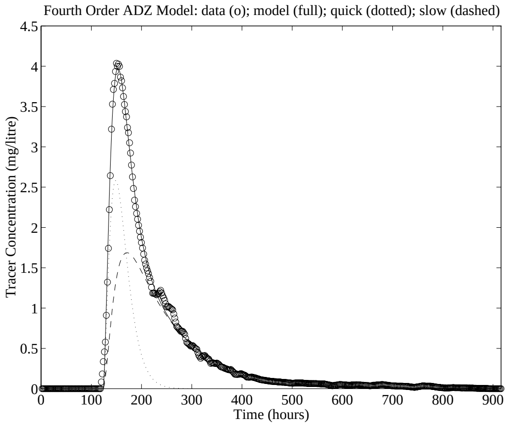

Fig. 2. Comparison of the DBM model output (full line) and tracer experiment data (circular points). Also shown are the inferred slow flow component (dashed line) and quick flow component (dotted line).

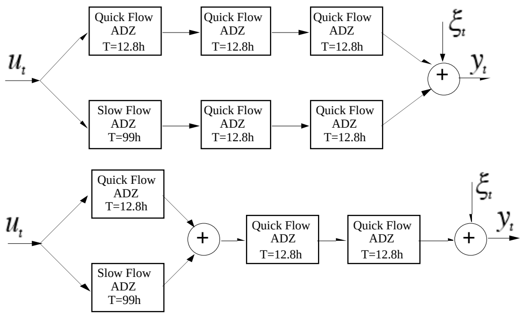

Fig. 3. Block diagram of transfer function decompositions that can be interpreted in physical terms: fully parallel decomposition (upper plot); equivalent parallel-serial decomposition (lower plot).

The 'quick-flow' pathway has three identical ADZ reaches connected in series, each with a residence time of 12.8 hours; while the 'slow-flow' pathway is similar but with one of the reaches replaced by the longer ADZ residence time of 99 hours associated with the other identified eigenvalue (0.98). The total *travel time* for this complete system is 181.5 hours (the sum of the 44 hour advective time delay and the cumulative overall time constant of 137.4 hours). This means that the '*dispersive fraction*' (see Wallis *et al.*, 1989; Young and Wallis, 1994; Young, 1999a) is 0.76 (i.e. 137.4 ÷ 181.5): in other words, 76% of the water appears to be effective in dispersing the solute. This is a very high proportion, reflecting the nature of the system in this case, with a much higher potential for dispersion of tracer than in normal, faster moving streams, where the dispersive fraction is normally in the range 0.3-0.4. The inferred responses of the two parallel pathways are plotted in figure 2: the dotted line shows the estimated concentration changes in the quick pathway, which accounts mainly for the initial response measured at the weir; the dashed line are the estimated changes in the slow pathway, and these are responsible for the raised tail of the measured response.

It is possible to compute estimates of other physical attributes associated with the model. First, the steady state gains associated with the two parallel pathways define the partitioning of the flow, with 33% of flow associated with the quick pathway and 67% with the slow pathway. And since the flow rate is known in this example, the *Active Mixing Volumes* (AMVs: Young and Lees, 1993), based on the estimated partitioned flow, are $361m^3$ in the quick pathway and $5,656m^3$ in the slow pathway. As a result, the total estimated AMV is $5,656 + 3\text{x}361 = 6739m^3$, which seems reasonable when compared with the $9,749m^3$ for the total volume of the wetland, estimated by physical measurement. This suggests that about 70% of the wetland is important in dispersing the tracer (and, therefore, the waste water) and compares reasonably with the dispersive fraction derived percentage of 76%.

Of course, all of the results above are statistical estimates and so they are inherently uncertain. The advantage of the DBM approach is that we can quantify and consider the consequences of this uncertainty (see Young, 1999a). For instance, based on the covariance matrix of the parameters produced by the SRIV estimation analysis, empirical probability distributions, in the form of histograms, can be computed for the 'derived' physical parameters, such as the residence times, partition percentages, AMVs, total AMV and steady state gain, using *Monte Carlo Simulation* (MCS) analysis. Figure 4 is a typical example of such analysis: it shows the normalized empirical distributions for the two residence times obtained by MCS using 10,000 random realizations (the procedure used here is discussed in Young, 1999a).

Of course, it should be noted that the parallel decomposition of the estimated TF used above is not unique: there are other decompositions that are just as valid and give precisely the same y_t response. For example, two other examples are: (i) a parallel decomposition of the two ADZs with residence times 99 and 12.8 hours, in series with two other identical ADZs, both with residence times 12.8 hours (see lower block diagram of figure 3); (ii) various decompositions including feedback processes. However the latter seem less supportable in physical terms and are rejected according to the DBM ethos.

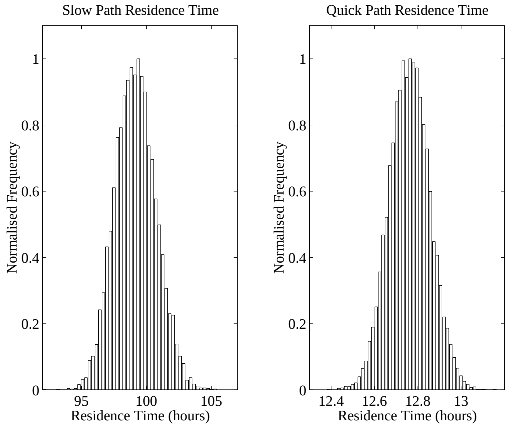

Fig. 4. MCS analysis results: normalized histograms of the slow (left panel) and quick (right panel) residence times.

Finally, how can decompositions of ADZ reaches, such as those shown in figure 3 be interpreted in terms of the wetland system? The most plausible mechanism is that the quick parallel pathway represents the 'main stream-flow' that is relatively unhindered by the vegetation; while the slow pathway represents the solute that is captured by the heavy vegetation and so dispersed more widely and slowly before reaching the weir. It is this latter pathway, which we have shown above accounts for some 67% of the flow, together with the large associated dispersive fraction of 76%, that is most useful in terms of nutrient removal, since it allows more time for the biological activity to take place.

5.2 *A Nonlinear Example: Rainfall-Flow Modelling*

This example is concerned with the analysis of daily rainfall, flow and temperature data from the 'ephemeral' Canning River in Western Australia which stops flowing over Summer, as shown in figure 5. These data have been analyzed before and reported fully in Young *et al.* (1997). The results of this previous analysis are outlined briefly below but most attention is focussed on more recent analysis that shows how the inductive DBM modelling can help to develop and enhance alternative conceptual ('grey-box') modelling that has been carried out previously in a more conventional hypothetico-deductive manner.

Young *et al.* (1997) show that, in this example, the most appropriate generic model form is the nonlinear SDP model (see above, section 4). Analysis of the rainfall-flow data in figure 5, based on this type of model, is accomplished in two stages. First, nonparametric estimates of the SDPs are obtained using the *State Dependent parameter Auto-Regressive eXogenous Vari-*

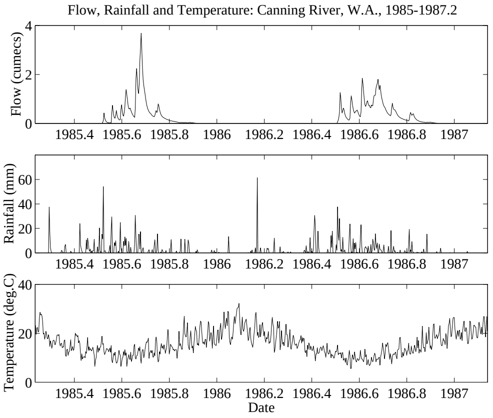

Fig. 5. Daily rainfall-flow and temperature data for the ephemeral Canning River in Western Australia for the period 23^{rd} March 1985 to 26^{th} February, 1987.

able (SDARX) model form (see Young, 2001a,b in which it is discussed at some length within a rainfall-flow context):

$$y_t = \mathbf{z}_t^T \mathbf{p}_t + e_t \qquad e_t = N(0, \sigma^2) \qquad (3)$$

where,

$$\begin{aligned} \mathbf{z}_t^T &= [y_{t-1} \quad y_{t-2} \quad \ldots \quad y_{t-n} \quad r_{t-\delta} \ldots \quad r_{t-\delta-m}] \\ \mathbf{p}_t &= [a_1(z_t) \quad a_2(z_t) \ldots a_n(z_t) \quad b_0(z_t) \ldots b_m(z_t)]^T \end{aligned}$$

where, in the present context, y_t and r_t are, respectively, the measured flow and rainfall and δ is a pure advective time delay. Here, $n = 2$, $m = 3$, $\delta = 0$ and the parameters are all assumed initially to be dependent on a state variable z_t. In this case, the SDP analysis then shows that the state dependency is apparently in terms of the measured flow variable (i.e. $z_t = y_t$: see later explanation) and is limited to those parameters associated with the rainfall r_t.

In the second stage of the analysis, the nonparametric estimate of the nonlinearity is parameterized in the simplest manner possible; in this case as a power law in y_t. The constant parameters of this parameterized nonlinear model are then estimated using a nonlinear optimization procedure (see Young, 2001b). The resulting model is the following simplified version of the nonlinear *SDP Transfer Function* (SDTF) model (e.g. Young, 2000):

$$y_t = \frac{\hat{B}(z^{-1})}{\hat{A}(z^{-1})} u_t + \xi_t \qquad (4)$$

where,

$$\hat{A}(z^{-1}) = 1 - 1.646z^{-1} + 0.658z^{-2}$$

$$\hat{B}(z^{-1}) = 0.0115 + 0.0185z^{-1} - 0.0028z^{-2}$$

and

$$u_t = c.y_t^{\hat{\beta}}.r_t \qquad (5)$$

with $\hat{\beta} = 0.85$. This shows that the input variable u_t is a nonlinear function in which the measured rainfall r_t is multiplied by the flow raised to a power $\hat{\beta}$, with the normalization parameter c simply chosen so that the steady state gain of the linear TF between u_t and y_t is unity[3]. In other words, the SDP analysis shows, in a relatively objective manner, that the underlying *dynamics* are predominantly linear but the overall response is made nonlinear because of a very significant input nonlinearity.

This model not only explains the data well ($R_T^2 = 0.958$) it is also consistent with hydrological theory, as required by the tenets of DBM modelling. This suggests that the changing soil-water storage conditions in the catchment reduce the '*effective*' level of the rainfall and that the relationship between the measured rainfall and this effective rainfall (or rainfall excess) u_t is quite nonlinear. For example, if the catchment is very dry because little rain has fallen for some time, then most new rainfall will be absorbed by the dry soil and little, if any, will be effective in promoting increases in river flow. Subsequently, however, if the soil-water storage increases because of further rainfall, so the '*run-off*' of excess water from the catchment rises and the flow increases because of this. In this manner, the effect of rainfall on flow depends upon the antecedent conditions in the catchment and a similar rainfall event occurring at different times and under different soil-water storage conditions can yield markedly different changes in river flow.

The linear TF part of the model conforms also with the classical 'unit hydrograph' theory of rainfall-flow dynamics: indeed, its unit impulse response at any time is, by definition, the unit hydrograph. And the TF model itself can be seen as a parametrically efficient method of quantifying this unit hydrograph. Additionally, as in the solute transport example, the TF model can be decomposed by partial fraction expansion into a parallel pathway form which has a clear hydrological interpretation. In particular, it suggests that the effective rainfall is partitioned into three pathways: the instantaneous effect, arising from the [2 3 0] TF model form which, as might be expected, accounts for only a small 5.8% of the flow; a fast flow pathway with a residence time of 2.65 days which accounts for the largest 53.9% of the flow; and a slow flow pathway of 19.86 days residence time accounting for the remaining 40.3% of the flow. It is this latter pathway that leads to an extended tail on the associated hydrograph and can be associated with the slowly changing

[3] This is an arbitrary decision in this case. However, if the rainfall and flow are in the same units, then this ensures that the total volume of effective rainfall is the same as the total flow volume.

baseflow in the river. (for a more detailed explanation and other examples, see Young, 1992, 1993, 1998, 2001b, Young and Beven, 1994; Young *et al.*, 1997).

The most paradoxical and, at first sight, least interpretable model characteristic is that the effective rainfall nonlinearity is a function flow. Although this is physically impossible, the analysis produces such a clearly defined relationship of this sort that it must have some physical connotations. The most hydrologically reasonable explanation is that the flow is acting as a surrogate for soil water storage. Of course, it would be better to investigate this relationship directly by measuring the soil-water storage in some manner and incorporating these measurements into the SDP analysis. Unfortunately, it is much more difficult to obtain such 'soil moisture' measures and these were not available in the present example.

The temperature measurements are available, however, and this suggests that we should explore the model (4) further, with the object of enhancing its physical interpretation using these additional data. Two interesting conceptual ('grey-box') models of rainfall-flow dynamics are the Bedford-Ouse River model (e.g. Whitehead and Young, 1975); and a development of this, the IHACRES model (Jakeman *et al.*, 1990). Both of these '*Hybrid-Metric-Conceptual*' (HCM) models (Wheater *et al.*, 1993) have the same basic form as (4), except that the nature of the effective rainfall nonlinearity is somewhat different. In the case of the IHACRES model, for instance, this nonlinearity is modelled by the following equations:

$$\tau_s(T_t) = \tau_s e^{\frac{\bar{T}_t - T_t}{f}} \tag{6a}$$

$$s_t = s_{t-1} + \frac{1}{\tau_s(T_t)}(r_t - s_{t-1}) \tag{6b}$$

$$u_t = c.s_t^{\beta}.r_t \tag{6c}$$

where T_t is the temperature; $\bar{T}_t$ is the mean temperature; s_t represents a conceptual soil-water storage variable; and c, τ_s, f and β are *a priori* unknown parameters. Comparing (6c) with (5), we see that the main difference between the two models is that the measured y_t in (5), acting as a surrogate for soil-water storage, has been replaced by a modelled (or latent) soil-water storage variable s_t. The model (6b) that generates this variable is a first order discrete-time storage equation with a residence time $\tau_s(T_t)$ defined as τ_s multiplied by an exponential function of the difference between the temperature T_t and its mean value $\bar{T}_t$, as defined in (6a).

In the original IHACRES model (e.g. Jakeman and Hornberger, 1993), $\bar{T}_t$ is normally set at 20°C, but the estimation results are not sensitive to this value. Also, s_t is not raised to a power, as in (6c). Some later versions of IHACRES have incorporated this parameter, but it has been added here so that the two nonlinearities in (5) and (6c) can be compared. More importantly, its introduction is practically important in this particular example since, without modification, the IHACRES model is not able to model the ephemeral Canning flow very well.

Using a constrained nonlinear optimization procedure procedure similar to that in the previous example, the parameters in this modified IHACRES model are estimated as follows:

$$\hat{A}(z^{-1}) = 1 - 1.737z^{-1} + 0.745z^{-2}$$

$$\hat{B}(z^{-1}) = 0.0285 + 0.140z^{-1} - 0.160z^{-2}$$

$$\hat{\tau}_s = 65.5, \hat{f} = 30.1; \hat{\beta} = 6.1 \text{ and } \bar{T}_t = 15.9$$

These parameters are all statistically well defined and the model explains 96.9% of the flow y_t ($R_T^2 = 0.969$), marginally better than the DBM model. Moreover, as shown in figure 6, it performs well in validation terms when applied, without re-estimation, to the data for the years 1977-78. The coefficient of determination in this case is 0.91 which is again better than that achieved by the DBM model (0.88). However, when validated against the 1978-79 data, the positions are reversed and the DBM model is superior. Overall, therefore, the two models are comparable in their ability to explain and predict the Canning River data.

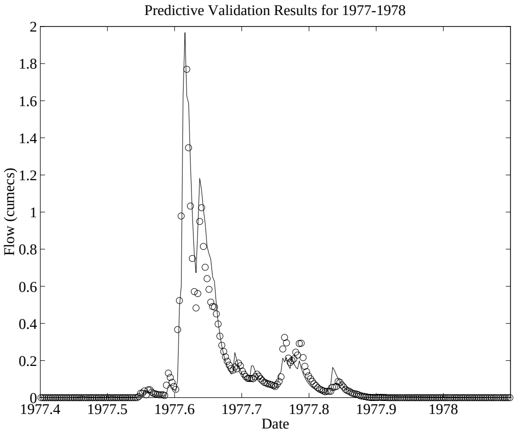

Fig. 6. Validation results on 1977-78 data.

Figure 7 compares the measured rainfall r_t (upper panel) with the effective rainfall (middle panel), as computed by equation (6c). It is clear that, not surprisingly, the nonlinear transformation produced by equations (6a)-(6c) has a marked effect: in particular, as shown in figure 8, the power law transformation in (6c), with $\beta = 6.1$ considerably modifies the soil-water storage s_t, effectively reducing it to zero, in relative terms, over the

Summer period, as required. The reason why the modified IHACRES and DBM models perform similarly becomes clear if we compare the normalized (since they differ by a scale factor) effective rainfall variables for both models, as shown in the lower panel of figure 7 (modified IHACRES, full line; DBM, dashed line). The similarity between these variables is obvious and the normalized impulse responses (unit hydrographs) of the models are also closely comparable.

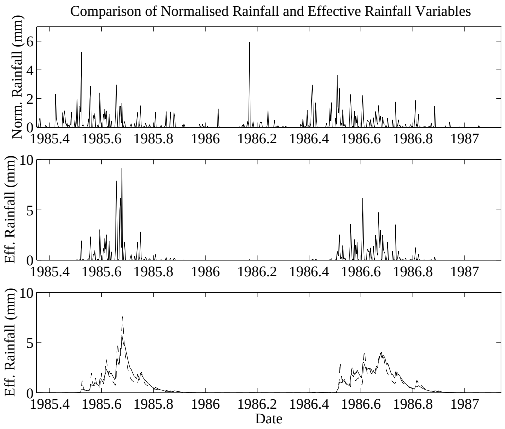

Fig. 7. Comparison of rainfall and effective rainfall measures.

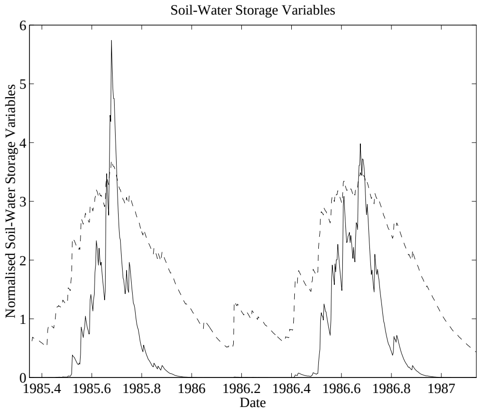

Fig. 8. Comparison of the estimated soil-water storage variable s_t (dashed line) and s_t^{β} (full line).

Finally, it must be emphasized that this example is purely illustrative and it is not suggested that the modified IHACRES model identified here cannot be improved upon by the introduction of some alternative nonlinear mechanism. For instance, the estimation of a power law nonlinearity with such a large power of 6.1 seems a rather odd way to handle this type of nonlinearity, although the reader will see from figures 7 and 8 that it is very effective. Nevertheless, the example illustrates well how DBM modelling can, in a reasonably objective manner, reveal the *nature* of the nonlinearity required to model the data well and then seek out a parameterization that achieves this. In this example, it clearly demonstrates that the standard IHACRES model nonlinearity cannot do this unless it is modified in some manner. Of course, the power law nonlinearity is not the only nor necessarily the best way of achieving this. For example, Ye *et al.* (1997) introduce a threshold-type nonlinearity on s_t, which makes good physical sense, and obtain reasonable results with R_T^2 values of around 0.88-0.89 for estimation and 0.82-0.88 for validation.

6. CONCLUSIONS

For too long in the environmental sciences, deterministic reductionism has reigned supreme and has had a dominating influence on mathematical modelling in almost all areas of the discipline. This paper has argued that the uncertainty which pervades most environmental systems demands an alternative approach, where stochastic models and statistical modelling procedures provide a means of acknowledging this uncertainty and quantifying its effects. But the conventional statistical approach to stochastic model building is too often posed in a 'black-box' manner that fails to produce models that can be interpreted in physically meaningful terms. The *Data-Based Mechanistic* (DBM) approach to modelling discussed in this paper tries to correct these deficiencies. It provides a modelling strategy that not only exploits powerful statistical techniques but also produces models that can be interpreted in physically meaningful terms that are normally more acceptable to environmental scientists and engineers.

7. REFERENCES

Akaike, H. (1974) A new look at statistical model identification, *I.E.E.E. Trans. Auto. Control*, **ACl9**, 716-722.

Beck, M. B. (1983) Uncertainty, system identification and the prediction of water quality, in *Uncertainty and Forecasting of Water Quality*, M.B. Beck and G. Van Straten (Eds.), Springer-Verlag: Berlin, 3-68.

Beck, M. B. and Young, P. C. (1975) A dynamic model for BOD-DO relationships in a non-tidal stream, *Water Research*, **9**, 769-776.

Beer, T. and Young, P. C. (1983) Longitudinal dispersion in natural streams, *Jnl. Env. Eng. Div., American Soc. Civ. Eng.*, **102**, 1049-1067.

Beven, K. J. (2000) Uniqueness of place and process representations in hydrological modelling, *Hydrology and Earth System Sciences*, **4**, 203-213.

Billings, S.A. and Voon, W.S.F. (1986) Correlation based model validity tests for nonlinear models, *Int. Journal of Control*, **44**, 235–244.

Box, G. E. P., and Jenkins, G. M., (1970) *Time-Series Analysis: Forecasting and Control*, San Francisco: Holden-Day.

Bryson, A. E. and Ho, Y-C. (1969) *Applied Optimal Control*, Blaisdell Publishing: Massachusetts.

Davis, P. M. and Atkinson, T. C. (2000) Longitudinal dispersion in natural channels: 3. An aggregated dead zone model applied to the River Severn, U.K., *Hydrology and Earth System Sciences*, **4**, 373-381.

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

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

Jarvis, A. J., Young, P. C., Taylor, C. J., and Davies, W. J. (1999). An analysis of the dynamic response of stomatal conductance to a reduction in humidity over leaves of *cedrella odorata*, *Plant Cell and Environment*, **22**, 913-924.

Jothityangkoon, C., Sivapalan, M. and Farmer, D. L. (2000) Process controls of water balance variability in a large semi-arid catchment: downward approach to hydrological model development, submitted for publication.

Konikow, L.F. and Bredehoeft, J.D. (1992) Ground water models cannot be validated, *Advances in Water Resources*, **15**, 75-83.

Lees, M. J, Taylor, J., Chotai, A., Young, P. C., and Chalabi, Z. S. (1996). Design and implementation of a Proportional-Integral-Plus (PIP) control system for temperature, humidity and carbon dioxide in a glasshouse. *Acta Horticulturae*, **406**, 115-123.

Lees, M., Young, P.C., Beven, K. J., Ferguson, S. and Burns, J. (1994). An adaptive flood warning system for the River Nith at Dumfries. In *River Flood Hydraulics*, (W. R. White and J. Watts eds.), Institute of Hydrology: Wallingford.

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

Oreskes, N., Shrader-Frechette, K. and Belitz, K. (1994) Verification, validation, and confirmation of numerical models in the earth sciences, *Science*, **263**, 641-646.

Popper, K. (1959) *The Logic of Scientific Discovery*, Hutchinson: London.

Price, L., Young, P. C., Berckmans, D., Janssens, K. and Taylor, J. (1999) Data-based mechanistic modelling and control of mass and energy transfer in agricultural buildings, *Annual Reviews in Control*, **23**, 71-82.

Silvert, W. (1993) Top-down modelling in ecology. In *Concise Encyclopedia of Environmental Systems*, P. C. Young (Ed.), Pergamon Press: Oxford, 605.

Taylor, C. J., Young, P. C., Chotai, A., and Whittaker, J. (1998). Nonminimal state space approach to multivariable ramp metering control of motorway bottlenecks. *IEE Proc., Control Theory Appl.*, **145**, 568-574.

Wallis, S.G., P.C. Young and Beven, K. J. (1989). Experimental investigation of the Aggregated Dead Zone (ADZ) model for longitudinal solute transport in stream channels. *Proc. Inst. of Civil Engrs*, Part 2, **87**, 1-22.

Wheater, H. S., Jakeman, A. J. and Beven. K. J. (1993). Progress and directions in rainfall-run-off modelling. Chapter 5 in A. J. Jakeman, M. B. Beck and M. J. McAleer (eds.), *Modelling Change in Environmental Systems*, Wiley: Chichester, 101-132.

Whitehead, P.G. and Young, P.C. (1975) A dynamic-stochastic model for water quality in part of the Bedford-Ouse River system. In *Computer Simulation of Water Resources Systems*, G.C. Vansteenkiste (Ed.), North Holland: Amsterdam, 417-438.

Ye, W., B.C. Bates, N.R. Viney, M. Sivapalan, and A.J. Jakeman (1997) Performance of conceptual rainfall-runoff models in low-yielding ephemeral catchments, *Water Resources Research*, **33**, 153-166.

Young, P.C. (1974) Recursive approaches to time-series analysis. *Bull. of Inst. Maths and its Applications*, **10**, 209-224.

Young, P.C. (1978) A general theory of modeling for badly defined dynamic systems. In *Modeling, Identification and Control in Environmental Systems*, G.C. Vansteenkiste (Ed.), North Holland, 103-135.

Young, P.C. (1983) The validity and credibility of models for badly defined systems. In *Uncertainty and Forecasting of Water Quality*, M.B. Beck and G. Van Straten (Eds.), Springer-Verlag: Berlin, 69-100.

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

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

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

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

Young, P.C. (1999a) Data-based mechanistic modelling, generalized sensitivity and dominant mode analysis, *Computer Physics Communications* **115** 1-17.

Young, P. C. (1999b). Nonstationary time series analysis and forecasting, *Progress in Environmental Science,* **1**, 3-48.

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

Young, P.C. (2001a) The identification and estimation of nonlinear stochastic systems. In *Nonlinear Dynamics and Statistics*, A.I. Mees (Ed.), Birkhauser: Boston. In press.

Young, P. C. (2001b). Data-based mechanistic modelling and validation of rainfall-flow processes. In M. G. Anderson (Ed.), Model Validation in Hydrological Science. J. Wiley: Chichester. In press.

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

Young, P.C. and Lees, M. J. (1993) The active mixing volume: a new concept in modelling environmental systems. In *Statistics for the Environment*, V. Barnett and K. F. Turkman (Eds.), J. Wiley: Chichester, 3-43.

Young, P. C. and D. J. Pedregal (1998). Recursive and en-bloc approaches to signal extraction. *Journal of Applied Statistics,* **26**, 103-128.

Young, P. C. and D. J. Pedregal (1999). Macro-economic relativity: government spending, private investment and unemployment in the USA 1948-1998. *Jnl. Structural Change and Economic Dynamics* **10**, 359-380.

Young, P. C. and Tomlin, C. M. (2000) Data-based mechanistic modelling and adaptive flow forecasting. In *Flood Forecasting: What Does Current Research Offer the Practitioner?*, M. J. Lees and P. Walsh (Eds.), BHS Occasional paper No 12, produced by the Centre for Ecology and Hydrology on behalf of the British Hydrological Society, 26-40.

Young, P. C. and Wallis S, G. (1994) Solute transport and dispersion in channels. Chapter 6 in *Channel Networks*, K.J. Beven and M.J. Kirkby (Eds.), Wiley: Chichester, 129-173.

Young, P.C., Jakeman, A. J. and Post, D. A. (1997) Recent advances in data-based modelling and analysis of hydrological systems, *Water Sci. Tech.*, **36**, 99-116.

Young, P. C., Parkinson, S.D., and Lees, M. (1996) Simplicity out of complexity in environmental systems: Occam's Razor revisited, *J. Appl. Stat.*, **23**, 165-210.

Young, P.C., Ng, C.N., Lane, K. and Parker, D. (1991). Recursive forecasting, smoothing and seasonal adjustment of nonstationary environmental data, *Jnl. of Forecasting*, **10**, 57-89.