19th International Congress on Modelling and Simulation, Perth, Australia, 12–16 December 2011
http://mssanz.org.au/modsim2011

Input variable selection for ecological modelling in inter-basin water transfer management

Roberta Fornarelli[a], Stefano Galelli[b], Jason P. Antenucci[c] and Andrea Castelletti[d]

[a] *Centre for Water Research, University of Western Australia, M023, 35 Stirling Hwy, Crawley 6009, Western Australia, Australia*

[b] *Singapore-Delft Water Alliance, National University of Singapore,21 Lower Kent Ridge Road, Singapore 119077*

[c] *Hatch Associates, 144 Stirling St, Perth, Western Australia, Australia*

[d] *Dipartimento di Elettronica e Informazione, Politecnico di Milano, Piazza Leonardo da Vinci, 32, 20133 Milano, Italy*

Email: fornarel@cwr.uwa.edu.au

Abstract: Inter-basin water transfers, usually driven by purely economic purposes can have complex physical and biological implications both in the upstream and downstream reservoirs. These transfers are well studied from a water quantity management point of view, but increasing pressure on water resources is encouraging decision-makers to move towards a more holistic management strategy, accounting for both quantity and quality issues. The integration of high fidelity water quality modelling within an optimization based decision-making framework calls for new mathematical tools that balance computational efficiency and accuracy. In particular, input variable selection techniques represent the first step towards low-order emulation of physically based model by selecting, among all the possible candidate inputs, the ones that better reproduce the behavior of a specified variable. In this study, a novel data-driven approach to input variable selection is applied to a 9-year dataset of phytoplankton data measured in the receiving end of two interconnected reservoirs. Preliminary analyses show that diatom growth in the receiving reservoir is influenced not only by internal processes (e.g. net growth as function of nutrients concentrations), but also by the input of algal cells from the upstream reservoir (e.g. seeding effect of particular algal species) via water transfers. To evaluate the relative importance of all the candidate input variables in explaining the output behavior, we resorted to the Iterative Input Selection (IIS) algorithm. The algorithm, which is based on Extremely Randomized Trees (Extra-Trees) as the underlying model class, incrementally builds the set of variables by adding the most significant ones according to a ranking procedure, based on a statistical measure of significance that accounts for non-linear dependencies. The IIS algorithm stops selecting new variables when the desired accuracy is achieved. The capability of IIS in selecting the most relevant input variables is shown to be superior to common methods based on cross-correlation analyses. This study represents the first application of the IIS algorithm to the interpretation of ecological data.

Keywords: *Water transfers, diatom biovolume, extra-trees, input variable selection*

1. INTRODUCTION

Phytoplankton biovolume is a fundamental variable to characterize the ecosystem status in lakes and reservoirs in terms of eutrophication trends and algal blooms. Phytoplankton biovolume is directly connected to phytoplankton growth, thus depends mainly on nutrient concentrations, water temperature and light availability. However, in a reservoir system, where water is transferred between two reservoirs, other variables can play a pivotal role in determining phytoplankton biovolume, such as nutrient and phytoplankton cell flux from the source reservoir. Inter-basin water transfers, usually managed according to water quantity interests (e.g. drinking water supply and hydropower production), have complex biological implications both in the upstream and downstream reservoir mainly due to physical and biological differences between the connected systems and to the magnitude, frequency and duration of transfers (Soulsby et al., 1999, Gibbins et al., 2000). Within this context, determining the main drivers of the observed phytoplankton biovolume is a fundamental step towards a comprehensive understanding, and subsequent modelling and management, of the reservoir system.

The selection of a reduced amount of input variables out of a larger set of candidate variables is a fundamental exercise to eventually explain a given output. Different approaches to input variable selection are found in literature ranging from expert based sorting, linear cross-correlation, heuristic techniques, sensitivity analysis, and information-theoretic measures (Bowden et al., 2005, and references therein). Lately, in the context of input variable selection in ecological studies, particular attention is given to data-driven models based on machine learning techniques, e.g. artificial neural network, genetic programming, tree-based algorithms (Kim et al., 2007, Muttil and Chau, 2007).

In this study, we considered a 9-year dataset of diatom biovolume data measured in the receiving end of two interconnected reservoirs. Diatoms were selected out of the overall phytoplankton community as they were the dominant group and were particularly linked to the management of water transfers. To evaluate the relative importance of different candidate input variables in explaining the diatom biovolume (i.e. the output), we resorted to the Iterative Input Selection (IIS) algorithm (Castelletti et al., 2010). The algorithm, which uses Extremely Randomized Trees (Extra-Trees) as the underlying model class, incrementally builds the set of variables by adding the most significant ones according to a ranking procedure, based on a statistical measure of significance that accounts for non-linear dependencies. This automatic input variable selection procedure probably outperforms implicit input selection as it can handle short time-series of data with potential redundant information (Castelletti et al., 2010). Further, Extra-Trees is particularly suitable for solving variable selection problems, but primarily represents a powerful tool in characterizing the strong non-linearities of ecological processes providing great flexibility, computational efficiency, and accuracy.

The objective of this study was therefore twofold: i) to identify and rank the importance of some selected variables that best explain the observed time trend of diatom biovolume (i.e. the output variable); ii) to test the suitability of IIS and Extra-Trees in characterizing the non-linear ecological processes, thus outperforming in variable selection. The IIS algorithm, tested for the first time on ecological data in the present study, represents the first step towards the implementation of simplified models, e.g. emulation models, to be eventually coupled with optimization techniques for water quality and water quantity integrated optimal management.

2. METHODS

2.1. Iterative Input Selection

Input variable selection is an initial and critical step of any modelling exercise. The selection of a reduced number of appropriate and significant input variables among a larger set of candidates to explain a given output might result in both a more accurate and compact model of the underlying process. The ideal selection algorithm should account for non-linear dependencies and redundancy between inputs, as hydrodynamic and ecological processes are usually characterized by non-linear dynamic modes with multiple coupled variables. Moreover, such algorithm must be computationally efficient, since the number of candidate input variables is generally large.

In this paper, we used a novel data-driven approach called Iterative Input Selection (IIS) that incrementally builds the set of selected variables by adding the best variable provided by a ranking algorithm, based on statistical measure of significance. This measure is computed using an underlying model that accounts for non-linear dependencies among variables. At the first iteration the ranking algorithm is run on a data set

composed of time series of all the candidate input variables and the associated output values. At the subsequent iterations, the original output values are replaced by the residuals of the underlying model built at the previous step. The reason behind this choice is that, once a variable is selected, all the variables that are highly correlated with that input may become useless and the ranking needs to be re-evaluated. The IIS algorithm terminates when the accuracy of the model built upon the selected variables, as evaluated with a *k*-fold cross-validation, starts decreasing. The two building blocks of the IIS algorithm are the ranking procedure and the underlying model. As for the latter, we use Extremely Randomized Trees (Extra-Trees), a tree-based regression method that provides a number of desirable features: modelling flexibility (i.e. the ability of accurately modelling strongly non-linear functions), scalability with respect to the input dimensionality (i.e. the ability of handling several input variables with a different range of variability), and computational efficiency. Moreover, as proposed in Fonteneau et al. (2008), Extra-Trees can also be used as a ranking procedure, since their particular structure and building algorithm can be exploited to infer the relative importance of the different candidate variables.

2.2. Extremely Randomized Trees

Extremely Randomized Trees (Extra-Trees, Geurts et al., 2006) are tree-like methods whose peculiarity is the (total or partial) randomization of both the input variable and the cut-point selection when splitting a node, in the process of building a tree, and the creation of an ensemble of trees to compensate for the randomization, via averaging of the constituent tree outcomes. The Extra-Trees building algorithm grows ensembles 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 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 falling in that leaf. The estimates produced by the M trees are finally aggregated with arithmetic average (aggregation rule). The rationale behind this approach is that the combined use of randomization and ensemble averaging provides more effective variance reduction than other randomized methods, while minimizing the bias of the final estimate. The values of the three hyper-parameters K, n_{min}, and M associated to Extra-Trees can be fixed on the basis of empirical evaluations (Geurts et al., 2006): K is set equal to number of candidate input variables, M is generally higher than 100 in order to reduce the effect of randomization and n_{min} usually ranges between 5 to 100 depending on the size of the data set.

2.3. Variable ranking

Apart 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 input variables in explaining the output behavior and then identify the most relevant variables among the candidate inputs. This approach 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. For further details on the methodology, please refer to Castelletti et al. (2010).

3. CASE STUDY

3.1. Description of the system and data availability

Fitzroy Falls Reservoir (Fig. 1) is a small, shallow reservoir, part of the Shoalhaven System (Sydney, Australia). It is artificially connected to Lake Yarrunga (Fig. 1), from which receives pumped water via inter-basin water transfers for water supply and hydropower generation. Bendeela pondage is a small reservoir, sited between Lake Yarrunga and Fitzroy Falls to aid hydropower generation: its effect on the water quality of Fitzroy Falls was neglected in this study because of its low retention time (2 days). The pumped inflow from Lake Yarrunga is Fitzroy Falls' biggest inflow: when water transfers between the two reservoirs are active, the natural catchment inflow of Fitzroy Falls is less than 10 % of the total inflow. Note that, in the following analysis, Fitzroy Falls was referred to also as the "receiving reservoir" and Lake Yarrunga as the "upstream reservoir".

Daily inflows and outflows in Fitzroy Falls were available from January 2001 to February 2010. Nutrient concentrations (total nitrogen, ammonia, oxidized nitrogen, total phosphorus, and dissolved silica) and water temperature at different depths were available at monthly intervals in one location in Fitzroy Falls close to

the dam wall (Fig. 1) from January 2001 to February 2010. Surface weekly diatom biovolume data were available at the same location. The same dataset was measured for the same period in Lake Yarrunga at a monitoring station 1 km downstream of the pumping station (Fig. 1), and was used to represent nutrient and diatom concentrations transported in Fitzroy Falls by the water transfers. All measurements were conducted using APHA standard protocols (American Public Health Association, 2005).

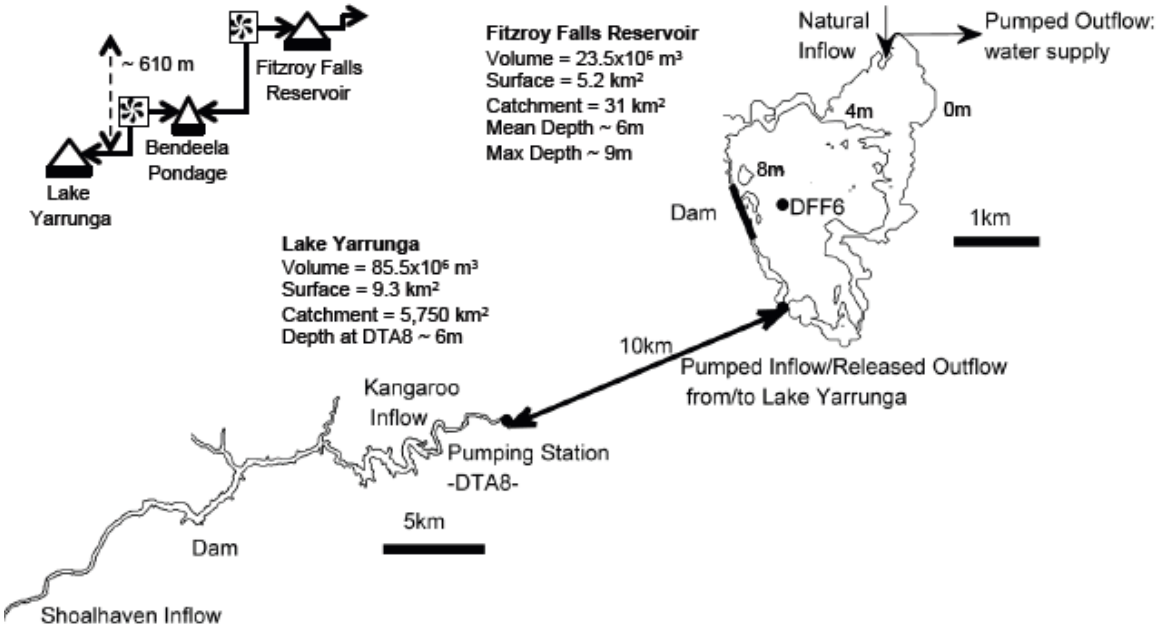

Figure 1. Lake Yarrunga and Fitzroy Falls Reservoir connection. Historical data were available at two monitoring stations (one in Lake Yarrunga, DTA8, and one in Fitzroy Falls, DFF6). Fitzroy Falls contours at 0, 4, 8 m depth. Elevation difference: 630 m; distance between the two reservoirs: 10 km. Figure not to scale.

3.2. Implementation of Iterative Input Selection

Depth average values of water temperature and nutrient concentrations were used in the two reservoirs, as they are both shallow and no persistent stratification was evident (note that we are considering only the portion of Lake Yarrunga close to the pumping station, Fig. 1). Monthly averages of nutrient concentrations, diatom biovolume and pumped flow rate were used in the following analysis. Nutrient (and diatom) fluxes towards Fitzroy Falls were calculated by multiplying the pumped flow rate by the nutrient concentrations (and diatom biovolume) measured in Lake Yarrunga.

Table 1 lists all the candidate inputs potentially useful in explaining the measured diatom biovolume in Fitzroy Falls at the time instant t (monthly time step). Beyond the variables explaining diatom growth in the reservoir (nutrient concentrations, water temperature, previous values of diatoms itself), the fluxes of nutrients and diatoms from the upstream to the receiving reservoir were added to the list of candidate inputs. The pumped flow rate was also added as a possible input because, from preliminary analyses, a strong correlation between transferred flow rate and diatom biovolume in Fitzroy Falls was found (Table 1). A cumulative pumped flow rate was calculated by summing up the flow rates at the time lag zero, one and two, and included in as a candidate input.

Note that the IIS algorithm is a data-driven approach applied on an initial set of candidate input variables that, in turn, were a-priori chosen (Table 1). This choice was based on preliminary screening of all the possible candidate input variables and was based on data availability, expert judgment, and results of cross-correlation analysis (for the choice of time lags). Therefore the initial set of candidate input variables (all the variables in Table 1) was empirically defined, while the final set of selected variables was determined in a fully data-driven way.

The hyper-parameters of the Extra-Tree model included in the IIS algorithm were set up on the basis of a trial-and-error procedure: the minimum cardinality for splitting a node, n_{min}, was set to 5 while the number of trees, M, was set to 500. The number of alternative cut-directions, K, was set equal to the number of inputs (Geurts et al., 2006). The input selection was conducted on the whole dataset of 110 observations using a 3-fold cross-validation procedure and two slightly different sets of candidate inputs. In Experiment 1, all the

variables listed in Table 1 were accounted in the variable selection exercise with the proper time lag, for a total of 39 candidate input variables. In Experiment 2, the pumped flow rate was excluded from the set of the candidate variables, therefore 33 input variables were considered. The choice of considering till time lag 2 followed the outcomes of cross-correlation analyses (Table 1).

3.3. Results and Discussion

Cross-correlation analysis (Table 1) shows that only some of the inputs were linearly correlated to the diatom biovolume at different time lags, while non-linear correlations might exist with the other inputs. Note that, despite a significant correlation with diatom biovolume in Fitzroy Falls existed till lag eight, only lag one, two and three were considered to explain the output behavior because, after lag three, the correlation coefficients dropped to values very close to the confidence bound.

Table 1. List of all the candidate input variables. Time lag and correlation coefficients as result of cross-correlation analysis. LY = Lake Yarrunga; FF = Fitzroy Falls; NS = not significant correlation. Dissolved inorganic nitrogen calculated as the sum of ammonia and oxidized nitrogen. Variables highlighted by the grey panel are the ones manually selected following the results of cross-correlation analysis.

Inputs Variables	Description	Unit	Time Lag - month -	Correlation Coefficient
Diat_FF	Diatom biovolume in FF	$mm^3\ L^{-1}$	1	0.6
Diat_Flux	Diatom flux from LY to FF	$mm^3\ d^{-1}$	0	0.5
TP_Flux	Total phosphorus flux from LY to FF	$kg\ d^{-1}$	2	0.45
DIN_Flux	Dissolved inorganic nitrogen flux from LY to FF	$kg\ d^{-1}$	2	0.4
TN_Flux	Total nitrogen flux from LY to FF	$kg\ d^{-1}$	2	0.4
Diat_FF	Diatom biovolume in FF	$mm^3\ L^{-1}$	2	0.4
Qcum	Cumulative pumped flow rate from LY to FF	ML	0	0.3
SiO_FF	Dissolved silica in FF	$mg\ L^{-1}$	2	0.3
SiO_Flux	Dissolved silica flux from LY to FF	$kg\ d^{-1}$	1	0.3
DIN_Flux	Dissolved inorganic nitrogen flux from LY to FF	$kg\ d^{-1}$	1	0.3
Diat_FF	Diatom biovolume in FF	$mm^3\ L^{-1}$	3	0.3
Q	Pumped flow rate from LY to FF	$ML\ d^{-1}$	1	0.3
Q	Pumped flow rate from LY to FF	$ML\ d^{-1}$	2	0.3
Qcum	Cumulative pumped flow rate from LY to FF	ML	1	0.3
Qcum	Cumulative pumped flow rate from LY to FF	ML	2	0.2
Q	Pumped flow rate from LY to FF	$ML\ d^{-1}$	0	0.2
SiO_FF	Dissolved silica in FF	$mg\ L^{-1}$	1	0.2
TN_Flux	Total nitrogen flux from LY to FF	$kg\ d^{-1}$	1	0.2
Diat_Flux	Diatom flux from LY to FF	$mm^3\ d^{-1}$	1 - 2	NS
TN_FF	Total nitrogen in FF	$mg\ L^{-1}$	0 - 1 - 2	NS
DIN_FF	Dissolved inorganic nitrogen in FF	$mg\ L^{-1}$	0 - 1 - 2	NS
TP_FF	Total phosphorus in FF	$mg\ L^{-1}$	0 - 1 - 2	NS
SiO_FF	Dissolved silica in FF	$mg\ L^{-1}$	0	NS
Temp_FF	Temperature in FF	^oC	0 - 1 - 2	NS
TN_Flux	Total nitrogen flux from LY to FF	$kg\ d^{-1}$	0	NS
DIN_Flux	Dissolved inorganic nitrogen flux from LY to FF	$kg\ d^{-1}$	0	NS
TP_Flux	Total phosphorus flux from LY to FF	$kg\ d^{-1}$	0 - 1	NS
SiO_Flux	Dissolved silica flux from LY to FF	$kg\ d^{-1}$	0 - 2	NS

Figure 2a shows the results of the IIS exercise in Experiment 1. By selecting 8 inputs out of 39, an R^2 of 0.61 was reached and the corresponding trajectories are shown in Figure 2c. The most relevant variable is the cumulative pumped flow rate that explains by itself more than half of the final result ($R^2 = 0.36$), followed by the flux of diatoms at lag zero ($R^2 = 0.16$). A linear, positive correlation between diatom biovolume and pumped flow rate was already identified by cross-correlation analysis (see Table 1). The way the pumped flow rate can positively affect diatom biovolume is related to the physical consequences caused by the inflow: higher turbulence/mixing and lower retention time in the reservoir due to inflow events are two factors leading to increase diatom growth and dominance over other phytoplankton groups (Reynolds, 2006). However, in our specific case, we are more interested in what the pumped flow rate could actually bring in the reservoir (i.e. nutrient and diatom cells), therefore Experiment 2 didn't consider flow rate as a candidate input variable. Interestingly, the same performance was reached ($R^2 = 0.61$, Fig. 2b and c) as in Experiment 1 by considering 10 inputs out of 33: diatom, silica and total phosphorus fluxes became the most important

variables in explaining the diatom biovolume in Fitzroy Falls (Fig. 2b). Silica and diatom fluxes were equally important (R^2 = 0.18 and 0.19 respectively), followed by the flux of phosphorus, the nutrient and diatom concentrations in Fitzroy Falls. Diatom flux influenced the diatom biovolume at lag 0, representing a direct effect of water transfers: within the same month, the amount of diatoms that was pumped in from Lake Yarrunga became part of the actual measured biovolume in Fitzroy Falls. On the contrary, nutrient fluxes (dissolved silica and phosphorus) exerted their control on diatom biovolume with a time lag of 1 and 2 months, respectively, thus modifying the internal processes of diatom growth. The relevance of the phosphorus loads in explaining diatom growth is in accordance with the phosphorus-limiting conditions of the reservoir (ratio of total nitrogen on total phosphorus higher than 20).

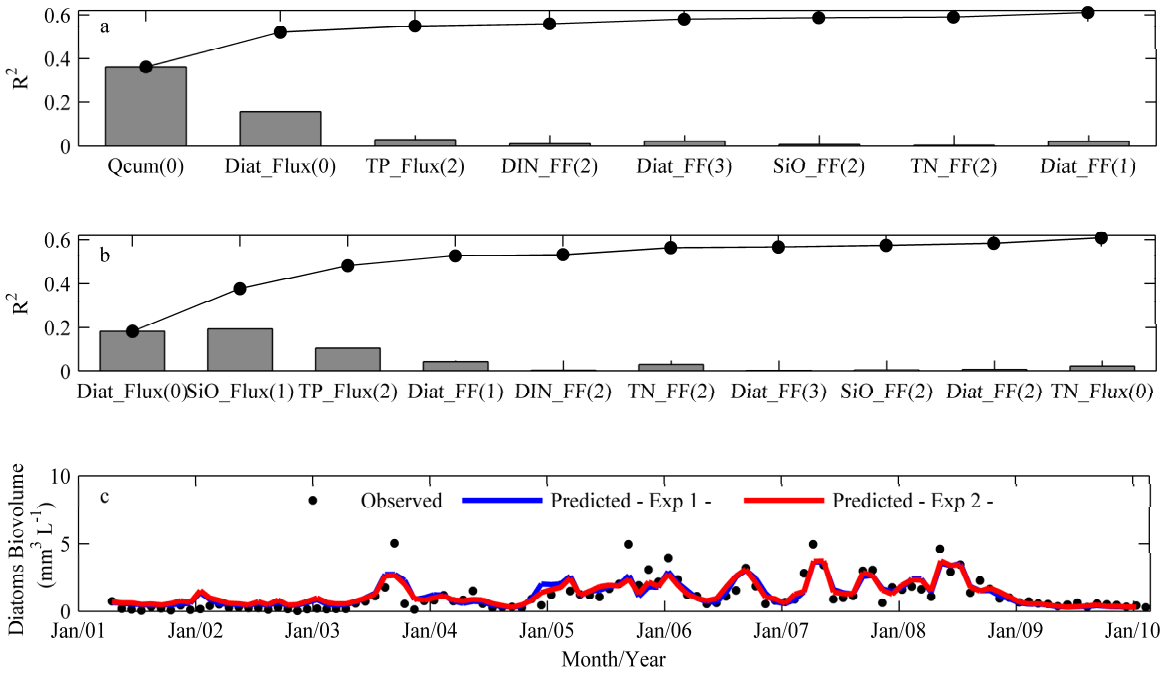

Figure 2. Input variables selected in the first (a) and second (b) experiment. Numbers in parenthesis indicate the time lag: (0) lag zero, (1) lag one, (2) lag two. Trajectories of observed and predicted diatom biovolume in Fitzroy Falls using the selected input variables (c).

These results demonstrate that the diatom biovolume was dependant primarily on the fluxes of nutrient and diatom cells transported from the upstream reservoir via the water transfers, and secondarily on the nutrient concentrations and internal dynamics in the reservoir. This represents a remarkable result in term of the management implications in reservoir systems: it is fundamental to evaluate and quantify the effects of seeding water coming from an upstream reservoir and take them into considerations when designing optimal control policies for water quality management.

The variable selection based on IIS was compared with a variable selection based on the results of cross-correlation analysis (Table 2). The first 11 variables with the highest correlation coefficient were selected (Table 1, variables highlighted) and simulated via Extra-Trees regression. Note that a number of inputs similar to the one derived from IIS algorithm was manually selected and the cumulative pumped flow rate at lag zero was chosen as a representative for the flow rate. Together with R^2, other performance indices were calculated: the "index of agreement" defined in Willmott (1982) is bounded between 0 and 1 and it can be suitable to make cross-comparisons between

Table 2. Comparison between cross-correlation based variable selection and iterative input selection. R^2 (-) = coefficient of determination; MAE (mm^3 L^{-1}) = mean absolute error; RMSE (mm^3 L^{-1}) = root mean squared error; D (-) = index of agreement.

Variable selection method	R^2	MAE	RMSE	D
Cross-correlation	0.54	0.45	0.67	0.88
Iterative Input selection	0.60	0.43	0.61	0.90

models (Willmott, 1982). All the performance indices (Table 2) show that IIS outperform the variable selection based on cross-correlation results (R^2 of 0.60 for IIS against 0.54 of the variable selection based on cross-correlation analysis).

4. CONCLUSIONS

This paper applied a novel data-driven tree-based input variable selection approach to the interpretation of diatom biovolume data measured in the receiving end of two interconnected reservoirs. Two main conclusions can be drawn from this study. i) In reservoir systems, it is fundamental to consider the connection between the reservoirs since internal dynamics can be not sufficient in explaining the ecological processes occurring in the system. ii) The Extra-Tree based IIS algorithm resulted to be successful in detecting and quantifying the importance of some specific variables in explaining the output behavior. The outcome and future applications of this analysis represent the first step towards the implementation of simplified models, e.g. emulation models, more convenient than process-based models to be coupled with optimization techniques for water quality and water quantity integrated optimal management. Further research will concentrate on the efficiency of Extra-Trees in understanding non-linear processes, when compared to other artificial intelligence techniques, e.g. artificial neural networks, genetic programming.

ACKNOWLEDGMENTS

The first author was the recipient of a Scholarship for International Research Fees from the University of Western Australia and of a University International Stipend from the Centre for Water Research. We gratefully acknowledge the support of the Sydney Catchment Authority for providing access to the data. The findings, opinions and conclusions expressed herein are those of the authors and do not represent the views or opinions of the Sydney Catchment Authority or any person employed in the Sydney Catchment Authority Division. This article represents Centre for Water Research reference 2485-RF.

REFERENCES

American Public Health Association (2005). Standard Methods for the Examination of Water and Wastewater, 21st edn. American Public Health Association (APHA), American Water Works Association (AWWA) & Water Environment Federation (WEF).

Bowden, G.V., Dandy, G.C., Maier, H.R., (2005). Input determination for neural network models in water resources applications. Part 1 – Background and methodology. *Journal of Hydrology*, 301, 75–92.

Castelletti, A., Galelli, S., Salvetti, A., Ventimiglia, A., (2010). Extremely randomized trees and feature ranking for streamflow modeling. In: *9th International Conference on Hydroinformatics*, HIC 2010, Tianjin, China.

Fonteneau, R., Wehenkel, L., Ernst, D., (2008). Variable selection for dynamic treatment regimes: a reinforcement learning approach. In: *Proceedings of the European Workshop on Reinforcement Learning*, Villeneuve d'Ascq, France.

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

Gibbins, C.N., Jeffries, M.J., Soulsby, C., (2000). Impacts of an inter-basin water transfer: distribution and abundance of Micronecta poweri (Insecta: Corixidae) in the River Wear, north-east England. *Aquatic Conservation: Marine and Freshwater Ecosystem*, 10, 103-115.

Kim, D.K., Jeong, K.S., Whigham, P.A., Joo, G.J., (2007). Winter diatom blooms in a regulated river in South Korea: explanations based on evolutionary computation. *Freshwater Biology*, 52, 2021-2041.

Muttil, N., Chau, K.W., (2007). Machine-learning paradigms for selecting ecologically significant input variables. *Engineering Application of Artificial Intelligence*, 20, 735-744.

Reynolds, C.S., (2006). The ecology of phytoplankton. Cambridge University Press.

Soulsby, C., Gibbins, C.N., Robins, T., (1999). Inter-basin water transfers and drought management in the Kielder/Derwent system. *Journal of the Chartered Institution of Water and Environmental Management*, 13, 213-223.

Willmott, C.J., (1982). Some comments on the evaluation of model performance. *Bulletin of American Meteorological Society*, 63, 1309-1313.