Data Assimilation and Uncertainty Analysis of Environmental Assessment Problems

Renata J. Romanowicz and Peter C. Young
(p.young@lancaster.ac.uk)

Centre for Research on Environmental Systems and Statistics, Lancaster University, United Kingdom

Abstract: Stochastic Transfer Function (STF) and Generalised Likelihood Uncertainty Estimation (GLUE) techniques are outlined and applied to an environmental problem concerned with marine pollution dose assessment. The methods are used to estimate the amount and associated probability distributions of radionuclides transferred to marine biota from a given source: the British Nuclear Fuel Ltd (BNFL) repository plant in Sellafield, U.K. The complexity of the processes involved, together with the large dispersion and scarcity of observations regarding radionuclide concentrations in the marine environment, require efficient data assimilation techniques. In this regard, the basic STF methodology searches for identifiable, linear, Gaussian model structures that capture the maximum amount of information contained in the data with an identified parsimonious parameterisation. The GLUE based-methods, on the other hand, formulate the problem of estimation using a more general Bayesian approach, usually without prior statistical identification of the model structure. As a result, they are applicable to almost any linear or nonlinear stochastic model, although they are much less efficient both computationally and in their use of the information contained in the observations. As expected in this particular environmental application, the STF approach yields much narrower confidence limits for the estimates due to their more efficient use of the information contained in the data. The STF and GLUE techniques are then used to combine information originating from different locations. A final aim of the paper is to use the results obtained in this particular example to explore the differences between the STF and GLUE methods.

***Keywords**:* Stochastic Transfer Function; Monte Carlo Simulation analysis; Generalised Likelihood Uncertainty Estimation; marine dose assessment; predictive uncertainty; data assimilation.

1. INTRODUCTION

The methods used for regulatory purposes in marine pollution are normally based on linear regression estimates (e.g. Hunt [1984]). In recent years compartment-type models have also been developed to describe the transfer of pollutants to marine biota (e.g. Nielsen [1995]). While the first approach is too simplistic, the second almost always leads to over-parameterised (i.e. poorly defined) problems. Moreover, despite the obvious uncertainties in the system, both methods are based on a philosophy of deterministic-reductionism (see e.g. Young [2002]). They assume complete, deterministic knowledge of the processes involved; and they do not take into account measurement uncertainties, which are naturally present in the observations because of their normally scattered and sparse nature. In this paper, we propose a different approach that involves statistical identification, estimation and prediction based on *Stochastic Transfer Function* (STF) modelling techniques. When applicable, these techniques yield an identifiable and parametrically efficient (parsimonious) model structure, as well as providing estimates of the modelling errors and the uncertainties on the model parameters (based on Gaussian probability assumptions). The resulting STF model is in an ideal form for use in model-based prediction and data assimilation exercises. The STF analysis conforms with the *Data-Based Mechanistic* (DBM) modelling philosophy (e.g. Young [1998]) and exploits the CAPTAIN Matlab Toolbox developed at Lancaster University (see http://www.es.lancs.ac.uk/cres/captain/).

Another objective of the paper is to compare the STF approach with an alternative *Generalised Likelihood Uncertainty Estimation* (GLUE) procedure of Beven and Binley [1992]. In particular, both methods are used to estimate how the marine pollution associated with discharges from the British Nuclear Fuel Ltd (BNFL) repository plant in Sellafield, U.K. affects fish in the marine

environment. Finally, the differences between the STF approach and GLUE techniques are explored when they are used to combine information derived from different sources.

2. METHODOLOGY: STF AND GLUE TECHNIQUES

The single input, single output STF model can be written in the following discrete-time (sampled data) equation form:

$$y_t + a_1 y_{t-1} + \ldots + a_n y_{t-n} = b_0 u_{t-\delta} + \ldots b_m u_{t-\delta-m} + \eta_t \qquad (1)$$

where y_t and u_t are, respectively, the dose and the concentration of the input pollutant at the t^{th} sampling instant; δ denotes any pure, 'advective' time delay; η_t represents the noise (not necessarily white); and m, n define the model order.

Here, the RIVID algorithm in the CAPTAIN Matlab toolbox and the associated Data Based Mechanistic (DBM) modelling concepts are used to identify the order of the STF model (the values of n, m and δ) and to statistically estimate the associated parameters (see e.g. Young [1984]). Depending on the identified order and the estimated values of the parameters, the STF description (1) can be decomposed into serial, parallel or feedback connections of first order systems that often have a direct physical interpretation (see e.g. Young [1998]). This demonstrates the analogy between transfer function and compartmental modelling techniques, as applied to transfer of a pollutant in the food chain. However, there are two main differences. First, the STF model (1) is identified and estimated statistically, so ensuring a parsimonious model structure and parametric identifiability. This means that there is no danger that STF models will be over-parameterised, in contrast to typical compartment models. Second, the STF model is stochastic, with all the uncertainty in the model quantified.

STF model estimation (calibration) is also able to exploit recursive estimation (see Young [1984], [1999a]), so the model parameters can be updated as new observational data are obtained. As a result, STF estimation, in its most common form, can be considered as a Bayesian approach to model estimation and data assimilation for linear models under Gaussian assumptions. Moreover, recent research has shown that it can be extended further to handle a widely applicable *State Dependent Parameter* (SDP) class of non-linear TF models (see later).

The Generalised Likelihood Uncertainty Estimation (GLUE) method introduced by Beven and Binley [1992] is overtly Bayesian in character. A statistical formulation of GLUE is given in Romanowicz *et al*. [1994]. It is particularly applicable to large, over-parameterised models, where there is no inverse solution; and, hence, the estimation of a unique set of parameters, which optimise goodness of fit criteria given the observations, is not normally possible. The method is a simple example of numerical Bayesian estimation that exploits *Monte Carlo Simulation* (MCS) based on sampling the parameter values from assumed prior probability distributions. The technique is based on the estimation of the probabilistic weights associated with different parameter sets, using arbitrary chosen goodness of fit criteria and the derivation of a posterior probability distribution function using the Bayes rule. This distribution function is subsequently used to derive the predictive probability of the output variables. The main advantage of GLUE in relation to the STF approach lies in the fact that it can be applied in a simple fashion to practically any linear or nonlinear model. As we shall see, however, it is much less efficient than STF, both numerically and statistically, when applied within linear-quadratic formulation of environmental problems.

3. THE STF APPROACH TO DOSE ASSESSMENT

The concentrations of radionuclide ^{137}Cs in the Irish Sea derive from low level liquid discharges made by the British Nuclear Fuel Ltd (BNFL) repository plant in Sellafield. The radionuclides are present in sediment, plankton and sea-water, whence they are transferred to the fish. The available data are sparse and very scattered. They include liquid discharges of ^{137}Cs from the Sellafield pipe-line (under the authorisation of BNFL) dating back to 1952 (Annual BNFL Reports). Also available are measurements of concentrations of radionuclide in fish flesh at different locations within the Irish Sea (Baxter and Camplin [1993], Camplin [1995], BNFL Reports [1962-1999]). For the present study, the observations from different sources have been combined for the same species and similar locations along the Cumbrian coast. For the purpose of comparing the model results with observations, two sets of non-uniformly sampled data for fish flesh concentrations for two sites (one for calibration and one for validation) were chosen for the radionuclide ^{137}Cs. These sets cover the same period of time (1970-1999). All the data were optimally interpolated to produce series sampled at a uniform monthly sampling interval, using the *Dynamic Harmonic Regression* (DHR) algorithm in the CAPTAIN toolbox (Young *et al*. [1999]).

Using the RIVID algorithm in the CAPTAIN Toolbox, the best identified and estimated (using the calibration data) model has the form:

$$y_t - 0.9212 y_{t-1} = 0.0107 u_{t-5} + \eta_t \qquad (2)$$

The simulated output of this model explains 94% of the variance in data (i.e. $R_T^2 = 0.94$). It has a 12 months time constant and a 5 months advective time delay. This means that the changes in liquid discharges are first detected in fish flesh concentrations after a period of 5 months. The effect then gradually increases thereafter, with an exponential rise time of 12 months, giving a total 'travel time' of 17 months.

This model was applied to validation data giving the results shown in the lower panel of figure 1. It can be seen that the high observation values do not fit within the confidence limits of the estimates obtained from the estimation data set. This indicates that, on the basis of the estimation data set, some features of the transfer process have not been sufficiently well captured by the model at the high concentration levels. This could be because of different behaviour for larger concentrations of radionuclides in the sea-water, requiring either a non-stationary (time variable parameter) or non-linear model. But it could also arise from factors such as input and output observational errors; or the influence of other, unaccounted for processes.

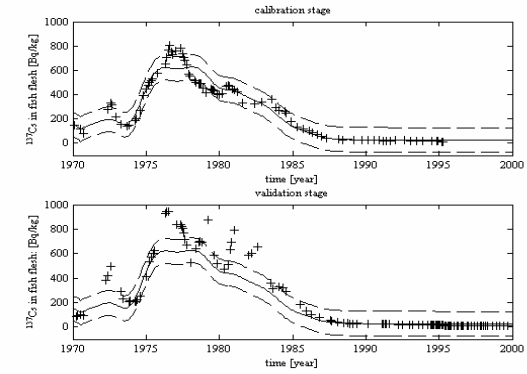

Figure 1. Simulated STF model output (full line) for the transfer of radionuclides from liquid discharges to fish flesh: calibration stage (upper panel); validation stage based on a different data set (lower panel). Crosses denote the observations and the standard error bounds (95% confidence intervals) are shown as dashed lines.

4. USING GLUE AND STF TECHNIQUES TO COMBINE INFORMATION FROM DIFFERENT SITES

In general, the use of additional data presents no problems within the STF approach, since the recursive estimation of the parameters combined with the Kalman Filter enable the effective combination of the information from different, simultaneous sources. However, this approach requires linearity of the relations within the process and introduces Gaussian error for estimates residuals. An alternative solution is to use the GLUE technique to update estimates of ^{137}Cs fish concentrations using data from the other observation sources (here the validation set was used for the updating).

The application of the GLUE updating procedure in combination with the STF model allows us to consider the influence of conditioning on the confidence limits of the predictions. In the first instance, following normal practice, the prior distributions for parameters in the MCS analysis are set to be uniform. Typically in GLUE analysis, the estimates of parameter ranges are based on the preliminary sensitivity analysis. In this example, this procedure may lead to the choice of parameter ranges that do not adequately represent the STF mean estimates and their Gaussian distribution. Here, therefore, the parameter ranges equal to about ten times STF-derived standard deviations were set to yield mean values only slightly different to STF estimates. In order to see how the information about the prior parameter distribution influences the posterior distribution, additional GLUE analysis was also carried out using the Gaussian priors for parameters with both mean *and* standard deviation values obtained from STF analysis.

In both of the above investigations, the results of the 1000 MCS over 30 year period starting from 1970 are used to derive the posterior distributions for the parameters (see Romanowicz et al. [1994]). Here, the exponent to the sum of square errors between simulated and observed concentrations is used as a 'likelihood' weighting for the parameters. The posterior marginal cumulative distributions for both model parameters obtained in this manner are shown in figure 2. This reveals that, as expected, the confidence limits for the parameters derived from the GLUE model with uniform priors are much wider than those obtained from the MCS analysis based on the STF estimates. What is more important, the posterior distributions obtained from GLUE procedure are flat, and different parameter ranges give different (not optimal) parameter estimates.

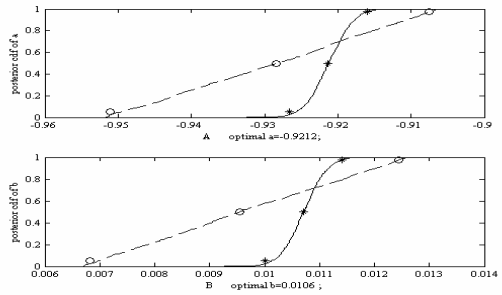

Figure 2 Estimates of the marginal cumulative posterior distribution function (cdf): parameter a_1 (upper panel); parameter b_0 (lower panel). The dashed lines are obtained from flat priors; while the full lines show cdfs for Gaussian priors, with the mean and standard deviation based on STF estimates.

The posterior distributions of parameters are used to derive the predictive distribution of the dose

$Y_t, P(Y_t < y \mid y_{obs,t})$. The predictions y are evaluated as in Romanowicz et al. [1994]:

$$y = y_{sim,t}(\boldsymbol{\theta}) + \varepsilon_t \qquad (3)$$

where $y_{sim,t}(.)$ denotes the model output (a function of the model parameter vector $\boldsymbol{\theta} = [a_1\ b_0]^T$ and the uncertainty of model input) obtained from the parameter simulations; and ε_t denotes the prediction error related to model structure and observation errors (not known in the prediction stage). In the STF approach, this error is assumed to be equal to prediction error and is estimated in the estimation (calibration) stage of the STF procedure. It is assumed here that the distribution of this error follows a Gaussian distribution with zero mean and the variance derived from the calibration stage in the analysis. In other words, it is assumed that $\varepsilon_t \cong y_{sim,t}^{cal} - y_t^{cal}$ (although, in more general applications, the mean and the variance are assumed unknown: see Romanowicz *et al*. [1994]). From (3) it follows that the standard error bounds on the predictions of y_t (i.e. the total prediction error bounds) should be derived from the sum of two stochastic processes: that arising from parametric uncertainty and that caused by observational errors. In this regard, it should be stressed that the typical GLUE approach considers *only* the uncertainty related to the posterior distribution of parameters contained in $y_{sim,t}(.)$, neglecting ε_t.

The posterior distributions of parameters derived from Gaussian priors are used to derive the predictive distribution of the concentrations. In order to analyse the influence of correlation between prior parameter distributions, two cases are considered: first, the case with no correlation between the parameters; and second, that with full information about the covariance structure of parameters, as obtained from the STF analysis. This latter case corresponds *exactly* to the solution derived using the normal STF analysis: i.e. STF estimation combined with associated MCS analysis (see Young [1999b]). As seen in the lower panel of figure 3, the variance related to "parameter" error is relatively small for STF models, as expected. Most of the predictive uncertainty results from the prediction error directly related to the observation variance. On the other hand, when the parameters were varied with the same Gaussian distributions around their mean, but without cross-correlation, the confidence limits related to the posterior distribution of the simulated model output were much wider, as shown by the inner dashed curve in the upper panel of figure 3.

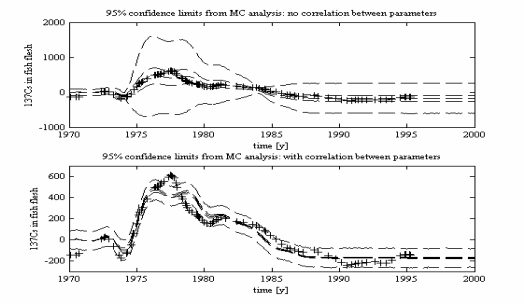

Figure 3. 95% confidence intervals for the concentration estimates based on MCS (GLUE-STF) analysis with no correlation between parameters (upper panel) and with correlation between parameters (lower panel). The inner intervals (also dashed but on the lower panel very small) correspond to the variance associated with parametric errors; the crosses denote the observations.

In this case, the confidence limits relating to the prediction errors (3) (including the estimate of error ε_t and simulated output uncertainty) were also much wider and the resulting "total" 95% predictive confidence limits are shown on the same upper panel as the outer dashed lines. These upper bounds of the confidence limits may be treated as the maximum possible error boundaries estimated from the data. On this basis, it follows that, in contrast to normal usage, the GLUE methodology should include the uncertainty of not only the output distribution based on the uncertainty in the parameter estimates but also that related to the prediction errors ε_t.

The possibility of updating the information from different sources is considered as one of the main advantages of the GLUE technique. Following the discussion on the prediction error (3), updating can be considered as conditioning the posterior parameter distribution, and consequently the posterior output distribution, on the basis of a new observation set. Here, the confidence limits for the output distribution *based on the parametric uncertainty alone* (as typically applied in GLUE), would be narrower. We shall show here, that this is not necessarily the case when the full uncertainty in the model (including the prediction uncertainty) is taken into account.

In order to update the estimates of the ^{137}Cs fish concentrations shown in figure 3, the "validation" observation set from the other site was used as a source of additional information for the GLUE technique. Figure 4 illustrates the influence of updating information from the two sites using the combination of the STF technique and the GLUE procedure. The 95% confidence intervals for the model updated by the additional observation set posterior predictive distribution (lower panel) are wider than the non-updated one (upper panel). This

results from the larger variance of the errors for the second observation set. Note that in both plots the confidence intervals based on posterior cdfs of parameters are very small and are not apparent in the figure.

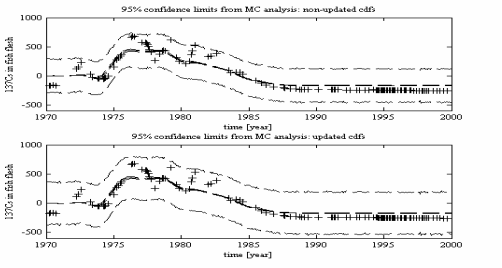

Figure 4. ^{137}Cs concentration estimates in fish flesh: the 95% confidence limits based on GLUE-STF model conditioned on first observation set (upper panel); 95% confidence limits based on GLUE-STF model conditioned on both observation sets (lower panel). Dashed lines are confidence limits; crosses denote observations.

5. MODEL UPDATING, DATA ASSIMILATION AND FORECASTING

As we have pointed out, the STF model parameters can be estimated recursively and so they can be either updated 'on-line', as additional data are received; or the recursive algorithm can be modified to allow for *Time Variable Parameter* (TVP) estimation. This yields an estimated parameter vector

$$\hat{\mathbf{a}}_{t|t} = [-\hat{a}_{1,t} \; -\hat{a}_{2,t} \; , \; \ldots \; , \; -\hat{a}_{n,t} \; \hat{b}_{0,t} \; \hat{b}_{1,t} \; , \; \ldots \; , \; -\hat{b}_{m,t}]$$

where the subscript $t|t$ means that the model parameters at any t^{th} sampling instant are estimated on the basis of all of the data up to time t (normally, this estimate is just denoted by $\hat{\mathbf{a}}_t$ for convenience). Moreover, in the TVP case, backward-recursive *Fixed Interval Smoothing* (FIS) can be introduced, where the parameter estimates are further updated so that, at each sample time t, the estimate $\hat{\mathbf{a}}_{t|N}$ is based on all N samples in the data set (see Young [1984] [1999a]). This is useful since it yields 'smoothed' estimates of the TVPs that have smaller estimation error variance than the 'forward pass' estimates $\hat{\mathbf{a}}_t$, and the algorithm can be used to interpolate very well over gaps in the data.

As mentioned previously, another advantage of the STF model is that it can be converted easily into a stochastic state space form and embedded within a KF framework. This has the advantage that it allows for state estimation and updating or, when combined with the recursive parameter updating, on-line 'data assimilation' and adaptive forecasting. Here, the model is continuously updated on-line, so that its parameters always reflect the latest data; and it can be used to forecast y_t into the future based on these latest updates (see e.g. Young [2002])[1].

Of course, these extensions of STF modelling only apply if the model is linear or linear with TVPs (non-stationary). Whilst this class of models, particularly in the TVP case, can be used to describe a wide class of practical environmental systems, it does not cover truly nonlinear processes. However, it is possible to extend the methodology to nonlinear stochastic systems by 're-linearisation', where the nonlinear model is linearised at each recursive update of the KF in order to allow for updating the covariance matrix associated with the parameter estimates. The best known algorithm of this type is the *Extended Kalman Filter* (e.g. Jazwinski [1970]). However, recent developments in *State Dependent Parameter* (SDP) estimation (Young [2000]; Young *et al*. [2001]) offer an alternative approach to handling nonlinear systems. Here, the TVPs are estimated in a special recursive manner that allows them to be interpreted in terms of the variations in other 'state' variables (e.g. u_t, y_t or other physically relevant variables). This very effectively extends the range of the recursive estimation and data assimilation to a widely applicable class of truly nonlinear systems that even includes chaotic processes. However, the forecast probability distributions are still in a Gaussian form (defined by their first two moments) and so further developments, using MCS methods, are currently being considered to eliminate this restriction.

6. CONCLUSIONS

The research presented in this paper describes the application of STF techniques and GLUE analysis to environmental assessment problems based on an example of ^{137}Cs radionuclide transfer from liquid discharges to fish flesh in Irish Sea near Sellafield, UK. The example has shown that a reasonable explanation of the data (model fit) and dose prediction is achieved by introducing the STF model. The differences between the GLUE and STF-based approaches have been discussed from the point of view of their applicability to modelling environmental processes. In the original form of Beven and Binley [1992], GLUE differs from Bayesian analysis "sensu stricto"; e.g. in the subjective choice of the goodness of fit criterion. It is a method applicable to any computationally tractable linear or nonlinear environmental problem. Unlike the STF alternative, however, it assumes that there is no unique solution to the inverse problem, i.e. the model is over-parameterised and not identifiable.

[1] Indeed, when used with FIS recursion, it can also be used to backcast into the past.

In contrast to GLUE, the STF methods used here identify and estimate, from the observational data, an identifiable and parsimonious model structure from amongst all linear structures that have an inverse solution. Moreover, the STF method yields quantitative information on the posterior probability distribution of the parameters in the model, as well as the variance of the prediction errors. Also, since they are based on recursive estimation, STF model parameter estimates and their associated covariance matrix can be updated on-line, as the data are received. Then, together with the Kalman Filter, the continuously updated model can be used to develop adaptive data assimilation and forecasting algorithms. Of course, as a result of the stochastic assumptions required in its derivation, the standard, constant parameter STF model form used in this paper, is restricted to the class of linear, identifiable models with Gaussian disturbances. We have pointed out, however, that more sophisticated time variable (TVP) and state-dependent parameter (SDP) forms of STF analysis are now available that remove the linearity and stationarity restrictions. Moreover, research is proceeding on the removal of the normality assumption.

Finally, we have shown that one disadvantage of the original GLUE method (Beven and Binley [1992]) is that it estimates the predictive uncertainty bounds based on the uncertainty of the output originating from the posterior distribution of parameters alone. In so-doing, it neglects the uncertainty related to the prediction errors and so applies only when the variance of prediction errors is small in comparison with the variance of parameter-related errors. In the case when information about the probability distribution of the parameters *and* prediction errors is available (as in the case of the STF model), this assumption may well not be fulfilled. The results of the present research confirm that, as expected in the case of linear, Gaussian models, the STF methods are much simpler and yield narrower confidence limits for the estimates than the GLUE method.

Acknowledgements
This work was supported by the IMPACT project IST-1999-11313.

7. REFERENCES

Baxter A.J. and W.C. Camplin, *Radiocaesium in the seas of northern Europe: 1985-89*, Fisheries Research Data Report, 32, 1993.

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

BNFL, *Radioactive Discharges and Monitoring of the Environment, 1996*, British Nuclear Fuels plc: Risley, 30, see also previous and subsequent reports, 1997.

Camplin W. C., Radioactivity in surface and coastal waters of the British Isles, 1994, Radiobiological Technical Report FRL46, (see also previous and subsequent reports), 1995.

Hunt G.J., Simple models for prediction of external radiation exposure from aquatic pathways, *Radiation Protection Dosimetry*, 8, 4 215-224, 1984.

Jazwinski, A.H., *Stochastic Processes and Filtering Theory*, Academic Press: New York, 1970.

Nielsen S.P., A box model for North-East atlantic coastal waters compared with radioactive tracers, *J. Marine Systems*, 6, 545-560, 1995.

Romanowicz, R., Beven, K., and Tawn, J., Evaluation of Predictive Uncertainty in Nonlinear Hydrological Models Using a Bayesian Approach, in V. Barnett and K. F. Turkman (eds.), *Statistics for the Environment 2,* 297-315, 1994.

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

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

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

Young, P.C., Data-based mechanistic modelling, generalised sensitivity and dominant mode analysis. *Computer Phys. Communications*, 117, 113-129, 1999b.

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

Young, P.C., Advances in real-time flood forecasting, *Philosophical Transactions of the Royal Society: Mathematical, Physical and Engineering Sciences*. In press, 2002.

Young, P.C., McKenna, P. and Bruun, J., Identification of nonlinear stochastic systems by state dependent parameter estimation. *Int. Jnl. Control*, 74, 1837-1957, 2001.

Young, P.C., Pedregal, D.J. and Tych, W., Dynamic harmonic regression, *Journal of Forecasting*, 18, 369-394, 1999.