*International Environmental Modelling and Software Society (iEMSs)*
*2010 International Congress on Environmental Modelling and Software*
*Modelling for Environment's Sake, Fifth Biennial Meeting, Ottawa, Canada*
*David A. Swayne, Wanhong Yang, A. A. Voinov, A. Rizzoli, T. Filatova (Eds.)*
*http://www.iemss.org/iemss2010/index.php?n=Main.Proceedings*

# A tree-based feature ranking approach to enhance emulation modelling of 3D hydrodynamic-ecological models

**A. Castelletti [a], S. Galelli[a] and R. Soncini-Sessa[a]**

[a]*Dipartimento di Elettronica e Informazione - Politecnico di Milano, P.zza Leonardo da Vinci 32, 20133 Milano, Italy (castelle@elet.polimi.it, galelli@elet.polimi.it, soncini@elet.polimi.it)*

**Abstract:** The paper presents an emulation modelling approach to reduce the computational burden associated with the use of process-based, distributed, dynamic models in planning and management of water resources. This approach, which relies on a procedure composed by few steps, leads to the identification of an emulation model, namely a simplified, computationally-efficient model built over a sample data-set produced via simulation of the original computationally expensive model. The core of the procedure is a feature ranking algorithm, based on Extremely Randomized Trees (Extra-Trees), through which the most relevant input variables to the emulation model are selected among the large set of candidate input variables associated with the original model. The calibration of the emulation model, in the form of an input-output relationship, is then performed using again Extra-Trees. The approach is demonstrated on a real-world case study (Googong Reservoir, AUS).

***Keywords:*** emulation modelling; feature selection; water resources management; hydrodynamic-ecological models

## 1 INTRODUCTION

Increasing human pressure on water bodies and reservoirs is often causing strong quality problems, which can be effectively mitigated by adopting a combination of different rehabilitation technologies (e.g. mixers, curtains, submerged pipes). Choosing the best technological mix and fixing how to manage it on the medium/long run to meet a given set of criteria is a rather difficult task, which involves combining simulation models and optimization techniques. In the water quality sector the former are usually process-based, distributed, dynamic models, whose high dimensional state prevent their inclusion in optimization frameworks even with the most advanced optimization techniques (see Castelletti *et al.* [2010] and references therein).
A potentially interesting approach to reduce the computational burden associated with the use of these models is to build a simplified, computationally efficient model that 'emulates' the relevant dynamics of the original one in a way that is accurate enough for discriminating between different planning/management options. The construction of an emulation model can be articulated in different steps, through which the information provided by the large model is re-elaborated to obtain a lower order, lumped-parameter model, which describes the output required to solve the planning/management problem. Emulation modelling has been widely adopted in mechanical and aerospace engineering (for a survey, see Queipo *et al.* [2005]), but it has only relatively attracted researchers in the context of water resources planning and management, and on a limited number of application fields. Yan and Minsker [2004] use emulation models to groundwater remediation, Broad *et al.* [2005] developed an emulation model of a water distribution network using artificial neural networks, and Galelli and Soncini-Sessa [2010] use emulation modelling to include the dynamics of water demand for irrigation into reservoir operation.

In this paper we present a procedure to identify, in a 'data-driven fashion', a dynamic emulation model of a 3D coupled hydrodynamic-ecological model (ELCOM-CAEDYM) for lakes and reservoirs. The core mechanism of the procedure is a feature ranking algorithm, based on Extremely Randomized Trees (Extra-Trees; Geurts *et al.* [2006]), through which the input variables to the emulation model are automatically selected. ELCOM-CAEDYM is used to generate, via simulation, synthetic I/O time-series of the variables of interest. Based on these I/O data, the relevant input variables are selected with the feature ranking algorithm and the emulation model, in the form of an I/O relationship, is then identified using Extra-Trees. The approach is demonstrated on a real world case study (Googong Reservoir, AUS).

## 2 EMULATION MODELLING

As the emulation model has to be used for the resolution of planning/management problems, it is assumed that the design objectives defining the problem and the associated step indicators (see Soncini-Sessa *et al.* [2007] for more details) have already been defined. The step indicators are the natural output of the emulation model, since they represent the most general input to any optimal planning/management algorithm. Once these outputs are fixed, the purpose of the emulation modelling exercise is clearly stated: to construct a simplified, computationally efficient model that accurately computes the step indicators. The model can be identified by employing the following procedure, composed of four steps.

Step 1. *Design Of Experiments (DOE)*. The Design Of Experiments (DOE) is the sampling plan in the space of the inputs (decisions and disturbances) of the process-based (PB) model, when this is non-dynamic; while it is the sampling plan in the space of the input trajectories in the opposite case. Since the number of samples is generally limited by the computational requirements of each simulation, proper techniques should be employed to select the samples as to effectively explore the state space of the PB model. DOE can be carried out either using statistical techniques (see, for instance, Queipo *et al.*, [2005]) or based on physical considerations and a-priori knowledge (Galelli and Soncini-Sessa [2010] and references therein). While the former are well established for steady-state planning problems (Queipo *et al.*, [2005]), the latter are, up to now, the only available solution in management problems.

Step 2. *Simulation of the PB model*. For each sample of the inputs (decisions and disturbances) as obtained in Step 1, the outputs (step indicators) and the state trajectories must be computed via simulation of the computationally expensive PB model, thus obtaining the data-set of tuples (inputs, states and outputs) for the subsequent identification of the emulation model.

Step 3. *Input and state variables selection*. Of the whole set of input and state variables of the PB model, the most informative subset with respect to the emulation model output has to be selected (spatial aggregation of the input and state variables can be performed prior to the input/state selection). This operation can be done on the base of a priori knowledge or, when the number of candidate input and state variables is too large, by adopting proper automatic techniques. Generally, considering the complexity and the underlying non-linearity of the PB model, it is advisable to leave simple statistical techniques (e.g. cross-correlation, which is only able to detect linear dependence between two variables; see Bowden *et al.* [2005]) and resort to more sophisticated tools, as feature selection algorithms commonly adopted for machine learning tasks (see Das [2001]). An example is given in Section 3.2, where the feature ranking algorithm adopted in this work is described. For each state variable included in the selected subset, the input and state variables selection must be repeated (assuming as new output the selected state variable), until only disturbances and decisions remain in the subset.

Step 4. *Model selection, calibration and validation*. This is a traditional model identification problem, composed of model selection, parameter estimation (calibration) and validation. As far as the class to which the emulation model will belong is considered, it is advisable to employ the same class of model adopted at Step 3 when selecting the input variables to the emulation model (in the present application, for example, Extra-Trees are adopted to build the emulation model,

since the feature ranking algorithm is based on Extra-Trees). Parameter estimation and validation can be finally performed by adopting suitable calibration/validation algorithms. A cascade of emulation models is identified if more than one state variable is selected at Step 3.

## 3 TREE-BASED MODELLING AND FEATURE RANKING

Tree-based methods are non-parametric supervised learning methods (see Breiman *et al.* [1984]) that can provide modelling flexibility, computational efficiency, interpretability and good accuracy in both regression and classification problems. They are all based on the idea of decision tree models, which are tree-like structures representing a cascade of rules leading to numerical (or class) values. These structures, composed of decision nodes, branches and leaves, are obtained by first partitioning at the top decision node (or root node), with a proper splitting criterion, the set of the input variables into two sub-sets (first branch); then repeating the splitting process recursively on each derived sub-set, until either the numerical (or class) values belonging to a sub-set vary just slightly or only few elements remain (termination test). When the splitting process is over, the branches represent the hierarchical structure of the sub-set partitions, while the leaves are the finest sub-sets associated to the terminal branches. Each leaf is finally associated with a numerical (or class) value.

### 3.1 Extra-Trees

In this study we explore a new tree-based method called Extremely Randomized Trees (Geurts *et al.* [2006]) that randomizes (totally or partially) both the input variable and the cut-point selection when splitting a node, and create an ensemble of trees to compensate for the randomization, via averaging of the constituent tree outcomes. Extra-Trees were empirically demonstrated (Geurts *et al.* [2006]) to outperform other deterministic (CART) and randomized (Tree Bagging, Random Forests; Breiman [2001]) methods in terms of both prediction accuracy and computational efficiency. Moreover, the particular structure of Extra-Trees can be exploited to infer the relative importance of the input variables in explaining the output and then identify the most relevant variables (features ranking) among a candidate input variables set. The Extra-Trees building algorithm grows ensemble of $M$ trees. Nodes are split using the following rule: $K$ alternative cut-directions (input variables) are randomly selected and, for each one, a random cut-point is chosen; a score (explained variance) is then associated to each cut-direction and the one maximizing the score is adopted to split the node. The algorithm stops partitioning a node if its cardinality is smaller than $n_{min}$ (termination test) and the node is therefore a leaf. To each leaf a value is assigned, obtained as the average of the outputs associated to the inputs that fall in that leaf. The estimates produced by the $M$ trees are finally aggregated with arithmetic average (aggregation rule). The rationale behind the approach is that the combined used of randomization and ensemble averaging provide more effective variance reduction than other randomization methods, while minimizing the bias of the final estimate. The values of the three parameters $k$, $n_{min}$ and $M$ associated to Extra-Trees can be fixed on the basis of empirical evaluations:

- $K$, the number of alternative cut-directions, can be chosen in the interval $[1, \ldots, n]$, where $n$ is the number of inputs. When $K$ is equal to $n$, the choice of the cut-direction is not randomized and the randomization acts only through the choice of the cut-point. On the contrary, low values of $K$ increase the randomization of the trees and weaken the dependence of their structure on the output of the training data-set. Geurts *et al.* [2006] have empirically demonstrated that, for regression problems, the optimal default value for $K$ is $n$.
- $n_{min}$, the minimum cardinality for splitting a node. Large values of $n_{min}$ lead to small trees (few leaves), with high bias and small variance. Conversely, low values of $n_{min}$ lead to fully-grown trees, which may over-fit the data. The optimal value of $n_{min}$ depends not only on the risk aversion to over-fitting, but also on the level of noise in the outputs of the training data-set: the noisier are the outputs, the higher should be the optimal value of $n_{min}$. As for $K$, Geurts *et al.* [2006] have empirically shown that, although possibly

slightly suboptimal, the default value of $n_{min}$ = 5 appear to be robust choices in a broad range of typical conditions in regression problems.

- $M$, the number of trees in the forest, influences the strength of the variance reduction and the behavior of the estimation error, which is a decreasing function of $M$ (Breiman [2001]). The estimation accuracy thus increases with $M$ and the choice of its value depends on a trade-off between the desired model accuracy and available computing power.

### 3.2 Feature ranking

A part from providing good performance in terms of bias-variance reduction, the particular structure of Extra-Trees can be exploited to rank the importance of the $n$ input variables in explaining the output behavior and then identify the most relevant variables among $n$ candidate inputs. This approach, as proposed by Fonteneau *et al.* [2008], is based on the idea of scoring each input variable by estimating the variance reduction it can be associated with by propagating the training data-set over the $M$ different tree structures composing the ensemble. More precisely, let us consider a regression problem with an output variable $y$, $n$ input variables $\{x_1, x_2, \ldots, x_n\}$ and a training data-set $S$, composed of $N$ input-output observations. The relevance $G(x_i)$ of each input variable $x_i$ in explaining the output $y$ can be evaluated as follow

$$G(x_i) = \frac{\sum_{\tau=1}^{M} \sum_{j=1}^{\Omega} \delta(\nu_j, x_i) \cdot \Delta_{var}(\nu_j)|S|}{\sum_{\tau=1}^{M} \sum_{\nu_j=1}^{\Omega} \Delta_{var}(\nu_j)|S|} \tag{1}$$

where $\nu_j$ is the $j$-th non-terminal node in the tree $\tau$, $\Omega$ is the number of non-terminal nodes in the tree $\tau$, $\delta(\nu_j, x_i)$ is equal to 1 if $x_i$ is used to split the node $\nu_j$ (and 0 otherwise), $|S|$ is the number of samples in the considered sub-set $S$, $\Delta_{var}(\nu_j)$ is the variance reduction when splitting the node $\nu_j$, namely

$$\Delta_{var}(\nu_j) = \text{var}\{y|S\} - \frac{|S_{i,l}|}{|S|}\text{var}\{y|S_{i,l}\} - \frac{|S_{i,r}|}{|S|}\text{var}\{y|S_{i,r}\} \tag{2}$$

where the terms $S_{i,l}$ and $S_{i,r}$ are the two sub-sets of $S$ satisfying the conditions $x_i < s_i$ and $x_i \geq s_i$ respectively. The input variables are finally sorted by decreasing values of their importance.

## 4 CASE STUDY

### 4.1 Googong reservoir and its management problem

Googong Reservoir is located in New South Wales and is one of the five sources supplying Canberra's water. The reservoir has a full-supply volume of 121GL (1.21 x $10^6$ $m^3$), a surface area at full supply of approximately 3.5 $km^2$, an average and maximum depth of 35 and 50 m respectively. Its waters are mainly used for potable water supply but the reservoir is also used for recreational purposes. The reservoir has a history of low to medium levels of Cyanobacteria and metal release (particularly of Manganese causing stain on clothes). Destratification was thought as a suitable way to solve the problem. This technique involves increasing rates of vertical mixing via mechanical means, with the objective of improving dissolved oxygen concentration at depth, which in-turn reduces the likelihood of nutrient and metal release from the sediments under anoxic conditions. Two pairs of 5 m diameter WEARS (brand) surface-mounted mixers have been installed for that purpose in March 2007. However, this intervention had only limited success as the Cyanobacteria increased and metal release is still significant, which suggests that in-depth oxygenation could be not enough.

### 4.2 The management problem

As demonstrated in a recent study (Castelletti *et al.* [2010]), the reallocation of existing mixers and the installation of new ones could improve the current solution. However, the problem was

Table 1: ELCOM-CAEDYM input and state variables giving a contribution larger than 4% in explaining $\Delta Mn_t^{2+,B}$ behavior.

| variable | $\mathbf{Fe}_{t-1}^{3+,B}$ | $Mn_{t-1}^{2+,B}$ | $\mathbf{T}_{t-1}^{E}$ | sol. rad. | env. flow |
|---|---|---|---|---|---|
| u. of meas. | [mg/L] | [mg/L] | [°C] | [W/m$^2$] | [m$^3$/s] |
| ind. score | 36.20 % | 5.39 % | 4.99 % | 4.80 % | 4.30 % |

addressed considering a fixed thrust, while it is interesting to explore if any advantage exists in varying the mixers' thrust in time (i.e. in adopting a control policy). For this purpose, an optimal control problem has to be formulated and solved, based on the step indicator $g_t^{Mn}$ (considering the methodological focus of this paper, just the case of manganese is issued) that is equal to the average concentration of Manganese in the benthic area $Mn_t^{2+,B}$ [mg/L] (temporal mean over the time step $[t-1,t]$). As the number of alternative policies is infinite, the problem can not be solved with a simulation-based exhaustive approach, and optimal control algorithms should be adopted. However, the use of such algorithms is precluded by the infinite dimensionality of the PB model state. To overcome this difficulty, an emulation model must be identified and calibrated on the data obtained via simulation of the 3D coupled hydrodynamic-ecological ELCOM-CAEDYM of Googong reservoir.

### 4.3 The process-based model

The Estuary, Lake and Coastal Ocean Model (ELCOM) is a three-dimensional hydrodynamics model used for predicting the velocity, temperature and salinity distribution in natural water bodies subjected to external environmental forcing such as wind stress and surface fluxes. The Computational Aquatic Ecosystem Dynamics Model (CAEDYM) consists of a series of mathematical equations representing the major biogeochemical processes influencing water quality, including primary production, secondary production, nutrient and metal cycling, and oxygen dynamics and the movement of sediment. For the purpose of producing sample data for the identification of the emulation model, the coupled ELCOM-CAEDYM model was run using a simulation step of 1 minute with a 60 x 60 m grid bathymetry with 1 m vertical grid resolution. This gives a total of about 59000 computational cells, which, for the 24 state variables/cell, gives a total of 1.42 million data points produced for every time step.

## 5 IDENTIFICATION OF THE EMULATION MODEL

The practical resolution of the optimal control problem (see Section 4.2) requires the identification of a Multi Input-Single Output (MISO) emulation model, with multiple inputs and the step indicator $g_t^{Mn}$ as output. Based on the definition of the step indicator, the concentration $Mn_t^{2+,B}$ is assumed as output of the MISO emulation model. The target is to reduce as much as possible the number of state variables involved in the PB sub-model concerning manganese in the sub-domain of the reservoir where the effect of mixers is more evident.

### 5.1 DOE and simulation of the PB model

To explore the behaviour of the reservoir under different management conditions, 10 samples of input trajectories were defined by combining five control policies with two disturbance scenarios (the realization of meteorological processes in 2002/2003 and 2004/2005). The control policies considered are: (1) mixers always switched off, (2) mixers always switched on and (3-5) a step-wise variation in time of the mixers condition (on or off) (the steps, which are 3 or more months long, are defined on the base of physical considerations).

### 5.2 Input and state variables selection

The PB model state variables were aggregated in space by considering their spatial mean over three layers defined on the basis of physical considerations: (B) the benthic layer, defined as the volume of water in the reservoir near the bottom, subject to anaerobic and anoxic processes leading to metals and nutrients releases; (E) the euphotic layer, defined as the volume reached by at least 1% of the incoming light (i.e. where the photosynthetic processes take place); (M) the middle layer, which is the remaining volume in the water body. A mean over 72 hours was considered for the aggregation in time. This post-processing of the PB simulations leads to a total of 87 variables to be considered when selecting the input variables to the emulation model: 72 are ELCOM-CAEDYM state variables (24 for each layer) and 15 are ELCOM-CAEDYM input variables (meteo and mixers' thrust). This number is still large and the most informative subset can not be identified simply relying on physical considerations. For this reason, the feature ranking algorithm based on Extra-Trees (see Section 3.2) was employed. Following Geurts *et al.* [2006], $K$ and $n_{min}$ were posed equal to 87 and 5 respectively, while for $M$ a value of 100 was adopted. As for the output to be considered when running the ranking algorithm, we did not consider $Mn_t^{2+,B}$, but $\Delta Mn_t^{2+,B}$, defined as $Mn_t^{2+,B} - Mn_{t-1}^{2+,B}$, to account for the high inertia characterizing the system's dynamics. In order to select the most informative variables, different criteria can be adopted to analyze the ranking results, e.g. choosing the variables ranked in the first $p$ positions (with $p$ equal to a pre-defined value), selecting the variables contributing to reach a (pre-defined) cumulated explained variance or the variables giving an individual score larger than a (pre-defined) value. In the present application the last criterion was adopted, employing a threshold on the individual score equal to 4 %.
The ranking of the variables giving a contribution larger than 4 % is reported in Table 1. The dynamics of $\Delta Mn_t^{2+,B}$ is mainly dominated by the concentration of iron (positive ion 3+) in the benthic layer ($Fe_{t-1}^{3+,B}$), the concentration of Manganese itself ($Mn_{t-1}^{2+,B}$) in the benthic layer, the euphotic temperature ($T_{t-1}^{E}$), the solar radiation (sol. rad.) and the release from the dam for environmental protection (env. flow.). Among these variables (and a part from the auto-regressive term $Mn_{t-1}^{2+,B}$) $Fe_{t-1}^{3+,B}$ and $T_{t-1}^{E}$ are ELCOM-CAEDYM state variables (bold symbols in Table 1), whose dynamics must thus be described. The ranking algorithm must thus be re-applied assuming as new output variables $\Delta Fe_t^{3+,B}$ and $\Delta T_t^E$. The ranking of the variables with a contribution larger than 4 % in describing $\Delta Fe_t^{3+,B}$ and $\Delta T_t^E$ behavior is shown in Table 2 (first two panels): the former variable, a part from the auto-regressive term, depends on the concentration of iron (positive ion 2+) in the benthic layer ($Fe_{t-1}^{2+,B}$) and the temperature in the euphotic and middle layer ($T_{t-1}^{E}$, $T_{t-1}^{M}$). This result requires to describe also the dynamics of $\Delta Fe_t^{2+,B}$ and $\Delta T_t^M$. As for $\Delta T_t^E$, it depends on many ELCOM-CAEDYM input variables and on $Fe_{t-1}^{2+,B}$ and $Mn_{t-1}^{2+,B}$. The ranking of $\Delta Fe_t^{2+,B}$ and $\Delta T_t^M$ (see the two lower panels of Table 2) permits to close the ranking exercise, since both $\Delta Fe_t^{2+,B}$ and $\Delta T_t^M$ depend on ELCOM-CAEDYM input variables and on state variables whose dynamics was already accounted during the previous ranking runs. At the end of this Step, 5 state variables and 8 input variables from the initial subset of 87 ELCOM-CAEDYM input/state variables were thus selected.

### 5.3 Model calibration and validation

To give consistency to the emulation modelling procedure, an ensemble of Extra-Trees was selected to model $\Delta Mn_t^{2+,B}$, $\Delta Fe_t^{3+,B}$, $\Delta Fe_t^{2+,B}$, $\Delta T_t^E$ and $\Delta T_t^M$. As for the Extra-Trees parameters, the same setting adopted during the ranking was adopted ($M$ = 100, $n_{min}$ = 5, $K$ = number of input). As for the input variables to the five Extra-Trees to be calibrated, the results obtained at the previous Step were slightly modified, on the basis of physical considerations, in order to avoid the presence of circular dependencies (e.g. $\Delta T_t^E$ depends on $Mn_{t-1}^{2+,B}$ and $\Delta Mn_t^{2+,B}$ depends on $T_{t-1}^{E}$). The final structure of the model is thus a cascade of emulation models (see Eqs. (3) that was calibrated and validated with a $k$-fold cross-validation (with $k$ = 10) by employing the 10 different simulations produced at Steps 1 and 2 (see Section 5.1).

$$\hat{T}_t^E = T_{t-1}^E + \hat{\Delta T}_t^E(\text{sol. rad., wind speed, wind dir., month, Qinfl, Tinfl}) \tag{3a}$$

Table 2: ELCOM-CAEDYM input and state variables giving a contribution larger than 4% in explaining $\Delta Fe_t^{3+,B}$, $\Delta T_t^E$, $\Delta Fe_t^{2+,B}$, $\Delta T_t^M$ behavior.

| variable | $Fe_{t-1}^{3+,B}$ | $\mathbf{Fe}_{t-1}^{2+,B}$ | $\mathbf{T}_{t-1}^{M}$ | $\mathbf{T}_{t-1}^{E}$ |
|---|---|---|---|---|
| u. of meas. | [mg/L] | [mg/L] | [°C] | [°C] |
| ind. score | 22.4 % | 7.42 % | 5.67 % | 4.36 % |

| variable | sol. rad. | $\mathbf{Fe}_{t-1}^{2+,B}$ | wind speed | $\mathbf{Mn}_{t-1}^{2+,B}$ | wind dir. | month | Qinfl | Tinfl |
|---|---|---|---|---|---|---|---|---|
| u. of meas. | [W/m$^2$] | [mg/L] | [m/s] | [mg/L] | [°] | [-] | [m$^3$/s] | [°C] |
| ind. score | 9.37 % | 8.71 % | 7.52 % | 6.95 % | 5.38 % | 4.91 % | 4.73 % | 4.12 % |

| variable | $\mathbf{Fe}_{t-1}^{3+,B}$ | $\mathbf{Mn}_{t-1}^{2+,B}$ | $Fe_{t-1}^{2+,B}$ | $\mathbf{T}_{t-1}^{M}$ | $\mathbf{T}_{t-1}^{E}$ |
|---|---|---|---|---|---|
| u. of meas. | [mg/L] | [mg/L | [mg/L | [°C] | [°C] |
| ind. score | 29.30 % | 9.66 % | 8.45 % | 6.02 % | 5.19 % |

| variable | $\mathbf{Fe}_{t-1}^{2+,B}$ | $\mathbf{Mn}_{t-1}^{2+,B}$ | mixer | month | wind speed | wind dir. | sol. rad. |
|---|---|---|---|---|---|---|---|
| u. of meas. | [mg/L] | [mg/L] | kW/h | [-] | [m/s] | [°C] | [w/m$^2$] |
| ind. score | 15.60 % | 11.70 % | 4.91 % | 4.66 % | 4.11 % | 4.09 % | 4.04 % |

$$\hat{T}_t^M = T_{t-1}^M + \hat{\Delta T}_t^M(\text{mixer, month, wind speed, wind dir., sol. rad.}) \tag{3b}$$

$$\hat{Fe}_t^{2+,B} = Fe_{t-1}^{2+,B} + \Delta\hat{Fe}_t^{2+,B}(Fe_{t-1}^{2+,B}, T_t^M, T_t^E) \tag{3c}$$

$$\hat{Fe}_t^{3+,B} = Fe_{t-1}^{3+,B} + \Delta\hat{Fe}_t^{3+,B}(Fe_{t-1}^{3+,B}, Fe_{t-1}^{2+,B}, T_t^M, T_t^E) \tag{3d}$$

$$\hat{Mn}_t^{2+,B} = Mn_{t-1}^{2+,B} + \Delta\hat{Mn}_t^{2+,B}(Fe_{t-1}^{3+,B}, Mn_{t-1}^{2+,B}, T_t^E, \text{sol. rad., env. flow}) \tag{3e}$$

The whole model shows good performances both in one-step ahead prediction and simulation (see Fig. 1), with a coefficient of determination $R^2$ equal to 0.998 in prediction and 0.904 in simulation for $Mn_t^{2+,B}$.

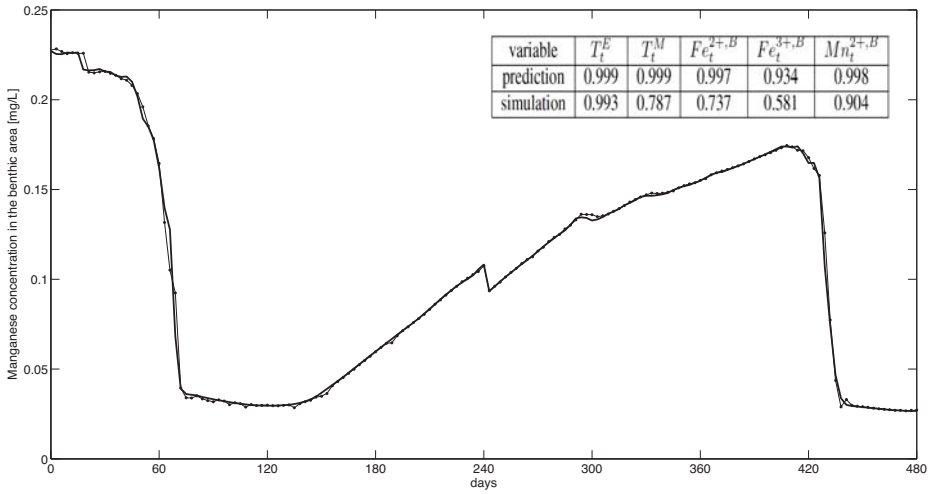


Figure 1: A specimen of the trajectories for the Manganese concentration in the benthic area $Mn_t^{2+,B}$ computed by ELCOM-CAEDYM (solid line) and the emulation model (one-step ahead prediction, dashed line). The figure also reports the coefficient of determination $R^2$ [-] in one-step ahead prediction and simulation (10-fold cross-validation) for the five state models of Eqs. (3).

## 6 CONCLUSIONS

The paper presents a four step procedure for the identification of a dynamic emulation model to reduce the complexity of process-based, distributed, dynamic models commonly adopted to describe the quality conditions of water bodies. The procedure, which relies on a feature ranking algorithm for the selection of the most informative variables, is tested on a real-world case study, namely to predict/simulate the Manganese concentration in Googong reservoir (Australia). The emulation model performances are quite satisfactory, and the complexity reduction with respect to the original model (ELCOM-CAEDYM) is remarkable (from 1.42 milion to 5 state variables). This suggests the approach can be successfully applied to any environmental management problem involving large process-based models, for which 'what-if' analysis over a very small number of alternative decisions was the only feasible way of supporting decision-making.
Further research will concentrate on the development of a more rigorous DOE on a larger data-set, on the analysis of the different criteria available to manage the ranking algorithm results and on a more accurate analysis of the circular dependencies identified by the ranking algorithm. Research effort will finally be devoted to very scope of this emulation modelling exercise, namely the design of an optimal control policy for the mixers.

## REFERENCES

Bowden, G., G. Dandy, and H. Maier. Input determination for neural network models in water resources applications. Part 1 - background and methodology. *Journal of Hydrology*, 301:75–92, 2005.

Breiman, L. Random forests. *Machine Learning*, 45(1):5–32, 2001.

Breiman, L., J. Friedman, R. Olsen, and C. Stone. *Classification and regression trees.* Wadsworth & Brooks, Pacific Grove, CA, 1984.

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

Castelletti, A., F. Pianosi, R. Soncini-Sessa, and J. Antenucci. A multi-objective response surface approach for improved water quality planning in lakes and reservoirs. *Water Resources Research*, 2010.

Das, S. Filters, wrappers and a boosting-based hybrid for feature selection. In *Proceedings of the Eighteenth International Conference on Machine Learning, 28 June - 1 July*, Williamstown, US, 2001.

Fonteneau, R., L. Wehenkel, and D. Ernst. Variable selection for dynamic treatment regimes: a reinforcement learning approach. In *European Workshop on Reinforcement Learning, 30 June - 4 July*, Villeneuve d'Ascq, FR, 2008.

Galelli, S. and R. Soncini-Sessa. Combining metamodelling and stochastic dynamic programming for the design of reservoirs release policies. *Environmental Modelling & Software*, 25 (2):209–222, 2010.

Geurts, P., D. Ernst, and L. Wehenkel. Extremely randomized trees. *Machine Learning*, 63(1): 3–42, 2006.

Queipo, N., R. Haftka, W. Shyy, T. Goel, R. Vaidyanathan, and P. Tucker. Surrogate-based analysis and optimization. *Progress in Aerospace Sciences*, 41:1–28, 2005.

Soncini-Sessa, R., A. Castelletti, and E. Weber. *Integrated and participatory water resources management. Theory*. Elsevier, Amsterdam, NL, 2007.

Yan, S. and B. Minsker. A dynamic meta-model approach to genetic algorithm solution of a risk-based groundwater remediation design model. In *Proc. of the 2004 World Water & Environmental Resources Congress*, 2004.