International Environmental Modelling and Software Society (iEMSs)
7th Intl. Congress on Env. Modelling and Software, San Diego, California, USA
Daniel P. Ames, Nigel W. T. Quinn, Andrea E. Rizzoli (Eds.)
http://www.iemss.org/society/index.php/iemss-2014-proceedings

# A new evaluation framework for input variable selection algorithms used in environmental modelling

**Greer B. Humphrey[a], Stefano Galelli[b], Andrea Castelletti[c], Holger R. Maier[a], Graeme C. Dandy[a] and Matthew S. Gibbs[a]**

[a] *School of Civil, Environmental, and Mining Engineering, University of Adelaide, SA 5005, Australia (greer.humphrey@adelaide.edu.au, holger.maier@adelaide.edu.au, graeme.dandy@adelaide.edu.au, matthew.gibbs@adelaide.edu.au)*

[b] *Pillar of Engineering Systems and Design, Singapore University of Technology and Design, 20 Dover Drive, 138684, Singapore (stefano_galelli@sutd.edu.sg)*

[c] *Department of Electronics, Information, and Bioengineering, Politecnico di Milano, Piazza L. da Vinci, 32, 20133, Milan, Italy (andrea.castelletti@polimi.it)*

**Abstract:** Input variable selection is an essential step in the development of statistical models and is particularly relevant in environmental modelling, where potential model inputs often consist of time lagged values of each different potential input variable. While new methods for identifying important model inputs continue to emerge, each has its own advantages and limitations and no method is best suited to all datasets and purposes. Nevertheless, rigorous evaluation of new and existing input variable selection methods, is largely neglected due to the lack of guidelines or precedent to facilitate consistent and standardised assessment. This rigorous evaluation would allow the effectiveness of these algorithms to be properly identified in various circumstances. In this paper, we propose a new framework for the evaluation of input variable selection methods which takes into account a wide range of dataset properties that are relevant to real world environmental data and assessment criteria selected to highlight algorithm suitability in different situations of interest. The framework is supported by a repository of datasets to enable standardised and statistically significant testing. It is hoped that this framework helps to promote the appropriate application and comparison of input variable selection algorithms and eventually serves to provide guidance as to which algorithm is most suitable in a given situation.

***Keywords:*** Input variable selection; Data-driven modelling; Environmental datasets

## 1 INTRODUCTION

In data-driven modelling, determining which inputs are most useful for predicting a variable of interest can be one of the most critical decisions in the model development process. The input variables contain the information necessary for defining, albeit, in a simplified manner, the underlying process that generated the data. However, the set of candidate inputs usually also includes variables which might be either irrelevant to the problem or redundant. Irrelevant input variables are uninformative about the underlying process and only serve to add noise and complexity into the model, while the inclusion of redundant, but relevant, inputs increases the dimensionality of the model identification problem without providing any additional predictive benefit. The omission of relevant input variables, on the other hand, leads to an inaccurate model, where part of the output behaviour remains unexplained by the selected input variables. Thus, the appropriate selection of both relevant and non-redundant

inputs can mean the difference between a reliable and parsimonious model, which generalises well to the underlying process, and a model that produces nonsensical outputs, is slower to run, and more difficult to interpret [Guyon and Elisseeff, 2003]. The challenge of Input Variable Selection (IVS) is, therefore, to select the fewest number of input variables that best characterise the underlying input-output relationship while minimising variable redundancy [May et al., 2011].

While the task of IVS is not unique to environmental modelling, it can be a particularly difficult one when it comes to environmental systems, since many of the underlying processes are often poorly understood. Furthermore, as environmental systems vary in space and time, potentially important inputs may include observations of causal variables at different locations and time lags, as well as lagged observations of the same dependent variable of interest [Maier and Dandy, 2000]. As a result, the number of potentially important inputs can be very large. To further complicate matters, the correlated nature of such input variables induces redundancy in the input pool, while the non-linearity and inherent complexity associated with environmental systems make it ineffective to apply well established analytical variable selection methods, such as correlation analysis [May et al., 2011]. As such, the development and adaptation of IVS methods for environmental modelling applications is an important and active field of research, which has further been stimulated by reviews of environmental modelling procedures discussing the need for improved and more conscientious IVS (see, for example, Maier et al. [2010]).

However, despite recent efforts to improve IVS in environmental modelling, studies in this field tend to draw overly general conclusions about the performance of the IVS approaches used. They are usually conducted with a single focus (e.g. to select the inputs for a particular case study of interest) and the evaluation of IVS methods is summarised accordingly (e.g. based on the predictive performance of the resulting model). These generally impaired evaluations make it difficult to determine how the performance of one IVS method, either new or existing, compares to that of another, and, ultimately, do not provide any useful information to drive the final user in selecting the method that is most appropriate for the problem at hand. This present lack of rigour in the assessment of IVS methods has been somewhat unavoidable given the absence of guidelines or precedents to facilitate consistent and standardised assessment. To fill this methodological gap, we introduce in this paper a generic framework for the standardised and rigorous comparative analysis of IVS algorithms applicable to environmental modelling datasets. The framework consists of three main components: (1) a set of benchmark data, (2) a recommended set of evaluation criteria, and (3) a website for sharing data and results. The datasets have been synthetically generated to have, to different degrees, the typical properties of real environmental data, while the evaluation criteria have been designed to quantitatively assess selection accuracy. Two IVS algorithms commonly adopted in environmental modelling exercises and representative of different IVS approaches are comparatively analysed.

## 2 IVS EVALUATION FRAMEWORK

### 2.1 Benchmark datasets

A total of 26 synthetic datasets, summarised in Table 1, were generated for benchmarking the performance of IVS algorithms. These datasets exhibit, to different degrees, the following properties: non-linearity in the underlying function, collinearity amongst the input variables, non-Gaussian input/output variables, noise in the output, incomplete input information and interdependence of the input variables. All of these properties are considered to reflect the features of real environmental data that might be most useful for assessing the relative performance of individual IVS algorithms. Furthermore, the benchmark datasets were generated to have different sample sizes and dimensionalities to enable an investigation of the sensitivities of IVS methods to the relative proportion of irrelevant candidate inputs and of the abilities of IVS methods to identify important input-output relationships within datasets of varying lengths. In Table 1, sample size is denoted by $N$, $K$ is the number of relevant inputs and $P$ is the total number of candidate inputs. The $P - K$ candidate inputs which are included in the datasets but contain no (or only redundant) information about the outputs are primarily lagged values of the true inputs or inputs drawn from distributions resembling those of the true inputs. The ratio $N/P$ is also

given in Table 1, as this value is indicative of the risk of retaining irrelevant or redundant inputs. This risk increases with increasing correlation between the candidate inputs.

While synthetic data may be considered somewhat unrealistic and lacking in substance, their use for IVS algorithm benchmarking is necessary since such data provide the only means for adequately assessing the performance of IVS algorithms using quantitative approaches. Firstly, and most importantly, the use of synthetic data enables selected inputs to be compared to the known set of "true" input variables. This allows 'selection accuracy' to be evaluated without relying on prediction accuracy, which can be complicated by a number of factors including the choice of model, calibration method, error model and calibration criteria, among others. Secondly, with synthetic data, it is relatively easy to systematically vary features such as those listed above in order to achieve a balanced design for the comprehensive evaluation of IVS techniques. With real data, this would be far more difficult and would rely on methods for quantifying the above properties without knowledge of the true underlying function. Finally, the use of synthetic data enables previously unanalysed datasets to be included in the benchmark set, whereas evaluating IVS methods on previously unanalysed real datasets would provide very limited information about algorithm performance. However, in order to ensure that the true characteristics of real environmental data were captured in the benchmark datasets at least to some extent, several of the benchmark sets are only partially synthetic, where the input data are real, while only the outputs are modelled. Whether a benchmark dataset is fully or partially synthetic is also indicated in Table 1. To account for any variability in algorithm performance that may result from variability in the data, 30 replicates of each benchmark dataset are created. This enables the statistical significance of comparison results to be considered.

### 2.2 Evaluation criteria

The most commonly used measure for assessing the selection accuracy of IVS methods is predictive performance on an independent validation data set. However, prediction accuracy is influenced not only by the inputs selected, but also by other factors, such as the choice of model, the calibration method and the experience of the modeller, for example. Furthermore, if a wrapper or embedded IVS method is used, the choice of model would be obvious; however, this may not be the case when evaluating filter approaches. As the relevant or true inputs are known for all of the benchmark datasets, there are more objective metrics of selection accuracy that can be employed.

A selection accuracy ($SA$) score which expresses the degree to which a selected input subset matches the true input subset is recommended for use in this framework. The proposed $SA$ score is based on the similarity score proposed by Molina et al. [2002], but unlike the original version, it makes no distinction between irrelevant and redundant inputs and simply treats all unnecessary inputs as extraneous. The proposed $SA$ score is given as follows:

$$SA = \gamma \frac{k}{K} + (1-\gamma)\left(1 - \frac{p}{P-K}\right) \tag{1}$$

where $K$ is the total number of relevant inputs; $k$ is the number of relevant inputs selected; $p$ is the number of extraneous (irrelevant or redundant) inputs selected; $P$ is the total number of inputs in the candidate input pool and $\gamma$ is a weight ranging from 0 to 1, which influences the penalty applied to the selection of extraneous inputs in relation to the gain achieved from each correctly selected input. This score ranges from 0 to 1, where $SA = 1$ corresponds to a correctly specified model, while $SA = 0$ corresponds to a completely misspecified model with no relevant inputs and all extraneous inputs selected. An advantage of this score is that information about the degree to which a model has been correctly or incorrectly specified is combined into a single metric, which makes for the straightforward comparison of IVS algorithm selection accuracy. As far as the value of $\gamma$ is concerned, all the results reported in this study are obtained with $\gamma = 0.7$. The rationale behind this choice is that this value reflects the fact that choosing an extraneous input is usually better than missing a relevant one.

While values of $SA < 1$ denote over- or under-specification, a limitation of the $SA$ score is that it does not adequately indicate where the selected input subset is deficient; for example, whether too many or

**Table 1.** Benchmark dataset properties

| | Dataset | $N$ | $K$ | $P$ | $N/P$ | Fully/Partially Synthetic | Non-Gaussian Output | Highly Nonlinear | High Noise | High Collinearity | Inter-dependency | Incomplete Information |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | AR1 | 500 | 1 | 15 | 33.3 | Fully | | | X | X | | |
| 2 | AR9_500 | 500 | 3 | 15 | 33.3 | Fully | | | X | X | | |
| 3 | AR9_70 | 70 | 3 | 15 | 4.7 | Fully | | | X | X | | |
| 4 | TAR1 | 500 | 1 | 15 | 33.3 | Fully | | | X | X | | |
| 5 | TAR2 | 500 | 2 | 15 | 33.3 | Fully | | | X | X | | |
| 6 | NL_500 | 500 | 3 | 15 | 33.3 | Fully | X | X | | | | |
| 7 | NL_70 | 70 | 3 | 15 | 4.7 | Fully | X | X | | | | |
| 8 | NL2 | 500 | 3 | 15 | 33.3 | Fully | X | X | X | X | | |
| 9 | Bank_fm | 400 | 8 | 32 | 12.5 | Fully | X | | | | | X |
| 10 | Bank_fh | 400 | 8 | 32 | 12.5 | Fully | X | | X | | | X |
| 11 | Bank_nm | 400 | 8 | 32 | 12.5 | Fully | X | X | | | | X |
| 12 | Bank_nh | 400 | 8 | 32 | 12.5 | Fully | X | X | X | | | X |
| 13 | Friedman_c0_10_m | 250 | 5 | 10 | 25 | Fully | | X | | | | |
| 14 | Friedman_c0_10_h | 250 | 5 | 10 | 25 | Fully | | X | X | | | |
| 15 | Friedman_c0_50_m | 250 | 5 | 50 | 5 | Fully | | X | | | | |
| 16 | Friedman_c0_50_h | 250 | 5 | 50 | 5 | Fully | | X | X | | | |
| 17 | Friedman_c25_10_m | 250 | 5 | 10 | 25 | Fully | | X | | X | | |
| 18 | Friedman_c25_10_h | 250 | 5 | 10 | 25 | Fully | | X | X | X | | |
| 19 | Salinity_5_l | 4120 | 3 | 80 | 51.5 | Partially | | | | X | | |
| 20 | Salinity_5_m | 4120 | 3 | 80 | 51.5 | Partially | | | | X | | |
| 21 | Salinity_5_h | 4120 | 3 | 80 | 51.5 | Partially | | | X | X | | |
| 22 | Salinity_10_l | 4115 | 3 | 160 | 25.7 | Partially | | | | X | | |
| 23 | Salinity_10_m | 4115 | 3 | 160 | 25.7 | Partially | | | | X | | |
| 24 | Salinity_10_h | 4115 | 3 | 160 | 25.7 | Partially | | | X | X | | |
| 25 | Kentucky | 4739 | 4 | 21 | 225.7 | Partially | X | | | X | | |
| 26 | Miller | 200 | 2 | 3 | 66.7 | Fully | X | | | | X | |

too few inputs have been selected. To overcome this limitation, the SA score given by eq. (1) can be broken into two sub-scores:

$$SA_c = \frac{k}{K} \tag{2}$$

$$SA_e = 1 - \frac{p}{P - K} \tag{3}$$

where $SA_c$ indicates the proportion of correct inputs that have been selected and $SA_e$ is based on the proportion of extraneous inputs that have been selected. Unlike the overall $SA$ score given by eq. (1), these sub-scores do not trade off one measure of accuracy against another; therefore, they do not require the $\gamma$ parameter. Both of these terms can range from 0 to 1, where a value closer to 1 denotes a better model. A value of $SA_c = 1$ together with a value of $SA_e < 1$ implies over-specification, while a value of $SA_c < 1$ indicates under-specification. The advantage of these scores is that they express the degree to which a model is over- or under-specified, which is important for differentiating between IVS algorithm results.

## 3 EXPERIMENTAL SETUP

The proposed IVS evaluation framework was applied for the evaluation and comparison of two IVS algorithms. The aim here is not to provide a definitive answer as to which of the algorithms has the best overall performance, but rather to demonstrate application of the proposed framework and how the results obtained may be used for evaluating and gaining greater insight into algorithm performance. The two IVS algorithms are the Partial Mutual Information (PMI) algorithm, developed by Sharma [2000] and later modified by Bowden et al. [2005] and May et al. [2008], and the Iterative Input variable Selection (IIS) algorithm, introduced by Galelli and Castelletti [2013]. Both of these algorithms are filter, forward selection algorithms that have been applied to environmental IVS problems previously. Each algorithm is run on 30 replicates of the 26 benchmark datasets, for a total of 780 runs per algorithm, with the scores value computed as the average over the 30 replicates.

## 4 RESULTS

The results obtained for the $SA$ score are summarised in Figure 1, where they can be analyzed by comparing either the performance of a specific algorithm on the different datasets or the performance of both algorithms on a specific dataset. The former analysis allows understanding of how different dataset properties impact on algorithm behaviour, while the latter is aimed at assessing algorithm performance under the same modelling conditions. Since the aim of this paper is to demonstrate the application of the proposed framework, rather than discussing the specific performance of the considered algorithms, we will mostly focus on the first analysis.

Figure 1 reveals that the PMIS and IIS algorithms share a similar range of variation of $SA$, which varies from 1 (corresponding to correctly specified models) to about 0.4. These extreme cases are represented by AR1 and Miller datasets: both algorithms are capable of selecting the only relevant variable in the first dataset without choosing any other extraneous input, while in the second dataset the overall performance decreases. The analysis of $SA$ intra-dataset variation shows that both algorithms have variable performances depending on the dataset properties. For instance, the AR1 and AR9_500 datasets are characterised by high noise and high collinearity, but the presence of few relevant inputs and the high $N/P$ ratio make the input selection exercise solvable with both algorithms. This condition can be totally affected by the variation of a single property, such as the ratio $N/P$ in the AR9_70 dataset, which differs from the AR9_500 dataset in the number of observations (70 instead of 500). The results in Figure 1 show that such variation is particularly affecting the IIS algorithm, with PMIS showing better performance. The extreme case of intra-dataset variation is represented by the Miller dataset, which is characterised by a strong inter-dependency between the two relevant inputs. In this case, the IIS algorithms is still capable of achieving good performance (with $SA$ equal to about 0.7), while decreasing values of $SA$ are found for PMIS.

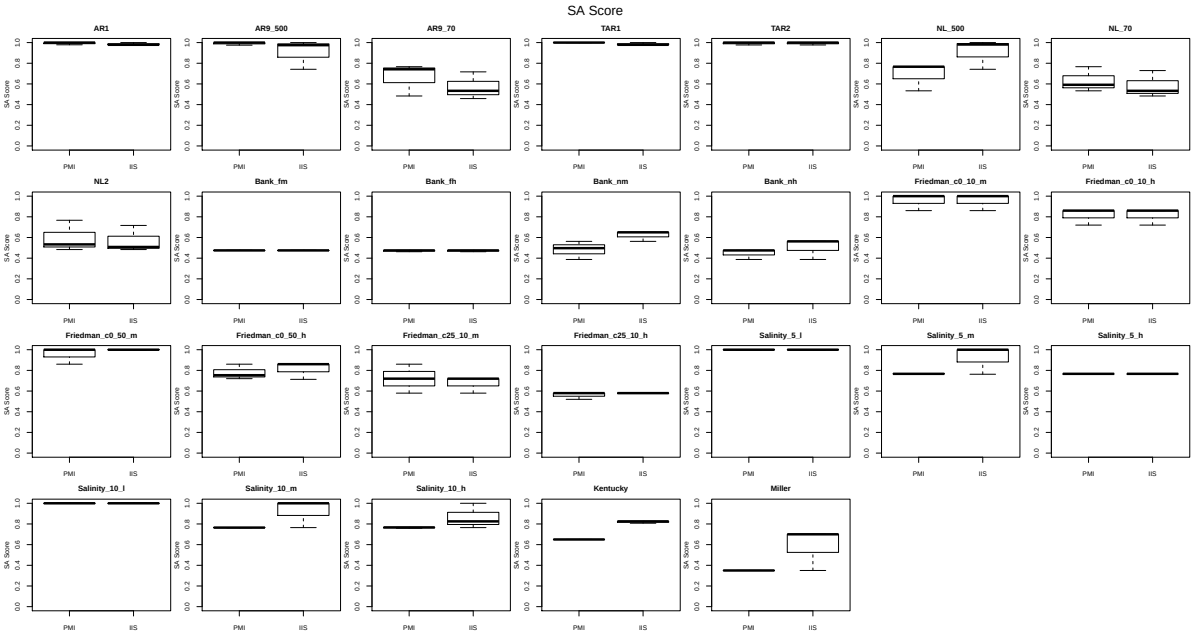


**Figure 1.** Boxplot representing the values of the $SA$ score obtained by running the PMIS and IIS algorithms on the 26 benchmark datasets, with crossbars, boxes and whiskers giving the mean, quartile range and range.

While $SA$ is used to assess the overall performance of an IVS algorithm, the sub-scores $SA_c$ and $SA_e$ specifically account for the proportion of correct and extraneous inputs and thus provide an insight into the over- and under-specification shown by a specific algorithm, as discussed previously. The results reported in Figure 2 show that all four combinations of $SA_c$ and $SA_e$ are possible. The case of perfect specification is found for different datasets, such as AR1, AR9_500, TAR1 and TAR2, for which both algorithms are capable of selecting the relevant variables only. As mentioned above, these datasets have high noise and high collinearity, which are somehow 'compensated' by the $N/P$ ratio. A decrease in the number of observations, as in the AR9_70 dataset, is found to affect the algorithms' performance, causing under-specification of relevant inputs or both under-specification and over-specification of some extraneous inputs. For example, the PMI algorithm does not select any extraneous inputs ($SA_e = 1$), but over the 30 replicates of the dataset it results in an average $SA_c$ score of about 0.65, meaning that the proportion of correct inputs that has been selected is 65%. This results in a $SA$ score of about 0.75, as shown in Figures 1 and 2. Also the IIS algorithm shows a $SA_c$ score of about 0.65, but its overall performance is reduced by the over-specification of some extraneous inputs, with $SA_e$ equal to 0.80, which means that the proportion of extraneous inputs that has been selected is 20%. The case of over-specification of some extraneous inputs is found for the Miller dataset, where the IIS algorithm selects all relevant variables ($SA_c = 1$), but at the same time, it always includes the only extraneous input, resulting in a value of $SA_e$ equal to 0. Worse results are found for PMIS, with $SA_c$ equal to 0.5 and $SA_e$ equal to 0.

## 5 CONCLUSIONS

This work proposes a novel framework for assessing and developing IVS algorithms used in environmental modelling problems. A further development of the framework will follow two directions. Firstly, the preliminary numerical results reported in this paper will be analysed in greater detail with the purpose of determining the effect of the dataset properties on algorithm performance, and defining a number of rules and guidelines for IVS users (and developers). Furthermore, other popular IVS algorithms will be evaluated under this framework. Secondly, the framework will be expanded in order to account for other relevant evaluation criteria, such as computational efficiency and the explanation

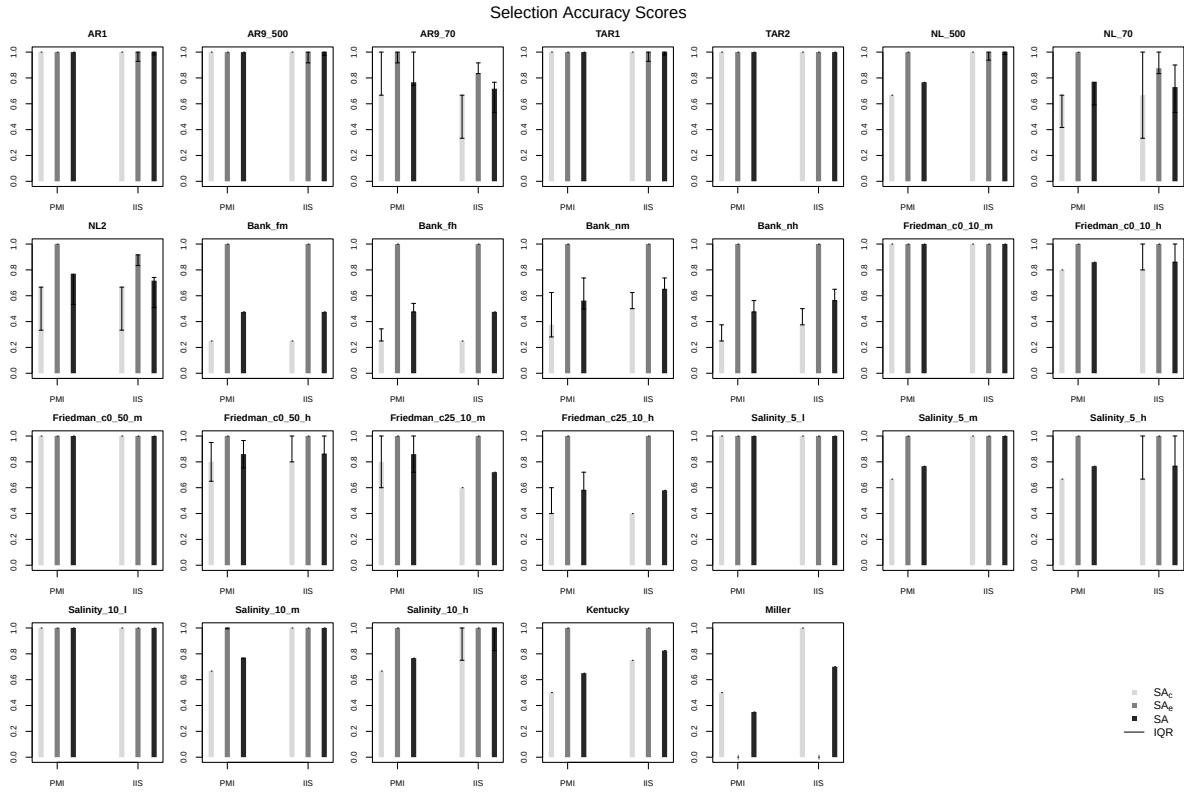


**Figure 2.** Bar charts representing the values of the scores $SA$, $SA_c$ and $SA_e$ obtained by running the PMIS and IIS algorithms on the 26 benchmark datasets.

capability of IVS algorithms.

## ACKNOWLEDGMENTS

This work was supported by the Goyder Institute for Water Research, Project E.2.4. The second author is currently supported by the SRG ESD 2013 061 Start-up Research project

## REFERENCES

Bowden, G. J., Maier, H. R., and Dandy, G. C. (2005). Input determination for neural network models in water resources applications. Part 1. Background and methodology. *Journal of Hydrology*, 301(1-4):75–92.

Galelli, S. and Castelletti, A. (2013). Tree-based iterative input variable selection for hydrological modelling. *Water Resources Research*, 49:4295 – 4310.

Guyon, I. and Elisseeff, A. (2003). An introduction to variable and feature selection. *The Journal of Machine Learning Research*, 3:1157–1182.

Maier, H. R. and Dandy, G. C. (2000). Neural networks for the prediction and forecasting of water resources variables: a review of modelling issues and applications. *Environmental Modelling and Software*, 15(1):101–124.

Maier, H. R., Jain, A., Dandy, G. C., and Sudheer, K. P. (2010). Methods used for the development of neural networks for the prediction of water resource variables in river systems: Current status and future directions. *Environmental Modelling & Software*, 25(8):891–909.

May, R. J., Dandy, G. C., and Maier, H. R. (2011). *Review of input variable selection methods for artificial neural networks*. InTech, Rijeka, Croatia.

May, R. J., Maier, H. R., Dandy, G. C., and Fernando, T. M. K. G. (2008). Non-linear variable selection for artificial neural networks using partial mutual information. *Environmental Modelling & Software*, 23(10–11):1312–1326.

Molina, L. C., Belanche, L., and Nebot, A. (2002). Feature selection algorithms: a survey and experimental evaluation. In *The 2002 IEEE International Conference on Data Mining*, pages 306–313.

Sharma, A. (2000). Seasonal to interannual rainfall probabilistic forecasts for improved water supply management: Part 1 - a strategy for system predictor identification. *Journal of Hydrology*, 239(1-4):232–239.