Xuyuan Li, Aaron C. Zecchin, Holger R. Maier

Improving partial mutual information-based input variable selection by consideration of boundary issues associated with bandwidth estimation

Environmental Modelling and Software, 2015; 71:78-96

Final publication at http://dx.doi.org/10.1016/j.envsoft.2015.05.013

Li X., Zecchin A.C. and Maier H.R. (2015) **Improving Partial Mutual Information-based input variable selection by consideration of boundary issues associated with bandwidth estimation**, *Environmental Modelling and Software*, **71**, 78-96, DOI: 10.1016/j.envsoft.2015.05.013

Improving Partial Mutual Information-based input variable selection by consideration of boundary issues associated with bandwidth estimation

Xuyuan Li[1], Aaron C. Zecchin[2], Holger R. Maier[3]

[1] Email: xliadelaide@gmail.com; Address: School of Civil, Environmental and Mining Engineering, The University of Adelaide, Adelaide, South Australia, 5005, Australia.

[2] CORRESPONDING AUTHOR Email: aaron.zecchin@adelaide.edu.au; Tel: +618 8303 3027 , Fax: +618 8303 4359; Address: School of Civil, Environmental and Mining Engineering, The University of Adelaide, Adelaide, South Australia, 5005, Australia.

[3]Email: holger.maier@adelaide.edu.au; Address: School of Civil, Environmental and Mining Engineering, The University of Adelaide, Adelaide, South Australia, 5005, Australia.

Abstract

Input variable selection (IVS) is vital in the development of data-driven models. Among different IVS methods, partial mutual information (PMI) has shown significant promise, although its performance has been found to deteriorate for non-Gaussian and non-linear data. In this paper, the effectiveness of different approaches to improving PMI performance is investigated, focussing on boundary issues associated with bandwidth estimation. Boundary issues, associated with kernel-based density and residual computations within PMI, arise from the extension of symmetrical kernels beyond the feasible bounds of potential inputs, and result in an underestimation of kernel-based marginal and joint probability distribution functions in the PMI algorithm. In total, the effectiveness of 16 different approaches is tested on synthetically generated data and the results are used to develop preliminary guidelines for PMI IVS. By using the proposed guidelines, the correct inputs can be identified in 100% of trials, even if the data are highly non-linear or non-Gaussian.

Key words

Artificial neural networks; data-driven models; partial mutual information; kernel density estimation; kernel bandwidth; boundary issues; hydrology and water resources; input variable selection

Software availability

Software name: IVS_PMI_2014

Developers: Xuyuan Li, Postgraduate Student, the University of Adelaide, School of Civil, Environmental & Mining Engineering, Adelaide, SA 5005, Australia

Email: xliadelaide@gmail.com

Hardware requirements: 64-bit AMD64, 64-bit Intel 64 or 32-bit x86 processor-based workstation or server with one or more single core or multi-core microprocessors; 256 MB RAM

Software requirements: All versions of Visual Studio 2012, 2010 and 2008 are supported except Visual Studio Express; PGI Visual Fortran 2003 or later version; Windows or Linux 2.6.32.2 operating system

Language: English

Size: 4.55MB

Availability: Free to download for research purposes from the following website: https://github.com/xuyuanli/IVS_PMI_2014

1 INTRODUCTION

Input variable selection (IVS) plays a vital role in the development of data driven environmental models, such as artificial neural networks (ANNs), as the performance of such models can be compromised significantly if either too few or too many inputs are selected (Galelli et al., 2014; Maier et al., 2010; Wu et al., 2014a,b). Although the task of IVS is not unique to environmental modelling, its application in an environmental modelling context is complicated by a lack of understanding of the underlying physical processes, the presence of significant temporal and spatial variation in potential input variables, the non-Gaussian, correlated and collinear nature of potential input variables, and the non-linearity and inherent complexity associated with environmental systems themselves, as emphasised in Galelli et al. (2014). Given the importance and challenges associated with the IVS problem, a large number of approaches, categorised as either model free (utilising a statistical measure of significance between the candidate inputs and the output) or model based (utilising an optimization algorithm for determining the combination of input variables that maximizes the performance of a pre-selected data-driven model), have been developed and refined for the purpose of more accurate IVS (e.g. Galelli and Castelletti, 2013; Galelli et al., 2014; Li et al., 2015; May et al., 2011; May et al., 2008b; Sharma, 2000), with the specific aim to determine the number of inputs that best characterise the input-output relationship with the least amount of variable irrelevance or redundancy (Galelli et al., 2014; Guyon and Elisseeff, 2003). Among existing IVS techniques, partial mutual information (PMI) based approaches are among the most promising model free techniques, as they account for both the significance and independence of potential inputs and have been successfully and extensively implemented in environmental modelling (e.g. Bowden et al., 2005a,b; Fernando et al., 2009; Galelli et al., 2014; Gibbs et al., 2006; He et al., 2011; Li et al., 2015; May et al., 2008a,b; Wu et al., 2014b; Wu et al., 2013).

The PMI IVS approach was introduced by Sharma (2000) and is based on Shannon's entropy(Shannon, 1948), which measures the Mutual Information (MI) between a random input variable X and a random output variable Yas the reduction in uncertainty of Y due to observation of X. As part of the PMI algorithm, inputs are chosen as part of a forward selection approach, during which one input variable is selected at each iteration of the algorithm (starting with an empty set), based on the amount of information a potential input provides (in addition to inputs selected at previous iterations), until certain stopping criteria are met. The amount of information provided by a potential input is given as a function of mutual information (MI) and the contribution of already selected inputs is accounted for by calculating the MI between potential inputs and the residuals of models between the already selected inputs and the desired output, referred to as PMI. Consequently, the performance of different implementations of the PMI algorithm, in terms of input variable selection accuracy and computational efficiency, is a function of the methods used for mutual information (MI) and residual estimation (RE), as highlighted in Li et al., (2015) and May et al. (2008b).

In previous studies on the use of PMI for IVS for data-driven environmental models, the requisite MI and RE are a function of marginal and joint PDFs estimated by kernel density and kernel regression (for the estimation of kernel density based weights) based methods (e.g. Bowden et al., 2005a,b; Gibbs et al., 2006; He et al., 2011; Li et al., 2015; May et al., 2008a,b). Kernel methods are an approach to constructing input/output (I/O) models from input and output data. The resulting I/O model is an ensemble of kernel functions, each centred about a data point in the input space, and returns a weighted average of the influence of all data points. The weight associated with each data point is dependent on the proximity of the input to that data point (i.e. closer points have more influence). Kernel methods are primarily controlled by a bandwidth parameter, which determines the extent to which a single kernel is spread throughout the input space (e.g. a small bandwidth means that data points will only have a localised influence). As such, the performance of PMI IVS is heavily influenced by the accuracy of the kernel density estimates required for MI and RE, which are a function of bandwidth (used interchangeably with smoothing parameter) selection and how well any boundary issues are addressed (Santhosh and Srinivas, 2013; Scott, 1992; Wand and Jones, 1995), as discussed below.

Determination of the optimal bandwidth (the bandwidth that provides the most accurate estimation of the density function) is not trivial, as there is no clear consensus as to which bandwidth estimator performs best for general cases. Overestimating the bandwidth can lead to an over-smoothing of the probability density function (PDF) or residual predictions, so that detailed local information will not be effectively captured. On the contrary, under-estimating the bandwidth can make the general trend become more vulnerable to localised features, or even noise (Li et al., 2014). Although many methods for bandwidth estimation exist in other disciplines (e.g. mathematics and statistics (e.g. Hall et al., 1992; Park and Marron, 1990; Rudemo, 1982; Scott, 1992; Scott and Terrell, 1987)), in almost all existing PMI IVS studies in environmental modelling (e.g. Bowden et al., 2005a,b; He et al., 2011; May et al., 2008a,b) the Gaussian reference rule (GRR) has been used predominately for bandwidth estimation due to its simplicity. However, as highlighted by Harrold et al. (2001) and Galelli et al. (2014), use of the GRR can result in less accurate estimation of MI and PMI for data that are highly non-Gaussian, which is generally the case in environmental and water resources modelling problems. In addition, Li et al. (2015) showed that PMI IVS performance can be improved if alternative bandwidth estimation methods are used for MI and RE for data that are non-Gaussian.

Another potential problem with kernel based methods is the so called ‘boundary issue’, which is associated with the inaccuracies in density estimation arising from the extension of symmetrical kernels beyond the feasible bounds of potential input variable values (e.g. densities associated with negative values of flow obtained using symmetrical kernels) (Wand and Jones, 1995) and generally results in an underestimation of MI or residuals near the boundary. This is commonly encountered in environmental and water resources modelling by the fact that data can be bounded due to their physical feasibility (e.g. rainfall-runoff data are bounded at zero). Although a number of potential methods have been proposed within the statistical literature for addressing this issue (e.g. Cowling and Hall, 1996; Dai and Sperlich, 2010; Fan, 1992; Fan and Gijbels, 1996; Gasser and Müller, 1979; Hall and Park, 2002; Marron and Ruppert, 1994; Schuster, 1985; Zhang and Karunamuni, 1998), their effectiveness has not yet been tested in the context of PMI-based IVS for data-driven environmental modelling. However, this is likely to be a significant problem, as environmental data can be highly skewed near variable boundaries. Consequently, there is a need to establish to what degree the performance of PMI IVS is influenced by the boundary issue, and which methods are the most effective in addressing this.

In order to address the aforementioned research needs, the objectives of the current study are: (i) to assess if, and to what degree, the performance of PMI IVS can be improved by various approaches to addressing boundary issues for data with different properties (i.e. degree of linearity and degree of normality); and (ii) to develop and test a set of preliminary empirical guidelines for the selection of the most appropriate methods for bandwidth estimation and addressing boundary issues for data with different properties. The remainder of this paper is organised as follows. An explanation of PMI IVS and boundary issues is provided in Section 2, followed by the methodology for fulfilling the outlined objectives in Section 3. The results are presented and analysed in Section 4. The proposed guidelines are validated on the semi-real studies in Section 5, before a summary and conclusions given in Section 6.

2 BACKGROUND ON PMI IVS AND BOUNDARY ISSUES

2.1 PMI IVS

Although details of the PMI IVS approach are provided in a number of papers (e.g. Sharma, 2000; Bowden et al., 2005a; May et al., 2008b; He et al., 2011; May et al. 2011; Li et al., 2015), a brief outline of the main steps in the process are given below for the sake of completeness:

Let: $\boldsymbol{X} = [X_1 \ldots X_m]^T$ be the input vector, where m is the number of inputs; y be the output; and $(\boldsymbol{X}^j, y^j)$ be the observed pairs of input and output data for $j = 1, \ldots, n$, where n is the number of observations.

Step 1: Procure candidate inputs $\boldsymbol{X}$ and the output y based on an understanding of the system to be modelled;

Step 2: Estimate the marginal PDF of each candidate input $f(X_i)$ and the output $f(y)$ through univariate kernel density estimation (KDE) (i.e. $K_{h_x}(X_i)$ and $K_{h_y}(y)$) (May et al., 2008b; Scott, 2004; Wand and Jones, 1995), where h_x and h_y are the univariate kernel bandwidths, which determine the accuracy of the kernel based marginal PDFs (Duong and Hazelton, 2003; Scott, 1992; Wand and Jones, 1995);

Step 3: Calculate the joint PDF $f(X_i, y)$ between each candidate input and the output through bivariate KDE (Cacoullos, 1966; Parzen, 1962). Calculation of the bivariate KDE requires the determination of a bandwidth matrix, which is formed by the univariate kernel bandwidths h_x and h_y as mentioned above;

Step 4: Approximate the MI $I_{X_i,y}$ between each candidate input X_i and the output y based on the estimated marginal ($f(X_i)$ and $f(y)$) and joint $f(X_i, y)$ PDFs in accordance with Shannon's entropy (Shannon, 1948), which measures the reduction in uncertainty in y due to an observation of X_i;

Step 5: Select the candidate input with the highest MI;

Step 6: Remove the redundant information provided by the selected input(s) through (i) development of input-output model(s) $\widehat{m}_y(X_{i^*})$ between the selected input(s) X_{i^*} and the output $y$and (ii) obtaining the residuals $(y - \widehat{m}_y(X_{i^*}))$ of these models (i.e. the components of the remaining input and output that are not captured by a conditional prediction by the selected input). In past studies, kernel regression models, such as generalised regression neural networks (GRNNs) (Specht, 1991), have been used for this purpose;

Step 7: Determine if the selected stopping criterion has been satisfied .Potential stopping criteria include bootstrapping, tabulated critical values, the Akaike information criterion (AIC), and the Hampel test, as discussed and tested in May et al. (2008b). If the stopping criterion has been satisfied, stop the process. If the stopping criterion has not been satisfied, proceed to step 8;

Step 8: Estimate the marginal PDF (i.e. $f(v_i)$ and $f(u)$) of each remaining candidate input $v_i = X_i - \widehat{m}_{X_i}(X_{i^*})$ and output residual $u = y - \widehat{m}_y(X_{i^*})$ obtained in Step 6 through univariate kernel density estimation (Wand and Jones, 1995; Scott, 1992; May et al., 2008b);

Step 9: Calculate the joint PDF $f(v_i, u)$ between each remaining candidate input v_i and the output residuals u through bivariate kernel density estimation (Cacoullos, 1966; Parzen, 1962);

Step 10: Approximate the MI $I_{v_i,u}$between each remaining candidate input $v_i$and the output residuals u based on the estimated marginal and joint PDFs in accordance with Shannon's entropy (Shannon, 1948). This is the PMI between the candidate input and output;

Step 11: Select the candidate input with highest PMI;

Step 12: Repeat Steps 7 to 12.

As can be seen, the performance of PMI IVS is a function of MI approximation (Steps 2 to 4 and 7 to 9) and RE (Step 6). As discussed previously, the accuracy of MI approximation is a function of the way the kernel density is estimated (KDE in Step 2 and Step 3), which is likely to be affected by boundary issues. In addition, based on the way residual have been estimated in previous studies (i.e. using kernel regression models in Step 6), the accuracy of RE is also affected by boundary issues. However, it should be noted that there is the possibility of avoiding any potential boundary issues associated with RE by using modelling approaches that are not reliant on kernel regression methods. Further details of the boundary issues in relation to the steps of PMI IVS are given in the following subsection.

2.2 Boundary issues in PMI IVS

Let $\hat{f}$ indicate a non-parametric estimation of the marginal($m = 1$) and joint($m > 1$)PDFs of the input $\boldsymbol{X}$with support $[-\boldsymbol{a}, \boldsymbol{a}]$, and $\boldsymbol{X} = [X_1 \ldots X_m]^T$be the input vector, where$m$ is the number of input variables (i.e., the number of elements in the input column vector$\boldsymbol{X}$); $\boldsymbol{X}^j = \left[X_1^j \ldots X_m^j\right]^T$ are the observed input data from which the non-parametric estimation is undertaken, for $j = 1, \ldots, n$, where n is the number of observations(data points). The conventional KDE (used in Steps 2, 3, and 6 in PMI IVS) PDF is given by

$$\hat{\boldsymbol{f}}(\boldsymbol{X}_i; \boldsymbol{H}) = \frac{1}{n}\sum_{j=1}^{n} \boldsymbol{K}_{\boldsymbol{H}}\left(\boldsymbol{X}_i - \boldsymbol{X}_i^j\right) \quad (1)$$

where $X_i$represents the i^{th} input vector andK_H denotes the kernel type, commonly selected as the Gaussian kernel (May et al., 2008b; Scott, 1992; Wand and Jones, 1995), which is expressed as

$$\boldsymbol{K}_{\boldsymbol{H}}(\boldsymbol{X}) = \frac{1}{(\sqrt{2\pi}|\boldsymbol{H}|)^m} \boldsymbol{exp}\left[-\frac{1}{2}\boldsymbol{X}^T\boldsymbol{H}^{-1}\boldsymbol{X}\right] \quad (2)$$

In Eq. (2), $\boldsymbol{H}$ is the kernel bandwidth matrixif $m > 1$ (or kernel bandwidth for univariate problems if $m = 1$). The commonly used $K_H$is symmetric, satisfies the following integral and moment conditions$\int K_H(\boldsymbol{X})d\boldsymbol{X} = 1, \int \boldsymbol{X}K_H(\boldsymbol{X})d\boldsymbol{X} = 0, \int \boldsymbol{X}\boldsymbol{X}^T K_H(\boldsymbol{X})d\boldsymbol{X} = m$, and has at least two continuous derivatives. According to Dai and Sperlich (2010), if the support $[-\boldsymbol{a}, \boldsymbol{a}]$of $\hat{f}$ is bounded, and in the absence of exponentially falling tails (e.g. support $[0, \boldsymbol{a}]$), strong under-estimation occurs for all data points in the boundary region, which is defined as a distance of the bandwidth h from the boundary, because of the nonzero KDE outside the support of $\hat{f}$. As a consequence, the corresponding bias of $\hat{f}$ is larger than expected. For example, the bias of $\hat{f}$ is of order $O(h)$, rather than $O(h^2)$, at the boundary point for the univariate case in accordance with Dai and Sperlich (2010), Karunamuni and Alberts (2005), and Wand and Jones (1995). These are the so-called 'boundary issues' associated with non-parametric kernel-based estimation. A graphical representation of boundary issue (in 2D) can be found in Hazelton and Marshall (2009).

As mentioned previously, for PMI IVS in environmental modelling, boundary issues can potentially be encountered in both MI (through KDE, in steps 2 and 3) and RE (through KDE, in step 6) when the observations are bounded and/or follow non-Gaussian distributions (e.g. with high skewness and kurtosis).

2.3 Potential options for solving boundary issues in PMI IVS

In order to address the impact of boundary issues, a number of methods have been suggested in the literature (e.g. Dai and Sperlich, 2010; Karunamuni and Alberts, 2005;Wand and Jones, 1995;Fan and Gijbels, 1996), which have been categorised in accordance with whether they can be used during MI estimation, RE, or both, as outlined in Fig. 1. Methods used to correct the boundary issue in MI estimation can be further divided into two groups based on whether they modify kernel functions or bandwidths. As can be seen from Fig.1:

1. Methods that consider modification of the kernel functions include:

- Reflection correction (RC) (Schuster, 1985; Silverman, 1986), which 'reflects' the data at the boundary and adds the density outside the support of $\hat{f}$ back to the boundary region;
- Boundary kernel (BK) (Gasser and Müller, 1979; Marshall and Hazelton, 2010; Zhang and Karunamuni, 2000), which replaces the conventional Gaussian kernel with a more adaptive kernel that is able to capture any shape of the density, although negative densities can be generated near the boundary;
- Pseudo-data approach (PA)(Cowling and Hall, 1996), which generates additional data based on the 'three-point-rule' and combines them with the original data before implementing kernel estimation;
- Kernel transformation (KT) (Marron and Ruppert, 1994), which requires (i) a transformation function g so that $g(X_i)$ has a first derivative of 0 at the boundary; (ii) a kernel estimator with reflection on $g(X_i)$; and (iii) a back-conversion through the change-of-variables formula to achieve $\hat{f}$. As a result of applying the transformation function g, the impact of the boundary issue becomes insignificant because the non-Gaussian data are transformed to a nearly Gaussian distribution prior to KDE;
- Local linear method (LLM) (Zhang and Karunamuni, 1998), which plugs a special case of the boundary kernel (with fixed bandwidth) into a local linear fitting function;
- Empirical translation correction (ETC) (Hall and Park, 2002; Jakeman et al., 2006), which removes boundary issues by introducing an additional empirical data perturbation term $\hat{\alpha}$, which is a translation term constructed specifically to adjust the bias of the density estimate to be within the boundary region, inside the kernel.

2. Methods that consider modification of the bandwidth include:

- Local bandwidth (reducing) (LBR) (Dai and Sperlich, 2010), which adopts a reduced local bandwidth within the boundary region;
- Local bandwidth (enlarging) (LBE) (Gasser et al., 1985; Hall and Wehrly, 1991; John, 1984), which uses a larger local bandwidth within the boundary region.

As can be seen from Fig.1, all of the methods used to correct the boundary issue in MI estimation are theoretically also applicable to RE in cases where kernel regression models are used for this purpose. However, in the case of RE, there are also other alternatives for addressing boundary issues, including modification of the kernel regression type and the use of kernel free modelling approaches. In relation to different kernel regression types, typical options include local linear, quadratic, and high order polynomial regression (LLP, LQP, and LHOP), all of which belong to the local polynomial family. Compared with the most commonly used univariate general regression neural network (GRNN) (which is equivalent to the Nadaraya-Watson estimator), the LLP (also known as the linear smoother), LQP, and LHOP regression types are much less influenced by boundary issues (Dai and Sperlich, 2010; Fan, 1992; Fan and Gijbels, 1996) because the weighted average of each estimating point is more adaptive to the actual observations. In relation to kernel free modelling approaches, multi-layer perceptron artificial neural networks (MLPANNs) provide an attractive option, as they are universal function approximators and have been applied successfully and extensively to environmental (Adeloye et al., 2012; Ibarra-Berastegi et al., 2008; Luccarini et al., 2010; Maier and Dandy, 1997; Maier et al., 2004; Millie et al., 2012; Muñoz-Mas et al., 2014; Ozkaya et al., 2007; Pradhan and Lee, 2010; Young II et al., 2011) and water resources (Abrahart et al., 2007; Abrahart et al., 2012; ASCE, 2000a, b; Dawson and Wilby, 2001; Maier and Dandy, 2000; Maier et al., 2010; Wolfs and Willems, 2014; Wu et al., 2014a; Wu et al., 2014b) problems. In addition, they are independent of boundary issues due to their kernel free features (Maier et al., 2010; Wu et al., 2014b), although a major drawback of MLPANNs is their high computational requirements. Even though there are a number of potential methods aiming to ameliorate boundary issues by means of modification of the kernel function, not all are suited to MI estimation from a practical perspective. This is because MI estimation requires application of these methods in a bivariate setting, but the performance of a number of the methods has not been verified under these conditions. Consequently, in this paper, only selected and appropriate approaches from the aforementioned methods (see Fig. 1) are implemented to fulfil the objectives of this paper, as detailed in the subsequent section.

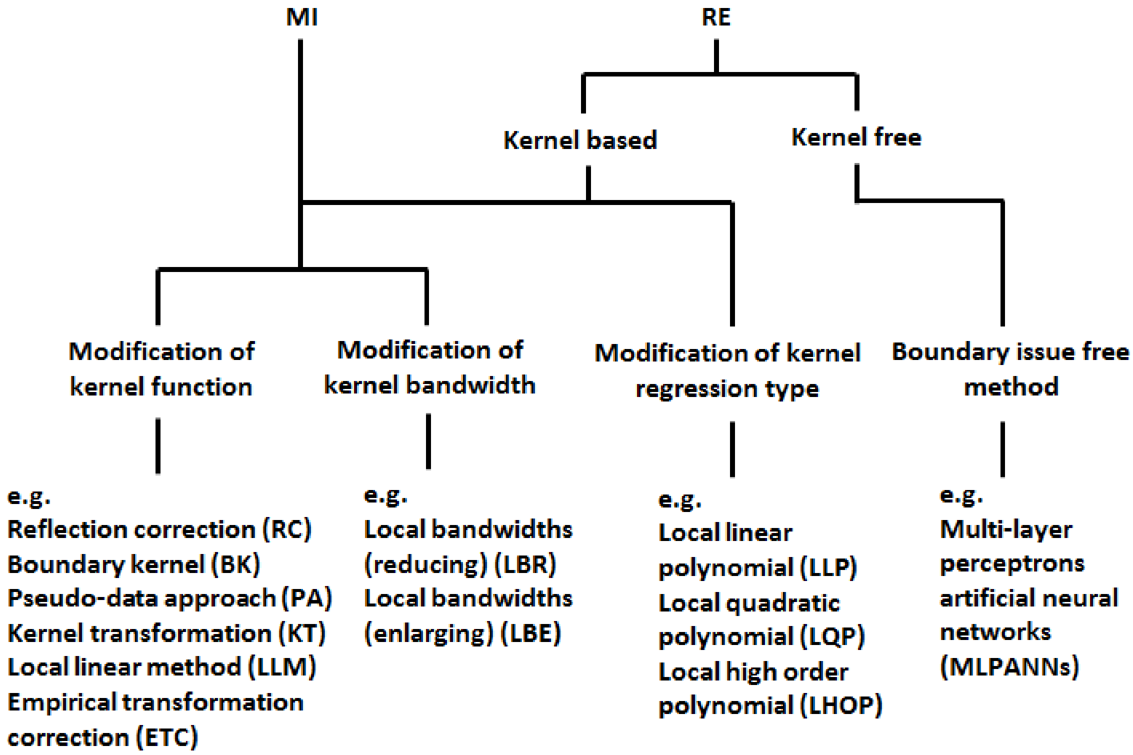


Fig.1. Taxonomy of methods for dealing with boundary issues in mutual information and residual estimation

3 METHODOLOGY

The approach adopted for the systematic assessment of methods for addressing boundary issues on the performance of PMI IVS is outlined in Fig. 2. As can be seen, the approach consists of four main steps, including: (i) generation of input/output data that follow a range of distributions (with different degrees of normality, measured by skewness and kurtosis, and severity of boundary issue, as classified by how the probability density was clustered near the boundary); (ii) estimation of MI using different approaches for dealing with boundary issues; (iii) estimation of residuals using different approaches for dealing with boundary issues; (iv) assessment of the performance of PMI IVS in terms of input variable selection accuracy and computational efficiency for different combinations of approaches for dealing with boundary issues for MI and RE. Details of each of these steps are given in the subsequent sections.

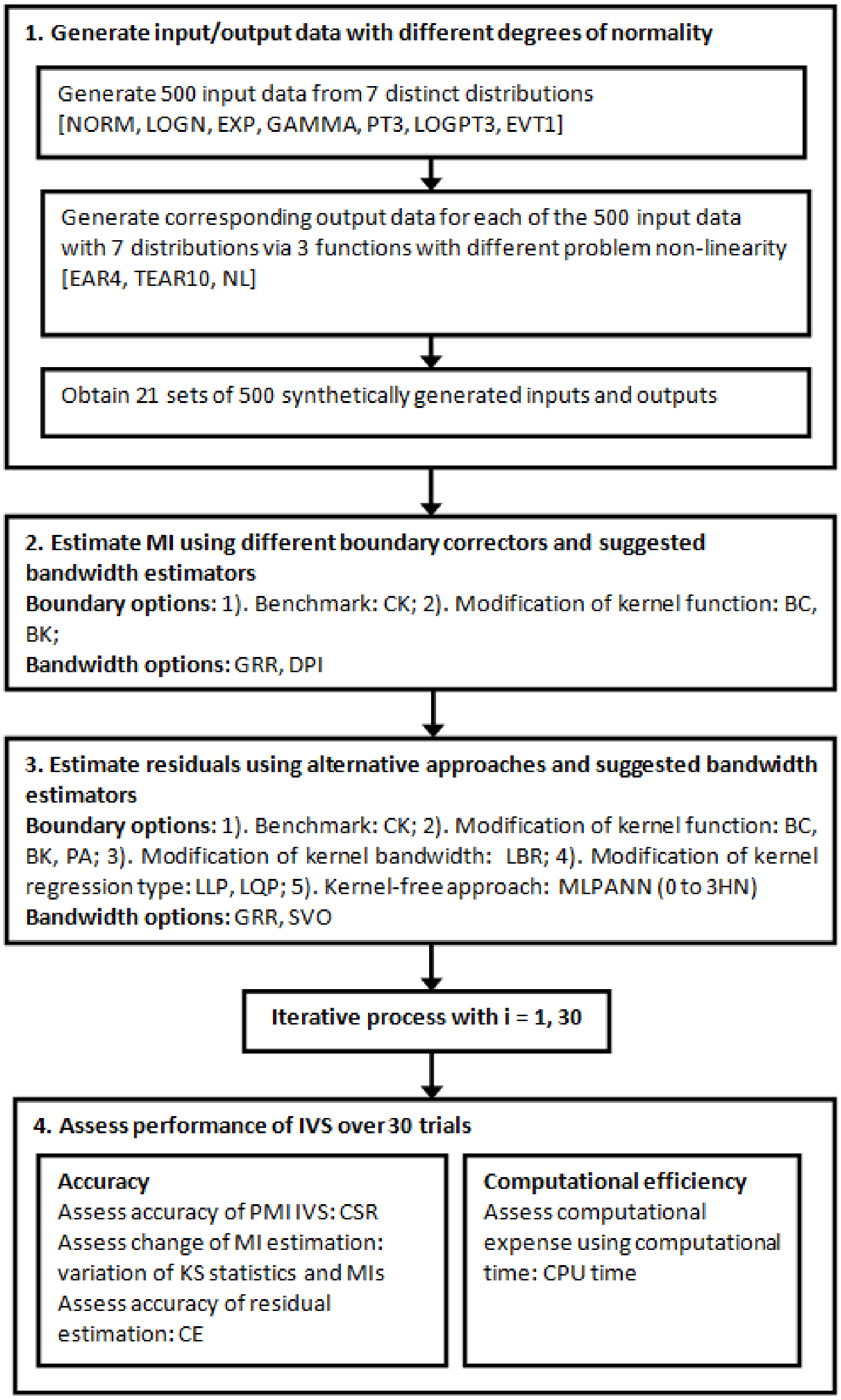


Fig.2. Overview of the proposed analysis for the PMI IVS influenced by bandwidth and boundary issues

3.1 Generate input/output data with different degrees of normality

As pointed out by Galelli et al. (2014), the accuracy of IVS algorithms can only be assessed in an objective and rigorous manner if the correct outputs are known. Consequently, input data are generated from distributions with differing degrees of normality, and the corresponding output data are obtained by substituting the generated inputs into mathematical models. The synthetic data are generated from seven distributions with different degrees of normality, including normal (NORM), log-normal (LOGN), exponential (EXP), gamma (GAMMA), Pearson type III (PT3), log-Pearson type III (LOGPT3), and extreme value type I (EVT1), as these are the most commonly adopted distributions in hydrological modelling (Chow et al., 1988) and result in boundary issues of varying severity. The degree of normality of the input/output data is measured using skewness and kurtosis based on Bennett et al. (2013). The properties of each distribution are listed in Tables 1 and 2. In total, 525data points are generated for each of the exogenous inputs for the three functions considered (details given below) and the first 25 points are rejected in order to prevent initialisation effects (May et al., 2008b), resulting in 500 data points to be used in the analysis.

Table 1Details of the distributions used to generate values of the exogenous input variables and the statistical properties of the generated data for all time series models (EAR4, TEAR10)

Distribution	Key Parameters	s	k	Normality	Boundary Issue
NORM	Mean=3.0; sd =1.0	0.000	-0.013	High	None
GAMMA	Shape=2.0; Scale=1.0	1.370	2.638	High	Low
LOGN	Mean=0.5; sd=1.0	5.326	53.694	Low	High
EXP	Rate=1.0	2.132	7.219	Moderate	Moderate
PT3	Shape=2.5; Scale=3.0; Location=2.0	1.251	2.381	High	Low
LOGPT3	Shape=0.5; Scale=0.2; Location=2.0	4.792	43.265	Low	High
EVT1	Shape=0.0; Scale=0.5; Location=10.0	1.198	2.880	High	Low

(The skewness and kurtosis shown in the table are the averaged values of all input and output data)

Table 2Details of the distributions used to generate values of the input variables and the statistical properties of the generated data for the non-linear model (NL)

Distribution	Key Parameters	s	k	Normality	Boundary Issue
NORM	Mean=3.0; sd =1.0	1.826	5.158	High	None
GAMMA	Shape=2.0; Scale=1.0	10.520	192.091	Low	High
LOGN	Mean=0.5; sd=0.4	5.389	47.767	Low	High
EXP	Rate=1.0	14.029	334.408	Low	High
PT3	Shape=0.5; Scale=1.0; Location=0.5	16.271	514.270	Low	High
LOGPT3	Shape=0.5; Scale=0.2; Location=0.5	14.261	390.522	Low	High
EVT1	Shape=0.1; Scale=0.0; Location=10.0	1.788	9.807	Moderate	Moderate

(The skewness and kurtosis shown in the table are the averaged values of all input and output data)

The output data are generated by substituting the generated input data into three synthetic models, including one linear exogenous auto-regressive time series model (EAR4), one threshold exogenous auto-regressive time series model (TEAR10), and one non-linear input-output function (NL), as they are representative of general water resource problem scenarios with increasing degrees of problem non-linearity. Similar models have also been used in previous IVS algorithm evaluation studies (Bowden et al., 2005b; Galelli and Castelletti, 2013; Li et al., 2014, 2015; May et al., 2008b).

The equation of the EAR4 model is given by

$$x_t = 0.6x_{t-1} - 0.4x_{t-4} + p_{t-1} + \varepsilon_t \tag{3}$$

where$x_t$denotes the output time series; x_{t-n} stands for the input time series with lag n; p_{t-n} represents the exogenous input with lag n; and ε_t is the introduced error term (explained below).The equation for the TEAR10 model is given by

$$x_t = \begin{cases} -0.5x_{t-6} + 0.5x_{t-10} - 0.3p_{t-1} + \varepsilon_t; x_{t-6} \leq 0 \\ 0.8x_{t-10} - 0.3p_{t-1} + \varepsilon_t;\ otherwise \end{cases} \tag{4}$$

The equation for NL is given by

$$y = (x_2)^3 + x_6 + 5\, sin(x_9) + \varepsilon_t \tag{5}$$

The first two time series models are modified from May et al. (2008b) by introducing an additional independent lagged input p_{t-1} into the exogenous AR models, and the third synthetic model is modified from the one used by Bowden et al. (2005a) through the slight adjustment of the significance (coefficient) of each input. The rationale behind these modifications is to create data sets with known distributions through the independent lagged input p_{t-1} and to generate known significance (relative ranking) of input variables through adjusting the coefficient of each input. All three synthetic models have also been used by Li et al. (2014, 2015). The error term ε_t follows a normal distribution $N(0,0.01)$, which introduces noise without obscuring the influence of the actual independent variables. In the present study, all data are scaled between 0 and 1.

3.2 Estimate MI using different boundary correctors and suggested bandwidth estimators

By recalling the fact that not all potential methods aiming to ameliorate boundary issues are suited to MI estimation from a practical point of view, as mentioned in Section 2.2,only three methods, including the conventional kernel (CK) (Bowden et al., 2005a; He et al., 2011; May et al., 2008b) without boundary correction, the reflection correction (RC) (Schuster, 1985; Silverman, 1986), and the boundary kernel (BK) (Gasser and Müller, 1979; Marshall and Hazelton, 2010; Zhang and Karunamuni, 2000) are applied in this study. The CK is selected as a benchmark model against which the performance of the other approaches can be compared; the RC is adopted because it can be extended into a bivariate setting with relative ease; while the BK is implemented because it has theoretically amenable derivations and successful applications to both univariate and bivariate cases. Details of these estimators are given in the following subsections. It should be noted that in each case, in order to minimise any impact due to bandwidth selection, the bandwidths are estimated based on the GRR (for data with Gaussian or nearly Gaussian distributions; e.g. NORM and EVT1 synthetic cases) and 2-stage direct plug-in (DPI) (for data with non-Gaussian distributions; e.g. LOGN and LOGPT3 synthetic cases), according to the empirical guidelines proposed by Li et al. (2015).

Conventional kernel (CK) The CK is the most commonly used approach for the estimation of the PDF and its expression is given in Eqs. (1) and (2). As mentioned in Section 2, this method does not provide any boundary correction, and is therefore used as a benchmark approach.

Reflection correction (RC) As described in Section 2, the motivation behind the RC approach is to 'reflect' data (add $-X_i^j, j = 1, \cdots, n$ to the original data set) so that the underestimated density within the boundary region can be added back based on these reflected data. The more adaptive approach is to only reflect the data within the boundary region (add $-X_i$ if $h_x \geq X_i \geq 0$) (Dai and Sperlich, 2010; Silverman, 1986) and the corresponding expression for the univariate RC becomes

$$\hat{f}(X_i; h_x) = \begin{cases} \frac{1}{n}\sum_{j=1}^{n}\left[K_{h_x}\left(X_i - X_i^j\right) + K_{h_x}(X_i + X_i^j)\right];\ h_x \geq X_i \geq 0 \\ \frac{1}{n}\sum_{j=1}^{n}\left[K_{h_x}\left(X_i - X_i^j\right)\right];\ X_i > h_x \\ 0; X_i < 0 \end{cases} \tag{6}$$

where h_x is the bandwidth for input X_i and the expression for the bivariate RC can be extended as

$\hat{f}(X_i, y; \boldsymbol{H}) =$

$$\begin{cases} \frac{1}{n}\sum_{j=1}^{n}\left[K_H\left(\begin{bmatrix} X_i \\ y \end{bmatrix} - \begin{bmatrix} X_i^j \\ y^j \end{bmatrix}\right) + K_H\left(\begin{bmatrix} X_i \\ y \end{bmatrix} - \begin{bmatrix} -X_i^j \\ -y^j \end{bmatrix}\right)\right]; h_x \geq X_i \geq 0, h_y \geq y \geq 0 \\ \frac{1}{n}\sum_{j=1}^{n}\left[K_H\left(\begin{bmatrix} X_i \\ y \end{bmatrix} - \begin{bmatrix} X_i^j \\ y^j \end{bmatrix}\right) + K_H\left(\begin{bmatrix} X_i \\ y \end{bmatrix} - \begin{bmatrix} -X_i^j \\ y^j \end{bmatrix}\right)\right]; h_x \geq X_i \geq 0,\ y > h_y \\ \frac{1}{n}\sum_{j=1}^{n}\left[K_H\left(\begin{bmatrix} X_i \\ y \end{bmatrix} - \begin{bmatrix} X_i^j \\ y^j \end{bmatrix}\right) + K_H\left(\begin{bmatrix} X_i \\ y \end{bmatrix} - \begin{bmatrix} X_i^j \\ -y^j \end{bmatrix}\right)\right]; X_i > h_x, h_y \geq y \geq 0 \\ \frac{1}{n}\sum_{j=1}^{n}\left[K_H\left(\begin{bmatrix} X_i \\ y \end{bmatrix} - \begin{bmatrix} X_i^j \\ y^j \end{bmatrix}\right)\right]; X_i > h_x,\ y > h_y \\ 0; X_i < 0\ , y < 0 \end{cases} \tag{7}$$

where$\boldsymbol{H}$ is the bandwidth matrix, defined as

$$\boldsymbol{H} = \begin{bmatrix} h_x^2 & \rho_{xy}h_xh_y \\ \rho_{xy}h_xh_y & h_y^2 \end{bmatrix} \tag{8}$$

(known as a hybrid class of bandwidth matrix), where h_y is the bandwidth for output y and$\rho_{xy}$is the correlation coefficient between input $X_i$and output y, in accordance with Li et al. (2015). The detailed explanation of the bivariate RC can be found in the Appendix A.1 and it should be noted that the conditional terms all correspond to different regions in the data space, as influenced by both boundaries, just *x*, just *y*, and neither.

Boundary kernel (BK) Compared with RC, BK is more flexible, as it is designed to automatically adapt to any shape of density within the boundary region. The motivation behind BK is that it is a type of linear boundary kernel for use with an adaptive density estimator (Abramson, 1982) and the adaptive density estimator adjusts the weight of each of the kernel functions in accordance with the actual distribution of the data. Consequently, no assumption is required about the distribution of the data (Marshall and Hazelton, 2010).

The expression of the univariate BK is given by

$$B(u; h_x) = \frac{\left[\left(a_3^{(1)}+4a_2\right)-\left(a_2^{(1)}+3a_1\right)u\right]K_{h_x}(u)}{\left(a_3^{(1)}+4a_2\right)a_0-\left(a_2^{(1)}+3a_1\right)a_1} \tag{9}$$

where $a_\alpha^{(\gamma)} = \int u^\alpha D^\gamma K_h(u)\, du$; $D^\gamma K_h(u) = (\partial^{\int uK_h(u)du}/\partial\, u^{\int uK_h(u)du})K_h(u)$; and $u = (X_i - X_i^j)/h_x$. The adaptive kernel estimator $B(u; h_x)$ results from a linear combination of kernel terms, combined with an adaptive bandwidth, dependent on the density function $f(x)$. This maintains the bias as $O(h^2)$ for thedensity estimation function $\hat{f}$ regardless of the boundary issue. . The scaled data result in two regions, including the boundary region $(u_{min}, 1)$ and the boundary free region $(1, u_{max})$. The univariate BK has an adaptive form for the scaled data within $(u_{min}, 1)$ and a fixed form for the scaled data within $(1, u_{max})$. By extending this concept into two dimensions, the expression of the bivariate BK is given as

$$B(u, v; \boldsymbol{H}) = \frac{b_0 K_H(u,v) + b_1 u K_H(u,v) + b_2 v K_H(u,v)}{b_0 a_{00} + b_1 a_{10} + b_2 a_{01}} \tag{10}$$

where

$$b_0 = \left(a_{30}^{(10)} + a_{21}^{(01)} + 5a_{20}\right)\left(a_{12}^{(10)} + a_{03}^{(01)} + 5a_{02}\right) - \left(a_{21}^{(10)} + a_{12}^{(01)} + 5a_{11}\right)\left(a_{21}^{(10)} + a_{12}^{(01)} + 5a_{11}\right);$$

$$b_1 = \left(a_{11}^{(10)} + a_{02}^{(01)} + 4a_{01}\right)\left(a_{21}^{(10)} + a_{12}^{(01)} + 5a_{11}\right) - \left(a_{20}^{(10)} + a_{11}^{(01)} + 4a_{10}\right)\left(a_{12}^{(10)} + a_{03}^{(01)} + 5a_{02}\right);$$

$$b_2 = \left(a_{20}^{(10)} + a_{11}^{(01)} + 4a_{10}\right)\left(a_{21}^{(10)} + a_{12}^{(01)} + 5a_{11}\right) - \left(a_{11}^{(10)} + a_{02}^{(01)} + 4a_{01}\right)\left(a_{30}^{(10)} + a_{21}^{(01)} + 5a_{20}\right);$$

and $v = (y - y^j)/h_y$. The bivariate BK is adaptive for the scaled data within the boundary region [i.e. $u \in (u_{min}, 1)$ and/or $v \in (v_{min}, 1)$], however, it becomes constant when the scaled data are within the boundary free region [i.e. $(1, u_{max})$and$(1, v_{max})$]. Further details can be found in Marshall and Hazelton (2010).

3.3 Estimate residuals using alternative approaches and suggested bandwidth estimators

In order to assess the effectiveness of different approaches to minimising the impact of any boundary issues in RE, selected approaches from those shown in Fig. 2 are implemented. In addition to the most commonly used GRNN with the CK (as a benchmark), seven alternative residual estimators are implemented. Of these, three are based on the modification of the kernel function (i.e. BC, BK, and PA); one is based on the modification of the kernel bandwidth (i.e. LBR); two are based on the modification of the regression type (i.e. LLP and LQP); and one is a kernel free approach (i.e. MLPANN). The selected approaches are not only representative of the different categories outlined in Fig. 2, but are also theoretically applicable to univariate approaches to RE. Details of these methods are given in the following subsections.

It should be noted that in each case, in order to minimise any impact due to bandwidth selection, where applicable, the bandwidths are estimated based on the empirical guidelines proposed by Li et al. (2014), as outlined in Table 3.

Table 3 GRNN bandwidth estimation techniques used for residual estimation during the PMI IVS

Synthetic data set 1				**EAR4**			
Data distribution	NORM	EVT1	PT3	GAMMA	EXP	LOGN	LOGPT3
Bandwidth estimator	GRR	GRR	GRR	SVO	SVO	SVO	SVO
Synthetic data set 2				**TEAR10**			

Data distribution	NORM	EVT1	PT3	GAMMA	EXP	LOGN	LOGPT3
Bandwidth estimator	GRR	GRR	GRR	SVO	SVO	SVO	SVO
Synthetic data set 3				**NL**			
Data distribution	NORM	EVT1	LOGN	PT3	EXP	LOGPT3	GAMMA
Bandwidth estimator	GRR	GRR	SVO	SVO	SVO	SVO	SVO

(GRR stands for the Gaussian reference rule; SVO denotes single variable optimisation)

GRNN with CK The GRNN with CK, developed by Specht (1991), is the univariate regression approach used for residual approximation in all previous studies of PMI IVS in environmental modelling. Its expression is given by (Li et al., 2014)

$$\hat{y}_{GRNN}(X_i, h) = \frac{\sum_{j=1}^{n} y^j exp\left[-\frac{\left(X_i - X_i^j\right)^2}{2h_x^2}\right]}{\sum_{j=1}^{n} exp\left[-\frac{\left(X_i - X_i^j\right)^2}{2h_x^2}\right]} \tag{11}$$

This method does not involve any boundary correction, therefore it is expected to be significantly influenced by boundary issues and is used as a benchmark approach.

GRNN with RC The motivation behind RC (Silverman, 1986) has been explained in Section 2.2 and Section 3.2. The RC method is implemented by replacing the symmetric kernel estimation part $exp\left[-\frac{\left(X_i - X_i^j\right)^2}{\mathbf{2}h_x^2}\right]$ in Eq. (11) with the RC in Eq. (6). The expression for the estimator then becomes

$$\hat{y}_{RC}(X_i, h) = \begin{cases} \dfrac{\sum_{j=1}^{n} y^j \left[exp\left(-\frac{\left(X_i - X_i^j\right)^2}{2h_x^2}\right) + exp\left(-\frac{\left(X_i + X_i^j\right)^2}{2h_x^2}\right)\right]}{\sum_{j=1}^{n} \left[exp\left(-\frac{\left(X_i - X_i^j\right)^2}{2h_x^2}\right) + exp\left(-\frac{\left(X_i + X_i^j\right)^2}{2h_x^2}\right)\right]}; h_x \geq X_i \geq 0 \\ \dfrac{\sum_{j=1}^{n} y^j \left[exp\left(-\frac{\left(X_i - X_i^j\right)^2}{2h_x^2}\right)\right]}{\sum_{j=1}^{n} \left[exp\left(-\frac{\left(X_i - X_i^j\right)^2}{2h_x^2}\right)\right]}; X_i > h_x \\ 0;\ X_i < 0 \end{cases} \tag{12}$$

GRNN with BK The motivation behind BK has also been explained in Section 2.2 and Section 3.2. Similar to the approach taken with the RC method, the boundary kernel [Eq. (9)] is plugged into Eq. (11), resulting in the following expression

$$\hat{y}_{BK}(X_i,h) = \frac{\sum_{j=1}^{n} y^j \left\{ \frac{\left[\left(a_3^{(1)}+4a_2\right)-\left(a_2^{(1)}+3a_1\right)u\right]K_h(u)}{\left(a_3^{(1)}+4a_2\right)a_0-\left(a_2^{(1)}+3a_1\right)a_1} \right\}}{\sum_{j=1}^{n} \left\{ \frac{\left[\left(a_3^{(1)}+4a_2\right)-\left(a_2^{(1)}+3a_1\right)u\right]K_h(u)}{\left(a_3^{(1)}+4a_2\right)a_0-\left(a_2^{(1)}+3a_1\right)a_1} \right\}} \tag{13}$$

GRNN with PA The implementation of PA is different from the above three methods. According to Cowling and Hall (1996), the motivation behind this approach is to generate pseudo-data beyond the boundary based on the existing data, so that the under-estimated kernel density near the boundary can be compensated by these additional data that contain the same trend. By using the PA, the bias does not increase significantly at the boundary, nor does the variance. The PA was implemented in three steps. Firstly, two additional data points are linearly interpolated in-between every two adjacent original data points and the pseudo-data are then generated by the 'three-point rule', which is

$$X^{(-j)} = -5X^{\left(\frac{j}{3}\right)} - 4X^{\left(\frac{2j}{3}\right)} + \frac{10}{3}X^{(j)}, j = 1, \cdots, n \tag{14}$$

where $X^{\left(\frac{j}{3}\right)}$ and $X^{\left(\frac{2j}{3}\right)}$ refer to the $\frac{j}{3}$th and $\frac{2j}{3}$th data points formed by the interpolated and original data points (Cowling and Hall, 1996), which effectively capture the features of the original data. Secondly, the corresponding density estimation is approximated as

$$\hat{f}(X_i) = \frac{1}{nh}\left\{\sum_{j=1}^{n} K_h\left[(X_i - X_i^j)/h\right] + \sum_{j=1}^{l} K_h\left[(X_i - X_i^{(-j)})/h\right]\right\} \tag{15}$$

where l is an integer less than n. When X_i^j is within the boundary region, the pseudo-data $X_i^{(-j)}$ contribute to the estimation of $\hat{f}$ by rendering the bias and variance to the minimal possible values $O(h^m)$ and $O[(nh)^{-1}]$ if l is a large integer close to n. However, when X_i^j is not in the vicinity of the boundary region, the correction due to the pseudo-data $X_i^{(-j)}$ is negligible with small integer l, as explained by Cowling and Hall (1996).Although l can significantly affect the performance of boundary correction, determination of this parameter is not trivial. In the present study, l is estimated through the golden section search (GSS) optimisation algorithm (Press et al., 1992) and the search is truncated using the ceiling function. Finally, by combining Eq. (11) and Eq. (15), the expression for GRNN(PA) is given by

$$\hat{y}_{PA}(X_i,h) = \frac{\sum_{j=1}^{n} y^j \left\{\sum_{j=1}^{n} K_h\left[(X_i - X_i^j)/h\right] + \sum_{j=1}^{l} K_h\left[(X_i - X_i^{(-j)})/h\right]\right\}}{\sum_{j=1}^{n} K_h\left[(X_i - X_i^j)/h\right] + \sum_{j=1}^{l} K_h\left[(X_i - X_i^{(-j)})/h\right]} \tag{16}$$

GRNN with LBR The concept behind the LBR is to adjust the bandwidth within the boundary region, rather than modifying the kernel. It is found that use of a smaller bandwidth within the boundary region can correct the density estimation affected by the boundary issue, therefore, according to Dai and Sperlich (2010), the bandwidth h used for $a \leq X_i^j \leq c$, where a and $c$are left and right boundaries determined based on the physical meaning of the variable (e.g. a case where the average daily rainfall varies between 0 and 20mm), is defined by

$$h_{X_i^j} = \begin{cases} \max\left(X_i^j - a, \varepsilon\right); if\ a \leq X_i^j < (h + a) \\ \max\left(c - X_i^j, \varepsilon\right); if\ (c - h) < X_i^j \leq c \\ h; otherwise \end{cases} \tag{17}$$

and $\varepsilon = 0.001$ is added to avoid zero bandwidth values and the regression model used is identical to Eq. (11).

Local linear polynomial regression (LLP) As mentioned in Section 2.2, the LLP regression model is theoretically more advanced than the GRNN in terms of its resistance to boundary issues (Dai and Sperlich, 2010; Fan, 1992; Fan and Gijbels, 1996). This is due to the fact that the LLP is a linear order polynomial regression, while the GRNN is a zero-order polynomial regression. Consequently, the estimates obtained from the former are more driven by the actual distribution of the data than those obtained from the latter since the estimated weight of each point is more sensitive to the actual data. As a result, the bias and variance of the estimates from the former are smaller than those from the latter. The general expression for models belonging to the local polynomial family is given by

$$\hat{y}_{LP}(X_i; p, h) = \boldsymbol{e}_1^T \begin{bmatrix} \hat{s}_0 & \cdots & \hat{s}_p \\ \vdots & \ddots & \vdots \\ \hat{s}_p & \cdots & \hat{s}_{2p} \end{bmatrix}^{-1} \begin{bmatrix} \hat{t}_0 \\ \vdots \\ \hat{t}_p \end{bmatrix} \tag{18}$$

where $\boldsymbol{e_1}$ is a vector having 1 in the first entry and 0 elsewhere, $\hat{s}_r = n^{-1} \sum_{j=1}^{n} (X_i^j - X_i)^r K_h\left(X_i^j - X_i\right)$ and $\hat{t}_r = n^{-1} \sum_{j=1}^{n} (X_i^j - X_i)^r K_h\left(X_i^j - X_i\right) y^j$ (Cigizoglu and Alp, 2006). The univariate LLP is obtained by substituting $p = 1$ into Eq. (18), giving

$$\hat{y}_{LLP}(X_i; 1, h) = n^{-1} \sum_{j=1}^{n} \frac{\left\{\hat{s}_2 - \hat{s}_1 (X_i^j - X_i)\right\} K_h\left(X_i^j - X_i\right) y^j}{\hat{s}_2 \hat{s}_0 - \hat{s}_1 \hat{s}_1} \tag{19}$$

Local quadratic polynomial regression (LQP) Although the general expression for the LQP and LLP is identical [Eq. (18)], the former is more flexible and adaptive than the latter because $\hat{s}_r$ and $\hat{t}_r$ are approximated based on a quadratic relationship ($p = 2$), rather than a linear relationship ($p = 1$). As a result, the LQP is theoretically more resistant to the boundary issue than the LLP because the density depends more on the actual distribution of the data, resulting in smaller values of bias and variance. By substituting$p = 2$ into Eq. (18), the univariate equation for the LQP is given as

$$\hat{y}_{LQP}(X_i; 2, h) = n^{-1} \sum_{j=1}^{n} \frac{\left[(\hat{s}_2\hat{s}_4 - \hat{s}_3\hat{s}_3) - (\hat{s}_1\hat{s}_4 - \hat{s}_2\hat{s}_3)\left(X_i^j - X_i\right) + (\hat{s}_1\hat{s}_3 - \hat{s}_2\hat{s}_2)\left(X^i - X\right)^2\right] K_h\left(X_i^j - X_i\right) y^i}{\left[\hat{s}_0(\hat{s}_2\hat{s}_4 - \hat{s}_3\hat{s}_3) - \hat{s}_1(\hat{s}_4\hat{s}_1 - \hat{s}_3\hat{s}_2) + \hat{s}_2(\hat{s}_1\hat{s}_3 - \hat{s}_2\hat{s}_2)\right]} \tag{20}$$

MLPANN The MLP models are developed using the systematic approach proposed by Wu et al. (2014b). A single hidden layer is used and the optimal number of hidden nodes is obtained by trial and error, considering a range of 0 to 4. The number of trials is considered to be sufficient for the three synthetic models (Eqs. (3) to (5)) used in this paper, as the coefficient of efficiency (CE) values (between estimated and actual residuals) of the selected MLPANN are all above 0.95, which indicates very good residual estimates in accordance with Bennett et al. (2013). Such trials also prevent training from over-fitting, as the maximum number of hidden nodes is 4. The back-propagation (BP) algorithm (with learning rate of 0.1 and momentum of 0.1, suggested by Wu et al. (2014b)) is used for calibration and the MLPANN with CE closest to 1.0 is selected as the best model. The optimal number of hidden nodes for the different models is 2 (EAR4), 2 (TEAR10), and 3 (NL). This is consistent with the procedure implemented by Li et al. (2015).

3.4 Test regime

As outlined in Fig. 2, 630 synthetic data sets are simulated, which include 30 replicates for each of the three synthetic models (Eqs. (3), (4) and(5), including25, 25, and 15 candidate inputs, respectively), for each of the seven distributions. For each of the 630 synthetic data sets, 16 distinct PMI IVS approaches are applied, consisting of a combination of the 3 methods used for MI estimation and the 8 regression approaches used for RE (as shown in Table 4), resulting in a total of 10,080 tests.

Of these 16 approaches, three are benchmark approaches without consideration of the boundary issue (B1 to B3), two aim to improve the boundary issue in MI estimation (M1 to M2), seven aim to minimise the effect of the boundary issue in RE (R1 to R7), and four take into account the boundary issue in both MI and RE (C1 to C4). The benchmark studies represent the most commonly used approach applied in previous studies (B1) and the proposed approaches for data with non-Gaussian distributions, in accordance with Li et al. (2014,2015) (B2 and B3). The methods that only address the boundary issue in MI estimation include the RC and BK based MI estimations, as mentioned in Section 3.2. The approaches that only investigate the boundary issue in RE contain kernel based (modification of kernel function, kernel bandwidth, and kernel type) and kernel free methods, as detailed in Section 3.3. The techniques that consider the boundary issue in both MI and RE are a combination of one boundary corrector used in MI (RK) and four boundary resistant algorithms from each category outlined in Sections 2.2 and 3.3. These 16 approaches cover the different combinations of approaches for dealing with the boundary issue in PMI IVS, although there are other combinations(combinations of bandwidth, kernel, and regression used in MI and RE excluded in Table 4) of methods that are likely to result in similar outcomes. In addition, the influence of the bandwidth selection issue in both MI and RE is minimised by following the guidelines proposed by Li et al. (2014, 2015), as specified in Sections 3.2 and 3.3, respectively.

Table 4 Different approaches used for PMI IVS by considering bandwidth and boundary issues

	MI		RE		
	Bandwidth	**Kernel**	**Bandwidth**	**Kernel**	**Regression**
B1	GRR	CK	GRR	CK	GRNN
B2	DPI	CK	GRR	CK	GRNN
B3	DPI	CK	SVO	CK	GRNN
M1	DPI	RC	SVO	CK	GRNN
M2	DPI	BK	SVO	CK	GRNN
R1	DPI	CK	SVO	RK	GRNN
R2	DPI	CK	SVO	BK	GRNN
R3	DPI	CK	SVO	PA	GRNN
R4	DPI	CK	SVO	CK	LBR
R5	DPI	CK	SVO	CK	LLP
R6	DPI	CK	SVO	CK	LQP
R7	DPI	CK	-	-	MLPANN
C1	DPI	RK	SVO	RC	GRNN
C2	DPI	RK	SVO	CK	LBR
C3	DPI	RK	SVO	CK	LLP
C4	DPI	RK	-	-	MLPANN

(B: benchmark approach; M: boundary correction in MI estimation; R: reducing boundary impact in residual estimation; C: combination of methods resistant to boundary issue, used in both MI and residual estimations)

The Akaike Information Criterion (AIC) (Akaike, 1974) is used as the PMI IVS algorithm stopping criterion because it provides a good balance between model accuracy and generalisation ability (Akaike, 1974; Bennett et al., 2013; Dawson et al., 2007; May et al., 2008b) and has been found to perform comparatively well with alternative criteria (May et al., 2008b). It has also been applied successfully by May et al. (2008a, b), He et al. (2011), Wu et al. (2013), and Li et al. (2015).

The software developed for conducting the numerical experiments is available for use by others (see Software Availability at the beginning of this paper), is coded in FORTRAN 90/95 and run on a Linux 2.6.32.2 operating system.

3.5 Assess performance of IVS over 30 trials

The performance of the PMI variants used in the tests is assessed in terms of selection accuracy and computational efficiency, as detailed below.

Selection Accuracy As shown in Fig. 2, the accuracy of PMI IVS is assessed by the correct selection rate (CSR) (Galelli and Castelletti, 2013; Li et al., 2015; May et al., 2008b), which measures the percentage of times the correct inputs are selected in the 30 independent trials (i.e. replicates). In order to better understand the relative impact of the different approaches to addressing the boundary issue on CSR, their impact on MI and RE is also assessed, as detailed below.

The impact of the different approaches to addressing the boundary issue on MI estimation is assessed by comparing both the variation of the Kolmogorov-Smirnov (K-S) statistic (Parsons and Wirsching, 1982) and the corresponding change in MI between two approaches, which is able to detect whether MI can be better estimated as a result of boundary correction in marginal or joint PDF estimates or not. The expression of the variation of the KS is expressed as follows

$$KS\ variation\ (\%) = \frac{KS_{A1} - KS_{A2}}{KS_{A1}} \times 100\% \quad (21)$$

where the K-S statistic measures the supremum distance between the empirical and estimated CDFs and the subscripts (A1, A2) refer to different approaches to addressing the boundary issue (see Table 4). A positive K-S variation indicates improvement of accuracy, and vice versa. As the performance of the empirical kernel based CDF is a function of bin width, a number of bin widths (from 0.001 to 1.0) are tested by means of sensitivity analysis. Bin widths of0.01 were found to be adequate for the purposes of this study, which is consistent with the tests conducted in Li et al. (2015). The corresponding expression measuring the change in MI is given by

$$MI\ variation (\%) = \frac{MI_{A1} - MI_{A2}}{MI_{A1}} \times 100\% \quad (22)$$

and indicates to what extent the improvement or deterioration in kernel density estimation can be propagated to the estimation of MI. When considering the outcomes of Eqs. (21) and

(22), high KS and MI variations indicate effective mitigation of the boundary issue in MI estimation as a result of boundary correction in the estimation of marginal PDFs. High MI variation but low KS variation indicates effective treatment of the boundary issue in MI estimation due to boundary correction in the estimation of joint PDFs, while low MI variation suggests insignificant impact of the boundary issue in MI estimation, regardless of the KS variation.

The impact of the different approaches to addressing the boundary issue on RE is assessed by using the coefficient of efficiency (CE) of the models from which the residuals are extracted. CE measures the difference in predictive performance of the model and a model that only contains the mean of the observations (Bennett et al., 2013) and ranges between 0 (poorest) and 1(Ozkaya et al., 2007).

Computational efficiency The computational efficiency of PMI IVS is evaluated by the computational time (CT), as measured by the average CPU time (measured on a dual processor 2.6 GHz Intel Machine).

4 RESULTS AND DISCUSSION

Within this section, the selection accuracy of the PMI IVS method with different approaches to addressing the boundary issue (see Table 4) and their corresponding computational efficiency are discussed in Sections 4.1 and 4.2, respectively. The resulting empirical guidelines for selecting the appropriate techniques for dealing with boundary and bandwidth issues are then summarised in Section 4.3.

4.1 Selection accuracy

The selection accuracy of the PMI IVS methods with the different approaches to addressing the boundary issue for the EAR4 model is summarised in Fig. 3. As can be seen, the benchmark approaches following the guidelines suggested by Li et al. (2015) (i.e. B2 and B3) have a CSR of 100% for the data that follow a Gaussian or nearly Gaussian distribution (i.e. NORM and EVT1), as these data are not expected to be impacted by any boundary issues. Consequently, there is no need for addressing boundary issues in these cases.

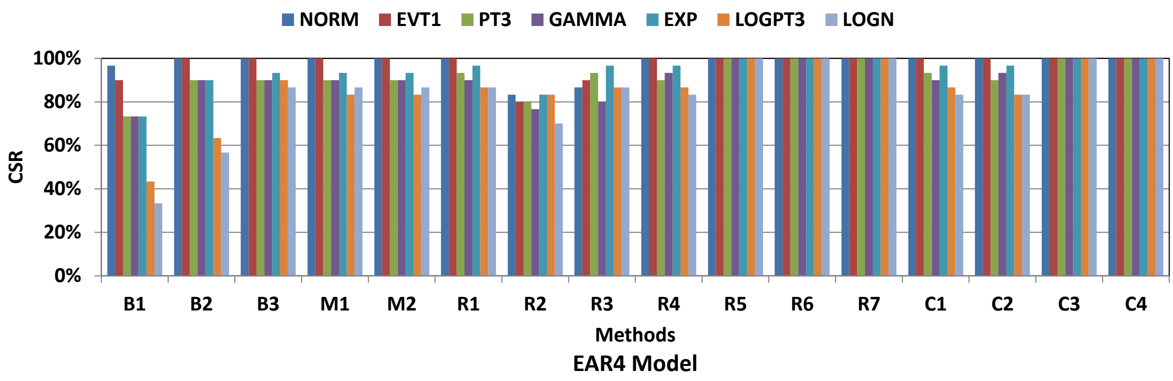


Fig.3. Selection accuracy of the PMI with suggested settings for EAR4 models

For the data that follow a moderately (i.e. PT3, GAMMA, EXP) or severely (i.e. LOGPT3, LOGN) non-Gaussian distribution and are therefore expected to be impacted by boundary issues, some improvement is observed when the benchmark approaches that utilise the guidelines proposed by Li et al. (2015) are implemented for MI estimation (B2) and both MI and RE (B3), compared with the most commonly used approach (B1), but generally CSRs do not exceed 90% (Fig. 3). However, these CSRs can be improved to 100% when some of the proposed approaches to addressing the boundary issue are used, including methods R5, R6, R7, C3 and C4, although not all of the approaches investigated exhibit the same level of success (i.e. methods M1, M2, R1, R2, R3, R4, C1, C2). Potential reasons for these differences in performance are discussed below.

The methods that only address boundary issues in MI estimation (i.e. methods M1 and M2) are not successful in improving CSR compared with the best-performing benchmark approach (i.e. B3). This is despite the fact that these methods are able to improve the accuracy with which the underlying distribution is estimated, as measured by changes in the K-S statistic between methods B3 and M1 (Fig 4a). The reason for this is that the improvements in the estimates in the underlying distributions do not translate into changes in MI estimates (e.g. an approximately 50% increase in the K-S statistic between methods B3 and M1 for the EXP distribution translates into a change in MI estimation that is close to 0%) (Figs.4a and 4b).This can be explained by considering the expression of MI (Shannon, 1948), which is given as

$$I_{X_i,y} \approx \frac{1}{n}\sum_{j=1}^{n} log[\frac{f\left(x_i^j, y^j\right)}{f\left(x_i^j\right)f\left(y^j\right)}] \quad (23)$$

When applying the boundary correction (e.g. RC in M1), estimation of $I_{X_i,y}$ becomes

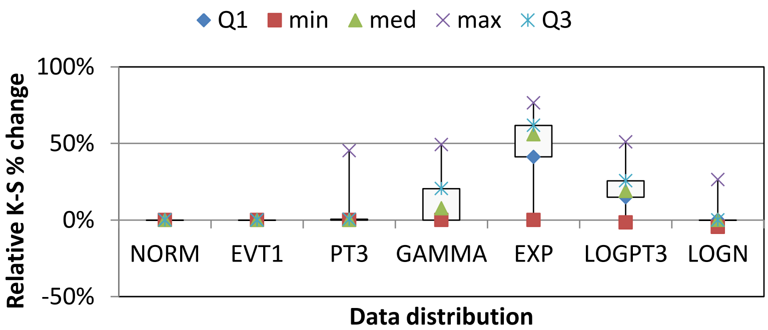
$$I_{X_i,y} \approx \frac{1}{n}\sum_{j=1}^{n} log\left\{\frac{f\left(X_i^j, y^j\right)\Delta f_{xy}}{\left[f\left(X_i^j\right)\Delta f_x\right]\left[f\left(y^j\right)\Delta f_y\right]}\right\} \tag{24}$$

where Δf_{xy}, Δf_x, and Δf_y indicate variations in the marginal and joint densities due to the boundary correction. This equation is equivalent to

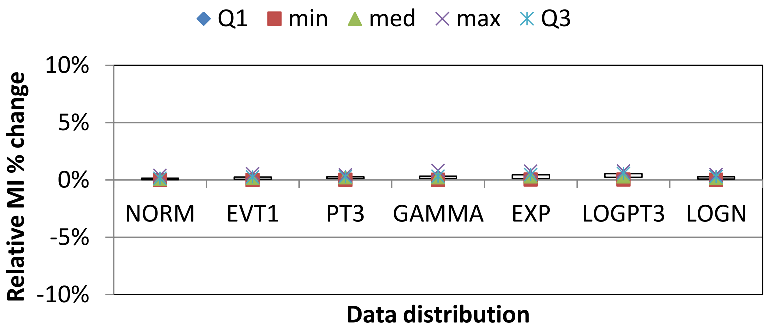
$$I_{X_i,y} \approx \frac{1}{n}\sum_{j=1}^{n} log[\frac{f\left(X_i^j, y^j\right)}{f\left(X_i^j\right)f\left(y^j\right)}] + \left\{log\left(\Delta X_i^j y^j\right) - log\left(\Delta X_i^j\right) - log\left(\Delta y^j\right)\right\} \tag{25}$$

In Eq. (25), the log terms (i.e. $log\left(\Delta X_i^j y^j\right)$, $log\left(\Delta X_i^j\right)$, and $log\left(\Delta y^j\right)$) can diminish the overall improvement of boundary correction (e.g. a change up to 50% in $f\left(X_i^j, y^j\right)$ only results in variation of 0.4 in $log\left(\Delta X_i^j y^j\right)$) and the overall sum of the term $\left\{log\left(\Delta X_i^j y^j\right) - log\left(\Delta X_i^j\right) - log\left(\Delta y^j\right)\right\}$ can be very small (close to zero), which yields a near negligible change in the resulting MI.

In contrast, the accuracy of the models from which the residuals are obtained has a significant impact on MI values. For example, the improved CSRs for methods R5, R6 and R7 (Fig.3) correspond to higher values of the Coefficients of Efficiency of these models compared with that for method B3 (Fig. 5). In contrast, there reverse applies for method R2. Similar results can also be found in Fig. A.2.3. The effectiveness of methods R5 and R6 can be explained by the fact that the bias of the Nadaraya-Watson Regression (equivalent to the univariate GRNN used in all three benchmark models) has an additional error term $\frac{m^{'}(x)f_x^{'}(x)}{f_x(x)}$ [$m(x)$ is the regression function; $f_x(x)$ is the probability density function with respect to x] than the local polynomial regression (e.g. LLP and LQP) used in R5 and R6, and this term increases as the boundary issue becomes severe (Fan, 1992; Masry, 1996; Ruppert and Wand, 1994). In contrast, the effectiveness of R7 can be ascribed to the kernel free feature of the MLPANN used for RE. Therefore, CSR is improved mainly through the adoption of boundary resistant methods in RE, rather than methods that focus on boundary correction.



(a) EAR4 K-S Variation (M1 vs. B3)



(b) EAR4 MI Variation (M1 vs. B3)

Fig.4. Relative change of K-S and MI between M1 and B3 for EAR4 model

The above results suggest that addressing boundary issues in RE is much more important than addressing these issues in MI estimation. This is also confirmed by the results for the combined methods, as the combined methods that resulted in a marked increase in CSR (i.e. C3 and C4) are those that used the most successful methods for addressing the boundary issue in RE (i.e. R5 and R7), and the methods that did not result in an increase in CSR (i.e. M1 and M2) are those that used methods for addressing the boundary issue in RE that are not successful (i.e. R1 and R4), irrespective of which methods are used for addressing the boundary issue in MI estimation.

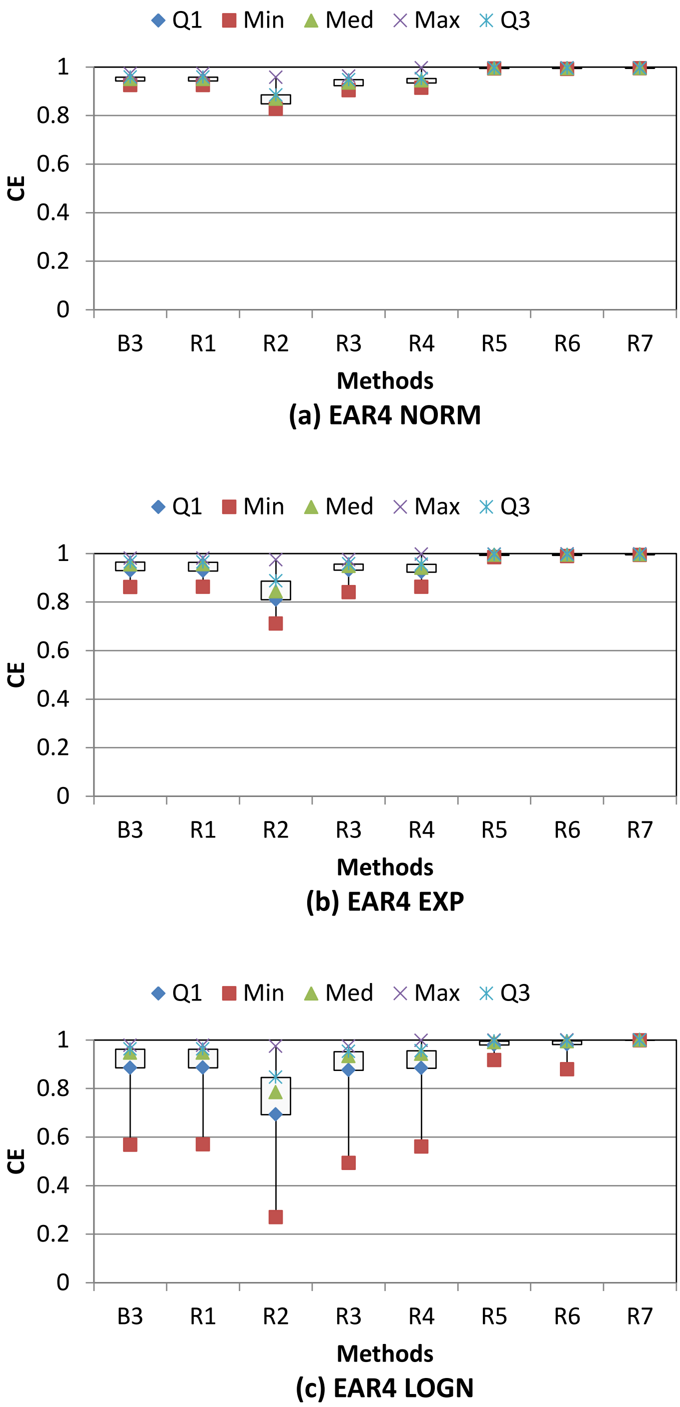


Fig.5. Accuracy of residual estimation with alternative estimators for EAR4 model (3 cases)

The general findings for the EAR4 model (addressing boundary issues in RE is more important than addressing boundary issues in MI estimation and that the use of boundary resistant methods is more effective than the use of boundary correction methods) are confirmed by the results for the TEAR10 (Fig. 6) and NL (Fig. 7) models, with additional supporting information provided in Figs. A.2.1 to A.2.5. However, it should be noted that compared with the results for the EAR4 model, the differences between the different methods are less pronounced for the TEAR10 and more pronounced for the NL model. This can be attributed to the relative predictive performance of the models from which the residuals are obtained for these two datasets, with much higher coefficients of efficiency for the TEAR10 model (Fig. 8) than the NL model (Fig. 9). This is most likely due to the different degrees of non-linearity of the data sets. In addition, benchmark method B1 is found to underestimate the correct number of significant inputs for the non-Gaussian cases (e.g. LOGN and LOGPT3), which can be ascribed to the underestimated bandwidth, as the severity of underestimating the correct number of significant inputs is proportional to the bandwidth ratio. Nevertheless, methods with effective improvement (e.g. R5, R6, R7, C3, and C4) tend to correct such errors with increased bandwidths, which is consistent with the finding in Harrold et al. (2001) and Li et al. (2015).

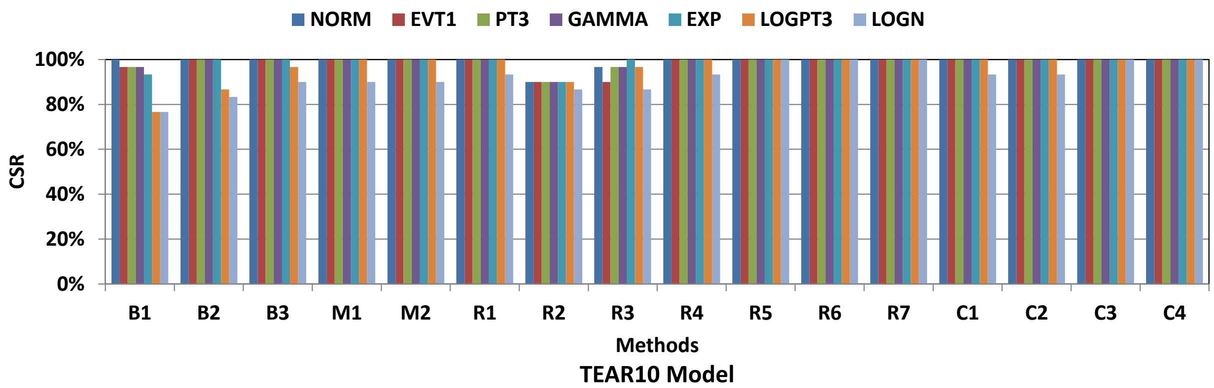


Fig.6. Selection accuracy of the PMI with suggested settings for TEAR10 models

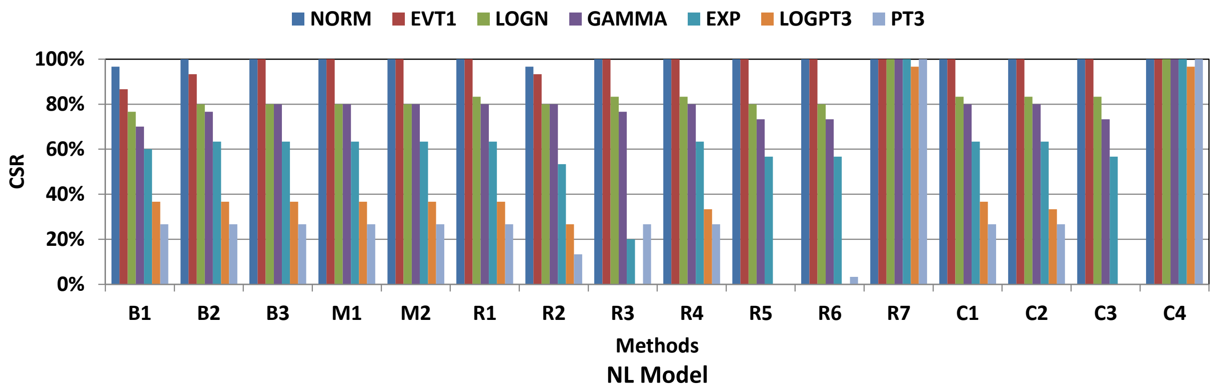
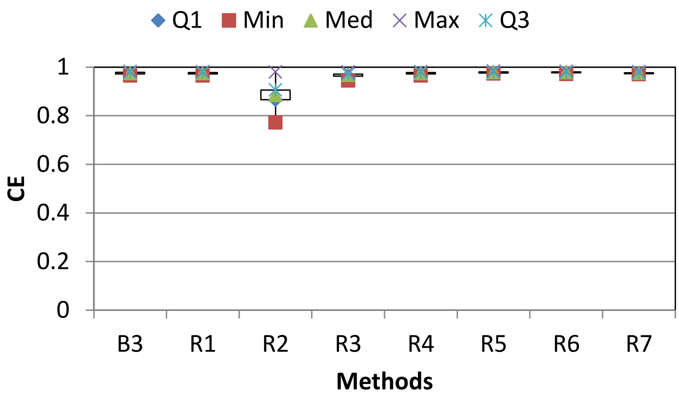
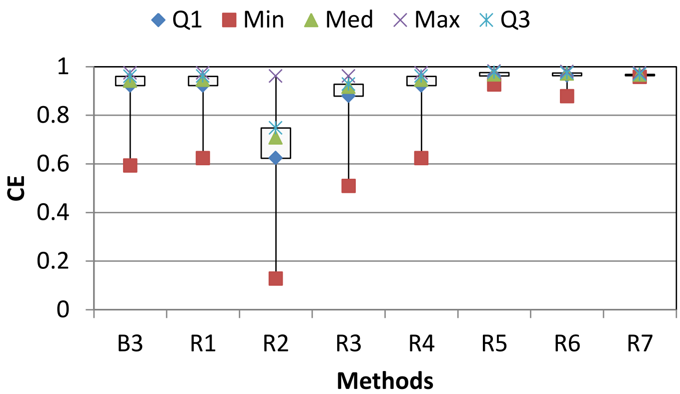


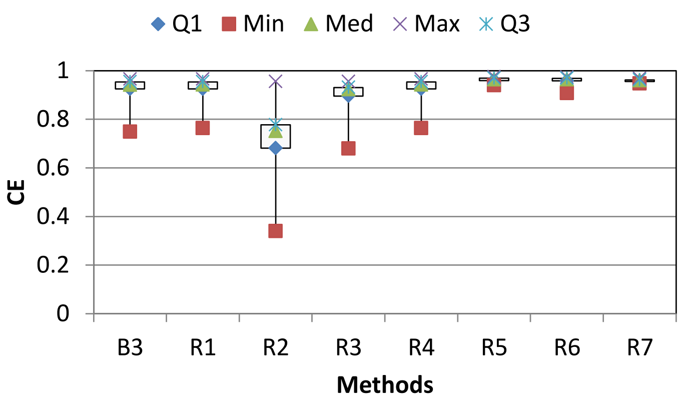
Fig.7. Selection accuracy of the PMI with suggested settings for NL models



(a) TEAR10 NORM

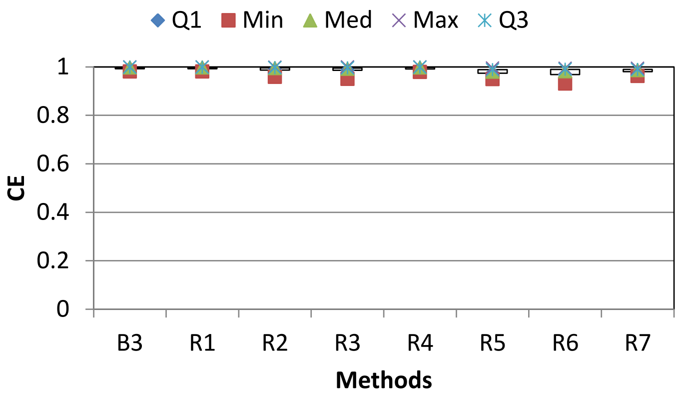


(b) TEAR10 LOGN

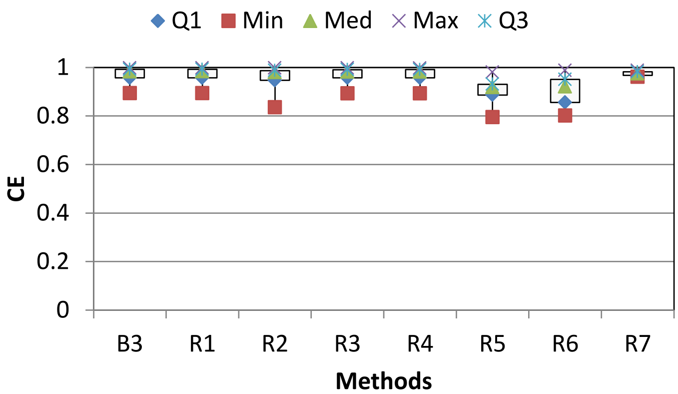


(c) TEAR10 LOGPT3

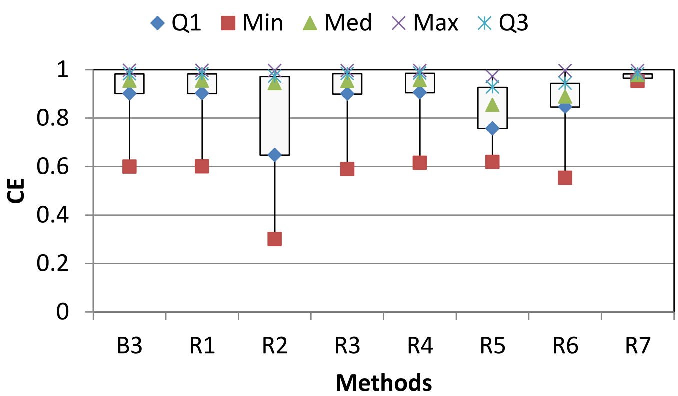
Fig.8. Accuracy of residual estimation with alternative estimators for TEAR10 model (3 cases)



(a) NL NORM



(b) NL LOGN



(c) NL LOGPT3

Fig.9. Accuracy of residual estimation with alternative estimators for NL model (3 cases)

While the TEAR10 model is a threshold function, and would therefore be expected to be more difficult to approximate than the EAR4 model, analysis of the data generated from the TEAR10 model indicates that the threshold function is not activated very often, thereby resulting in quasi-linear model behaviour. In contrast, the high degree of non-linearity of the NL model makes it more difficult to develop the single-input, single-output models from which the residuals are obtained, reducing the effectiveness of some of the methods for dealing with the boundary issue.

This effect is particularly marked for the local polynomial regression based approaches (R5 and R6), which are very effective for the EAR4 and TEAR10 models, with a 100% CSR for all distributions (Figs. 3 and 6), but much less effective for the NL model, for data that are moderately or severely non-Gaussian. This can be attributed to the fact that the RE of non-linear problems, as influenced by both the boundary issue and problem nonlinearity, cannot be effectively improved by using local linear (1st order) or quadratic (2nd order) regression. It should be noted that higher order polynomials ($p > 2$) could be introduced to potentially overcome these issues. The effectiveness of using models that are better able to deal with higher degrees of nonlinearity is confirmed by the 100% CSRs for almost all cases when approach R7 is used (Fig. 7), which uses a MLPANN as the RE model. In this setting, the use of MLPANNs might prove advantageous over using higher-order polynomials, as they are universal function approximators and do not require the functional form of the model to be selected *a priori*.

4.2 Computational efficiency

The computational efficiency of the different PMI IVS approaches investigated is displayed in Fig. 10. As can be seen, the conventional benchmark approach (B1) is the most efficient overall due to the simplicity of the GRR and GRNNs. B2 was the second most efficient approach, as the additional computational cost associated with improving the bandwidth (i.e. DPI) in MI estimation is minimal, followed by B3, which uses a more computationally expensive bandwidth estimator (i.e. SVO) in RE than B2. The efficiency of M1, M2 and C1 is similar to that of B3, indicating an insignificant increase in computational effort when applying boundary correction in MI estimation. On the contrary, the methods for addressing the boundary issue in RE (i.e. R1, R2, R3, R5, R6, R7, C3 and C4) have a marked negative impact on computational efficiency (please note the log-scale on the y-axis of Fig. 10), except for the modification of kernel bandwidth (R4 and C2), as these methods require the implementation of optimisation procedures. This reduction in computational efficiency is particularly prominent for the two approaches that performed best in terms of CSE (i.e. approaches R7 and C4), with an average runtime of 1122s, which is over 227 times greater than that of the most efficient approach (B1). This is mainly due to the time taken for the development of the MLPANNs.

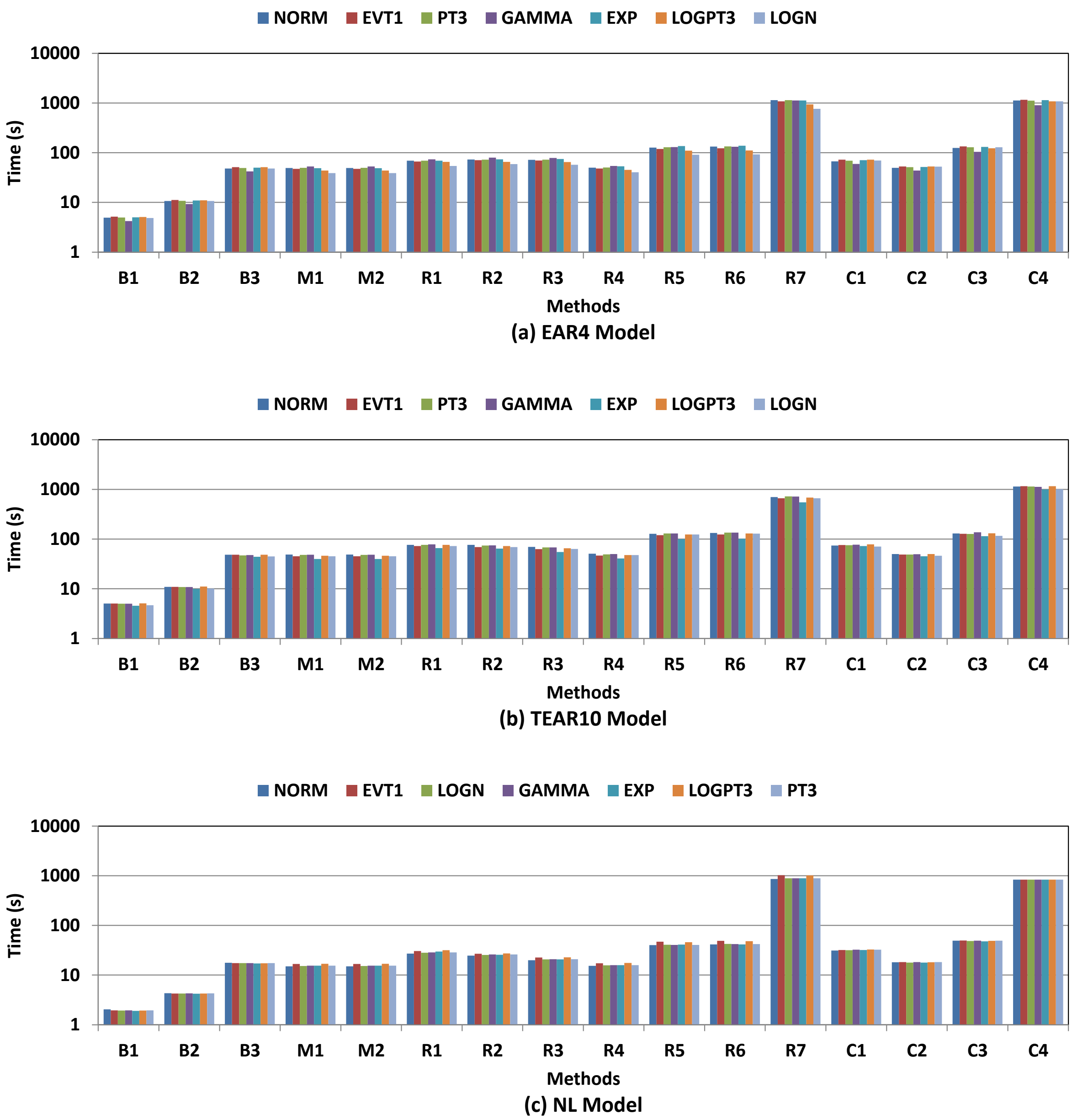


Fig.10. Selection efficiency of the PMI IVS with tested methods for EAR4 models

4.3 Suggested rules and guidelines

Based on the results presented in Sections 4.1 and 4.2, as well as the findings of previous studies by Li et al. (2014,2015), a set of empirical guidelines for determining the best composition of the PMI IVS approaches for a range of data distribution types and system input/output mappings have been developed, as shown in Fig. 11. It should be noted that reasonable trade-offs between selection accuracy and efficiency are considered in the development of these guidelines. However, it is acknowledged that the relative importance of CSR and computational efficiency is also a function of case-study dependent features and user preferences.

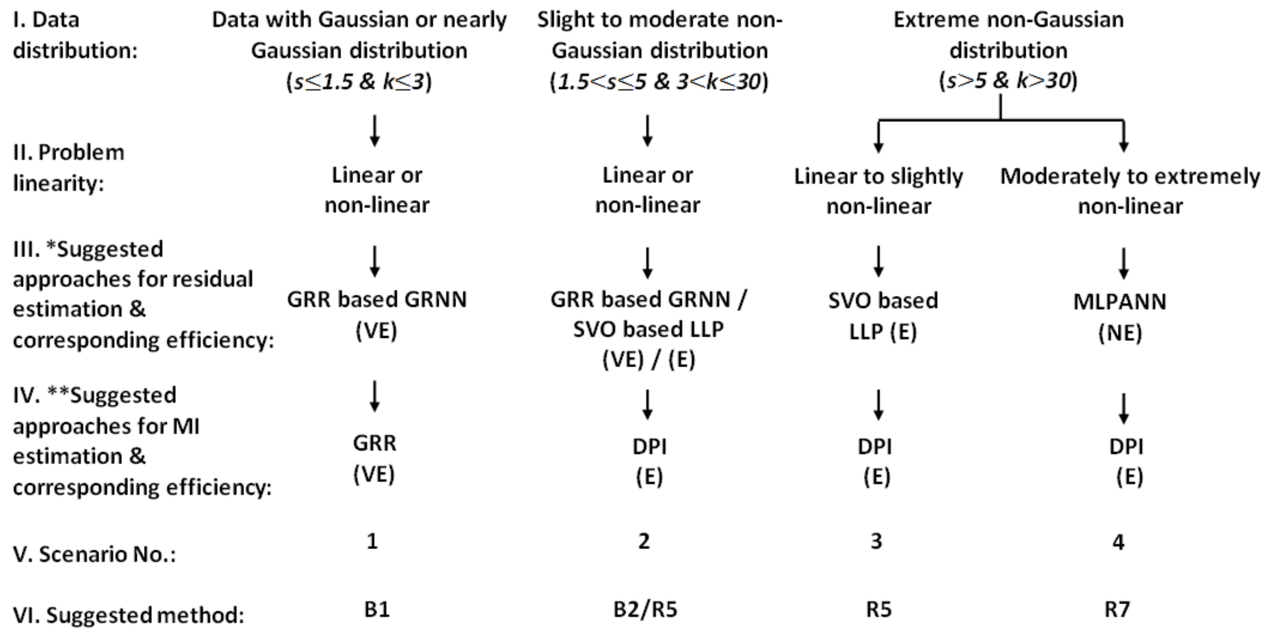


Fig.11. Suggested PMI IVS approaches under distinct scenarios (VE = comparatively very computationally efficient, E = comparatively computationally efficient, and NE = comparatively not computationally efficient; *recommendation based on Li et al. (2014) and present study; **recommendation based on Li et al. (2015)

Overall, four distinct scenarios are identified, as described below:

Scenario 1: If the input/output data are mainly, or nearly, Gaussian (average $s \leq 1.3$and$k \leq 3$), approach B1 (with the GRR based GRNN for RE and the GRR for MI estimation) is recommended, as this combination is able to provide good selection accuracy at the best possible computational efficiency.

Scenario 2: If the input/output data follow moderately non-Gaussian (average $1.3 < s \leq 5$and$3 < k \leq 30$) distributions, approach B2 (with the GRR based GRNN for RE and the DPI for MI estimation) is suggested, so that CSR can be improved with only a very small reduction in computational efficiency. In addition, if the boundary issue is anticipated to be significant (i.e. for cases where the input/output data are clustered near the physical bounds of the data variables), approach R5 (with the SVO based LLP for RE and the DPI for MI estimation) is proposed for IVS.

Scenario 3: If most of the input/output data follow extremely non-Gaussian (average $s > 5$and$k > 30$) distributions and the problem is linear or slightly non-linear, approach R5 (with the SVO based LLP for RE and the DPI for MI estimation) should be implemented, as the combined impact of bandwidth and boundary issues can be effectively overcome at a good trade-off between selection accuracy and efficiency when this approach is implemented.

Scenario 4: If the same conditions as in Scenario 3 apply, except that the problem becomes moderately to extremely non-linear, approach R7 (with the MLPANN for RE and the DPI for MI estimation) is proposed. Although this PMI IVS approach will decrease computational efficiency significantly, it is the only approach that results in reliable selection accuracy under these conditions.

5 VALIDATION ON MURRAY BRIDGE AND KENTUCKY RIVER BASIN CASE STUDIES

5.1 Background

The rules and guidelines proposed in Section 4.3 are tested on two semi-real case studies, including the estimation of salinity in the River Murray in South Australia 14 days in advance (Bowden et al., 2005b; Fernando et al., 2009; Kingston et al., 2005; Li et al., 2014, 2015; Maier and Dandy, 1996) and the prediction of flow in the Kentucky River Basin in the USA one day in advance (Bowden et al., 2012; Jain and Srinivasulu, 2004; Li et al., 2014,2015; Srinivasulu and Jain, 2006; Wu et al., 2013).

River salinity at Murray Bridge 14 days in advance (MBS+13) is a function of the salinity at Mannum, Morgan, Waikerie and Loxton, and the river level at Lock 1, given a specified lag time (i.e. river salinity: MAS-1, MOS-1, WAS-1, WAS-5, LOS-1 and river level: L1UL-1) (Galelli et al., 2014; Maier and Dandy, 1996). However, for the purposes of assessing the effectiveness of PMI IVS, an additional 24 redundant or irrelevant candidate inputs are introduced, as shown in Table 5.

Table 5 Candidate inputs and output used to forecast salinity at Murray Bridge 14 days in advance

<table>
<tr><th colspan="4">Candidate Inputs</th><th colspan="4">Output</th></tr>
<tr><th>Location</th><th>Variable</th><th>Abbreviation</th><th>Lags</th><th>Location</th><th>Variable</th><th>Abbreviation</th><th>Forecasting Period</th></tr>
<tr><td>Mannum</td><td>Salinity</td><td>MAS</td><td>1,3,5,7,9</td><td rowspan="6">Murray Bridge</td><td rowspan="6">Salinity</td><td rowspan="6">MBS</td><td rowspan="6">14</td></tr>
<tr><td>Morgan</td><td>Salinity</td><td>MOS</td><td>1,3,5,7,9</td></tr>
<tr><td>Waikerie</td><td>Salinity</td><td>WAS</td><td>1,2,3,4,5</td></tr>
<tr><td>Loxton</td><td>Salinity</td><td>LOS</td><td>1,2,3,4,5</td></tr>
<tr><td>Murray Bridge</td><td>Salinity</td><td>MBS</td><td>1,3,5,7,9</td></tr>
<tr><td>Lock 1 Upper</td><td>River level</td><td>L1UL</td><td>-3,-1,1,3,5</td></tr>
</table>

The average daily runoff in the Kentucky River Basin one day in advance is influenced by previous values of average daily effective rainfall and runoff (i.e. average daily effective rainfall: P(t), P(t-1) and average daily runoff: Q(t-1), Q(t-2)) (Galelli et al., 2014; Jain and Srinivasulu, 2004). For this case study, the effectiveness of PMI IVS is investigated by introducing another 17 redundant or irrelevant candidate inputs, as shown in Table 6.

Table 6 Candidate inputs and outputs used to forecast flow at Kentucky River Basin 1 day in advance

Candidate Inputs				Output			
Location	**Variable**	**Abbreviation**	**Lags**	**Location**	**Variable**	**Abbreviation**	**Forecasting Period**
Manchester	Average daily effective rainfall	P	0 to 10	Lock & Dam 10	Average daily runoff	Q	1
Hyden							
Jackson							
Heidelberg							
Lexington Airport							
Lock & Dam 10	Average daily runoff	Q	1 to 10				

5.2 Experimental procedure

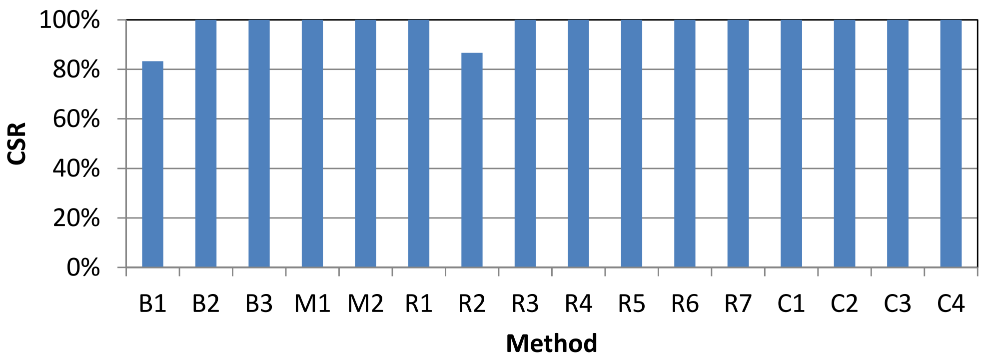
Both case studies are semi-real in the sense that actual input data are used, but that the corresponding output data are generated using a trained ANN model. The adoption of semi-real case studies enabled the benefits of utilising measured input data (i.e. not generated from a known distribution) to be combined with those of having known inputs, thereby enabling the performance of IVS methods to be tested in an objective and rigorous manner, as suggested by Galelli et al., (2014) and Humphrey et al. (2014).

For both case studies, standard MLPs are developed using the approach proposed by Wu et al. (2014b). The DUPLEX method (May et al., 2010) is implemented to split the historical records into training (60%), testing (20%) and validating (20%) sets. By using a single hidden layer and empirically trying between 0 and 6 hidden nodes (in increments of 1), the optimal model structures are found to be 6-4-1 and 4-4-1 for the salinity and rainfall-runoff cases, respectively. Model calibration is conducted using the back-propagation algorithm (with learning rate of 0.1 and momentum of 0.1). The input data used in the PMI IVS are re-simulated 30 times based on the observations, so that the data sets contain random variations while maintaining the major time patterns. Finally, the corresponding output data are obtained by substituting the re-simulated inputs into the trained ANN model. This procedure has also been successfully applied in Li et al. (2015).

5.3 Results and discussion

The salinity case study is categorised as a strong linear problem with mildly non-Gaussian input and output distributions (not significantly affected by bandwidth and boundary issues) (Bowden, 2003; Galelli et al., 2014; Li et al., 2014,2015; Wu et al., 2013).Consequently, these data correspond to Scenario 2 in Fig. 11. Given this, the performance of PMI IVS using approach B2 is expected to be superior in terms of a desirable trade-off between selection accuracy and efficiency.

The results presented in Fig. 12 are consistent with this expectation. The CSR associated with using approach B2 is 100% (estimated in 107s), compared with a CSR of less than 84% (estimated in 47s) when approach B1 is used. CSRs of 100% are also achieved by the alternative approaches (except R2), however, at additional computational cost (487s to 7565s). Consequently, the best trade-off between selection accuracy and efficiency is given by approach B2, as suggested by the proposed guidelines (Fig. 11). This is also consistent with the study carried by Li et al. (2015), which suggested that the DPI/BCVDPI based method provided the best overall performance against other tested methods.



(a) River Salinity at Murray Bridge

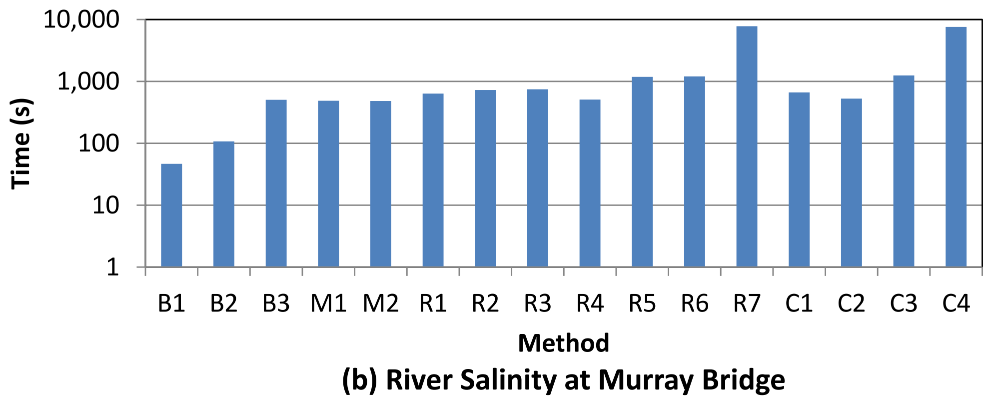


Fig.12. Selection accuracy and efficiency of the PMI IVS with suggested settings for Murray Bridge case

As the rainfall-runoff case is categorised as a strong non-linear problem with extremely non-Gaussian distributions (significantly influenced by bandwidth and boundary issues) (Galelli et al., 2014; Li et al., 2014,2015; Wu et al., 2013), it corresponds to Scenario 4 in Fig. 11. Given this, the performance of PMI IVS using approach R7 is expected to be superior in terms of a balance between selection accuracy and efficiency.

Based on the results in Figs.13 (a) and 13 (b), this is indeed the case. The CSRs associated with using approaches R7 and C4 are 100%, followed by those of approaches B3, M1, M2, R1, R4, C1, C2 (all around 93%), B2, R3 (both approximately 87%), R2 (83%), R6, B1 (both near 77%), R5 and C3 (both about 73%). While the use of approach R7 increased CSR at significant computational cost (at around 45856s; over 162 times B1's runtime), as shown in Fig. 13 (b), this provide the most robust selection accuracy, as suggested by the proposed guidelines (Fig. 11). Compared with the results of Li et al. (2015), selection accuracy is further improved to 100% with R7 (boundary issue free approach), which suggests that both boundary and bandwidth selection issues need to be considered during IVS for data with extremely non-Gaussian distributions.

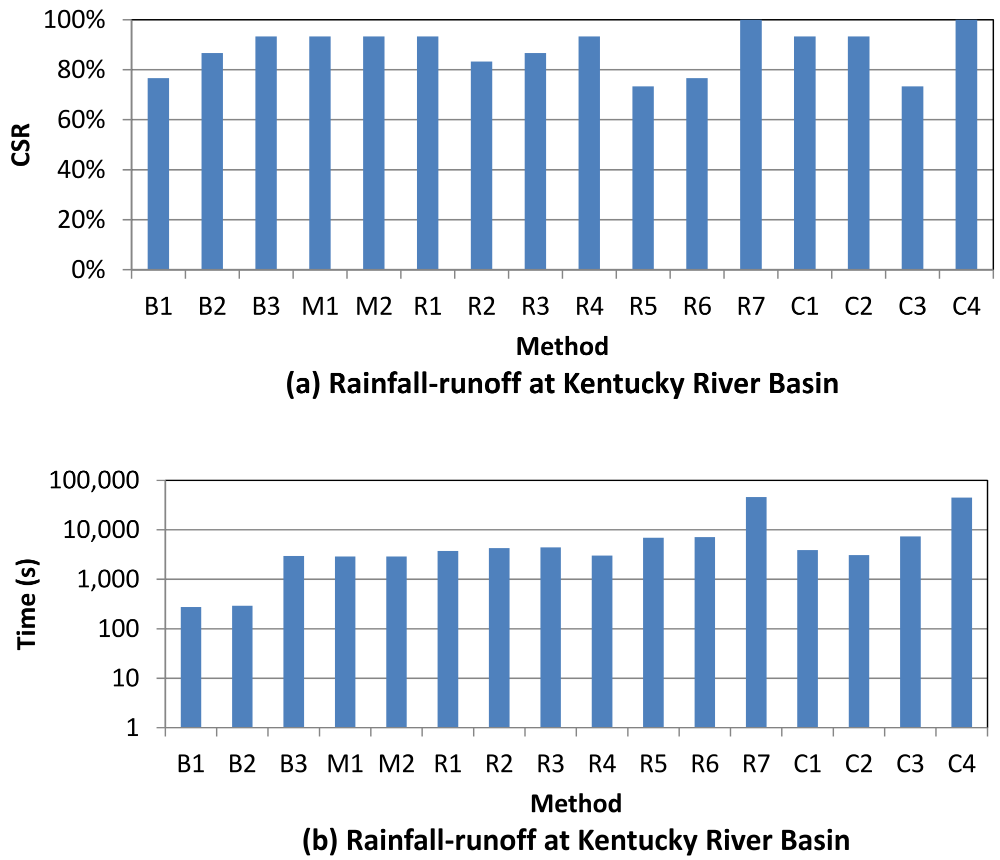


Fig.13. Selection accuracy and efficiency of the PMI IVS with suggested settings for Kentucky River basin case

6 SUMMARY AND CONCLUSIONS

Partial mutual information (PMI) has been successfully and extensively implemented in environmental and water resources modelling, as it considers both the significance and independence of candidate inputs. Given that PMI input variable selection (IVS) is a function of kernel based MI and RE, the performance of PMI IVS is influenced by the determination of an appropriate bandwidth (otherwise termed the smoothing parameter) and boundary issues. Although the impact of bandwidth selection on correct selection rate (CSR) and computational efficiency of PMI IVS has been studied previously, the impact of the boundary issue has not yet been addressed, making it difficult to know to what degree the performance of PMI IVS can be compromised by such issues and which methods can effectively address this impact.

In order to develop a more reliable PMI IVS algorithm for problems with boundary issues, in conjunction with bandwidth issues, the CSR and computational efficiency of PMI IVS were assessed for16 different approaches to addressing these issues on synthetic data sets with different degrees of normality and non-linearity. Of these 16 methods, three are benchmark approaches without explicitly considering the boundary issue (B1 to B3), two aim to improve the boundary issue in MI estimation (M1, M2), seven ameliorate the boundary issue in RE (R1 to R7), and four are combined approaches that take into account the boundary issue in both MI and RE (C1 to C4). The results from 10,080 trials with the synthetic data contributed to the establishment of preliminary empirical guidelines for the selection of the most appropriate PMI IVS approach, for data with different degrees of normality and non-linearity. The validity of the developed guidelines was then tested on two semi-real data sets.

Results of the synthetic studies suggest that methods that address boundary issues in MI estimation do not result in improvements in CSR. This can be ascribed to the fact that changes in the joint and marginal distributions, resulting from the boundary correction,have a diminished influence on PMI due to the appearance of these terms in log functions in the PMI calculation. In contrast, methods that address boundary issues in RE are able to increase CSR to 100% (or very close to 100%) for even the most non-Gaussian and non-linear datasets tested. However, this is not the case for all methods, with boundary resistant methods exhibiting greater success than methods focussed on boundary correction. In particular, the use of MLPANNs for RE results in the most robust selection accuracy, although at a significant decrease in computational efficiency.

Based on the empirical guidelines for the selection of the most appropriate PMI IVS approaches developed in Fig. 11, the most commonly used combination of GRR-based kernel bandwidth selection and GRNN-based RE only results in reliable IVS if the input/output data follow Gaussian or nearly Gaussian distributions and do not have any boundary issues. If the data are moderately or highly non-Gaussian, the DPI should be used for MI bandwidth estimation, regardless of the degree of non-linearity in the data. However, as the data become more non-Gaussian and non-linear, RE approaches should move from GRNNs to LLPs to MLPANNs in order to achieve CSRs near 100%, with associated decreases in computational efficiency. It should be noted that although the empirical guidelines can only be applied to datasets in which all variables have a similar distribution, this does not limit the methodological contribution of this research.

The accuracy of the proposed guidelines was supported by the results of the two semi-real case studies. For the salinity case study, for which the data were close to linear and followed a mildly non-Gaussian distribution, method B2 (Table 4), which used the DPI for MI bandwidth estimation and the GRNN with the GRR for bandwidth estimation, resulted in 100% CSR while being very computationally efficient. For the rainfall runoff case study, for which the data were highly nonlinear and followed an extremely non-Gaussian distribution, MLPANNs had to be used for RE in order to achieve 100% CSRs.

When applying the proposed guidelines to different water resources and environmental modelling problems, it is recommended to first consider the distribution statistics (i.e. skewness and kurtosis) of the input and output variables and then categorise the problem into the most suitable scenario. In general, most water quantity models contain input and output variables that are bounded by their physical meaning and form highly skewed distributions (e.g. average daily rainfall-runoff data), thereby selection of the most appropriate bandwidth and boundary corrector should be considered in accordance with scenarios 3 and 4 in Fig.11. In contrast, water resource models that mainly include input and output variables that follow Gaussian or nearly Gaussian distributions (e.g. concentrations of dissolved oxygen in rivers) should implement scenarios 1 and 2 in Fig. 11 for the sake of good selection accuracy at the best computational efficiency. However, it is acknowledged that the application of proposed guidelines is also a function of case-study dependent features and user preferences.

Overall, the results show that by using methods for MI and RE that are tailored to the input-output data under consideration, CSRs of 100% (or close to 100%) can be achieved when using PMI IVS, even for data that are highly non-linear and highly non-Gaussian. This is in contrast to PMI IVS methods that use “standard” approaches to MI and RE, which have been shown to perform poorly under such circumstances in this and previous studies (e.g. Li et al., 2015; Galelli et al., 2014). However, alternative methods for dealing with non-Gaussian data in the context of PMI IVS, such as transforming the input data to normality (e.g. Bowden et al., 2003) and estimating the required densities using histogram-based methods (e.g. Fernando et al., 2009), require further investigation, as does the impact of the stopping criterion (see May et al., 2008a) on the results obtained in this study. Although the objective of the present study is to improve PMI IVS itself, the ultimate goal of improving IVS is to improve the performance of the MLPANNs (or other data-driven environmental and water resource models), which requires assessment and quantification of the improvement in terms of MLPANN model performance using the proposed PMI IVS in the future research. In addition, the findings of this work should be tested more broadly, including for data sets with a wider range of attributes, such as different degrees of noise, collinearity and interdependency, as well as incomplete information (see Galelli et al., 2014).

ACKNOWLEDGEMENTS

This research was aided by the code and suggestions from Dr. R.J. May (GRR based MI/PMI) and Dr. G.B. Humphrey (GRR based GRNN). The authors would also like to thank the anonymous reviewers, whose inputs contribute to the improvement of this paper.

APPENDIX

A.1 Explanation of Bivariate Reflection Correction

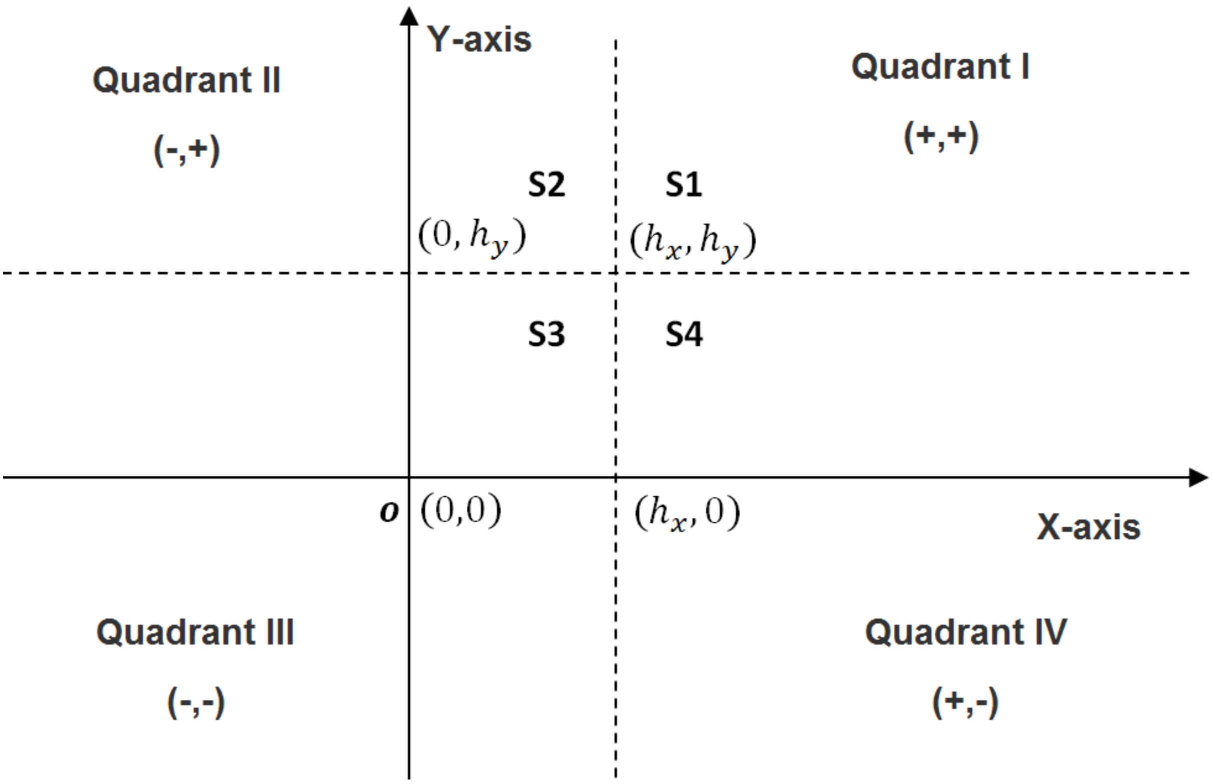


Fig. A.1.1 Quadrants of Bivariate Reflection Correction

As mentioned in Section 2, let: $\boldsymbol{X} = [X_1 \ldots X_m]^T$ be the input, where m is the number of inputs; $(\boldsymbol{X}^j, y^j)$ be the observed pairs of input and output data for $j = 1, \ldots, n$, where n is the number of observations, $\boldsymbol{X}^j = [X_1^j \ldots X_m^j]^T$ are the observed input data and y^j are the observed output data. $\boldsymbol{H}$ is the bandwidth matrix, defined as $\boldsymbol{H} = \begin{bmatrix} h_x^2 & \rho_{xy} h_x h_y \\ \rho_{xy} h_x h_y & h_y^2 \end{bmatrix}$, where h_x and h_y are the estimated bandwidths for input X_i and output y, respectively, and ρ_{xy} is the correlation coefficient between input X_i and output y. Four quadrants are created by the x-axis and y-axis, as shown in Fig. A.1.1. Within Quadrant I, four regions (S1 to S4) are further generated by the lines passing through $x = h_x$ and $y = h_y$.

After scaling all data within [0,1] in both x-axis and y-axis, all points fall into Quadrant I. Points falling into S1 ($X_i^j > h_x, y^j > h_y$) are not influenced by the boundary issue, therefore the density can be estimated based on Eqs. (1) and (2), as outlined in Section 2, which is expressed as

$$\hat{f}(X_i, y; \boldsymbol{H}) = \frac{1}{n}\sum_{j=1}^{n}\left[K_H\left(\begin{bmatrix} X_i \\ y \end{bmatrix} - \begin{bmatrix} X_i^j \\ y^j \end{bmatrix}\right)\right]; X_i \ > h_x, y > h_y$$

Points falling into S2 ($h_x \geq X_i^j \geq 0, y^j > h_y$) are only influenced by the boundary issue on the x-axis, therefore reflection correction is required only on the x-axis. By implementing the reflection kernel on the x-axis, the kernel density is given as

$$\hat{f}(X_i, y; \boldsymbol{H}) = \frac{1}{n}\sum_{j=1}^{n}\left[K_H\left(\begin{bmatrix} X_i \\ y \end{bmatrix} - \begin{bmatrix} X_i^j \\ y^j \end{bmatrix}\right) + K_H\left(\begin{bmatrix} X_i \\ y \end{bmatrix} - \begin{bmatrix} -X_i^j \\ y^j \end{bmatrix}\right)\right]; h_x \geq X_i \ \geq 0, y > h_y$$

where points in S2 are ‘reflected’ into Quadrant II, so that the underestimated density near the boundary (y-axis) can be compensated for.

Points falling into S3 ($h_x \geq X_i^j \geq 0, h_y \geq y^j \geq 0$) are affected by the boundary issue in both x-axis and y-axis, consequently, reflection correction is required in both dimensions, which then results in

$$\hat{f}(X_i, y; \boldsymbol{H}) = \frac{1}{n}\sum_{j=1}^{n}\left[K_H\left(\begin{bmatrix} X_i \\ y \end{bmatrix} - \begin{bmatrix} X_i^j \\ y^j \end{bmatrix}\right) + K_H\left(\begin{bmatrix} X_i \\ y \end{bmatrix} - \begin{bmatrix} -X_i^j \\ -y^j \end{bmatrix}\right)\right]; h_x \geq X_i \ \geq 0, h_y \geq y \geq 0$$

Where points in S3 are ‘reflected’ into Quadrant III, and hence the problem associated with underestimated density near the boundary (x-axis and y-axis) can be addressed.

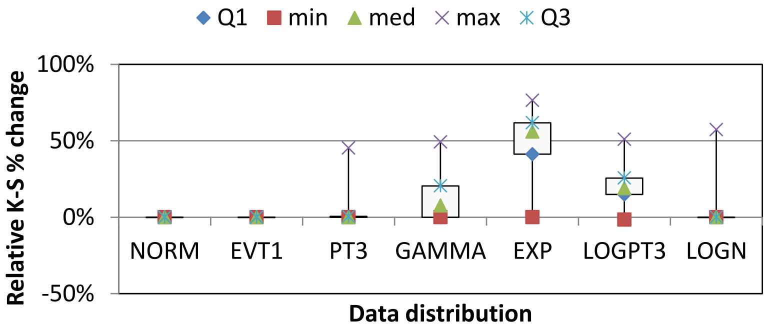
Points falling into S4 ($X_i^j > h_x,\ h_y \geq y^j \geq 0$) have identical circumstances tothose in S2, however, the impact due to the boundary issue is only on they-axis, therefore the corresponding expression is

$$\hat{f}(X_i, y; \boldsymbol{H}) = \frac{1}{n}\sum_{j=1}^{n}\left[K_H\left(\begin{bmatrix} X_i \\ y \end{bmatrix} - \begin{bmatrix} X_i^j \\ y^j \end{bmatrix}\right) + K_H\left(\begin{bmatrix} X_i \\ y \end{bmatrix} - \begin{bmatrix} X_i^j \\ -y^j \end{bmatrix}\right)\right]; X_i \ > h_x, \qquad h_y \geq y \geq 0$$

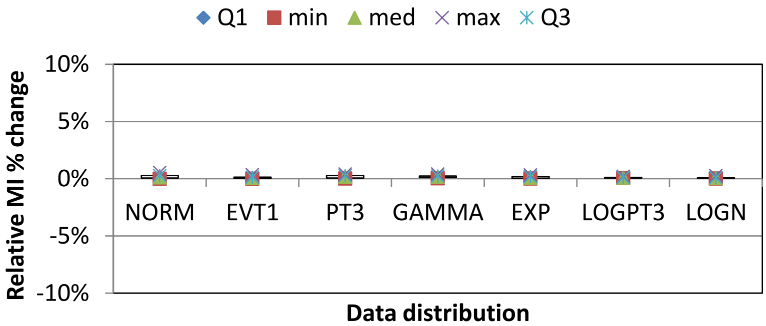
where points in S4 are 'reflected' into Quadrant IV, so that the underestimated density near the boundary (x-axis) can be ameliorated.

In addition, any points outside of Quadrant I result in a density of zero. By summarising all scenarios described above, the bivariate reflection correction can be derived as shown in Eq. (7).

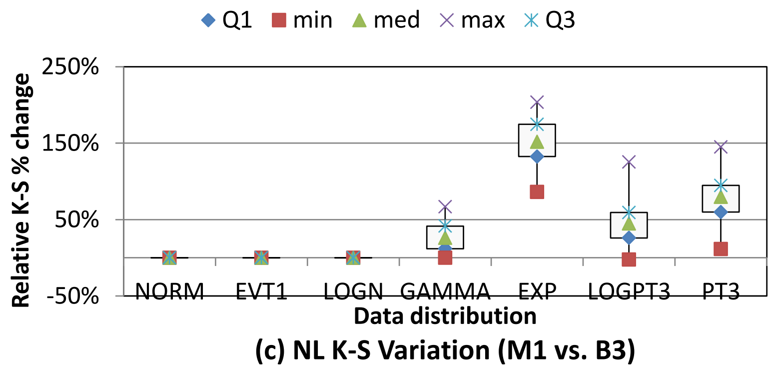
A.2 Supplementary figures and tables



(a) TEAR10 K-S Variation (M1 vs. B3)



(b) TEAR10 MI Variation (M1 vs. B3)



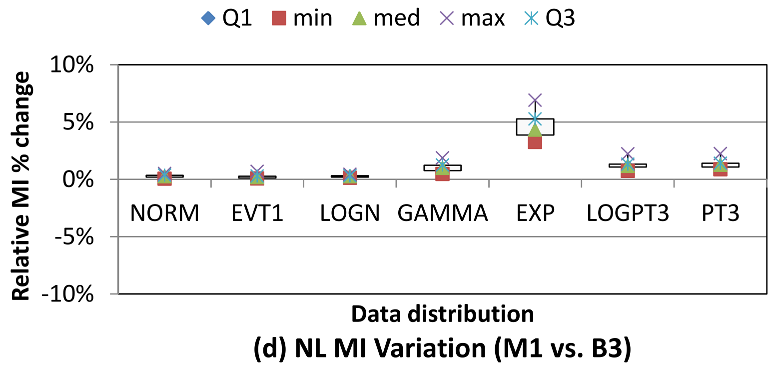
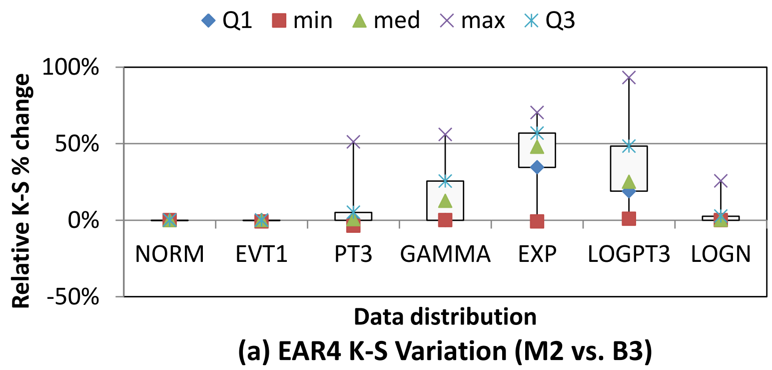
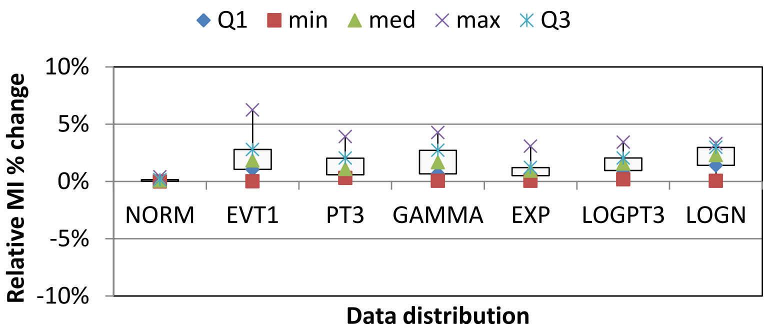
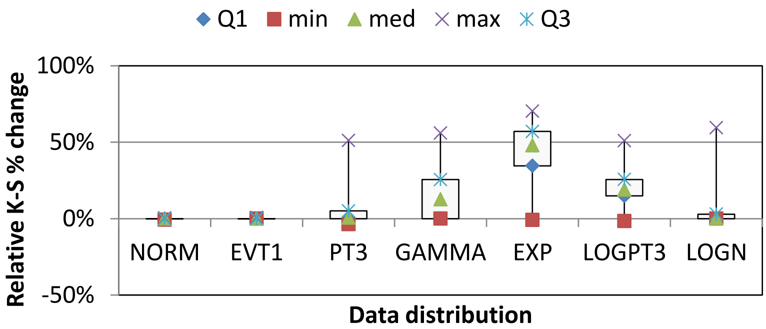


Fig. A.2.1. Relative change of K-S and MI in-between M1 and B3 for TEAR10 and NL models

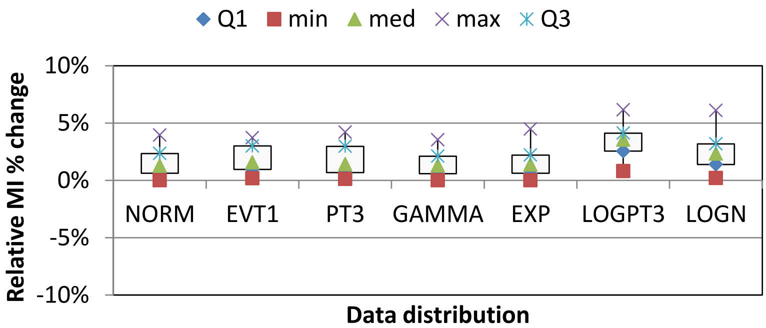




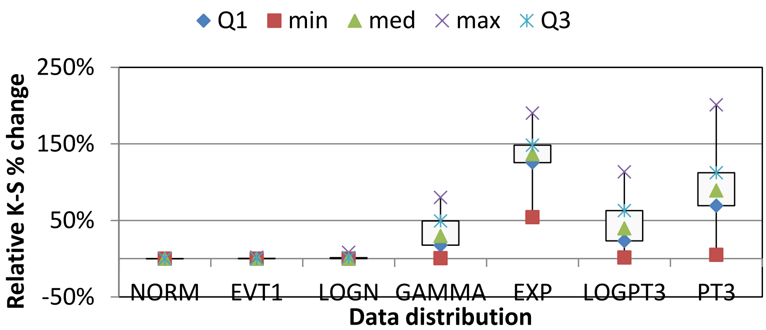
(b) EAR4 MI Variation (M2 vs. B3)



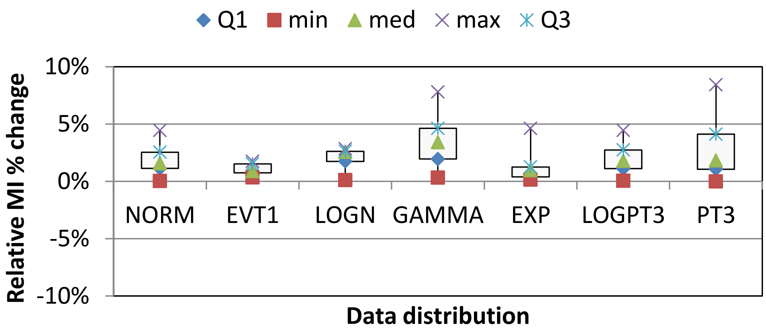
(c) TEAR10 K-S Variation (M2 vs. B3)



(d) TEAR10 MI Variation (M2 vs. B3)

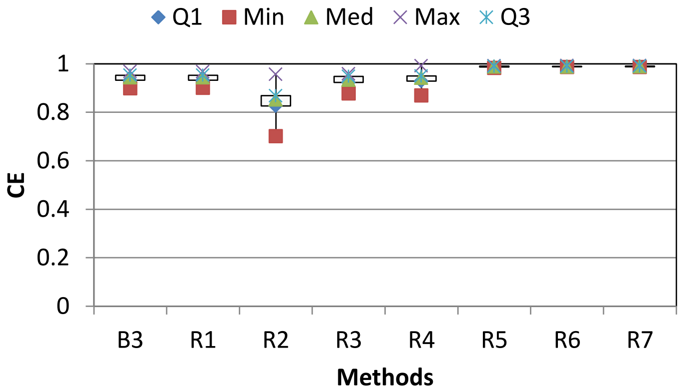


(e) NL K-S Variation (M2 vs. B3)

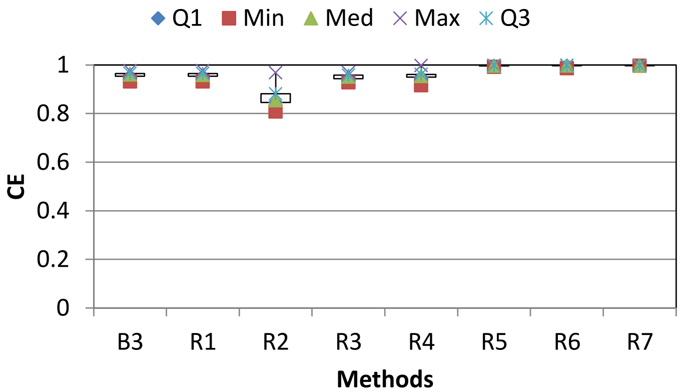


(f) NL MI Variation (M2 vs. B3)

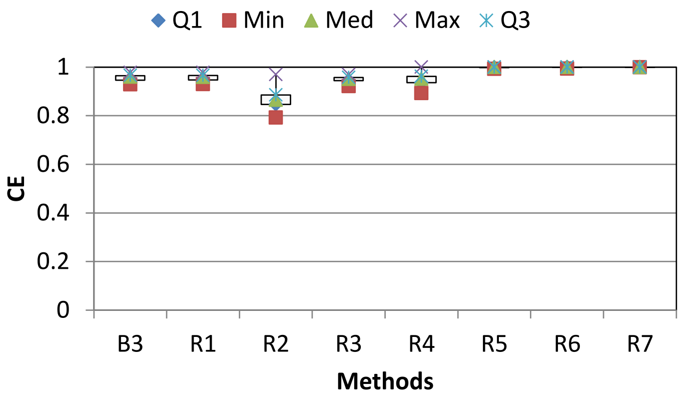
Fig. A.2.2. Relative change of K-S and MI in-between M2 and B3 for EAR4, TEAR10 and NL models



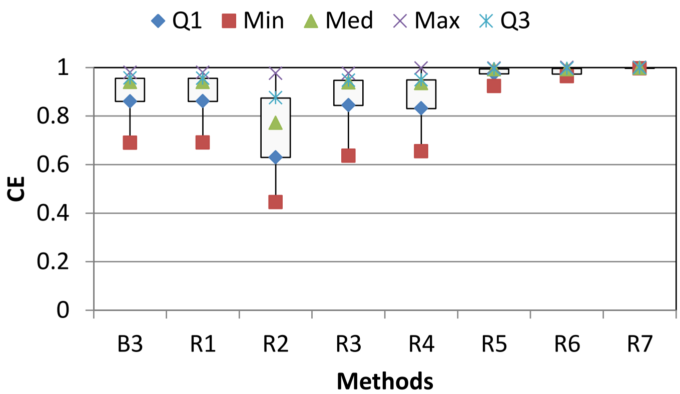
(a) EAR4 EVT1



(b) EAR4 GAMMA

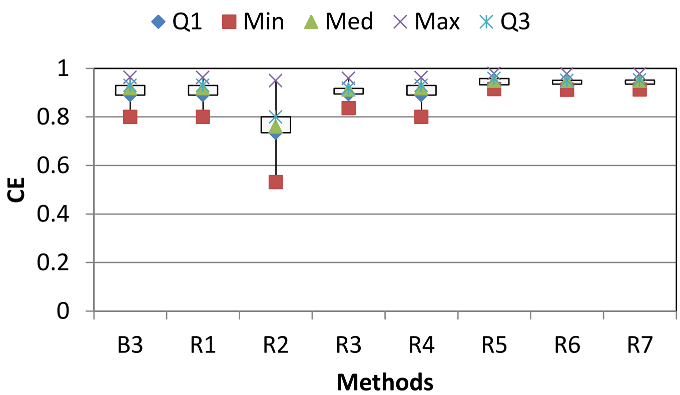


(c) EAR4 PT3

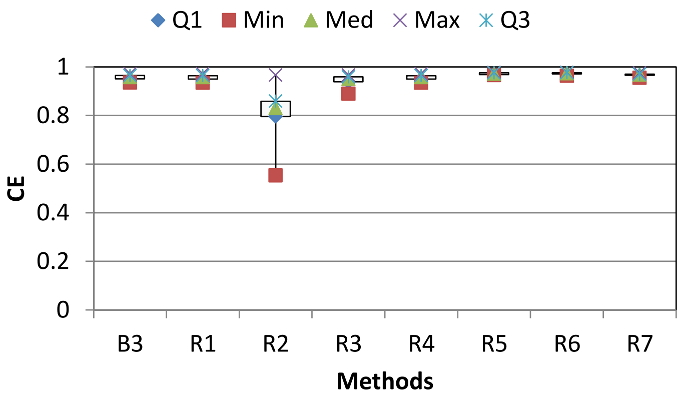


(d) EAR4 LOGPT3

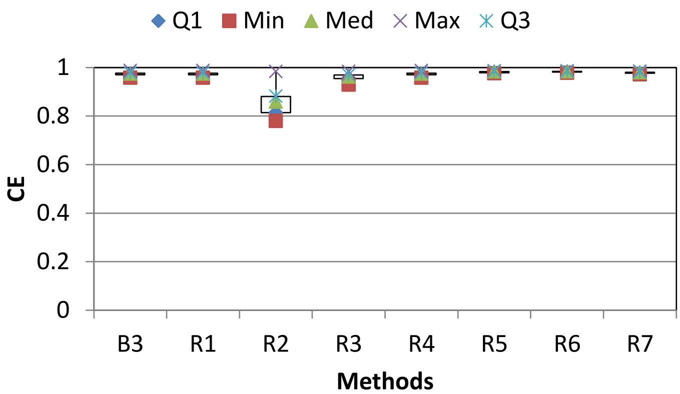
Fig. A.2.3. Accuracy of residual estimation with alternative estimators for EAR4 model (other 4 cases)



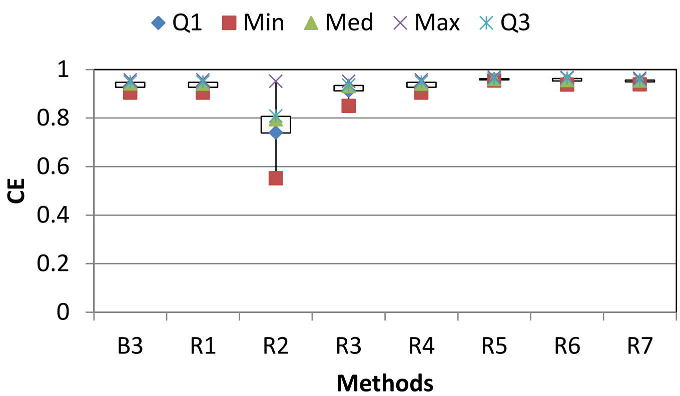
(a) TEAR10 EVT1



(b) TEAR10 GAMMA

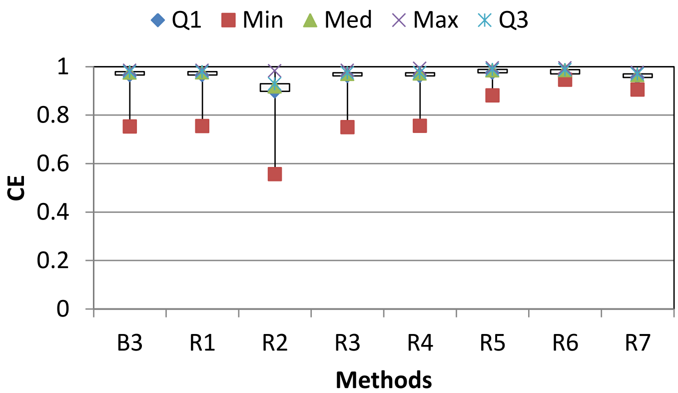


(c) TEAR10 PT3

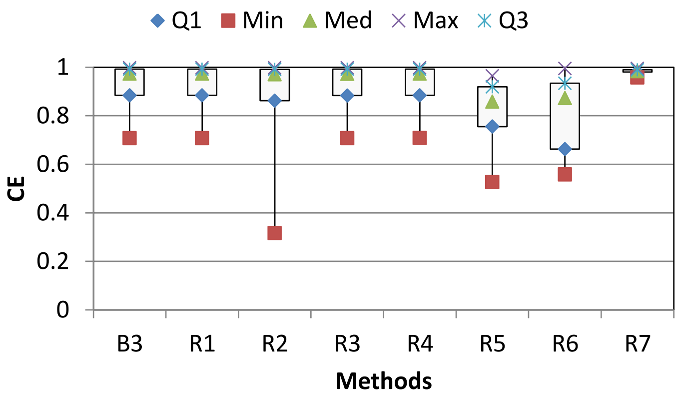


(d) TEAR10 EXP

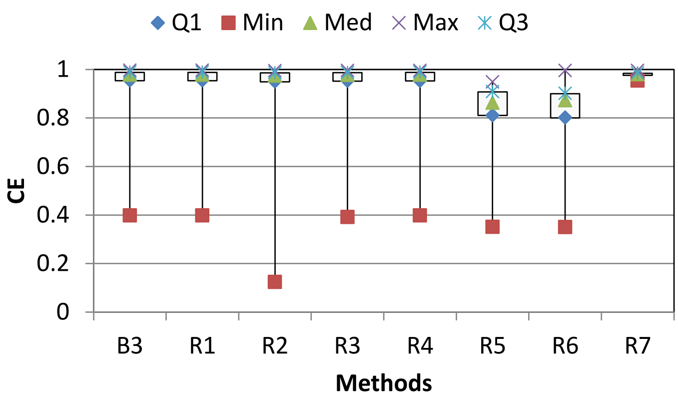
Fig. A.2.4. Accuracy of residual estimation with alternative estimators for TEAR10 model (other 4 cases)



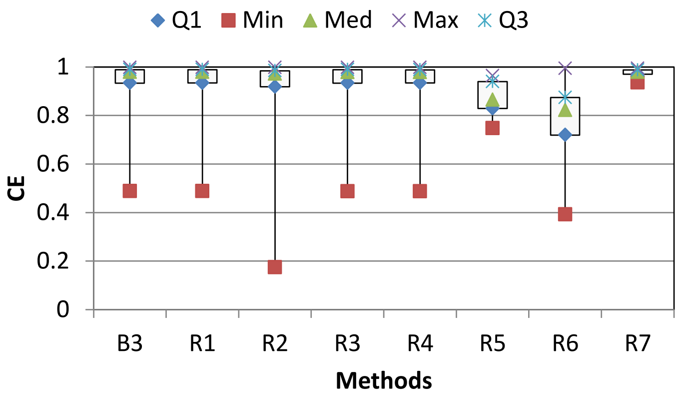
(a) NL EVT1



(b) NL GAMMA



(c) NL EXP



(d) NL PT3

Fig. A.2.5. Accuracy of residual estimation with alternative estimators for NL model (other 4 cases)

REFERENCES

Abrahart, R., Heppenstall, A.J., See, L.M., 2007. Timing error correction procedure applied to neural network rainfall—runoff modelling. Hydrological Sciences Journal 52(3) 414-431.

Abrahart, R.J., Anctil, F., Coulibaly, P., Dawson, C.W., Mount, N.J., See, L.M., Shamseldin, A.Y., Solomatine, D.P., Toth, E., Wilby, R.L., 2012. Two decades of anarchy? Emerging themes and outstanding challenges for neural network river forecasting. Progress in Physical Geography 36(4) 480-513.

Abramson, I.S., 1982. On bandwidth variation in kernel estimates-a square root law. The Annals of Statistics 10(4) 1217-1223.

Adeloye, A.J., Rustum, R., Kariyama, I.D., 2012. Neural computing modeling of the reference crop evapotranspiration. Environmental Modelling and Software 29(1) 61-73.

Akaike, H., 1974. A new look at the statistical model identification. Automatic Control, IEEE Transactions on 19(6) 716-723.

ASCE, 2000a. Artificial neural networks in hydrology II: hydrology applications. Hydrologic Engineering 5(2) 124-137.

ASCE, 2000b.Artificial neural networks in hydrology. I: Preliminary concepts. Hydrologic Engineering 5(2) 115-123.

Bennett, N.D., Croke, B.F., Guariso, G., Guillaume, J.H., Hamilton, S.H., Jakeman, A.J., Marsili-Libelli, S., Newham, L.T., Norton, J.P., Perrin, C., 2013.Characterising performance of environmental models. Environmental Modelling and Software 40 1-20.

Bowden, G.J., 2003. Forecasting Water Resources Variables Using Artificial Neural Networks, School of Civil, Environmental & Mining, Doctor of Philosophy Thesis. The University of Adelaide.

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

Bowden, G.J., Maier, H.R., Dandy, G.C., 2005b. Input determination for neural network models in water resources applications. Part 2. Case study: forecasting salinity in a river. Journal of Hydrology 301(1-4) 93-107.

Bowden, G.J., Maier, H.R., Dandy, G.C., 2012. Real - time deployment of artificial neural network forecasting models: Understanding the range of applicability. Water Resources Research 48(10) DOI:10.1029/2012WR011984.

Cacoullos, T., 1966.Estimation of a multivariate density. Annals of the Institute of Statistical Mathematics 18(1) 179-189.

Chow, V.T., Maidment, D.R., Mays, L.R., 1988. Applied Hydrology. McGraw-Hill Inc., New York.

Cigizoglu, H.K., Alp, M., 2006.Generalized regression neural network in modelling river sediment yield. Advances in Engineering Software 37(2) 63-68.

Cowling, A., Hall, P., 1996.On pseudodata methods for removing boundary effects in kernel density estimation.Journal of the Royal Statistical Society. Series B (Methodological) 58(3) 551-563.

Dai, J., Sperlich, S., 2010. Simple and effective boundary correction for kernel densities and regression with an application to the world income and engel curve estimation. Computational Statistics and Data Analysis 54(11) 2487-2497.

Dawson, C.W, Wilby, R., 2001. Hydrological modelling using artificial neural networks. Progress in Physical Geography 25(1) 80-108.

Dawson, C.W., Abrahart, R.J., See, L.M., 2007. HydroTest: a web-based toolbox of evaluation metrics for the standardised assessment of hydrological forecasts. Environmental Modelling and Software 22(7) 1034-1052.

Duong, T., Hazelton, M., 2003. Plug-in bandwidth matrices for bivariate kernel density estimation. J. Nonparametr. Stat. 15 (1) 17–30.

Fan, J., 1992.Design-adaptive nonparametric regression.Journal of the American Statistical Association 87(420) 998-1004.

Fan, J., Gijbels, I., 1996. Local polynomial modelling and its applications: monographs on statistics and applied probability 66. CRC Press, London, UK.

Fernando, T.M.K.G., Maier, H.R., Dandy, G.C., 2009. Selection of input variables for data driven models: An average shifted histogram partial mutual information estimator approach. Journal of Hydrology 367(3) 165-176.

Galelli, S., Castelletti, A., 2013. Tree‐based iterative input variable selection for hydrological modeling. Water Resources Research 49(7) 4295-4310.

Galelli, S., Humphrey, G.B., Maier, H.R., Castelletti, A., Dandy, G.C., Gibbs, M.S., 2014.An evaluation framework for input variable selection algorithms for environmental data-driven models. Environmental Modelling and Software 62 33-51.

Gasser, T., Müller, H.-G., 1979. Kernel estimation of regression functions. Springer, Berlin.

Gasser, T., Müller, H., Mammitzsch, V., 1985.Kernels for nonparametric curve estimation.Journal of the Royal Statistical Society. Series B (Methodological) 47(2) 238-252.

Gibbs, M.S., Morgan, N., Maier, H.R., Dandy, G.C., Nixon, J., Holmes, M., 2006.Investigation into the relationship between chlorine decay and water distribution parameters using data driven methods.Mathematical and Computer Modelling 44(5) 485-498.

Guyon, I., Elisseeff, A., 2003. An introduction to variable and feature selection. The Journal of Machine Learning Research 3 1157-1182.

Hall, P., Marron, J.S., Park, B.U., 1992. Smoothed cross-validation. Probability Theory and Related Fields 92(1) 1-20.

Hall, P., Park, B.U., 2002. New methods for bias correction at endpoints and boundaries. Annals of Statistics 30(5) 1460-1479.

Hall, P., Wehrly, T.E., 1991. A geometrical method for removing edge effects from kernel-type nonparametric regression estimators. Journal of the American Statistical Association 86(415) 665-672.

Harrold, T., Sharma, A., Sheather, S., 2001. Selection of a kernel bandwidth for measuring dependence in hydrologic time series using the mutual information criterion. Stochastic Environmental Research and Risk Assessment 15(4) 310-324.

Hazelton, M., Marshall, J., 2009. Linear boundary kernels for bivariate density estimation. Stat. Probab. Lett. 79, 999–1003.

He, J., Valeo, C., Chu, A., Neumann, N.F., 2011. Prediction of event-based stormwater runoff quantity and quality by ANNs developed using PMI-based input selection. Journal of Hydrology 400(1-2) 10-23.

Humphrey, G.B., Galelli, S., Castelletti, A., Maier, H.R., Dandy, G.C., Gibbs, M.S., 2014. A new evaluation framework for input variable selection algorithms used in environmental modelling, In: D.P. Ames, N.Q. (Ed.), 7th International Congress on Environmental Modelling and Software: San Diego, California, USA.

Ibarra-Berastegi, G., Elias, A., Barona, A., Saenz, J., Ezcurra, A., Diaz de Argandoña, J., 2008. From diagnosis to prognosis for forecasting air pollution using neural networks: Air pollution monitoring in Bilbao. Environmental Modelling and Software 23(5) 622-637.

Jain, A., Srinivasulu, S., 2004. Development of effective and efficient rainfall-runoff models using integration of deterministic, real-coded genetic algorithms and artificial neural network techniques. Water Resources Research 40(4) W04302.

Jakeman, A., Letcher, R., Norton, J., 2006. Ten iterative steps in development and evaluation of environmental models. Environmental Modelling and Software 21(5) 602-614.

John, R., 1984. Boundary modification for kernel regression. Communications in Statistics-Theory and Methods 13(7) 893-900.

Karunamuni, R.J., Alberts, T., 2005.A generalized reflection method of boundary correction in kernel density estimation. Canadian Journal of Statistics 33(4) 497-509.

Kingston, G.B., Lambert, M.F., Maier, H.R., 2005. Bayesian training of artificial neural networks used for water resources modeling. Water Resources Research 41(12) W12409.

Li, X., Maier, H.R., Zecchin, A.C., 2015. Improved PMI-based input variable selection approach for artificial neural network and other data driven environmental and water resource models. Environmental Modelling and Software65 15-29 DOI: 10.1016/j.envsoft.2014.11.028.

Li, X., Zecchin, A.C., Maier, H.R., 2014. Selection of smoothing parameter estimators for general regression neural networks - Applications to hydrological and water resources modelling. Environmental Modelling and Software 59 162-186 DOI: 110.1016/j.envsoft. 2014.1005.1010.

Luccarini, L., Bragadin, G.L., Colombini, G., Mancini, M., Mello, P., Montali, M., Sottara, D., 2010. Formal verification of wastewater treatment processes using events detected from continuous signals by means of artificial neural networks. Case study: SBR plant. Environmental Modelling and Software 25(5) 648-660.

Maier, H.R., Dandy, G.C., 1997. Modelling cyanobacteria (blue-green algae) in the River Murray using artificial neural networks. Mathematics and Computers in Simulation 43(3) 377-386.

Maier, H.R., 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., Dandy, G.C., 1996. The use of artificial neural networks for the prediction of water quality parameters. Water Resources Research 32(4) 1013-1022.

Maier, H.R., Jain, A., Dandy, G.C., 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 and Software 25(8) 891-909.

Maier, H.R., Morgan, N., Chow, C.W., 2004. Use of artificial neural networks for predicting optimal alum doses and treated water quality parameters. Environmental Modelling and Software 19(5) 485-494.

Marron, J.S., Ruppert, D., 1994.Transformations to reduce boundary bias in kernel density estimation.Journal of the Royal Statistical Society. Series B (Methodological) 56(4) 653-671.

Marshall, J.C., Hazelton, M.L., 2010. Boundary kernels for adaptive density estimators on regions with irregular boundaries. Journal of Multivariate Analysis 101(4) 949-963.

Masry, E., 1996. Multivariate local polynomial regression for time series: uniform strong consistency and rates. Journal of Time Series Analysis 17(6) 571-599.

May, R., Dandy, G., Maier, H., 2011. Review of input variable selection methods for artificial neural networks, In: InTech (Ed.), Artificial neural networks—methodological advances and biomedical applications: Rijeka, Croatia, pp. 19-44.

May, R.J., Dandy, G.C., Maier, H.R., Nixon, J.B., 2008a.Application of partial mutual information variable selection to ANN forecasting of water quality in water distribution systems. Environmental Modelling and Software 23(10) 1289-1299.

May, R.J., Maier, H.R., Dandy, G.C., 2010. Data splitting for artificial neural networks using SOM-based stratified sampling. Neural Networks 23(2) 283-294.

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

Millie, D.F., Weckman, G.R., Young II, W.A., Ivey, J.E., Carrick, H.J., Fahnenstiel, G.L., 2012. Modelingmicroalgal abundance with artificial neural networks: Demonstration of a heuristic ‘Grey-Box’todeconvolve and quantify environmental influences. Environmental Modelling and Software 38 27-39.

Muñoz-Mas, R., Martínez-Capel, F., Garófano-Gómez, V., Mouton, A., 2014. Application of Probabilistic Neural Networks to microhabitat suitability modelling for adult brown trout (Salmotrutta L.) in Iberian rivers. Environmental Modelling and Software 59 30-43.

Ozkaya, B., Demir, A., Bilgili, M.S., 2007. Neural network prediction model for the methane fraction in biogas from field-scale landfill bioreactors. Environmental Modelling and Software 22(6) 815-822.

Park, B.U., Marron, J.S., 1990. Comparison of data-driven bandwidth selectors. Journal of the American Statistical Association 85(409) 66-72.

Parsons, F., Wirsching, P., 1982.A Kolmogorov-Smirnov goodness-of-fit test for the two-parameter weibull distribution when the parameters are estimated from the data. Microelectronics Reliability 22(2) 163-167.

Parzen, E., 1962. On estimation of a probability density function and mode. Annals of Mathematical Statistics 33(3) 1065-1076.

Pradhan, B., Lee, S., 2010. Landslide susceptibility assessment and factor effect analysis: backpropagation artificial neural networks and their comparison with frequency ratio and bivariate logistic regression modelling. Environmental Modelling and Software 25(6) 747-759.

Press, W.H., Flannery, B.P., Teukolsky, S.A., Vetterling, W.T., 1992.Numerical Recipes in FORTRAN 77. In: Fortran Numerical Recipes: The Art of Scientific Computing. vol. 1.Cambridge university press.

Rudemo, M., 1982.Empirical choice of histograms and kernel density estimators. Scandinavian Journal of Statistics 9(2) 65-78.

Ruppert, D., Wand, M.P., 1994. Multivariate locally weighted least squares regression. The Annals of Statistics 22(3) 1346-1370.

Santhosh, D., Srinivas, V., 2013. Bivariate frequency analysis of floods using a diffusion based kernel density estimator. Water Resources Research 49(12) 8328-8343.

Schuster, E.F., 1985. Incorporating support constraints into nonparametric estimators of densities. Communications in Statistics-Theory and Methods 14(5) 1123-1136.

Scott, D.W., 1992. Multivariate density estimation and visualization.Handbook of Computational Statistics.Springer, New York, USA.

Scott, D.W., 2004. Multivariate density estimation and visualization.Handbook of Computational Statistics. New York: Springer 517-538.

Scott, D.W., Terrell, G.R., 1987. Biased and unbiased cross-validation in density estimation. Journal of the American Statistical Association 82(400) 1131-1146.

Shannon, C.E., 1948. A mathematical theory of communication. The Bell System Technical Journal 33(27) 379-423 & 623-656.

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.

Silverman, B.W., 1986. Density estimation for statistics and data analysis.CRC press, London, UK.

Specht, D.F., 1991. A general regression neural network. Neural Networks, IEEE Transactions on 2(6) 568-576.

Srinivasulu, S., Jain, A., 2006.A comparative analysis of training methods for artificial neural network rainfall–runoff models.Applied Soft Computing 6(3) 295-306.

Wand, M.P., Jones, M.C., 1995.Kernel smoothing.Chapman & Hall, London, UK.

Wolfs, V., Willems, P., 2014.Development of discharge-stage curves affected by hysteresis using time varying models, model trees and neural networks. Environmental Modelling and Software 55 107-119.

Wu, W., Dandy, G., Maier, H., 2014a.Optimal Control of Total Chlorine and Free Ammonia Levels in a Water Transmission Pipeline Using Artificial Neural Networks and Genetic Algorithms. Journal of Water Resources Planning and Management DOI: 10.1061/(ASCE)WR.1943-5452.0000486.

Wu, W., Dandy, G.C., Maier, H.R., 2014b. Protocol for developing ANN models and its application to the assessment of the quality of the ANN model development process in drinking water quality modelling. Environmental Modelling and Software 54 108-127.

Wu, W., May, R.J., Maier, H.R., Dandy, G.C., 2013. A benchmarking approach for comparing data splitting methods for modeling water resources parameters using artificial neural networks. Water Resources Research 49(11) 7598-7614.

Young II, W.A., Millie, D.F., Weckman, G.R., Anderson, J.S., Klarer, D.M., Fahnenstiel, G.L., 2011. Modeling net ecosystem metabolism with an artificial neural network and Bayesian belief network. Environmental Modelling and Software 26(10) 1199-1210.

Zhang, S., Karunamuni, R.J., 1998.On kernel density estimation near endpoints. Journal of Statistical Planning and Inference 70(2) 301-316.

Zhang, S., Karunamuni, R.J., 2000.On nonparametric density estimation at the boundary. Journal of Nonparametric Statistics 12(2) 197-221.