Predictions for groundwater well fields using stochastic modelling

F. THORDARSON[1], **H. MADSEN**[1] **& H. MADSEN**[2]

1 *DTU Informatics, Technical University of Denmark; Richard Petersens Plads 305, DK-2800 Kgs. Lyngby, Denmark*

ft@imm.dtu.dk

2 *DHI; Agern Allé 5, DK-2970 Hørsholm, Denmark*

Abstract A stochastic modelling framework for identifying groundwater well fields is presented, which combines prior physical knowledge of dynamic groundwater well field systems with available information embedded in data. The model is a conceptual stochastic model, formulated in continuous-discrete state space form that facilitates a direct physical interpretation of the estimated parameters. The parameter estimation method is a maximum likelihood method, and the model parameters are validated by applying statistical methods using all the available data. The statistical tools are used to identify the deficiencies in a model that is considered too simple. Even though the predictions seem adequate, statistical methods show that the model needs to be extended to be able to provide accurate predictions for the groundwater level in all wells.

Keywords groundwater; well field model; stochastic differential equations; grey-box model; prediction; parameter estimation; maximum likelihood method

INTRODUCTION

It is essential to ensure high quality drinking water in the future, which then calls for reliable operation and management of the groundwater resources at well fields. One of the foundations of the groundwater resource management is the mathematical model that describes the behavior of the aquifer penetrated by one or several wells. For control, optimization and forecasting, the complexity of the mathematical expressions needs to be reduced to enable more rigid stochastic representation of the dynamics.

The groundwater elevation in the well field varies in both time and space and is traditionally described by the governing equation for groundwater flow, which most frequently is facilitated by a deterministic partial differential equation (Anderson & Woessner, 2002). With multiple discharge locations in the well field the utility of the governing equation becomes highly complex. A popular approach for simplification is to consider a lumped parameter model where the partial differential equation is replaced by a finite set of ordinary differential equations in state-space form, which then introduces a set of state-space variables describing the dynamics of the well field. The state-space model is formulated using all the available prior physical knowledge, which include the known physical characteristics of the considered system and any auxiliary processes connected to the well field. This approach disregards any stochasticity related to the variation in time and space with a serious drawback of obtaining a reasonable parameterization. The total model is often characterized by having a large number of parameters and due to unavoidable idealizations, simplifications and unknown parameters, it is difficult to predict the accuracy of the total model. This modelling approach is often referred to as a white-box approach, since the model structure is completely transparent and the variation in the available data is neglected.

On the contrary is the black-box approach where the model is formulated by only considering the available well field data and statistical methods are applied to reduce

and validate the structure and the parameterization for the groundwater well field. The used of statistical methods enables a possibility for using rigorous stochastic dynamical models which then provide methods for predicting the uncertainty of the model predictions. However, the data is sampled at discrete time and a drawback of the discrete time formulation is that information about the physical parameters is partially hidden, and due to measurement errors or limitations in model flexibility, a reasonable continuous time model cannot be obtained.

It is desirable to obtain a modelling approach that reduces the gap between the conventional models based on physical characteristics and the pure statistical discrete time approach. Using formulation and estimation method, where the parameterization is kept in continuous time, a continuous time stochastic model is obtained where the estimated parameters do have a direct physical interpretation. Hence, in relation to the well field model any knowledge of physical constants and water balance relations can be exploited to improve the parameterization. This modelling approach is referred to as the grey-box approach, since being a combination of the other two approaches.

This paper presents a formulation and estimation of a simple continuous time stochastic model for the groundwater well field that explicitly describes how the measurements and model errors enter into the model, and, due to continuous time formulation, the model facilitates a direct physical interpretation of the estimated parameters. Based on the proposed method it is demonstrated that the rather simple continuous time stochastic model constitutes an operational description of the spatio-temporal variation for simulations and predictions for the considered groundwater well field.

CONTINUOUS-TIME STOCHASTIC MODEL FOR GROUNDWATER WELL FIELD

By considering the lumped parameter approach in state-space form, represented by a finite set of ordinary differential equations (ODEs), the translation into a set of stochastic differential equations (SDEs) is often a rather straightforward procedure. This is usually obtained by replacing the ODE models with the SDE models, which in addition also includes one or more algebraic equations describing how measurements are obtained at discrete time instants. Most often the models are formulated as continuous-discrete time state-space models and in its most general form it is written as

$$dx_t = f(x_t, u_t, t; \theta)dt + \sigma(t, u_t; \theta)d\omega \tag{1}$$

$$y_k = h(x_k, u_k, t_k; \theta) + e_k \tag{2}$$

where $t \in \Re$ is time (t_k, $k=1,...,N$ are sampling instants); $x_t \in \Re^n$ is a vector of state variables; $u_t \in \Re^m$ is a vector of input variables; $y_k \in \Re^l$ is a vector of output variables; $\theta \in \Re^p$ is a vector of possibly unknown parameters; $f(\cdot) \in \Re^n$, $\sigma(\cdot) \in \Re^{n \times n}$ and $h(\cdot) \in \Re^l$ are nonlinear functions; $\{\omega\}$ is a n-dimensional standard Wiener process, and $\{e_k\}$ is a l-dimensional white noise process with $e_k \in N(0, S(u_k, t_k, \theta))$. The standard Wiener process is a continuous stochastic process with stationary and

independent Gaussian time increments, which have the mean value zero and a covariance S equal to the magnitude of the increments (Jazwinski, 1970). Equation (1) is called the system equation whereas equation (2) is the observational equation.

The first term on the right side of the system equation is usually called the drift term, since it represents the physical structure of the system, determined and formed from the system of ordinary differential equations. Hence, any prior physical knowledge regarding the physical structure is included in the drift term where the parameters provide some physical interpretation of the system. Furthermore, the physical characteristics of the drift term are expressions most hydrogeologists are familiar with from formulating the traditional groundwater flow models.

The second term on the right side of the system equation is the diffusion term of the SDE model, which provides a suitable interpretation of the errors that exist due to the fact that the mathematical model is often not describing the true process exactly. However, the gap between the true process and the model should be reduced and by estimating the diffusion in the model, any unrecognized phenomena or unmodelled inputs can be detected and directly or indirectly considered in the model. Frequently is this discrepancy related to some specific state description in the model, and by extending this particular state description by additional state variables, more generic methods is obtained for systematic improvement of the model.

The observation equation (2) then relates the discrete time observations to the state variables at time points where observations are available. When determining unknown parameters of the model from a set of data, the model equations in (1) and (2) enables flexible estimation that can account for varying sample times and missing observations in the data series. The model provides a separation between the process noise and the measurement noise, which allow the parameters to be estimated in a prediction error setting, using statistical methods, the maximum likelihood method.

PARAMETER ESTIMATION

A solution to the well field prediction problem is to optimize a set of parameters, such that the model for the groundwater levels in the well field sufficiently fits the available data. The most direct terminology is to minimize the error between the model output and the observed output for the well field. For such an objective, mainly two estimation methods have been applied for optimizing the parameters in hydrological studies; the Output Error method (OE) and the Prediction Error method (PE).

The OE method minimizes the sum of squared simulation error and is applied for white-box models with well described physical characteristics, without considering variation in the available data. The parameters estimated by the OE method are, in general, not provided with any uncertainty. Furthermore, the OE method can only be considered for offline estimation, i.e. the estimates are only depending on the initial values; for online estimation the state estimates are updated for every time instants. The PE method seeks for minimizing the sum of squared one-step prediction error to obtain the best fitted model for the groundwater level in the well field, and the PE method includes both offline and online estimation. Moreover, the PE method also provides an uncertainty for the estimates, which is well suited for short-term predictions.

Given the model structure in (1) and (2), the unknown parameters can be determined by finding the parameters that maximize the likelihood function of a given sequence of measurements, i.e. by the Maximum Likelihood (ML) method. From probability theory the rule of independent probabilities can be applied to express the

likelihood function as a product of conditional densities, and by representing the measured sequence by $Y_k = [y_k, \mathsf{K}, y_0]$ the likelihood function is the joint probability density

$$L(\theta; Y_k) = P(Y_K|\theta) = \left(\prod_{k=1}^{N} P(y_k|Y_{k-1}, \theta) \right) P(y_0|\theta)$$

To obtain an exact estimation of the likelihood function, the continuous-discrete filtering problem needs to be solved, and the initial probability density function $P(y_k|\theta)$ must be known and parameterized, and all subsequent conditional densities must be determined to successively solve Kolmogorov's forward equation (Kloeden & Platen, 1999) In practice, however, this approach is not computationally feasible and an alternative is required. Since the SDE's in (1) are driven by a Wiener process, which has Gaussian increments, the conditional densities can be approximated by Gaussian densities. For linear models the Kalman filter provides the exact solution for the filtering problem, and for nonlinear models the problem is approximated by applying the extended Kalman filter (Madsen *et al.*, 2004).

The Gaussian density is completely characterized by its mean and covariance, which are denoted by $\hat{y}_{k|k-1} = E\{y_k|Y_{k-1}, \theta\}$ and $R_{k|k-1} = V\{y_k|Y_{k-1}, \theta\}$, respectively, and by introducing an expression for the innovation $\varepsilon_k = y_k - \hat{y}_{k|k-1}$ the likelihood function can be rewritten as

$$L(\theta; Y_N) = \left(\prod_{k=1}^{N} \frac{\exp\left(-\frac{1}{2}\varepsilon_k^T R_{k|k-1}^{-1} \varepsilon_k\right)}{\sqrt{\det(R_{k|k-1})}(\sqrt{2\pi})^l} \right) P(y_0|\theta)$$

and thereof, the parameter estimates can be determined by conditioning on the initial values and solving the optimization problem

$$\hat{\theta} = \arg\min_{\theta \in \Theta} \{-\ln(\theta; Y_N|y_0)\}$$

With the unknown parameters of the model estimated by the ML method, along with corresponding standard deviations, statistical tests can be performed to check if the parameters are significantly different from zero, which then indicates that some improvement is needed for the model structure. The parameters of the diffusion term in equation (1) are included in the ML estimation.

One of the main aspects of the modelling framework is its predictive ability, which implies that the output errors are examined for any systematic pattern for further extension of the model, as well as investigation of the sample autocorrelation function and sample partial autocorrelation function of the residuals to detect if two or more consecutive residuals are dependent or, in contrast, can be regarded as white noise (Kristensen *et al.*, 2004). Correlation between the residuals indicates that the model is not adequate for prediction, since systematic errors are detected in the model that can affect the model prediction drastically. An adequately parameterized model is characterized by having uncorrelated residuals (Madsen, 2008).

AN EXAMPLE

The following is an example to illustrate the important features of the continuous time stochastic model described above; the lumped model for the well field, the parameter estimation and model prediction. The well field has three pumping wells, which all pump from the same aquifer. These three wells are a part of water distribution network with 21 operating wells attached, where all wells are pumping from the same aquifer. The total well field is divided into three groups due to geographical location. Here, one of these is studied.

The conceptual model is sketched in Fig. 1a, showing the three wells located on a straight line, that is, well No. 2 is located on the line between well No. 1 and well No. 3. This simplifies the model by assuming that drawdown in well No. 3 when pumping from No. 1 is detected in well No. 2 as well. This assumption is also valid when the water level changes in well No. 1 when pumping from well No. 3.

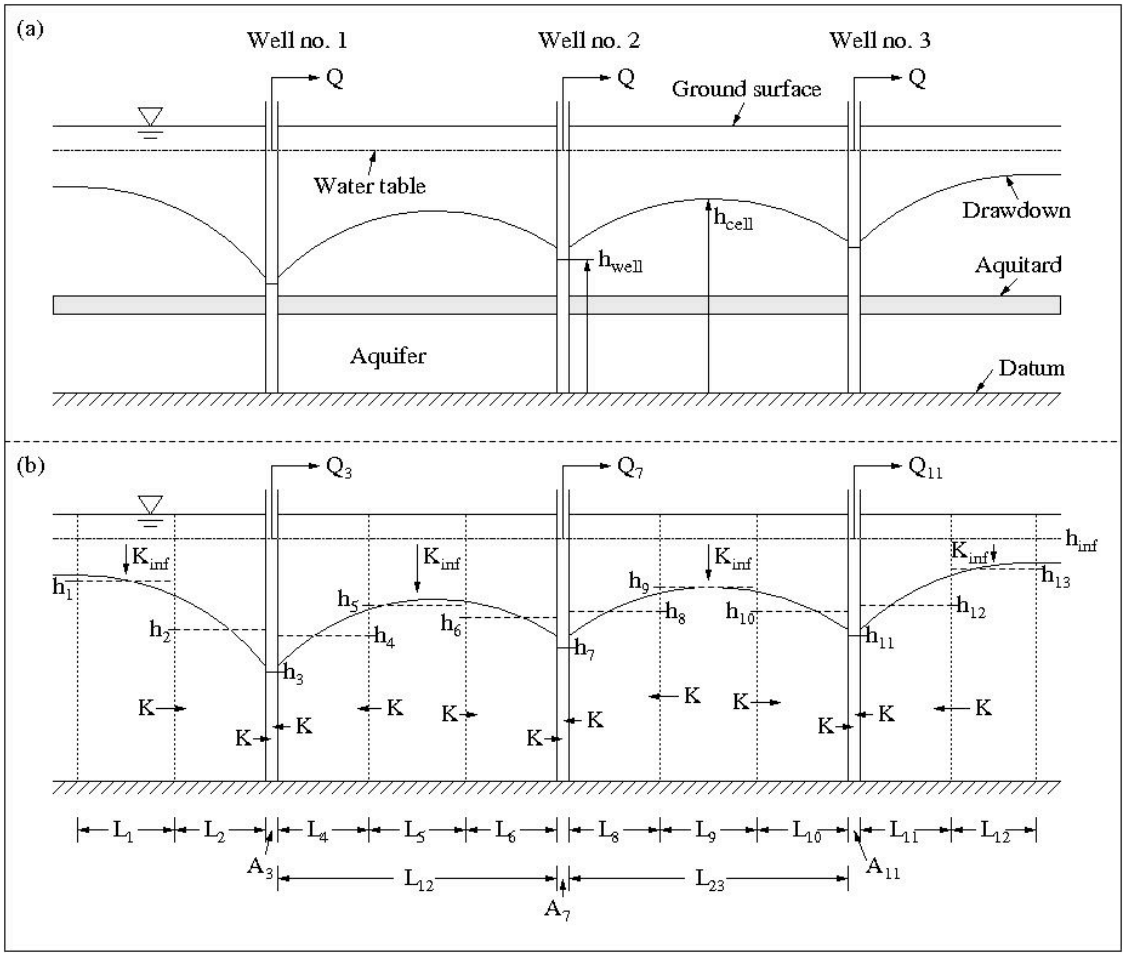

Fig. 1 Conceptual model for a well field with 3 operating wells: (a) The classical illustration of the model. (b) The lumped model represented as number of linear reservoirs.

The objective is to predict the piezometric heads in the wells when pumping from a confined aquifer. However, since the lumped parameter model is considered for the model structure, the parameters are lumped vertically, from datum to the piezometric head, and the suggested model for the groundwater well field is expected to consist of a number of reservoirs where the water levels in the reservoirs are the state variables in the state-space representation (Jacobsen *et al.*, 1997). As illustrated in Fig. 1b, the only measured state variables are the water-levels in the wells. The water levels between

any two wells, and at the boundaries, are unobserved state variables, which will be estimated in relation to the observations in its two adjacent operating wells. The behaviour of the water table between two operating wells is nonlinear, but by assuming several linear reservoirs for the water table to represent the flow from one well to another, the water table can be linearly approximated. The water level, or the reservoirs, in the unobserved states does never dry out, indicating that at least one of the unobserved reservoirs between every two observed wells is infiltrated with additional water.

Considering the states as given in Fig. 1b, and with the index i indicating the state of interest, the suggested stochastic state space model, as in equation (1), is represented as follows: The pumping wells are the observed states (h_3, h_7 and h_{11} in Fig. 1b) and their dynamics are described as

$$dh_i = \left[\frac{K}{A_i}h_{i-1}(t) - \frac{2K}{A_i}h_i(t) + \frac{K}{A_i}h_{i+1}(t) - \frac{1}{A_i}Q_i(t)\right]dt + \sigma_i d\omega_i(t)$$

with K assumed to be the lumped hydraulic conductivity and A_i is considered as the areal closest to the well directly affected by the discharge. Here, and in all the following system equations for the well field, the σ_i values represent the variation of the system noise for state description i , where $i=1,...,13,$ and corresponding noise term $d\omega_i$ is assumed to be an independent standard Wiener process, and also assumed independent from the measurement noise in the observation equation.

The state variables illustrating the recharge to the aquifer between operating wells (h_5 and h_9) become

$$dh_i = \left[\frac{K}{SL_i}h_{i-1}(t) - \frac{1}{SL_i}\left(2K + \frac{1}{K_{\inf}}\right)h_i(t) + \frac{K}{SL_i}h_{i+1}(t) - \frac{h_{\inf}}{SL_i K_{\inf}}Q_i(t)\right]dt + \sigma_i d\omega_i(t)$$

The same goes for the recharged boundary states (h_1 and h_{13}), except for one term is neglected in each case; for h_1 the first term in the square brackets is omitted, and for h_{13} the last term inside the square brackets. S is the storage coefficient for the lumped flow and L_i is the estimated size of the reservoir i. h_{inf} is the estimated boundary condition, i.e. the water level approaches the undisturbed water table if no pump is active in the well field for a reasonably long time. The term K_{inf} is the estimated resistance for the flow from the boundaries to the reservoirs.

For all the remaining states, the intermediate states of the water level in the reservoirs is represented as

$$dh_i = \left[\frac{K}{SL_i}h_{i-1}(t) - \frac{2K}{SL_i}h_i(t) + \frac{K}{SL_i}h_{i+1}(t)\right]dt + \sigma_i d\omega_i(t)$$

There are three observation equations since there are three measured water levels in the wells, i.e.

$$Y_1(k) = h_3(k) + e_1(k)$$

$$Y_2(k) = h_7(k) + e_2(k)$$

$$Y_3(k) = h_{11}(k) + e_3(k)$$

where the e_1, e_2, e_3, correspond to the measurement noises.

The parameter estimation are shown in Table 1. The estimation for the hydraulic conductivity and the storage coefficient are reasonably estimated, but compared to results from a pumping test for the aquifer the estimates are orders of magnitudes higher. This is explained by the fact that these two estimated parameters are lumped vertically and correspond to routing of water and storage in the aquifer, as well as all the layers above it. Therefore, it is impossible to compare results from pumping tests and the lumped estimates. The two estimated values, K and S, correspond to the individual reservoir in the lumped model, where K is assumed as routing coefficient per length unit and the storage S is considered as the total storage per length unit in each reservoir. Model extension to take consideration to the different layers in the conceptual model can be implemented into the introduced stochastic model, but no such attempt is made in this particular study.

Table 1 Estimated values for several parameters in the stochastic well field model. K:[m min^{-1}], S [-]; A_i [m^2]; h_{inf} [m].

Parameter	Pumping Test	Estimate	Std. dev.	Significant
K	0.0420	1.090	0.0254	YES
S	0.0012	2.083	0.3625	YES
A_3	-	10.253	0.7147	YES
A_7	-	5.481	0.2893	YES
A_{11}	-	6.264	0.5335	YES
h_{inf}	-	7.141	0.3642	YES
σ_1	-	0.038	0.0335	NO
σ_3	-	0.357	0.0485	YES
σ_5	-	0.030	0.0159	YES
σ_7	-	0.291	0.0247	YES
σ_9	-	0.165	0.0186	YES
σ_{11}	-	0.215	0.0192	YES
σ_{13}	-	0.018	0.0105	NO
S_1	-	0.000	0.0003	NO
S_2	-	0.001	0.0005	NO
S_3	-	0.002	0.0009	NO

By performing t-tests the parameters can be checked for being significantly different from zero. Status for significance of each parameter is displayed in the last column in Table 1, and it shows that the variances for the system noises, regarding the boundary conditions, are not significant (σ_1 and σ_{13}). For extending the model further, focus should be on the state descriptions for the boundary conditions, since from the parameter estimation it can be concluded that these states are not entirely fulfilled with the present description. The most probable cause is related to the other two groups of wells in the total well field and to get a better understanding of the boundary conditions for this small group of three wells, correlation to the other groups need to be exploited.

It is interesting to see how adequate the model is to predict the water level in the three wells. Fig. 2 displays a comparison between the observations (solid line) and corresponding model output (dashed line) for the pumping wells. Although it appears as the prediction follows the observations rather well, there is a clear difference for all three wells where the greatest deviation is in relation to abrupt changes in the water level, i.e. when a pump is switched on or off. Fig. 3 shows that the difference between the model and the observations is serially correlated, which indicates that an improved model should be obtained by addition of a reservoir between operating wells.

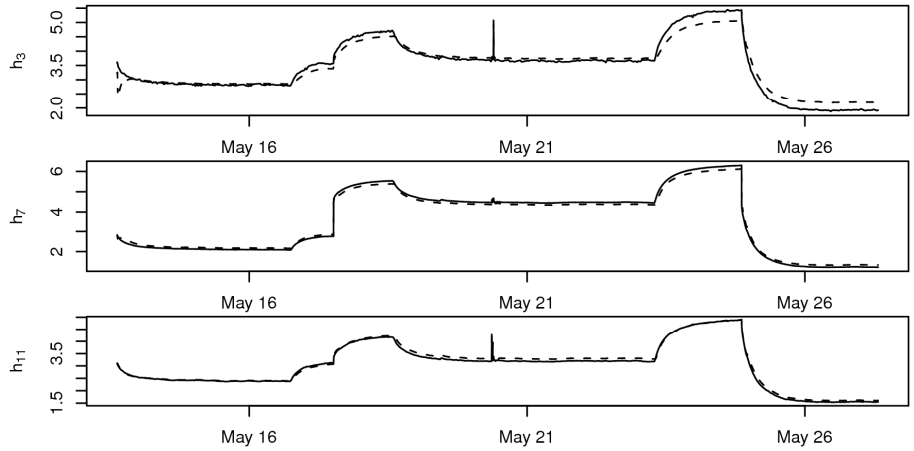

Fig. 2 Comparison between measurements (solid line) and predictions (dashed line) for all operating wells.

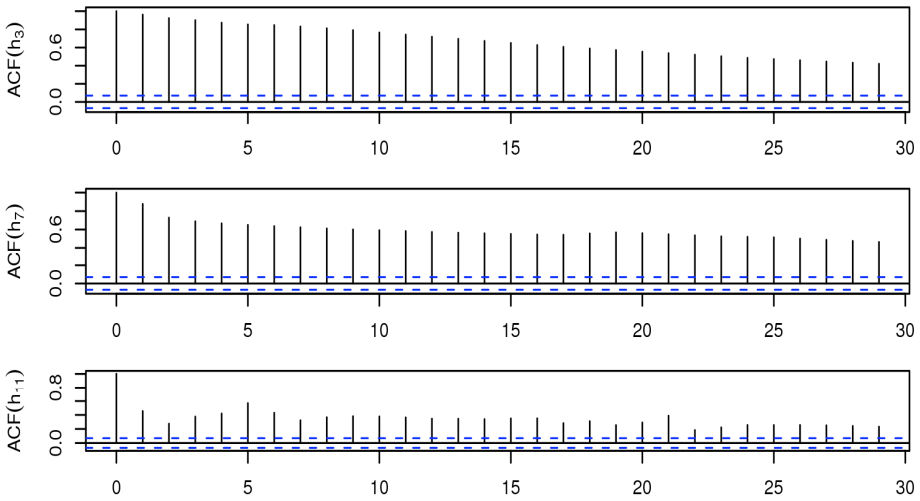

Fig 3 Autocorrelation functions for the residuals for all operating wells

This example shows how the presented lumped stochastic model can be used for parameter estimation and prediction for a groundwater well field. It is also shown how statistical methods can be applied to detect deficiencies in a model, as well as suggest which state descriptions require improvement. By optimizing the parameters with the prediction error method, the model is able to predict the water levels in the wells within the limited region, but from a statistical point of view an improved model is need to obtain more adequate results.

CONCLUSION

A continuous time stochastic model for a groundwater well field has been presented. This modelling framework combines the best from deterministic and stochastic modelling for identification of models, for model-based control of groundwater well fields. The model basis are the state descriptions in the stochastic state-space model, derived from stochastic differential equation models, which are just as appealing as ordinary differential equation models from an engineering point of view. The maximum likelihood method provides uncertainty to the estimates, which is highly important for performing model validation by means of statistical tests and residual analysis. Based on these methods it has been demonstrated that the rather simple stochastic model can be constructed to give sufficient results for the physically interpretable parameters. However, statistical tests showed that the model requires an extension to compose an operational description of the spatio-temporal variation of the groundwater well field, which eventually will improve the groundwater level predictions in the well field.

ACKNOWLEDGEMENTS

This work was partly funded by the Danish Strategic Research Council, Sustainable Energy and Environment Programme. For more information visit http://wellfield.dhigroup.com/.

REFERENCES

Anderson M. P., Woessner W. W. (2002) Applied groundwater modeling: simulation of flow and advective transport. Academic Press.

Jacobsen J. L., Madsen H. & Harremoes P. (1997) A stochastic model for two-station hydraulics exhibiting transient impact. *Water Science and Technology,* **36**(5), 19-26.

Jazwinski A. H. (1970) *Stochastic processes and filtering theory*. Academic Press, New York, USA.

Kirstensen N. R., Madsen H. & Jørgensen S. B. (2004) A method for systematic improvement of stochastic grey-box models. *Computers and Chemical Engineering*, **28**,1431-1449.

Kloeden P. E. & Platen (1999) *Numerical solution of stochastic differential equations*. Springer.

Madsen H., Nielsen J. N., Lindström E., Baadsgaard M. & Holst J. (2004) *Statistics in finance*. Lund University, Centre for mathematical sciences.

Madsen H. (2008) *Time series analysis*. Chapman&Hall/CRC