# Modelling Interval of Transfer Function-Noise Models in Groundwater Hydrology

W.L. Berendrecht[a], A.W. Heemink[a], F.C. van Geer[b] and J.C. Gehrels[c]

[a] *Delft University of Technology, Faculty of Information Technology and Systems, Department of Applied Mathematical Analysis, Delft, the Netherlands (w.berendrecht@nitg.tno.nl)*

[b] *Netherlands Institute of Applied Geoscience TNO – National Geological Survey, Delft, the Netherlands*

[c] *Delft University of Technology, Faculty of Civil Engineering and Geosciences, Department of Watermanagement, Delft, the Netherlands*

**Abstract:** Transfer function-noise (TFN) modelling is often applied to analyse groundwater time series. A TFN model relates a specific output variable (in this case groundwater head) to one or more input variables (e.g. precipitation), using some linear relation (transfer function). During the identification and construction stage of model development, an appropriate modelling interval has to be chosen. Up to now, the standard procedure has been to set the modelling interval equal to or larger than the measuring interval. However, it is probably more accurate to choose a smaller modelling interval, depending on the time scale of the hydrological processes. This paper investigates the influence of the modelling interval as well as the measuring interval (i.e. the interval of the output series) on the performance of TFN models, using a state space representation of a TFN model. For this purpose, groundwater time series are generated and modelled several times, varying the measuring interval as well as the modelling interval. The results of this study show the relationships between the time scale of several hydrological processes on the one hand and the level of detail and accuracy of the time series model on the other hand.

***Keywords:*** Groundwater; Time series analysis; Modelling interval; Sparse observations; Kalman filter

## 1. INTRODUCTION

Transfer function-noise (TFN) models [Box and Jenkins, 1970] of hydrological systems have been used for many years. Some important applications are modelling of river flow [Young et al., 1997], trend assessment of groundwater time series [Gehrels et al., 1994], and intervention modelling for use in environmental impact assessment [Hipel and McLeod, 1994]. A TFN model relates a specific output variable to one or more input variables, using some linear relation (transfer function). The residual series is described by a univariate time series model, e.g. an autoregressive moving-average (ARMA) model. A major advantage of TFN models over physical models is that, apart from the hydrological time series, no other data is needed. Besides, physical models are often based on subjective assumptions such as model schematisations and are thus less objective than stochastic TFN models.

An important aspect in the construction of TFN models is the choice of the modelling interval (i.e. the interval of the input series). Up to now, the standard procedure has been to set the modelling interval equal to or larger than the measuring interval of the output variable [van Geer and Zuur, 1997]. Such a model is satisfactory as long as the time scale of the underlying process is at least of the same order of magnitude as the modelling interval. However, this is often not the case. For example, in the Netherlands groundwater head is measured twice monthly, while in several parts of the country the response time of the groundwater system to precipitation excess is less than 14 days. Consequently, for many purposes such models will not be able to describe the hydrological system satisfactorily. Therefore, this study seeks to analyse the influence of a decrease in the modelling interval on the performance of TFN models. In addition, the relation between the number of data and the model performance is analysed. In this paper, TFN models are used for describing groundwater systems, but the results are valid for all systems that have identical response characteristics.

In order to model time series independently of the measuring interval (i.e. the interval of the output series), Bierkens et al. [1999] suggest writing the model in state space form, which allows the use of the well-known Kalman filter. The Kalman filter can then be combined with a maximum likelihood criterion to estimate the parameters.

An objective analysis of the relationship between the modelling interval and model performance requires the use of a time series that is completely known, which can be achieved by using a generated series. Another advantage of using generated time series is that the effect of variable time steps can be completely isolated from other real-world influences, such as distance to the nearest meteorological station or formulation of precipitation excess. Therefore, in this study a number of representative time series were generated by transferring input series into output series, assuming the transfer function follows the curve of a lognormal distribution. A stochastic component was added to the series, which makes the time series more similar to real time series. Finally, the time series were resampled at several intervals in order to determine the influence of the measuring interval and the amount of measurement data on the performance of the model.

The results of the calculations described in this paper show that the choice of the modelling interval of a TFN model should be based on the time scale of the system response and not on the interval of the output time series. If the modelling interval is larger than the time scale of the system response, decreasing the modelling interval improves the performance of the model.

## 2. METHODOLOGY

### 2.1 Introduction

The standard transfer function-noise model introduced by Box and Jenkins [1970] is written as

$$y_t = \frac{\Omega(B)}{\Delta(B)} u_t + \frac{\Theta(B)}{\Phi(B)} a_t \qquad (1)$$

where $y_t$ is the output variable at time $t$; $u_t$ is the input variable at time $t$; $a_t$ is a zero mean white noise process with variance $\sigma_a^2$; $B$ is a backward shift operator defined by $B^k x_t = x_{t-k}$; $\Omega(B) = \omega_0 + \omega_1 B + \delta_2 B^2 + \ldots + \omega_s B^s$ is the moving average (MA) operator of the transfer model; $\omega_i$ are the MA parameters up to order $s$; $\Delta(B) = 1 - \delta_1 B - \delta_2 B^2 - \ldots - \delta_r B^r$ is the auto-regressive (AR) operator of the transfer model; $\delta_k$ are the AR parameters up to order $r$; $\Theta(B)$ and $\Phi(B)$ are the MA and AR operator of the noise model respectively, both defined similarly to $\Delta(B)$.

The major drawback of (1) is that the modelling interval has to be equal to or a multiple of the interval of the output series. Therefore, the next subsection describes how (1) can be rewritten into state space form. Subsection 2.3 shows that this representation of a TFN model enables the use of the Kalman filter, which opens the way to reducing the interval of a TFN model. Finally, subsection 2.4 describes how the groundwater time series used in this paper were generated.

### 2.2 State Space Representation of TFN Model

The state space form is a powerful tool that allows handling a wide range of time series models. The general state space form is applied to a univariate time series, $y_t$. This observable variable is related to the $m\times1$ vector $\mathbf{x}_t$, known as the state vector, via the measurement equation:

$$y_t = \mathbf{C}_t \mathbf{x}_t + v_t, \quad t = 1, \ldots, T \qquad (2)$$

where $\mathbf{C}_t$ is a $1\times m$ matrix and $v_t$ is a scalar, representing measurement noise, which is uncorrelated in time with mean zero and variance $r$. In general, the elements of $\mathbf{x}_t$ are not measured, but are assumed to be described as a first-order Markov process:

$$\mathbf{x}_t = \mathbf{A}_t \mathbf{x}_{t-1} + \mathbf{B}_t u_t + \mathbf{G}_t w_t, \quad t = 1, \ldots, T \qquad (3)$$

where $\mathbf{A}_t$ is an $m\times m$ matrix; $\mathbf{B}_t$ is an $m\times1$ vector; $u_t$ is a scalar representing the input; $\mathbf{G}_t$ is an $m\times1$ vector; and $w_t$ is a scalar, representing the system noise, which is uncorrelated in time with mean zero and variance $q$. Equation (3) is referred to as the state equation. For time-invariant systems, the indices of the system matrices $\mathbf{A}_t$, $\mathbf{B}_t$, $\mathbf{G}_t$ and $\mathbf{C}_t$ disappear.

Using (2) and (3), the TFN model of (1) can be written in the following state space form:

$$\mathbf{x}_t = \begin{pmatrix} \mathbf{A}_s & 0 \\ 0 & \mathbf{A}_n \end{pmatrix} \mathbf{x}_{t-1} + \begin{pmatrix} \mathbf{B}_s \\ 0 \end{pmatrix} u_t + \begin{pmatrix} 0 \\ \mathbf{G}_n \end{pmatrix} w_t \qquad (4)$$

$$y_t = \begin{pmatrix} \mathbf{C}_s & \mathbf{C}_n \end{pmatrix} \mathbf{x}_t + v_t \qquad (5)$$

where the system matrices are defined as

$$\mathbf{A}_s = \begin{pmatrix} \delta_1 & 1 & 0 & \cdots & 0 \\ \delta_2 & 0 & \ddots & \ddots & \vdots \\ \vdots & \vdots & \ddots & \ddots & 0 \\ \delta_{r-1} & \vdots & & \ddots & 1 \\ \delta_r & 0 & \cdots & \cdots & 0 \end{pmatrix}, \quad \mathbf{A}_n = \begin{pmatrix} \phi_1 & 1 & 0 & \cdots & 0 \\ \phi_2 & 0 & \ddots & \ddots & \vdots \\ \vdots & \vdots & \ddots & \ddots & 0 \\ \phi_{p-1} & \vdots & & \ddots & 1 \\ \phi_p & 0 & \cdots & \cdots & 0 \end{pmatrix}$$

$$\mathbf{B}_s = (\omega_0 \quad \omega_1 \quad \cdots \quad \omega_{r-2} \quad \omega_{r-1})^T$$
$$\mathbf{G}_n = (1 \quad \theta_1 \quad \cdots \quad \theta_{p-2} \quad \theta_{p-1})^T$$
$$\mathbf{C}_s = \mathbf{C}_n = (1 \quad 0 \quad \cdots \quad \cdots \quad 0)$$

where the parameters in $\mathbf{A}_s$, $\mathbf{A}_n$, $\mathbf{B}_s$ and $\mathbf{G}_n$ are unknown and defined as in (1). For reasons of convenience, this parameter set is termed $\alpha$.

### 2.3 Kalman Filter and Parameter Estimation

Once a model has been written in state space form, it is possible to use the Kalman filter. The Kalman filter is a recursive procedure for computing the optimal estimator of the state vector at time $t$, based on the information available at time $t$ [Jazwinsky, 1970]. This information consists of the observations up to and including $y_t$. The Kalman filter algorithm for state equation (3) consists of the following equations:

Initial conditions

$\hat{\mathbf{x}}_0$ and $\mathbf{P}_0$

Time update

$$\bar{\mathbf{x}}_t = \mathbf{A}\hat{\mathbf{x}}_{t-1} + \mathbf{B}u_t \tag{6}$$

$$\mathbf{M}_t = \mathbf{A}\mathbf{P}_{t-1}\mathbf{A}^T + \mathbf{G}q\mathbf{G}^T \tag{7}$$

Measurement update

$$n_t = y_t - \mathbf{C}\bar{\mathbf{x}}_t \tag{8}$$

$$f_t = \mathbf{C}\mathbf{M}_t\mathbf{C}^T + r \tag{9}$$

$$\mathbf{K}_t = \mathbf{M}_t\mathbf{C}^T f_t^{-1} \tag{10}$$

$$\hat{\mathbf{x}}_t = \bar{\mathbf{x}}_t + \mathbf{K}_t n_t \tag{11}$$

$$\mathbf{P}_t = (\mathbf{I} - \mathbf{K}_t\mathbf{C})\mathbf{M}_t \tag{12}$$

where $\hat{\mathbf{x}}_t$ is the measurement update; $\bar{\mathbf{x}}_t$ is the time update; $\mathbf{P}_t$ is the covariance matrix of the error in the measurement update: $\mathrm{cov}(\mathbf{x}_t - \hat{\mathbf{x}}_t)$; $\mathbf{M}_t$ is the covariance matrix of the error in the time update: $\mathrm{cov}(\mathbf{x}_t - \bar{\mathbf{x}}_t)$; $n_t$ is the innovation, which is the difference between observation and time update; $f_t$ is the innovation variance; $\mathbf{K}_t$ is the Kalman gain; and $\mathbf{I}$ is a unity matrix. If at time $t$ no observation is available, $\hat{\mathbf{x}}_t = \bar{\mathbf{x}}_t$ and $\mathbf{P}_t = \mathbf{M}_t$. The variance of the measurement noise, $r$, is assumed to be known, while the variance of the system noise, $q$, results from the Kalman filter algorithm when both $q$ and $r$ are scaled by

$$\sigma_*^2(\boldsymbol{\alpha}) = \frac{1}{N-d}\sum_{t=d+1}^{N}\frac{n_t^2(\boldsymbol{\alpha})}{f_t(\boldsymbol{\alpha})} \tag{13}$$

so that $q = \mathrm{var}(w_t)/\sigma_*^2 = 1$ and $r = \mathrm{var}(v_t)/\sigma_*^2$. The parameter set $\alpha$ can then be estimated by evaluating the scaled log-likelihood function [Harvey, 1989]:

$$\log L(\boldsymbol{\alpha}) = -\frac{N-d}{2}(\log 2\pi + 1) - \frac{1}{2}\sum_{t=d+1}^{n}\log f_t(\boldsymbol{\alpha}) \; - \frac{N-d}{2}\log\sigma_*^2(\boldsymbol{\alpha}) \tag{14}$$

where $N$ is the number of time steps; and $d$ is the dimension of the state. The covariance matrix of parameter estimation errors is assumed to approach the Cramer-Rao lower bound [Schweppe, 1973]:

$$\mathbf{R}_C^{-1} = \frac{\partial^2 \log L(\boldsymbol{\alpha})}{\partial\boldsymbol{\alpha}\partial\boldsymbol{\alpha}^T} \tag{15}$$

where $\mathbf{R}_C$ is the Cramer-Rao lower bound.

### 2.4 Generation of Groundwater Time Series

Groundwater time series were generated by transferring an input series, using a pre-defined transfer function. In this study the input data (precipitation excess) were obtained from daily observations of the Royal Netherlands Meteorological Institute at De Bilt, the Netherlands. The time series ranges from July 1, 1957 to December 31, 1999.

The most flexible approach for defining a transfer function is to use a continuous function. While an exponential function is often used to describe the response of a hydrological system, we used the probability density function of a lognormal distribution to transfer precipitation excess into groundwater recharge:

$$\Psi_t = \frac{c}{t\sigma\sqrt{2\pi}}e^{-\frac{1}{2}\left(\frac{\ln t-\mu}{\sigma}\right)^2} \qquad \sigma > 0;\ 0 < t < \infty \tag{16}$$

where $\mu$ and $\sigma$ are the geometric mean and standard deviation of the distribution, respectively, and $c$ is a constant. An advantage of using (16) over other exponential functions is the small number of parameters.

Based on the input series and (16) many different time series were generated. These time series have different response times (i.e. in (16) different values for $\sigma$ and $\mu$ are used). Because the conclusions for these time series are identical, in this paper we only use one output time series to demonstrate the influence of the modelling interval. This output series has been generated by (16), with $c$ = 10, $\sigma$ = 0.5, and $\mu$ = 4. Figure 1 shows the response function. In addition, a stochastic component was added, which can be

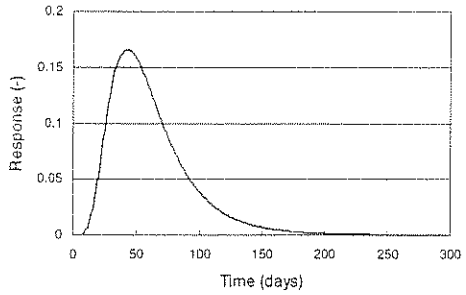


**Figure 1.** Transfer function for generation of groundwater time series.

described by the following autoregressive (AR) model:

$$n_t = \phi n_{t-1} + a_t \tag{17}$$

where $n_t$ is the value of the stochastic component at time $t$ (days); $\phi = 0.99$ is the AR-parameter; and $a_t$ is a white (uncorrelated) noise process with mean $\mu_a = 0$ and variance $\sigma_a^2 = 1.64$, which is 1% of the variance of the deterministic component (transferred precipitation excess). Figure 2 shows the resulting groundwater time series. Subsequently, a second output series was composed using the same transfer function and (17), but now $\phi = 0$. Finally, the generated groundwater time series was split into two parts: a calibration period from July 1, 1957 until December 31, 1989 and a validation period from January 1, 1990 until December 31, 1999.

## 3. RESULTS

### 3.1 Introduction

In order to evaluate the performance of the TFN models two criteria are used:

- mean absolute error (*MAE*) of the deterministic component, described as:

$$MAE = \frac{1}{N-d}\sum_{t=d+1}^{N}|\hat{z}_t - z_t| \tag{18}$$

where $\hat{z}_t$ is the estimated deterministic component; $z_t$ is the true deterministic component; $N$ is the number of model time steps; and $d$ is defined as in (14). *MAE* represents the 'fit' of the deterministic component of the TFN model and is thus implicitly a measure of 'fit' of the transfer function.

- Variance of the gain var[$G$], where the gain $G$ is represented by

$$G = \frac{\sum_i \omega_i}{1 - \sum_k \delta_k} \tag{19}$$

The gain is equal to the value of the unit step response at time $t = \infty$. The variance of the gain is a measure of the overall parameter accuracy. If $c = 10$ in (16), $G$ will approach 10 as well.

Section 3.2 describes the relation between the modelling interval and both criteria by evaluating the results of TFN modelling of the groundwater time series for four different cases:

1. Output series A ($\phi = 0$), using Model 1: modelling interval $dt_{mod}$ equals measuring interval $dt_{meas} = \{10, 20, \ldots, 70\}$ days;
2. Output series A ($\phi = 0$), using Model 2: $dt_{mod} = 10$ days, $dt_{meas} = \{10, 20, \ldots, 70\}$ days;
3. Output series B ($\phi = 0.99$), using Model 1: $dt_{mod} = dt_{meas} = \{10, 20, \ldots, 70\}$ days;
4. Output series B ($\phi = 0.99$), using Model 2: $dt_{mod} = 10$ days, $dt_{meas} = \{10, 20, \ldots, 70\}$ days.

Section 3.3 describes the relation between the amount of measurement data and the performance of the model. This relation is important because, due to the limited length of the time series used in this paper, a larger modelling interval implies a smaller amount of measurement data.

### 3.2 Modelling Interval

Figure 3 relates the criteria for model performance (*MAE* and var[$G$]) to the measuring interval for Model 1 as well as for Model 2. Figure 3a and 3b

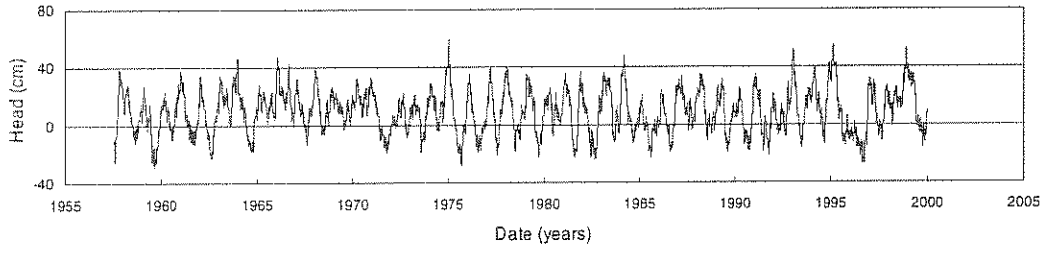


**Figure 2.** Generated time series of groundwater head ($\phi = 0.99$).

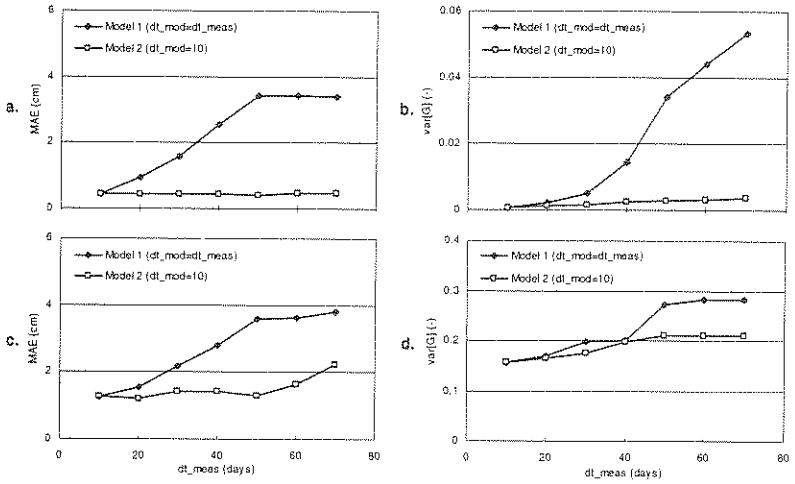


**Figure 3.** Relation between measuring interval and (a) *MAE*, case 1 and 2; (b) var[$G$], case 1 and 2; (c) *MAE*, case 3 and 4; and (d) var[$G$], case 3 and 4.

are based on Output series A ($\phi = 0$), while Figure 3c and 3d are based on Output series B ($\phi = 0.99$). All diagrams clearly show that the performance of Model 2 is always better than that of Model 1. Especially when the noise is small with respect to the deterministic component (3a and 3b), *MAE* and var[$G$] benefit from the small modelling interval. The reason for the better performance is that due to the smaller modelling interval the real response curve can be approximated better. Figure 3 also shows that the performance of Model 2 is much less sensitive to $dt_{meas}$ than the performance of Model 1. The curve of Model 1 flattens from the moment that the measuring interval becomes larger than the time of maximum response.

The better performance of Model 2 is visualised in another way in Figure 4, where the real and the modelled deterministic component of Output series A are shown (4a and 4c) as well as the error of the deterministic component (4b and 4d). The most important conclusion from Figure 4 is that Model 1 is often not able to model the peaks of the output series resulting in relatively large errors, whereas Model 2 models the peaks very well. The same holds for Output series B.

### 3.3 Amount of Measurement Data

The results reported in the previous section on the relation between the modelling interval and model performance are influenced by the amount of available measurement data *ndat*. Therefore, this section seeks to determine the relation between the amount of data and the performance of the model.

The relation between the amount of measurement data and the performance of the model is shown in Figure 5, where *MAE* and var[$G$] are plotted against *ndat* for Output series A ($\phi = 0$). Figure 5a clearly shows that *MAE* is rather insensitive to *ndat*. Only when *ndat* is small (about 100-200 measurements) *MAE* changes. As a result, the

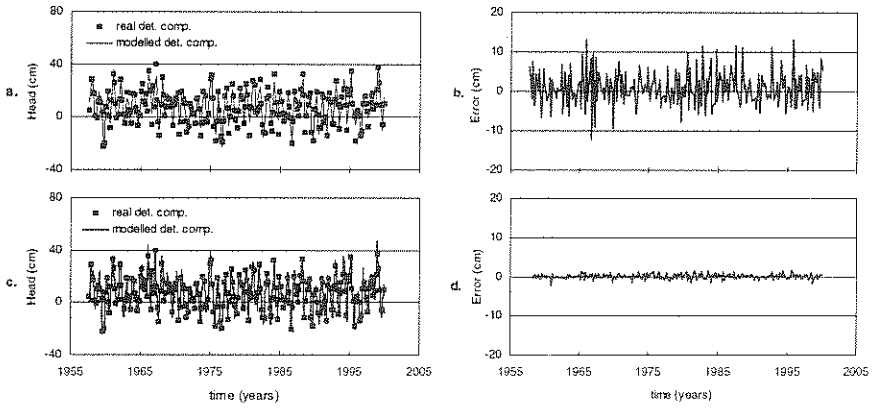


**Figure 4.** (a) Real and modelled deterministic component, Model 1, $dt_{meas} = 70$; (b) Deterministic error, Model 1, $dt_{meas} = 70$; (c) Real and modelled deterministic component, Model 2, $dt_{meas} = 70$; (d) Deterministic error, Model 2, $dt_{meas} = 70$.

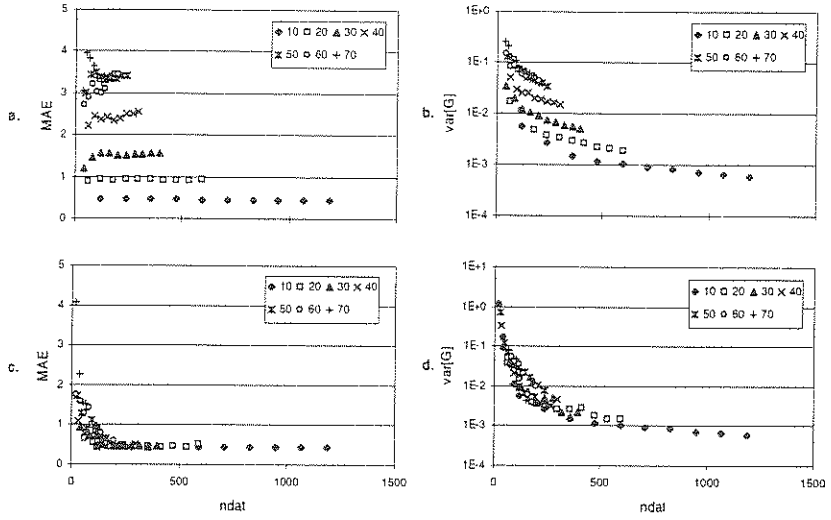


**Figure 5.** Influence of amount of measurement data *ndat* on model performance for different measurement intervals, using Model 1 (a and b) and Model 2 (c and d), based on Output series A.

increase of *MAE* in Figure 3a is not significantly influenced by the amount of data and is thus completely caused by an increase of the measuring interval. The same holds for Model 2 (Figure 5c). Again, *MAE* is more or less constant for *ndat* > 200 indicating that adding more data to the output series does not cause the model fit to increase. On the other hand, var[*G*] does decrease when *ndat* increases. Therefore, part of the increase of var[*G*] in Figure 3b and 3d is caused by the difference in the amount of available data. These conclusions are also valid for Output series B. However, due to the noise the relation between model accuracy and the amount of data is less clear.

## 4. CONCLUSIONS

The most important conclusion from this paper is that, if the time scale of the underlying process is smaller than the measuring interval, the performance of TFN models increases significantly by reducing the modelling interval. The reason for this is that a more detailed model, combined with a more detailed input series, is able to describe the underlying process more precisely. Consequently, even if the output time series is sparse, the use of a small modelling interval makes it possible to model the time series very well.

The amount of output measurements influences the accuracy of the TFN model in the sense that the parameter uncertainty increases when the amount of output measurements decreases. In contrast, the 'fit' of the model is less sensitive to the amount of measurements.

## 5. REFERENCES


Bierkens, M.F.P., M. Knotters, and F.C. van Geer, Calibration of transfer function-noise models to sparsely or irregularly observed time series, *Water Resources Research, 35*(6), 1741-1750, 1999.

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

Gehrels, J.C., F.C. van Geer, and J.J. de Vries, Decomposition of groundwater level fluctuations using transfer modelling in an area with shallow to deep unsaturated zones, *Journal of Hydrology, 157*, 105-138, 1994.

Harvey, A.C., Forecasting, *Structural Time Series Models and the Kalman Filter*, Cambridge University Press, Cambridge, 1989.

Hipel, K.W., and A.I. McLeod, *Time Series Modelling of Water Resources and Environmental Systems, Developments in Water Science Series*, vol. 45, Elsevier Science, Amsterdam, 1994.

Jazwinsky, A.H., *Stochastic Processes and Filtering Theory*, Academic Press, 1970.

Schweppe, F.C., *Uncertain Dynamic Systems*, Prentice-Hall, New Jersey, 1973.

van Geer, F.C., and A.F. Zuur, An extension of Box-Jenkins transfer/noise models for spatial interpolation of groundwater head series, *Journal of Hydrology, 192*, 65-80, 1997.

Young, P.C., A.J. Jakeman, and D.A. Post, Recent advances in the data-based modelling and analysis of hydrological systems, *Water Science and Technology, 36*(5), 99-116, 1997.