Errors in Hydrological Variables and Their Effect on Model Parameters

Shahadat Chowdhury[1], **Ashish Sharma** [2]

[1,2]**School of Civil and Environmental Engineering, University of NSW, Sydney NSW 2052**
[2]a.sharma@unsw.edu.au

Keywords: *SIMEX, Sacramento, measurement error, sea surface temperature, southern oscillation index.*

EXTENDED ABSTRACT

True measurements of hydrological variables across full time and space domain are rarely available. Rather than the true values the best we often have are measured values with associated error variances. Model parameters are then estimated based on these measured values. This paper demonstrates that ignoring the structure of the errors introduce a systematic bias to estimated parameter values.

We present a method to ascertain optimal parameter values in the presence of a known measurement error distribution. This method is known as simulation extrapolation (SIMEX). Say X is an input variable with no error and Y is the response variable. Consider instead of X our input variable is W where, W= X+ $N(0, \sigma^2)$. The calibrated parameter with W as input is called the naïve parameter estimate. At the start of this method we generate a series of alternate realisations for W (denoted W*) by artificially adding white noise to W in increasing multiples of the error variance σ^2. Given the new estimate of the parameter (using realisations W*), it now becomes possible to speculate on what the parameter would be were it to have no additive noise (or an error variance of 0). The trend in altered parameter values provides a basis to extrapolate the parameters to a situation where no error exists.

This paper explores the strength of the SIMEX method using two hydrological case studies. A synthetic example is presented in the first case study where three parameters of the Sacramento rainfall runoff model are ascertained to investigate the effect of errors in the rainfall input data. SIMEX is able to mitigate the bias from naïve parameters of Sacramento caused due to error in original rainfall.

Hydrological models frequently use input variables like rainfall, evaporation, solar radiation and temperature. Measurement errors may arise due to instrumentation, interpolation or extrapolation of data in space and time or transformation of point measurement into areal values. In some cases error variance can be estimated from the statistical inference of the interpolation schemes. An example of a variable with known error variance is globally distributed sea surface temperature anomaly (SSTA) data. SSTA is widely used in climate prediction models. The earlier part of the SSTA data set contains higher magnitude of error due to reduced sampling frequency coupled with poorer technology. Our hypothesis is that the higher magnitude of noise in the earlier period introduces bias to model parameters when SSTA is used as an input variable.

The relatively error invariant Southern Oscillation Index (SOI) is regressed over SSTA and calibrated using a subset of the series from 1900 to 1960. We use SSTA of winter, spring and summer as predictors of the SOI. The model forecasts SOI of next two seasons and also of next three months. We choose thirteen linear prediction models that show reasonable correlation between SSTA and SOI during calibration period. We validate the model during 1961 to 2003. Overall, the application of SIMEX has reduced the residual errors of nine out of thirteen predictions during validation periods with further two predictions remaining unchanged.

1. INTRODUCTION

Hydrological modelling involves estimating model parameters that describes the relationship between one or more response variables and associated covariates. In many instances the true value of the covariates are not known and the best that is available is an estimate along with a characterisation of associated errors. In hydrology these errors may arise due to instrumentation, interpolation or extrapolation of data in space and time or conversion of point measurement into areal values.

The sources of total errors can be traced into model structure, calibration data and input variables. The errors in model structure and calibration data are mainly investigated during calibration (Khadam and Kaluarachchi 2004). Competing hydrological models have been developed with an aim to minimise model uncertainty for various applications. A comprehensive list of models used in Australia is prepared by Boughton (2005). The conventional calibration practises, based on least square fit, assume white noise in calibration data. The least square method does not consider errors in input variables (Kavetski et al. 2002).

Early research into errors dealt with linear models (Fuller 1987), and more recently non linear regression (Carroll et al 1995). Most of this research was conducted in the field of biometrics.

The combined effect of errors in model structure, calibration data and input variables in hydrological models has been addressed by Kavetski et al. (2002). They introduce Bayesian Total Error Analysis method to simultaneously address the model uncertainty and input errors. Our study is limited to the effect of input error only. We offer a simulation based method with low computational burden since it does not attempt to address the model uncertainty. The error model is called SIMulation EXtrapolation or SIMEX.

2. SIMEX

2.1 Foreword

SIMEX is a simple simulation algorithm that graphically shows the effect of measurement error on parameter estimates. The technique was introduced by Cook and Stefanski (1994) and Stefanski and Cook (1995). Biometricians have been applying SIMEX for a variety of problems since the mid nineties. In contrast, we found no references of its use in hydrological publications. This method has been extended to non parametric models in recent years (Carroll et al. 1999; Lin and Carroll 2000).

2.2 Method

For simplicity, we introduce the method in this paper using a linear regression model with additive error as shown in the following Equation (1).

$$Y = \beta_x X + \varepsilon_x \qquad (1)$$

Where, X = independent variable or covariate,
Y= response variable,
β_x= parameter inclusive of the intercept,
ε_x = error.

The error term ε_x has a zero mean and reflects the uncertainty associated with Y and the model structure. We like to emphasise that ε is not a consistence estimator of error in X. Consider a case where we do not observe true X but observe W instead, as expressed in Equation (2):

$$W = X + U \qquad (2)$$

Where U is white Gaussian error, independent of (X, Y), and has zero mean and variance of σ_u^2, which in statistical notation U~ $N(0, \sigma_u^2)$. Note that we do not define the distribution of X, as no prior knowledge of that distribution is required for the validity of the SIMEX method. In practice, instead of Equation (1) we regress Y over W.

$$Y = \beta_w W + \varepsilon_w \qquad (3)$$

The error in W introduces a bias in parameter estimate β_w, hence $\beta_w \neq \beta_x$. In the statistical literature β_w is defined as the naïve estimate and β_x as the true estimate of the regression. The SIMEX method attempts to remove the bias from β_w by determining the trend via simulation.

The SIMEX method starts with the notion that the estimate of σ_u^2 is known. Subsequently, the method involves introducing a variable λ which successively generates higher magnitude of variances σ_i^2 according to Equation (4).

$$\sigma_i^2 = \lambda_i \sigma_u^2 , \qquad i = 1,2,\ldots n \qquad (4)$$

Where, $\lambda_1 < \lambda_2 < \lambda_3 \ldots . \lambda_n$.

Based on this, one can generate a synthetic series of $U_i^* \sim N(0, \sigma_i^2)$, Gaussian random deviates with

zero mean and σ_i^2 variance. Each set of the generated U_i* has a higher variance than the preceding set ie, $\sigma_1^2 < \sigma_2^2 < \sigma_3^2 < ..... \sigma_n^2$. The errors are then added to the original record W and hence artificially generating covariate W* with increased additive errors:

$$W_i^* = W + U_i^* \quad , \text{where} \quad i = 1, 2, \ldots n \qquad (5)$$

There will be n numbers of W* which can regress Y and solve β_i*

$$Y = \beta_i^* W_i^* + \varepsilon^* \qquad (6)$$

It is important to note that each U_i* series is replicated a few hundred times and the β_i* represents the average estimates out of those few hundred replications of Equation (6).

The combination of Equations (4) to (6) reveals that the estimates $\{\beta_1^*, \beta_2^*, \beta_3^* \ldots. \beta_n^*\}$ are directly related to $\{\lambda_1, \lambda_2, \lambda_3, \ldots. \lambda_n\}$. So we can write Equation (7).

$$\beta^* = \mathcal{F}(\lambda) \qquad (7)$$

Like any modelling case, the curve $\mathcal{F}(.)$ needs to be properly specified. The SIMEX estimate β_{simex}, a surrogate of true estimate β_x, can be found by extrapolating β* back to the notional no error zone:

$$\beta_{simex} = \mathcal{F}(\lambda = -1) \qquad (8)$$

$$\beta_x \approx \beta_{simex} \qquad (9)$$

2.3 Algorithm

We start with known Y, W and σ_u^2. The algorithm depends on the model structure and thus the parameter to be estimated. This section illustrates the linear model that is presented in the last section.

1. Initialise $i = 1$, and $\lambda_i = 0.2$.
2. Generate a random normal series $U^*_i \sim N(0, \lambda_i \sigma_u^2)$.
3. Compute the synthetic covariate, $W_i^* = W + U^*_i$
4. Fit the following linear model and estimate β_i^*, $\hat{Y} = \beta_i^* W_i^*$.
5. Repeat the steps from 2 to 5, say for 500 times, and accept the mean estimate as the expected value of β_i^*.
6. Repeat steps 1 to 5 for the following values. $i = \{2,3,..10\}$, $\lambda_i = \{0.4, 0.6 \ldots 2.0\}$
7. Draw (λ_i, β_i^*) and fit a line $\mathcal{F}(.)$ to $\beta_i^* \sim \lambda_i$.
8. Extrapolate to SIMEX estimate of the parameter β using Equation (8).

3. SYNTHETIC EXAMPLE

The application of SIMEX in a linear setting is demonstrated using a synthetic example by Chowdhury and Sharma (2005). We explore the validity of SIMEX in a non linear hydrological model here.

Sacramento is one of the two USA originated water balance models that are widely used in Australia (Boughton 2005). We select this model to demonstrate the effect of SIMEX on a parameters of non linear model. The model is named after the Sacramento River in California, USA, where it was first applied (Burnash 1975).

A conceptual rainfall-runoff model, where flow at time t is Q_t , can be expressed as:

$$Q_t = S(I_t , E_t; \theta_p) + \varepsilon_t \qquad (14)$$

Where S(.) is the corresponding model, I_t and E_t are model inputs at time t, θ_p is the set of unknown parameter values and ε_t is an error term. In our case S(.) is the Sacramento model which uses rainfall and evaporation data (I_t , E_t) to generate flow(Q_t).

The Sacramento model has five soil moisture storages. The model essentially operates based on water movements between storages, loss and routing as shown in Figure 1. It has 16 parameters, $\{\theta_p \; ; \; p = 1,2 \ldots 16\}^T$. The transpose superscript T, in statistical notation indicates that the set of θ remains the same for all the time steps denoted by the time subscript t.

We use daily rainfall and evaporation at Golspie, NSW in the Upper Lachlan Catchment from 1980 to 1992 as notional true estimate of catchment rainfall and evaporation, $\{I_t , E_t; t = 1,2.. 12x365\}$. The years and location are in fact irrelevant here except to keep the example in a realistic numerical domain. We generate flow based on a set of given values for all 16 parameters. This becomes our synthetic true flow.

The scope for error in a rain reading during a drier time is low (a dry day reading is error free). On the contrary a storm may completely miss the rain gauge. So rainfall errors are assumed to be multiplicative in this study (Kavetski et al. 2002). Now we artificially corrupt the rainfall series I_t by multiplying it by a Gaussian series $U \sim N(1, \sigma^2)$. The corrupted series, W_t becomes the notional

recorded rainfall. For simplicity we assume the evaporation estimate to be error free.

$$W_t = I_t * U \quad (10)$$

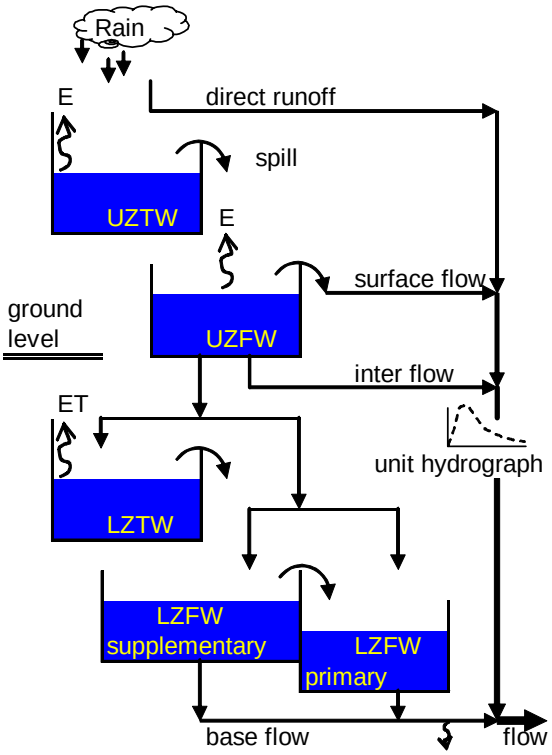

Figure 1. Sacramento Model

Three parameters are allowed to vary keeping the remaining 13 parameters constant. They are UZTW (upper zone tension water storage capacity), UZFW (upper zone free water storage capacity) and LZTW (lower zone tension water storage capacity).

All 16 parameters of the Sacramento model can simultaneously be calibrated using the shuffle complex evolution search optimisation scheme (Kuczera 1997). However due to the simpler synthetic setting with fewer parameters, we use a modification of BFGS quasi-Newton method (Byrd et.al. 1995) readily available as a function in software R (Team 2004). The calibrated parameters using W_t as input is known as naïve estimate (θ^{naive}).

$$Q^{est}_t = S\,(W_t, E_t; \theta^{naive}_p | \theta_q) \quad (11)$$

$$p \equiv \{1,2,3\} \text{ and } q \equiv \{4,5..\ 16\}$$

Now we generate replicates of W, W*, with increasing amount of error $\{\lambda_t\sigma^2\}$ and thus estimate the increasingly biased parameter θ*. The simulation enables us to setup the following regression relationship.

$$\theta^*_p = \mathcal{F}_p(\lambda) \quad (12)$$

The strength of any non linear SIMEX depends on our ability to assign a structure to $\mathcal{F}(.)$. We rely on the synthetic study, data structure and our experience to decide on a suitable model for $\mathcal{F}(.)$. The current synthetic study on Sacramento has promisingly found that $\mathcal{F}$ follows a clear structure in at least three parameter spaces. Hence the extrapolation to $\mathcal{F}_p(\lambda = -1)$ is possible as shown in Figure 2 and listed in Table 1.

Table 1. Sacramento Parameters before and after SIMEX (all values are in mm)

Parameters	True Value	Naïve Estimate	After SIMEX
UZTW	60	63	59
UZFW	150	146	149
LZTW	38	41	38

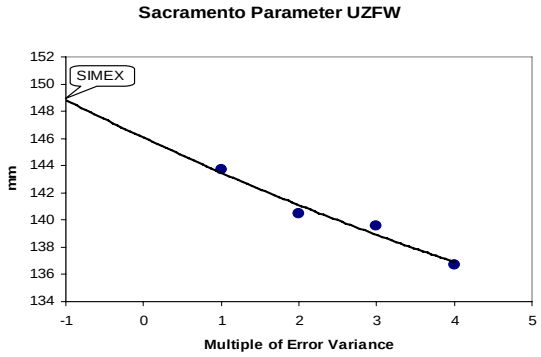

Figure 2. The SIMEX estimate of UZFW is 149 mm which is close to the true value of 150 mm.

4. PRACTICAL APPLICATION

4.1 General

We have demonstrated that traditional modelling practice of least square fit to the response data does not negate the errors of input variables. The motivation to estimate error in hydrological input variable has been limited due to lack of any procedure to incorporate them in the model structure. Error variance can be estimated from the statistical inference of the interpolation schemes used to fill the data in both space and time dimensions. Rainfall runoff models rely on this transformation of point rainfall into areal values prior to commencing any modelling exercise. The transformation introduces interpolation or extrapolation error.

4.2 Climate Prediction Models

Sea surface temperature data has been extensively used to formulate various prediction models (Fu

et. al 1987; Sharma 2000; Drosdowsky and Lynda 2001). Reconstructed, gridded, monthly sea surface temperature anomaly (SSTA) data, available from the Climate Data Library of the Lamont-Doherty Earth Observatory of Columbia University, New York, is used in the example discussed here. The SSTA data set was reconstructed based on point measurements of sea surface temperature using the Kaplan Optimal Smoother (OS) interpolation algorithm (Kaplan et al 1997; Kaplan et al 1998). The reconstructed data set and associated error characteristics are available from 1856 onwards, at a resolution of 5° latitude by 5° longitude. The interpolation procedure allows for estimation of the error variance at each time step/grid location. The interpolation in space and time for the missing data and the weights reflecting the reliability of the records contributed towards the final estimates of the error variances.

Figure 3. The error variance of monthly NINO3 anomaly during the month of October every year.

The arithmetic average of SSTA over the equatorial pacific region (5° N to 5° S and 150°W to 90°W) is known as NINO3. NINO3 has been extensively used in climate prediction models as a reduced form of the entire set of SSTA (Ruiz et al. 2005). The error variance of NINO3 data is shown in Figure 3, the overall variance of NINO3 is 0.60. The variance is higher during early periods of data and the two world wars. Error variance of similar magnitude is estimated in other locations of SSTA data.

The El Nino Southern Oscillation (ENSO) warm phase, associated with warming up of sea surface water in the equatorial eastern Pacific Ocean, is a predictor of dry weather in Australia, Southern Asia and South Africa. The strength of ENSO can be estimated from a sea level pressure anomaly based index known as the Southern Oscillation Index (SOI). This index is the standardised pressure difference between Darwin and Tahiti with records extending back to 1876. The SOI is estimated from two controlled weather stations without any need of spatial extrapolation. Hence the index is less exposed to measurement error compared to the globally distributed grided SSTA.

The SIMEX method deals with input error only, it does not account for structural uncertainty of the model. Accordingly the method yields superior result where the model structure is strong and well established. We do not attempt to identify the best candidate model here for SIMEX application. We choose a simple linear model that uses specified SSTA grid cells as predictors. Our aim is to demonstrate the potential, not quantum, of SIMEX improving the predictions.

Consider a setting where we need to fill in the SOI data based on a regression relationship using relevant predictors or we attempt to model persistence of SOI. As SOI and warming up or cooling down of central pacific SST are reflective of the common ENSO conditions, it would be intuitive to use the SSTA at selected location as the predictors of the SOI response variable. We use the relative low error time period of 1961 to 2003 to validate the regression derived using 1900 to 1960 error prone data.

We approach the study by first exploring the correlation between the SSTA at NINO3 region to SOI. We expect that any significantly higher errors in SSTA pre 1960 would result in a drop in correlation relative to post 1960.

The lagged correlation of seasonal SSTA to the SOI of next two seasons and the following three months are investigated. As an example, the spring (September to November) SSTA is correlated to SOI in the following summer, autumn, December, January and February. These lead times are common in developing forecasts (Sharma 2000; Drosdowsky and Lynda 2001; Ruiz et al. 2005).

A consistent drop in correlation in pre 1960 data is experienced in all seasons except when autumn SSTA is used. We discard the results affected by the autumn predictability barrier. The remaining three seasons produce 15 correlations of which 2 have very small values flagging the unsuitability

of linear dependence structure. We further discard these two models. We limit our investigation in these 13 prediction models. Ten out of the 13 models show reduction in correlation in earlier set of data, see Figure 4, validating our preliminary assumption of higher error in earlier part of the SSTA data.

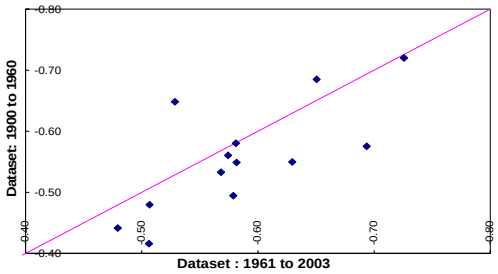

Figure 4. Correlation of SST to SOI data pre and post 1960 period. The points below 1:1 line are showing drop in pre 1960 correlation.

We proceed to investigate the 13 selected prediction models where three seasonal SSTA are the predictors and the predictands are the SOI of next two seasons and the next three months. The parameters are calibrated for the period of 1900 to 1960. We validate the model using post 1960 data and measure the sum of mean squares of errors of the prediction or also known as residual variance of prediction. Later we have altered the parameter estimate by SIMEX using the known error variance of SSTA during the calibration period. We re-compute the residual variance of prediction during validation period. Nine out of thirteen validation fits show reductions in residual variance with further two predictions remaining unchanged, see Figure 5 and Table 2. The improvements are minor due to the inherent structural uncertainty of the chosen models, nevertheless the trend is validating the potential of SIMEX.

The regression of SOI over SSTA is only presented here to investigate the sensitivity of the error variance in some controlled manner. In practise this regression has limited use. The application would be useful in the field of seasonal rainfall prediction that use SSTA.. The research on the effect of SSTA errors on rainfall prediction is ongoing.

Table 2. Sum of mean square of errors of SOI prediction where SSTA is the predictor.

SSTA	WINTER	
SOI	Before SIMEX	After SIMEX
Spring	9.6	9.4
Summer	9.5	9.5
September	12.0	11.7
October	13.3	13.2
November	15.6	15.3
SSTA	SPRING	
SOI	Before SIMEX	After SIMEX
Summer	7.1	7.1
Autumn	12.2	12.1
December	10.9	11.3
January	11.6	11.3
February	13.0	13.1
SSTA	SUMMER	
SOI	Before SIMEX	After SIMEX
Autumn	9.6	9.4
March	12.0	11.7
April	13.3	13.2

.

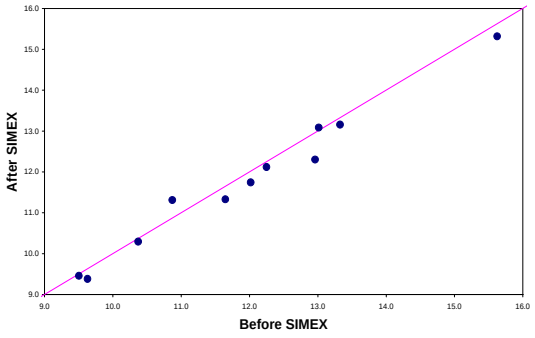

Figure 5. Sum of mean squares of errors during validation period before and after applying SIMEX. The points below 1:1 line are showing improvements due to SIMEX.

5 CONCLUSIONS

Errors in input data introduce bias in parameter estimates. SIMEX, a simulation based method, allows mitigating that bias from the parameters. We demonstrated by using synthetic rainfall with known error in the Sacramento model that

SIMEX can mitigate the bias for the three parameters tested. We explored errors in SSTA data widely used in climate prediction models. We find that the error in earlier sea surface temperature data results in a systematic drop in correlation with SOI. We demonstrate improvements in predicting SOI from SSTA after applying SIMEX. We conclude that hydrologist should check the effect of the errors on parameter estimates before using parameter values in validation.

REFERENCE

Boughton, W. (2005). "Catchment water balance modelling in Australia 1960–2004." *Agricultural Water Management*, 77, 91-116.

Burnash, R. J. (1975). "The NWS river forecast system–catchment modelling." Computer Models of Watershed Hydrology, V. P. Singh, ed., Water Resources Publications, Littleton, Colo.

Byrd, R. H., Lu, P., Nocedal, J. and Zhu, C. (1995). " A limited memory algorithm for bound constrained optimization." *Scientific Computing*, 16, 1190-1208.

Carroll, R. J., Maca, J. D., and Ruppert, D. (1999). "Nonparametric regression in the presence of measurement error." *Biometrika*, 86(3), 541-554.

Carroll, R. J., Ruppert, D., Stefanski. (1995). *Measurement Error in Nonlinear Models*, Chapman and Hall.

Chowdhury, S., and Sharma, A. (2005), "Measurement Errors in Hydrological Data and Their Effect on Water Resources Modelling." *MTERM International Conference*, AIT, Thailand.

Cook, J. R., and Stefanski, L. A. (1994). "Simulation-Extrapolation Estimation in Parametric Measurement Error Models." *Journal of the American Statistical Association*, 89(428), 1314-1328.

Drosdowsky, W. a. C., Lynda E. (2001). "Near-Global Sea Surface Temperature Anomalies as Predictors of Australian Seasonal Rainfall." *Journal of Climate*, 14, 1677-1687.

Fu, C., H. F. Diaz, and J. O. Fletcher. (1987). "Characteristics of the response of sea surface temperature in the Central Pacific associated with El Niño/Southern Oscillation." The Climate of China and Global Climate, C. F. D. Ye, J. Chao, and M. Yoshino, ed., China Ocean Press and Springer-Verlag, 177-201.

Fuller, W. A. (1987). *Measurement Error Models*, John Wiley & Sons, New York.

Kaplan, A. C., MA; Kushnir, Y; Clement, AC; Blumenthal, MB; Rajagopalan, B. (1998). "Analyses of global sea surface temperature 1856-1991." *Journal of Geophysical Research (C. Oceans)*. 103(C9), 8,567-18,589.

Kaplan, A. K., Y; Cane, MA; Blumenthal, MB. (1997). "Reduced space optimal analysis for historical data sets: 136 years of Atlantic sea surface temperatures." *Journal of Geophysical Research (C. Oceans)*. 102(C13), 27,835-27,860.

Kavetski, D., Franks, S., and Kuczera, G. (2002). "Confronting Input Uncertainty in Environmental Modelling." Calibration of Watershed Models, H. V. Gupta, Sorooshian, S., Rousseau, A. N. and R. Turcotte, ed., Duan, 49-68.

Khadam, I. M., and Kaluarachchi, J. J. (2004). "Use of soft information to describe the relative uncertainty of calibration data in hydrologic models." *Water Resources Research*, 40, W11505.

Kuczera, G. (1997). "Efficient subspace probabilistic parameter optimization for catchment models." *Water Resources Research*, 33(1), 177-185.

Lin, X. H., and Carroll, R. J. (2000). "Nonparametric function estimation for clustered data when the predictor is measured without/with error." *Journal of the American Statistical Association*, 95(450), 520-534.

Ruiz, J. E., Cordery, I., and Sharma, A. (2005). "Forecasting streamflows in Australia using the tropical Indo-Pacific thermocline as predictor." *Journal of Hydrology*, in press.

Sharma, A. (2000). "Seasonal to interannual rainfall probabilistic forecasts for improved water supply management: Part 3 - A nonparametric probabilistic forecast model." *Journal of Hydrology*, 239(1-4), 249-258.

Stefanski, L. A., and Cook, J. R. (1995). "Simulation-extrapolation: The measurement error jackknife." *Journal of the American Statistical Association*, 90(432), 1247-1256.

Team, R. D. C. (2004). "R: A language and environment for statistical computing. R Foundation for Statistical Computing." R Foundation for Statistical Computing, Vienna, Austria.