18th World IMACS / MODSIM Congress, Cairns, Australia 13-17 July 2009
http://mssanz.org.au/modsim09

# Deriving an emulation model of a rectangular-basin two-layer numerical model

**Galelli, S. [1], T. Shintani [2], F. Pianosi[1], J. Imberger [3] and R. Soncini-Sessa[1]**

[1] *Dipartimento di Elettronica e Informazione, Politecnico di Milano, Milan, Italy*
*Email: soncini@elet.polimi.it*
[2] *Department of Civil Engineering, Tokyo Metropolitan University, Tokyo, Japan*
[3] *Centre for Water Research, University of Western Australia, Perth, Western Australia*

**Abstract:** In the last decades, rapid improvement in processor speed has encouraged the development of large, distributed, process-based models, which are implemented in more and more complex computer codes. However, despite the increased computing power, the use of such models is far from being inexpensive: obtaining output trajectories over a time horizon of few years can require many days of simulation. As a consequence, the application for planning and management purposes is still very limited. What-if analysis, for example, i.e. the evaluation of the behavior of a system against a set of possible scenarios, can be applied to process-based models only when the number of scenarios is very small, as each model simulation can be prohibitively time consuming. Furthermore, the integration of process-based models into an optimization scheme is, at the state of the art, essentially impracticable.

In recent years, emulation modeling emerged as a promising technique to overcome these limitations. An emulation model is a simple, usually lumped model, which is identified from synthetic data generated via simulation of a computationally inefficient model, and that can be used in its place to run fast simulations and optimization. Emulation modeling is largely employed in aerospace and mechanical engineering and is an emerging issue in environmental modeling, especially in the field of air quality. Applications to water systems mainly concern modeling and control of diffuse pollution in groundwater and soils. In this paper, the application of emulation modeling techniques is extended to hydrodynamics, namely modeling of stratified lakes. The ultimate scope is to allow for an indirect use of large, distributed, process-based hydrodynamic models for optimization purposes, closing the gap between scientific-oriented research and decision-making practice.

The applicability of the emulation modeling approach is tested over a simplified but rather realistic case study: a one-dimensional, nonlinear, two-layer model of a rectangular basin. The model is simulated to generate synthetic time series of the thermocline displacement from the equilibrium position, as a function of the wind action, in different conditions of stratification (i.e. layers density). These data are used to develop a very simple emulation model. Precisely, an AutoRegressive-eXogenous (ARX) emulation model is identified, with different parameter settings estimated for each stratification condition. Although very simple, the emulation model provides an accurate estimate of the selected output variable. Moreover, its parameters can be provided with a physically meaningful interpretation. These promising results motivate further research to extend the application of emulation modeling to more sophisticated hydrodynamics models and to use this approach for planning and management of water resources.

***Keywords:*** *Emulation modeling, two-layer model, lakes stratification, water resources planning and management*

## 1. INTRODUCTION

An emulation model (also known as surrogate model or meta-model) is a simplified and computationally efficient model identified using statistical identification techniques over synthetic data generated via simulation with a large, distributed, process-based model. The emulation model aims at reproducing only the dominant modes of the large original model that are significant for the purpose of the emulation modeling exercise. Kleijnen and Sargent (2000) discuss four possible goals: understanding of the process-based model, prediction of future outputs, validation, and optimization, which is the most common goal in emulation modeling. As for the structure of the emulation models, the literature shows a wide variety of model types, e.g. polynomials, splines, kriging or neural networks (for a complete review, see, for example, Simpson et al., 2001) depending on the emulation model goal. Emulation modeling techniques are largely employed in aerospace and mechanical engineering and some applications to water systems have recently appeared (see, for example, Aly and Peralta, 1999; or Broad et al., 2005).

In this paper, the applicability of emulation modeling in the field of hydrodynamics will be explored. The emulation modeling exercise will focus on a one-dimensional nonlinear two-layer model of a stratified lake, as an example of simplified but rather realistic process-based hydrodynamic model. To derive the emulation model, the following identification procedure (based on Kleijnen and Sargent, 2000) was adopted.

- The process-based model is critically analyzed to construct the knowledge base of the physical system that will serve throughout all the subsequent emulation modeling exercise.
- The input and output of the emulation model are chosen. The output is one of the variable that can be simulated by the process-based model and whose knowledge is needed for the simulation/optimization purpose. The input variables are selected among the inputs of the process-based model; the selection is based on physical considerations about the behavior of the system and on data analysis (e.g. correlation analysis).
- Simulations of the process-based model are run to generate the I-O data necessary for the identification of the emulation model. This phase, known in the literature as Experimental Design (ED) (for a review of the most common ED techniques, see Kleijnen et al., 2005), is of crucial importance as the set of simulated data will directly affect the range of validity of the emulation model.
- The type of emulation model to be identified is chosen.
- The emulation model is calibrated and validated using the data-set generated in the ED. If results are satisfactory the procedure ends, otherwise it is necessary to return to the previous step and select a different type of emulation model.

The paper is organized as follows. In section 2 the process-based (two-layer) model of the lake is described. The procedure for identifying the emulation model and its application are presented in section 3 while section 4 provides a physically meaningful interpretation of the parameters of the emulation model. Section 5 draws the conclusions about the presented application.

## 2. THE TWO-LAYER MODEL

In a strongly stratified lake, the flow and mixing characteristics of the water column are mainly governed by the wind-induced response of the stratified layers. The idealized form of this situation is well described by a two-layer assumption consisting from lighter upper layer and denser bottom layer. In the present application we considered a rectangular basin (see Figure 1), with a section length of 10 Km.

By introducing the two-layer assumption, the equations of the mass and momentum conservation are greatly simplified and the solutions are easily obtained by a simple numerical procedure compare to the full 3D hydrodynamic model such as POM (Mellor, 1996) and ELCOM (Hodges, et al., 2000). In this paper, we employ a one-dimensional nonlinear two-layer model, shown as below,

$$\frac{\partial H_1}{\partial t} + \frac{\partial q_1}{\partial x} = 0 \tag{1}$$

$$\frac{\partial H_2}{\partial t} + \frac{\partial q_2}{\partial x} = 0 \tag{2}$$

$$\frac{\partial q_1}{\partial t} + \frac{\partial \left(u_1 q_1\right)}{\partial x} = -gH_1 \frac{\partial \eta_1}{\partial x} + u_*^2 - \frac{\tau_i}{\rho_0} \tag{3}$$

$$\frac{\partial q_2}{\partial t}+\frac{\partial(u_2 q_2)}{\partial x}=-gH_2\left(\frac{\rho_1}{\rho_2}\frac{\partial \eta_1}{\partial x}+\frac{\rho_2-\rho_1}{\rho_2}\frac{\partial \eta_2}{\partial x}\right)+\frac{(\tau_i-\tau_b)}{\rho_0} \tag{4}$$

where subscripts 1 and 2 denote variables in the upper and lower layer, respectively, $H$ is the layer thickness, $q$ is the volume transport, $u$ is the velocity, $\eta$ is the interface displacement from equilibrium position, $u_*$ is the friction velocity of the surface in the water phase, $\tau_i$ and $\tau_b$ are the shear stresses at the interface and bottom, respectively, $\rho$ is the density and $\rho_0$ is the reference density. The instantaneous layer thickness $H$ and the volume transport $q$ are defined for each layer as,

$$H_1 = h_1 + \eta_1 - \eta_2 \tag{5}$$

$$H_2 = h_2 + \eta_2 \tag{6}$$

$$q_1 = H_1 u_1 \tag{7}$$

$$q_2 = H_2 u_2 \tag{8}$$

The numerical scheme for solving the governing equations is similar to those developed by Oguz et al. (1990). The set of equations (1) to (4) were integrated explicitly using a predictor-corrector scheme. The horizontal advection terms were evaluated using the ULTIMATE-QUICKEST method (Leonard, 1991) to obtain stable and accurate solutions with relatively large time steps. To avoid the CFL (Courant-Friedrich-Levy) limit due to the celerity of surface gravity waves, an implicit method was applied to the barotropic force term following Casulli and Cheng (1992). Therefore, the time step was limited only by the CFL condition due to the celerity of internal gravity waves. The numerical discretization used a staggered grid (Arakawa and Lamb, 1977) so that the thicknesses and velocities in the layer were determined at consecutive grid points.

## 3. IDENTIFICATION OF THE EMULATION MODEL

In the following we will show how to approximate the two-layer model with an emulation model.

### 3.1. Output-Input selection

In the present application the variable that we want to approximate is the maximum thermocline displacement $\eta_2$ from the equilibrium position that is observed at the two extremes of the basin section (with equal amplitude and opposite sign, see Figure 1).

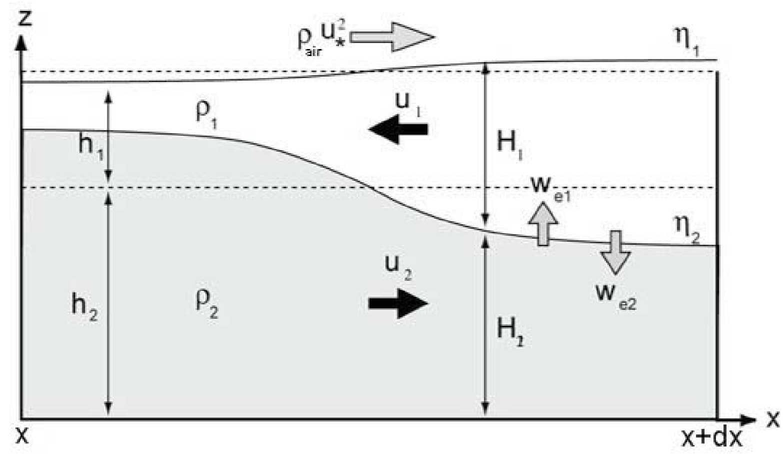


**Figure 1.** Graphical representation of the two-layer rectangular basin. The meaning of the variables is explained within the text.

The candidate input of the emulation model is the velocity (modulus and direction) $u_*$ of the wind, which is the physical phenomenon causing the thermocline displacement (and the only input to the two-layer model). However data analysis revealed that a better choice is the squared wind velocity, namely

$$u_*^2 = |u_*| \cdot u_* \tag{9}$$

This is also consistent with the physics of the system, since the shear stress $\tau_s$ at the free surface induced by the wind action is proportional to the square of the wind speed.

### 3.2. Experimental design

To get the data for identifying the emulation model and testing its degree of accuracy, a set of simulation experiments of the two-layer model were designed. For each simulation it was necessary (i) to specify the parameters (the upper and bottom layer temperature $T_1$ and $T_2$ and the initial thickness $h_1$ and $h_2$) characterizing the stratification condition of the basin; and (ii) to define the trajectory of the wind velocity at the basin surface over the simulation horizon.

In order to analyze the behavior of the emulation model for different stratification conditions, five different stratification scenarios were defined. The two-layer model parameters characterizing these five scenarios are shown in Table 1, together with the fundamental period[1] $T$ of each configuration. Scenarios characterized by strong conditions of stratification (a large difference between the layers temperature) present a short fundamental period and a consequent strong resistance to the wind action, while scenarios with weak stratification are characterized by a larger fundamental period and a lower resistance. The trajectory of the wind speed employed for the ED is shown in the bottom panel of Figure 2.

Table **1**. Values of the two-layer model parameters characterizing the different stratification scenarios.

| - | scenario 1 | scenario 2 | scenario 3 | scenario 4 | scenario 5 |
|---|---|---|---|---|---|
| $h_1$ [m] | 6.00 | 8.00 | 10.00 | 12.00 | 14.00 |
| $h_2$ [m] | 30.00 | 30.00 | 30.00 | 30.00 | 30.00 |
| $T_1$ [°C] | 20.00 | 22.00 | 25.00 | 27.00 | 30.00 |
| $T_2$ [°C] | 18.00 | 19.00 | 20.00 | 21.00 | 22.00 |
| $T$ [h] | 38.50 | 27.05 | 18.47 | 15.35 | 12.19 |

The simulation horizon is of 28 days and the result of the ED is a data-set of five vectors (one for each scenario) containing the simulated maximum thermocline displacement. An example of the ED results is given in Figure 2, where the trajectories of the wind action and of the corresponding thermocline displacement for scenarios 3 and 4 are reported (for the first 14 days of simulation). Note that in case of stronger stratification (e.g. scenario 4) the thermocline displacement has lower amplitude and higher frequency. Moreover, note that

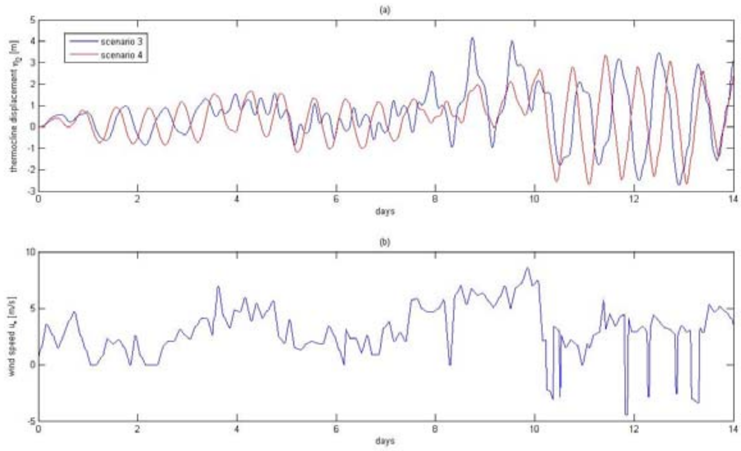


**Figure 2. (a)** Trajectories of the thermocline displacement for scenario 3 and 4 computed by the two-layer model for the first 14 days of simulation. (b) Trajectories of the wind speed.

1 The fundamental period is defined as $T = \dfrac{2 \cdot L}{\sqrt{g' \cdot \dfrac{h_1 \cdot h_2}{h_1 + h_2}}}$, where $g'$ is the reduced gravity, namely $g' = g \cdot \dfrac{\rho_2 - \rho_1}{\rho_2}$. The layers densities $\rho_1$ and $\rho_2$ are computed as functions of the layers temperature $T_1$ and $T_2$.

both oscillation trajectories are characterized not only by a fundamental mode (corresponding to the fundamental period), but also by other modes of smaller amplitude and shorter period.

### 3.3. Choice of the emulation model type

To approximate the thermocline displacement computed by the two-layer model, the following Auto-Regressive eXogenous (ARX) model was used

$$\eta_{2,t} = \sum_{i=1}^{n_a} a_i \cdot \eta_{2,t-i} + \sum_{j=1}^{n_b} b_j \cdot u_{*,t-j+1}^2 \tag{10}$$

where $\eta_{2,t}$ is the thermocline displacement at time $t$, $u_{*,t}^2$ is the square of the wind speed at time $t$, $a_i$ and $b_j$ are the emulation model parameters to be determined, and $n_a$ and $n_b$ represent the order of the ARX model. By introducing the backshift operator ($z^{-1} \cdot \eta_{2,t} = \eta_{2,t-1}$), eq. (10) can be rewritten as

$$\left(1 - \sum_{i=1}^{n_a} a_i \cdot z^{-i}\right) \cdot \eta_{2,t} = \sum_{j=1}^{n_b} b_j z^{-j} u_{*,t}^2 \tag{11a}$$

and finally in the Transfer Function (TF) form

$$\eta_{2,t} = \frac{B(z^{-j})}{A(z^{-i})} \cdot u_{*,t}^2 \tag{11b}$$

where $A(z^{-i}) = 1 - \sum_{i=1}^{n_a} a_i \cdot z^{-i}$ and $B(z^{-j}) = \sum_{j=1}^{n_b} b_j \cdot z^{-j}$. The analysis of the TF will provide a physically meaningful interpretation of the emulation model, thus maintaining a relation with the two-layer model, that would be precluded with other types of models (e.g. neural networks). We will return to this issue in Section 4.

### 3.4. Calibration and validation

The emulation model was calibrated based on synthetic data generated via simulation of the two-layer model (Section 3.2). The first 14 days of the data set was used for the parameters estimation while the latter 14 days for the model validation. Five emulation models of the form (10) were identified, one for each stratification scenario. For each model, the order $[n_a, n_b]$ was chosen by trial-and-error, while the parameters were estimated with the Refined Instrumental Variable method (Young, 1984). As shown in Table 2, the number $n_a$ of auto-regressive terms strongly influences the emulation model performances: if only one auto-regressive term is used, the model cannot be identified or it has very low

Table **2**. Coefficient of determination $R_T^2$ over the calibration data-set for different orders of model (10).

| $n_a$ | $n_b$ | scenario 1 | scenario 2 | scenario 3 | scenario 4 | scenario 5 |
|---|---|---|---|---|---|---|
| 2 | 3 | 0.927 | 0.989 | 0.949 | 0.994 | 0.981 |
| 2 | 2 | 0.911 | 0.985 | 0.941 | 0.991 | 0.977 |
| 2 | 1 | 0.914 | 0.985 | 0.941 | 0.991 | 0.977 |
| 1 | 2 | 0.363 | not identifiable | 0.312 | 0.259 | 0.063 |
| 1 | 1 | 0.359 | 0.215 | 0.127 | not identifiable | 0.113 |

Table **3**. Coefficient of determination $R_T^2$ over the validation data-set for the model (10) of order [2,1].

| $n_a$ | $n_b$ | scenario 1 | scenario 2 | scenario 3 | scenario 4 | scenario 5 |
|---|---|---|---|---|---|---|
| 2 | 1 | 0.983 | 0.928 | 0.951 | 0.987 | 0.987 |

coefficient of determination[2] $R_T^2$, while with two autoregressive terms performances are significantly improved. On the other hand, increasing the number $n_b$ of exogenous terms has no significant effect on the model performances. In the perspective of building the most simple (and effective) models and of analyzing the models parameters for each scenario, the order [2,1] was chosen for all the emulation models.

Despite their simple structure, the emulation models maintain good performances also on the validation data-set (see Table 3). The small differences in the coefficient of determination for different scenarios can be explained via graphical analysis of the trajectories simulated by the two-layer model and the emulation model. For instance, Figure 3 compares the trajectories of the maximum displacement on the validation data-set for scenario 3 and 4. In scenario 3, the fit of the emulation model is smaller (see for examples days 4-6), which reflects into a smaller $R_T^2$ (Table 3). The reason is that the emulation models are second-order regressors and as such they can capture only the oscillations corresponding to the fundamental mode but not the higher frequency oscillations. In other word, the emulation models behave like low-pass filters, whose cut-off frequency corresponds to the inverse of the fundamental period.

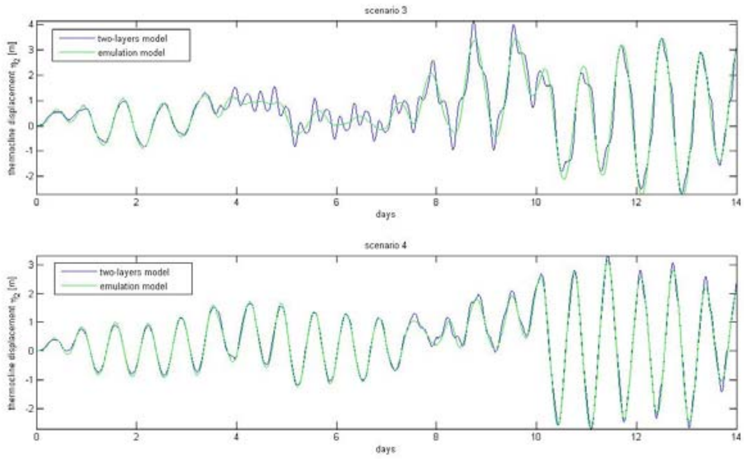


**Figure 3.** Comparison of the thermocline displacement simulated by the two-layer model and the emulation model for scenario 3 and 4 (upper and lower panel respectively).

## 4. ANALYSIS OF THE EMULATION MODEL

Before concluding, it is worthwhile to analyze, from a system identification perspective, the structure of the five emulation models and its relation with the modeled system. The poles of TF models, i.e. the roots of the polynomial $A(z^{-i})$ in (11b), which hold the information about the model stability, are reported in the second column of Table 4. First of all, note that the norm of all poles (column three) tend to one in all scenarios and thus all the emulation models are (marginally) stable. This is consistent with the two-layer model, which assumes that there is

Table **4**. Values of the poles and the corresponding norms of the TF denominator for the five scenarios. The last column shows the value of the exogenous parameter $b_1$.

| scenario | *pole* | *norm* | $b_1$ [s$^2$/m] |
|---|---|---|---|
| 1 | $0.999977 \pm 0.005221i$ | 0.999991 | $5.26 \cdot 10^{-6}$ |
| 2 | $0.999937 \pm 0.007476i$ | 0.999965 | $5.06 \cdot 10^{-6}$ |
| 3 | $0.999908 \pm 0.010953i$ | 0.999968 | $4.85 \cdot 10^{-6}$ |
| 4 | $0.999885 \pm 0.013292i$ | 0.999973 | $4.57 \cdot 10^{-6}$ |
| 5 | $0.999833 \pm 0.016806i$ | 0.999975 | $4.30 \cdot 10^{-6}$ |

[2] The coefficient of determination is defined as $R_T^2 = 1 - \frac{\mathrm{cov}(\overline{\eta}_2 - \eta_2)}{\mathrm{cov}(\overline{\eta}_2)}$, where $\overline{\eta}_2$ is the trajectory of the thermocline displacement simulated by the two-layer model, while $\eta_2$ is the trajectory simulated by the emulation model (10).

no friction at the interface between the two layers and between the lower layer and the bottom. As such, the two-layer model does not assume the dissipation of the energy provided by the wind action, with a resulting (marginally) stable system. Moreover, notice that, when considering stable scenarios, the value of the real part of the poles decreases, while the value of the imaginary part increases. This indicates that the frequency of the system considered by the emulation model becomes higher (and the corresponding period lower), coherently with the system behavior (see the values of the fundamental period in Table 1).

Finally, the last column of Table 4 reports the value of the exogenous parameter $b_1$ (see equation (10)). Its value is larger for larger values of the fundamental period $T$, meaning that the effect of the wind action increases with $T$. This is consistent with the system behavior, since the effect of wind action is stronger when the stratification is weak (and the corresponding fundamental period is large).

## 5. CONCLUSIONS

The paper shows the development of an emulation model to approximate the output of a one-dimensional nonlinear two-layer model, which simulates the wind-induced response of a stratified lake. Five different emulation models were identified in correspondence to different stratification scenarios, each providing the maximum displacement of the thermocline as a function of the squared wind velocity. Despite their simple structure the emulation models show good performances on both the calibration and validation data-set and a physically meaningful interpretation of their parameters can be given. Future research will concentrate on the emulation of more sophisticated hydrodynamics models (e.g. ELCOM), with the ultimate scope of employing emulation models for optimization, thus allowing for the indirect use of process-based models in decision making practice.

## ACKNOWLEDGMENTS

The authors are grateful to Prof. P.C. Young who provided the Captain Toolbox used to develop this work. For further information about the toolbox, see the web site http://www.es.lancs.ac.uk/cres/captain. This work was financially supported by the Centre for Water Research (CWR).

## REFERENCES

Aly, A., and Peralta, R. (1999), Optimal design of aquifer cleanup systems under uncertainty using a neural network and a genetic algorithm. *Water Resources Research*, 35(8), 2523–2532.

Arakawa, A. and Lamb, V. R., (1977), Methods of computational physics, Academic Press, 174-265.

Broad, D., Dandy, G., and Maier, H. (2005), Water distribution system optimization using metamodels. *Journal of Water Resources Planning and Management*, 131(3), 172–180.

Casulli, V., and Cheng, R.T. (1992), Semi-implicit finite difference methods for three-dimensional shallow water flow. *International Journal for Numerical Methods in Fluids*, 15(6), 629–648.

Hodges, B. R., Imberger, J., Saggio, A. and Winters, K.B. (2000), Modeling basin-scale internal waves in a stratified lake, *Limnology and Oceanography*, Vol.45(7), 1603-1620.

Kleijnen, J.P.C. and Sargent, R.G. (2000), A methodology for fitting and validating metamodels in simulation. *European Journal of the Operational Research*, 120(1), 14–29.

Kleijnen, J.P.C., Sanchez, S.M., Lucas, T.W. and Cioppa, T.M. (2005), A user's guide to the brave new world of designing simulation experiments. *INFORMS Journal of Computing*, 17(3), 263–289.

Leonard, B.P., (1991), The ULTIMATE conservative difference scheme applied to unsteady one-dimensional advection. *Computational Method in Applied Mechanical Engineering*, 88, 17–74.

Mellor, G. L. (1996), Users guide for a three-dimensional, primitive equation, numerical ocean model, Princeton University.

Oguz, T., Özsoy, E., Latif, M.A., Sur, H.I. and Ünlüata, Ü. (1990), Modelling of hydraulically controlled exchange flow in the Bosphorus Strait. *Journal of Physical Oceanography*, 20, 945–965.

Simpson, T.W., Peplinski, J.D., Koch, P.N. and Allen, J.K. (2001), Metamodels for computer based engineering design: survey and recommendations. *Engineering with Computers*, 17, 129–150.

Young, P.C., (1984), Recursive estimation and time series analysis. *Springer-Verlag*, Berlin.