THE UNIVERSITY of ADELAIDE

Improving Partial Mutual Information Based Input Variable Selection for Data Driven Environmental and Water Resources Models

XUYUAN LI

B.E. (Hons)

Thesis submitted in fulfilment of the requirements for the degree of Doctor of Philosophy

The University of Adelaide
Faculty of Engineering, Computer and Mathematical Sciences
School of Civil, Environmental and Mining Engineering

Improving Partial Mutual Information Based Input Variable Selection for Data Driven Environmental and Water Resources Models

By:

Xuyuan Li, *B.E. (Hons)*

Supervised by:

Professor Holger R. Maier, *B.E. (Hons), Ph.D., MIEAust, CPEng (NPER) Professor of Integrated Water Systems Engineering, Associate Editor, Water Resources Research, Member of Editorial Board, Environmental Modelling and Software, School of Civil, Environmental & Mining Engineering, The University of Adelaide*

Dr. Aaron C. Zecchin, *Ph.D., B.E. (Hons), B.Sc. (Math & Comp. Sci.), Senior lecturer, School of Civil, Environmental & Mining Engineering, The University of Adelaide*

Thesis submitted in fulfillment of the requirements for the degree of
Doctor of Philosophy

School of Civil, Environmental & Mining Engineering
Faculty of Engineering, Computer and Mathematical Sciences
The University of Adelaide
North Terrace, Adelaide, SA 5005, Australia
Phone: +61 8 8313 1575
Fax : +61 8 8303 4359
Email: xli@civeng.adelaide.edu.au, xliadelaide@gmail.com

TABLE OF CONTENT

TABLE OF CONTENT

TABLE OF CONTENT

LIST OF FIGURES

LIST OF FIGURES

LIST OF FIGURES

LIST OF TABLES

LIST OF TABLES

LIST OF TABLES

NOMENCLATURE & ABBREVIATIONS

Symbols

$\widehat{m}_{v_i}(X_{i^*}^j)$: residual estimate of v_i based on X_{i^*}

$(\boldsymbol{X}^j, y^j)$: observed pairs of input and output data

$\hat{\varphi}_4(g) = n^{-1}\sum_{i=1}^{n}\hat{L}^{(4)}(X^i; g)$: fourth order integrated squared density derivative

$F_{emp}(X_i^j)$: empirical CDF of the input variable estimated by a histogram

$F_{est}(X_i^j)$: estimated kernel based CDF of the input variable

$I_{X_i,y}$: mutual information

$I_{v_i,u}$: partial mutual information

$\widehat{ISB}(h)$: estimation of the integrated squared bias

$K^{(n)}$: n^{th} derivative of kernel function K

$\widetilde{R(f'')}$: approximation of the integrated squared second derivative of f

$S_{x,i}^2$: sample variance of the input X_i

$S_{xy,i}$: covariance between input X_i and output y

S_y^2: sample variance of output y

X_{i^*} selected inputs

$X^{(j)}$: *j-th* data point formed by the interpolated and original data points

X_s: significant input set

$\hat{f}(\boldsymbol{X}, y)$: estimation of the joint probability density function between inputs $\boldsymbol{X}$ and output y

p_{t-n}: exogenous input with lag n

$\hat{y}$: estimation of the actual output y

$\bar{y}$: sample mean of the observations

$\boldsymbol{e_1}$: vector having 1 in the first entry and 0 elsewhere

ε_t: introduced error term

$\mu_2(K) = \int x^2 K(x)dx$: second moment of K

$\mu_k(L) = \int u^k L(u)du$: *k-th* moment of L

$\mu_n(K)$: n^{th} moment of kernel function K

$\rho_{xy,i}$: correlation coefficient between input X_i and output y

σ_i: sample standard deviation of the X_i^j

$*$: convolution operation

h: kernel bandwidth

K: kernel function

$B(u; h_x)$: univariate boundary kernel with bandwidth h_x and variable $u = (X_i - X_i^j)/h_x$

$B(u, v; \boldsymbol{H})$: bivariate boundary kernel with bandwidth matrix $\boldsymbol{H}$ and vector (u, v) where $u = (X_i - X_i^j)/h_x$ and $v = (y - y^j)/h_y$

$E[y|\boldsymbol{X}]$: conditional expectation of output y given input $\boldsymbol{X}$

L: pilot kernel

$O(h)$: bias of density function

$P(t)$: lagged effective rainfall

$Q(t-1)$: lagged runoff

$R(K) = \int [K(x)]^2 dx$: integrated square of kernel function

a: left boundary of kernel density

c: right boundary of kernel density

e: number of effective inputs

$f(\boldsymbol{X}, y)$: joint probability density function between inputs $\boldsymbol{X}$and output y

g: pilot bandwidth

k: kurtosis

k: order of pilot kernel L

$m(x)$: regression function

m: number of inputs

n: number of observations

r: stage number of L

s: skewness

sup: supremum function

$\boldsymbol{H} = h_i^2 \begin{bmatrix} S_{x,i}^2 & S_{xy,i} \\ S_{xy,i} & S_y^2 \end{bmatrix}$: bivariate bandwidth matrix

$\boldsymbol{X} = [X_1 \ldots X_m]^T$: input variables

$\boldsymbol{h} = [h_1 \quad \cdots \quad h_n]^T$: kernel bandwidth vector

Abbreviation

ACF: auto-correlation function

AIC: Akaike information criterion

AMISE: asymptotic mean integrated squared error

ANNs: artificial neural networks

BCV: biased cross validation

BCVDPI: a combination of BCV and DPI

BE: backward elimination (pruning)

BJ: Box-Jenkins

BK: boundary kernel

BP: back-propagation algorithm

CE: coefficient of efficiency

CK: conventional kernel

CSR: correct selection rate

CT: computational time

DELSA: distributed evaluation of local sensitivity analysis

DPI: 2-stage direct plug-in

EAR4: exogenous auto-regressive time series model (with time order up to 4)

EMISE: exact mean integrated squared error

ES: exhaustive search

ETC: empirical translation correction

EVT1: extreme value type I distribution

EXP: exponential distribution

FS: forward selection

GAMMA: gamma distribution

GRNN: general regression neural network

GRR: Gaussian reference rule

GSS: golden section search

HS: heuristic search

ICAIVS: hybrid independent component analysis and input variable selection filter

IIS: tree-based iterative input variable selection

IoAd: index of agreement

IVS: input variable selection

KDEs: kernel density estimates

K-S: Kolmogorov-Smirnov statistic

KT: kernel transformation

L1UL: Lock 1 upper river level

LBE: local bandwidth (enlarging)

LBR: local bandwidth (reducing)

LHOP: local high order polynomial

LLM: local linear method

LLP: local linear polynomial

LOGN: log-normal distribution

LOGPT3: log-Pearson type III distribution

LOS: Loxton river salinity

LQP: local quadratic polynomial

LSCV: least squared cross validation

MAE: mean absolute error

MAS: Mannum river salinity

MBS: Murray Bridge river salinity

MCE: modified coefficient of efficiency

MI: mutual information

MIoAd: modified index of agreement

MLPANNs: multi-layer perceptron artificial neural networks

MLPs: multi-layer perceptrons

MOS: Morgan river salinity

MPI: modified persistence index

MVC: multi-variable calibration

MVCA: multi-variable calibration with mean absolute error as the objective function

MVCS: multi-variable calibration with squared error as the objective function

NL: nonlinear input-output function

NORM: normal distribution

NS: normal scale

OM: optimisation method

PA: pseudo-data approach

PACF: partial auto-correlation function

PC: partial correlation

PCA: principal component analysis

PDF: probability density function

PI: persistence index

PMI: partial mutual information

PNNs: probabilistic neural networks

PSO: particle swarm optimisation

PT3: Pearson type III distribution

RBFs: radial basis functions

RC: reflection correction

RE: residual estimation

RMSE: root mean squared error

RNNs: recurrent neural networks

RVSDEM: recursive variable selection embedded in dynamic emulation models

SCV: smoothed cross validation

SOM-GAGRNN: self-organising map genetic algorithm general regression neural network

SVC: single variable calibration

SVCA: single variable calibration with mean absolute error as the objective function

SVCS: single variable calibration with squared error as the objective function

SVO: single variable optimisation

SVR: single variable regression

TEAR10: threshold exogenous auto-regressive time series model (with time order up to 10)

WAS: Waikerie river salinity

ABSTRACT

Artificial neural networks (ANNs), as one of the most commonly used data driven models for environmental and water resources problems, have been applied successfully and extensively over the last two decades and are still gaining in popularity. Consideration of the methods used in the steps in the development of ANNs, which consist of data collection, data processing, input variable selection, data division, calibration and validation, are vitally important, as ANN model development is based on data, rather than understanding of the underlying physical processes.

Among these methods, input variable selection (IVS) plays a significant role, as the performance of the developed model can be compromised if inputs having a pronounced relationship with the modelled output are omitted. In contrast, calibration becomes extremely challenging and modelling validation, as well as knowledge extraction, are problematic if redundant or superfluous inputs are included. Given the facts explained above, various techniques have been developed for the sake of more accurate IVS.

Partial mutual information (PMI) is one of the most promising approaches to IVS, as it has a number of desirable properties, such as the ability to account for input relevance, the ability to cater to both linear and non-linear input-output relationships and the ability to check the redundancy of selected inputs. PMI is a stepwise input selection algorithm, which only selects one variable per iteration, as part of which the strength of the relationship between each potential input and the output is quantified using mutual information (MI) and input redundancy is accounted for by removing the influence of already selected inputs. This is achieved by developing models between the selected input and the output and assessing the strength of the relationship (in terms of MI) between the remaining potential inputs and the residuals of these models during the next iteration, which is referred to as PMI.

Although PMI IVS has already been applied successfully to a number of studies in hydrological and water resource modelling, present implementations predominantly depend on the assumption that the data used to develop the model follow a Gaussian distribution. This assumption has the

potential to affect two steps in the PMI process, including the estimation of MI/PMI and the estimation of the residuals. In terms of MI/PMI estimation, this requires kernel density estimates of the modelling data to be obtained for the estimation of marginal and joint probability density functions (PDFs), which rely on estimates of kernel bandwidths (or smoothing parameters) and in most studies, the Gaussian reference rule is used for this purpose, which only results in optimal bandwidth estimates if the modelling data follow a Gaussian distribution. However, this is unlikely to be the case when dealing with water resources and other environmental data. In terms of residual estimation (RE), this has generally been done using general regression neural networks (GRNNs), which also require estimates of kernel bandwidths to be obtained and therefore suffers from the same issues as MI/PMI estimation.

The purpose of this thesis is to assess the impact the assumption that the data follow a Gaussian distribution has on the performance of PMI IVS and the efficacy of potential methods for overcoming any problems associated with this assumption. In order to achieve this, a large number of numerical tests are conducted on synthetic data with different degrees of normality and non-linearity, investigating the effectiveness of a range of options for (i) bandwidth estimation (caused by making Gaussian assumptions for non-Gaussian circumstances when adopting kernel based estimations in both MI/PMI and RE) and (ii) for dealing with boundary issues (caused by using a symmetrical kernel for bounded/unsymmetrical data when implementing kernel based estimations in both MI/PMI and RE), as well as methods for RE that do not require kernel density estimates. The results from these tests are used to develop preliminary guidelines for the selection of the most appropriate bandwidth and the most effective treatment of the boundary issue, which are then validated for two water resources case studies with different data properties and problem linearity, including forecasting of river salinity in the River Murray, Australia, and rainfall-runoff modelling in the Kentucky River, USA.

The major research contributions are presented in three journal publications. The motivations underlying these publications include: 1) the development and testing of rigorous and novel analytical procedures for assessing if, and to

what degree, the performances of residual and MI estimates are affected by bandwidth selection and boundary issues; 2) clear explanation of the inaccurate performance of conventional PMI IVS under the influence of bandwidth selection and boundary issues; 3) the development of effective preliminary guidelines based upon synthetic studies dealing with both bandwidth selection and boundary issues under different scenarios categorised by data normality and problem linearity; 4) the development of more robust and reliable PMI IVS software for realistic environmental and water resource problems. Overall, the research outcomes suggest that the performance of PMI IVS is significantly influenced by bandwidth selection and boundary issues and can be effectively improved by following the proposed empirical guidelines, although the findings of this work could 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.

STATEMENT OF ORIGINALITY

I **Xuyuan Li** hereby certify that this work contains no material which has been accepted for the award of any other degree or diploma in any university or other tertiary institution and, to the best of my knowledge and belief, contains no material previously published or written by another person, except where due reference has been made in the text. In addition, I certify that no part of this work will, in the future, be used in a submission for any other degree or diploma in any university or other tertiary institution without the prior approval of the University of Adelaide and where applicable, any partner institution responsible for the joint-award of this degree.

I give consent to this copy of my thesis when deposited in the University Library, being made available for loan and photocopying, subject to the provisions of the Copyright Act 1968.

I acknowledge that copyright of published works contained within this thesis resides with the copyright holder(s) of those works.

I also give permission for the digital version of my thesis to be made available on the web, via the University's digital research repository, the Library catalogue and also through web search engines, unless permission has been granted by the University to restrict access for a period of time.

A list of works contained within this thesis is given in Section 5.3.

Signature_____ Date_____

ACKNOWLEDGEMENTS

I would take this opportunity to gratefully acknowledge my principal supervisor Professor Holger R. Maier and co-supervisor Dr. Aaron C. Zecchin, who have supervised me with invaluable support, encouragement, and patience. I really appreciate the critical and rigorous attitude, timely and effective feedback, great foresight, and comfortable research environment provided by both of my supervisors. I have enjoyed studying under their supervision and learnt a lot from them during my candidature. Without their excellent supervision, the present thesis would hardly be possible.

I would like to sincerely thank Prof. A. Sharma, Dr. G.J. Bowden, Dr. R.J. May and Dr. G.B. Humphrey who brought me into the present research topic and kindly provided their suggestions and research materials.

I would express my appreciation to my fellow postgraduates within the School for sharing their wisdom, experiences, successes, and lessons with me. It was my pleasure to meet Dr. Xun Sun, Dr. Wenyan Wu, Dr. Feifei Zheng, Dr. Liang Huang, Dr. Jeffrey P. Newman, Dr. Christopher Stokes, and Dr. Tao Zhang and have had a number of in-depth academic discussions with them. Special thanks are also given to School Editor Barbara Brougham, Computer Technician Dr. Stephen Carr, all School Administrators, and the research scholarship (AGRS) provided by the University of Adelaide.

Last but not least, I would like to give an immense gratitude to my parents, Mr Xuelong Li and Ms Yongjing Liu, for their altruistic care and support from beginning to end (*ab ovo usque ad mala*).

CHAPTER 1 INTRODUCTION

1.1 Background

1.1.1 ANNs in environmental and water resources modelling

Over the last two decades, artificial neural networks (ANNs) have been applied successfully and extensively to environmental (e.g. Adeloye et al., 2012; Ibarra-Berastegi et al., 2008; Luccarini et al., 2010; Maier and Dandy, 1997b; 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 (e.g. Abrahart et al., 2007; Abrahart et al., 2012; ASCE, 2000a, b; Dawson and Wilby, 2001; Maier and Dandy, 2000b; Maier et al., 2010; Wolfs and Willems, 2014; Wu et al., 2014b) problems, and their popularity is still increasing. The methods used for the development of ANNs are vitally important, as their establishment is based on data rather than underlying physical meaning. Consequently, investigating the methodological issues associated with their development, including data collection, data processing, input variable selection, data division, calibration, validation, and application on real problems (as can be illustrated in Fig. 1.1 *Development of ANNs*), is particularly vital, as suggested and emphasized by Abrahart et al. (2012), Maier et al. (2010), and Wu et al. (2014b).

1.1.2 IVS

Among the steps in the development procedure of ANNs, input variable selection (IVS) plays a crucial role, as the performance of such models can be compromised significantly if either too few or too many inputs are selected (Galelli et al., 2014; May et al., 2010; Wu et al., 2014b). 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 challenge of the IVS problem, a large number of approaches, categorised as either model free or model based, have been developed and refined for the purpose of more accurate IVS (e.g. Galelli et al., 2014; Galelli and Castelletti, 2013; Li et al., 2015; May et al., 2008, 2011; Sharma, 2000), aiming to determine the smallest number of inputs that best charaterise the input-output relationship with the least amount of variable irrelvance or redundancy (Galelli et al., 2014; Guyon and Elisseeff, 2003). Model free approaches determine the significant inputs on the basis of a statistical measure of significance between the candidate inputs and the output, while model based techniques depend on the adoption of an optimization algorithm that is used to determine the combination of input variables that maximizes the performance of a pre-selected data-driven model, in accordance with Maier et al. (2010), Wu et al. (2014), May (2010), and Castelletti et al. (2012b). Reviews of the typically applied IVS methods for ANN based environmental and water resources problems are summarised in Table 1.1 and each approach is categorised and evaluated in the aspects of type, criterion, linearity, computational cost, redundancy check, and optimum, as these are the critical attributes of the IVS approaches. In Table 1.1, the 'type' indicates whether the IVS method is a model based or a model free approach. The 'criterion' identifies the basis on which significant inputs are selected. The 'linearity' reflects whether the IVS approach can be used for linear problems under the linear assumption (which assumes linear input-output relationships) or for non-linear circumstance without the linear assumption. The 'computation cost' quantifies the efficiency of each IVS method. The 'redundancy check' gives an indication of whether the IVS approach removes redundant input variables, which contain useful but repeating information to the output. The 'optimum' demonstrates the convergence of the IVS method and shows whether the selected significant input variables are a result of local optima (the combination of input variables

that only outperforms some of other possible combinations in terms of describing the output) or global optima (the combination of input variables that outperforms all other possible combinations in terms of describing the output). Details of each IVS approach listed in Table 1.1 can be obtained in the corresponding reference provided in the table. As can be illustrated in Table 1,1, among the various 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 (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., 2013, 2014).

1.1.3 PMI IVS

The partial mutual information based input variable selection (PMI IVS) was introduced by Sharma (2000) and is based on Shannon's principle (Shannon, 1948). As illustrated in Fig. 1.1 (PMI-based IVS), the **first step** is to procure candidate inputs $\boldsymbol{X}$ and output(s) y from the available data in accordance with an understanding of the system. Let: $\boldsymbol{X} = [X_1 \dots X_m]^T$ 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 $j = 1, \dots, n$, where n is the number of observations, $\boldsymbol{X}^j = \left[X_1^j \dots X_m^j\right]^T$ are the observed input data and y^j are the observed output data.

The **second step** is to estimate the marginal PDF of each individual input $f(X_i)$ and the output $f(y)$. The PDF is approximated by kernel density estimation (KDE) in accordance with

$$\hat{f}(X_i) = \frac{1}{n}\sum_{j=1}^{n} K_h(X_i - X_i^j) \tag{1.1}$$

The kernel type K_h used in Eq. (1.1) is the most commonly used Gaussian kernel since the selection of kernel type has negligible impact on the accuracy of KDE (May et al., 2008b; Scott, 1992; Wand and Jones, 1995). The expression of the 1D Gaussian kernel is

$$K_h(\boldsymbol{X}) = \frac{1}{(\sqrt{2\pi}|h|)} exp\left(-\frac{\boldsymbol{X}^2}{2h^2}\right) \quad (1.2)$$

In Eq. (1.2), h is the univariate kernel bandwidth, which determines the accuracy of the KDE (Duong and Hazelton, 2003; Scott, 1992; Wand and Jones, 1995). This single dimensional bandwidth, used for the marginal PDF estimation, directly contributes to the bandwidth matrix used for the joint PDF estimation (as explained later).

The **third step** is to calculate the joint PDF $f(X_i, y)$ between the *i*-th input and the output, which requires the development of a 2-D bandwidth matrix for the joint KDE. The currently used bivariate bandwidth matrix for standardised data is

$$\boldsymbol{H} = h_i^2 \begin{bmatrix} S_{x,i}^2 & S_{xy,i} \\ S_{xy,i} & S_y^2 \end{bmatrix} \quad (1.3)$$

where $S_{x,i}^2$ is the sample variance of the input X_i; $S_{xy,i}$ is the covariance between input $X_i$and output y, S_y^2 is the sample variance of the output y, and $h_i(h_i = h_{x,i} = h_y)$ is the estimated 1-D kernel bandwidth if the data are standardised, or for non-standardised data

$$\boldsymbol{H} = \begin{bmatrix} h_{x,i}^2 & \rho_{xy,i} h_{x,i} h_y \\ \rho_{xy,i} h_{x,i} h_y & h_y^2 \end{bmatrix} \quad (1.4)$$

(known as a hybrid class of bandwidth matrix), where $\rho_{xy,i}$ is the correlation coefficient between input X_i and output y. According to Wand and Jones (1993), the diagonal terms of the bandwidth matrix adjust the shape of the joint PDF, while the off-diagonal terms control the orientation. The empirical joint density of the i-th X_i input and the output y can be estimated by the Gaussian kernel-based estimator as

$$\hat{f}(X_i, y) = \frac{1}{n}\sum_{j=1}^{n} K_{\boldsymbol{H}}\left(\begin{bmatrix} X_i \\ y \end{bmatrix} - \begin{bmatrix} X_i^j \\ y^j \end{bmatrix}\right) \quad (1.5)$$

where the multivariate kernel is given by

$$K_H(\boldsymbol{X}) = \frac{1}{(\sqrt{(2\pi)^m|\boldsymbol{H}|})} \exp\left[-\frac{1}{2}\boldsymbol{X}^T\boldsymbol{H}^{-1}\boldsymbol{X}\right] \quad (1.6)$$

It should be noted that this approximation is commonly known as the Parzen window density estimation (Cacoullos, 1966; Parzen, 1962). This is valid, however, only if the underlying density is continuous and the first partial derivative at any $\boldsymbol{X}$ is small.

According to Shannon (1948), MI, which quantifies the reduction in uncertainty with respect to *y* due to observation of X_i, is then approximated as

$$I_{X_i,y} \approx \frac{1}{n}\sum_{j=1}^{n} log[\frac{f(X_i^j,y^j)}{f(X_i^j)f(y^j)}] \tag{1.7}$$

(marginal PDFs $f(X_i^j)$ and $f(y^j)$ are as defined in Eq. (1.1)) in the **fourth step**. The input with the greatest MI value is the most significant input among the candidate inputs. The significant inputs are selected by means of these four steps during the first run of the algorithm and added to the significant input set X_s, that is, the set is updated to include $X_{i^*} \in X_s$ where $i^* = argmax\{I_{v_i,u}\}$.

In order to remove any redundant information, RE is required in the **fifth step**. RE is at the core of the 'partial' aspect of PMI IVS and the mutual information shared between the residual inputs and output is called PMI (the term used after the 1st iteration of the PMI IVS). Typically, a general regression neural network (GRNN) (Specht, 1991) is used as the residual estimator in PMI IVS (e.g. May et al., 2008b; He et al., 2011). The residual estimator is used to update the inputs and output by removing the influence of the selected input variables. The updated input is defined as the difference between the current value of the unselected inputs v_i and the estimation of v_i based on the selected input X_{i^*} and is given by

$$v_i^j \leftarrow v_i^j - \hat{m}_{v_i}(X_{i^*}^j) \tag{1.8}$$

where $\hat{m}_{v_i}(X_{i^*}^j)$ is the residual estimate of v_i based on X_{i^*}which removes the shared information between the selected input $X_{i^*}^j$ and the remaining inputs v_i. Kernel residual estimator (e.g. General regression neural network) is the most commonly used approach to estimate residual, therefore the performance of

RE is significantly affected by the determination of kernel bandwidth (for uni-dimensional variables) and bandwidth matrix (for multi-dimensional variables) (Bowden et al., 2005a; Bowden et al., 2005b; He et al., 2011; May et al., 2008a; May et al., 2008b; Li et al., 2015)

Similarly, the updated output is

$$u^j \leftarrow u^j - \hat{m}_u(X_{i^*}^j) \tag{1.9}$$

where $\hat{m}_u(X_{i^*}^j)$ is the residual estimate of u based on X_{i^*}, which again eliminates the shared information between the selected inputs X_{i^*} and the output u.

The **sixth step** is to judge the selected input against the chosen stopping criterion. 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). After updating the input and output variables based on the selected input variable, the corresponding PMI is estimated as

$$I_{v_i,u} \approx \frac{1}{n}\sum_{j=1}^{n} \log[\frac{f\left(v_i^j,u_i^j\right)}{f\left(v_i^j\right)f\left(u_i^j\right)}] \tag{1.10}$$

based on Eqs. (1.7), (1.8), and (1.9). If the PMI value of the selected input is still significant according to the applied termination criterion, the above steps are repeated, as shown in Fig. 1.1, until all significant inputs X_s have been determined. In this way, the algorithm can accommodate a large number of potential input variables, as demonstrated in Fernando et al. (2009).

1.1.4 Bandwidth issue in PMI IVS

In Eqs. 1.7 and 1.10, KDE is used to approximate both marginal and joint PDFs (Eqs. 1.2 and 1.6) by the fact that simple methods exist for KDE that are a function of only a single parameter, the kernel bandwidth, otherwise termed the smoothing parameter (Scott, 1992; Wand and Jones, 1995). Nevertheless, determination of the optimal bandwidth 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 PDF, so that detailed local information (useful information that is not significantly different from others nearby) will not be effectively captured. On the contrary, under-estimating the bandwidth can make the general trend become more vulnerable to localised features (redundant or irrelevant information that is significantly different from others nearby), or even noise (Li et al., 2015). While many methods exist for estimating the bandwidth, in almost all existing PMI IVS studies dealing with environmental and water resources problems (e.g. Bowden et al., 2005a,b; May et al., 2008a,b; He et al., 2011) the Gaussian reference rule (GRR) is used for this purpose. The inherent limitation of this implementation of the PMI algorithm is that the input/output data are assumed to follow a Gaussian distribution. However, this is unlikely to be the case, as the distribution of most environmental and water resources data is generally far from normal. This results in the so called 'bandwidth selection issue'. Such issue impacts both MI and RE by the fact that the MI is a function of KDE based marginal and joint PDFs while the RE is approximated by the kernel based regression models (e.g. General regression neural network), which also depends on KDE, as can be illustrated in Fig. 1.1 (unsolved issues for PMI-based IVS).

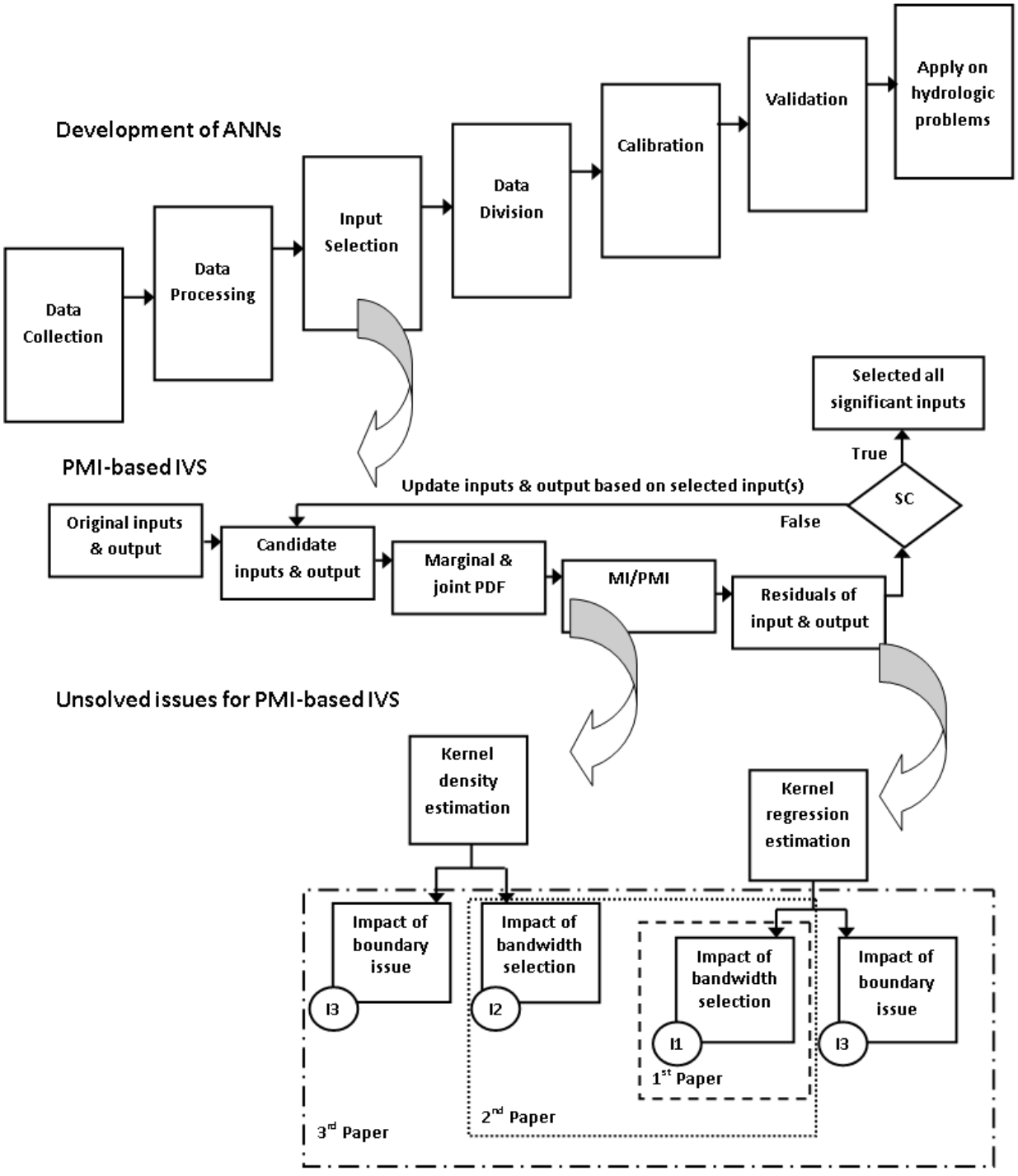
1.1.5 Boundary issue in PMI IVS

In Eqs. (1.6), $\boldsymbol{H}$ is the kernel bandwidth matrix. 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. If the support of $\hat{f}$ is bounded, and in the absence of exponentially falling tails (e.g. support $[0, a]$), strong under-estimation occurs for all data points closer to the boundary, within a distance of the bandwidth h from the boundary. This region is also named the boundary region (Dai and Sperlich, 2010) because of the nonzero kernel density estimation 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 (2005a),

and Wand and Jones (1995). These are the so-called 'boundary issues' associated with (non-parametric) KDE. As explained previously, KDE is used in both of the MI and RE, thereby the impact of boundary issue is expected to contribute to both MI and RE, as displayed in Fig. 1.1 (unsolved issues for PMI-based IVS).

Although many methods for bandwidth estimation exist in other disciplines (e.g. mathematics and statistics (Hall et al., 1992; Park and Marron, 1990; Rudemo, 1982; Scott, 1992; Scott and Terrell, 1987), as shown in Table 1.2), and 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; as shown in Table 1.3), their effectiveness has not yet been tested in the context of PMI IVS for data-driven environmental modelling.

Consequently, the critical issues for PMI IVS in hydrological and environmental applications mainly consist of bandwidth selection and how to deal with boundary issues (as shown in Fig. 1.1). These issues are the primary focus of this thesis.



Publications

1st paper (journal paper): Address I1 to improve GRNNs based residual estimations;
2nd paper (journal paper): Address I2, in conjunction with I1, to improve PMI-based IVS;
3rd paper (journal paper): Address I3, in conjunction with I1 and I2, to improve PMI-based IVS;

Figure 1.1 Framework of thesis

Table 1.1 Review of input variable selection methods for ANNs applied to environmental and water resources problems (developed based on May, 2010)

Algorithm	Type		Criterion	Linearity	Computational cost	Redundancy check	Optimum	Author	Year
Ad-Hoc	Model based/Model free		Prior knowledge	Non-linear	Very low	None	Local	(Hu et al.)	2001
FS	Wrapper	Model based	Accuracy & Complexity	Non-linear	High	None	Local	(Maier et al.)	1998
BE	Wrapper	Model based	Accuracy & Complexity	Non-linear	High	None	Local	(Castellano and Fanelli)	2000
ES	Wrapper	Model based	Accuracy & Complexity	Non-linear	Extremely high	None	Global	(Jain et al.)	1999
HS	Wrapper	Model based	Accuracy & Complexity	Non-linear	Very high	None	Global	(Wei et al.)	2010
SVR	Wrapper	Model based	Accuracy	Non-linear	Very high	None	Local	(Bowden et al.)	2006
SOM-GAGRNN	Wrapper	Model based	Accuracy	Non-linear	High	None	Global	(Bowden et al.)	2005
RVSDEM	Wrapper	Model based	Significance	Non-linear	Fair	Good	Global	(Castelletti et al.)	2012a,b
IIS	Wrapper/ Filter	Model based/ Model free	Significance	Non-linear	Fair	Good	Global	(Galelli and Castelletti)	2013
RC	Filter	Model free	Correlation	Linear	Low	None	Local	(Chua and Wong)	2010
PC	Filter	Model free	Correlation	Linear	Low	Good	Local	(Yang et al.)	2011
PCA	Filter	Model free	Covariance	Linear	High	None	Global	(Hu et al.)	2007
BJ	Filter	Model free	ACF/PACF	Linear	Low	Good	Local	(Box et al.)	2013
GAMMA	Filter	Model free	MSE	Non-linear	Low	None	Local	(Agalbjörn et al.)	1997
MI	Filter	Model free	MI	Non-linear	Low	Fair	Local	(De Vos and Rientjes)	2007
PMI	Filter	Model free	PMI	Non-linear	Fair	Very good	Local	(Bowden et al.)	2005a
ICAIVS	Filter	Model free	MI	Non-linear	Very high	Good	Global	(Trappenberg et al.)	2006
DELSA	Filter	Model free	Sensitivity & Gradients	Non-linear	Fair	Fair	Global/Local	(Rakovec et al.)	2014

FS: forward selection (constructive)
BE: backward elimination (pruning)
ES: exhaustive search
HS: heuristic search
SVR: single variable regression (correlation based)
SOM-GAGRNN: self-organising map genetic algorithm general regression neural network
RVSDEM: recursive variable selection embedded in dynamic emulation models
IIS: tree-based iterative input variable selection
RC: rank correlation (Pearson correlation or linear correlation or cross-correlation)
PC: partial correlation
PCA: principal component analysis
BJ: Box-Jenkins method
GAMMA: Gamma test
MSE: mean squared error
MI: mutual information
PMI: partial mutual information
ICAIVS: hybrid independent component analysis and input variable selection filter
DELSA: distributed evaluation of local sensitivity analysis
ACF: auto-correlation function
PACF: partial auto-correlation function

Table 1.2 Bandwidth estimators applied within the statistics literature

Bandwidth estimator	Author	Year	Fitness function	Dependence on	
				Normality	Error basis
GRR	(Scott)	1992	AMISE	High	Mean
BCV	(Scott and Terrell)	1987			
DPI	(Park and Marron)	1992		Low	
SCV	(Hall et al.)	1992	EMISE		
LSCV	(Rudemo)	1982		None	
OM	(Gibbs et al.)	2006	MAE/RMSE		Mean/Squared

GRR: Gaussian reference rule
BCV: biased cross validation
DPI: 2-stage direct plug-in
SCV: smoothed cross validation
LSCV: least squared cross validation
OM: optimisation method
AMISE: asymptotic mean integrated squared error
EMISE: exact mean integrated squared error
MAE: mean absolute error
RMSE: root mean squared error

Table 1.3 Boundary correctors proposed within the statistics literature

Boundary corrector	Author	Year	Modification
RC	(Schuster)	1985	Kernel function
KT	(Marron and Ruppert)	1994	Kernel function
BK	(Gasser and Müller)	1979	Kernel function
LLM	(Zhang and Karunamuni)	1998	Kernel function
PA	(Cowling and Hall)	1996	Kernel function
ETC	(Hall and Park)	2002	Kernel function
LBE	(Gasser et al.)	1985	Local bandwidth
LBR	(Dai and Sperlich)	2010	Local bandwidth
LLP	(Wand and Jones)	1995	Regression type
LQP	(Fan)	1992	Regression type
LHOP	(Fan and Gijbels)	1996	Regression type

RC: reflection correction
KT: kernel transformation
BK: boundary kernel
LLM: local linear method
PA: pseudo-data approach
ETC: empirical translation correction
LBE: local bandwidth (enlarging)
LBR: local bandwidth (reducing)
LLP: local linear polynomial
LQP: local quadratic polynomial
LHOP: local high order polynomial

1.2 Objectives

According to the aforementioned critical issues for PMI IVS, the ultimate objective of this thesis is to improve the performance of PMI IVS by investigating the impact of bandwidth selection and boundary issues for data driven environmental and water resources models, such as multi-layer perceptron artificial neural networks (MLPANNs). In order to achieve this overall objective, a framework that addresses the influence of bandwidth selection and boundary issues from residual and MI estimates to the overall performance of PMI IVS is developed, as highlighted by the series of dashed line boxes in Fig. 1.1 *Unsolved issues for PMI-based IVS*, and the three corresponding objectives are explained in detail in below.

Objective 1: The motivation underpinning this objective is the fact that the bandwidth (or smoothing parameters) of general regression neural networks (GRNNs), used for RE in PMI IVS, is still predominantly based on the GRR, which only results in optimal density estimates if the Gaussian assumption is valid. However this is not the general case for environmental and water resource data. As a consequence, this objective is concerned with assessing the impact of data with different distributions on the performance of GRNNs and the effectiveness of alternative kernel density estimation techniques in improving GRNN performance. Specifically, the sub-objectives are: (1) to compare the performance (accuracy and efficiency) of GRNN models for which bandwidths (or smoothing parameters) have been estimated using a range of methods, as well as that of a benchmark MLPANN model, for case studies with data that have varying degrees of normality, linearity and different modelling objectives (e.g. matching average or extreme events) (**I1 in Fig. 1.1**); (2) to develop and test empirical guidelines for the selection of the most appropriate methods for GRNN models that are a function of the properties of the available data (i.e. degree of normality and problem non-linearity) and the modelling objective (**Chapter 2**). In the context of PMI IVS, this develops and tests guidelines for the best approach to estimating residuals using GRNNs for data with different degrees of normality and non-linearity.

Objective 2: This objective builds on Objective 1 by using the guidelines developed in Objective 1 for RE to investigate the impact of data with different distributions on PMI IVS, as well as the effectiveness of alternative bandwidth estimators in improving PMI IVS performance, focussing on the best approaches for MI/PMI estimation. The specific objectives are: (1) to assess if, and to what degree, the performance of PMI IVS can be improved for data with different degrees of normality by using alternative bandwidth estimators with reduced reliance on the Gaussian assumption (GRR) (**I2 in Fig. 1.1**); (2) to develop and test a set of preliminary guidelines for selecting the most appropriate bandwidth estimator for data with different degrees of normality, which combines the outcomes of the studies addressing **Objectives 1 and 2** (**Chapter 3**). In the context of PMI IVS, this develops and tests guidelines for the best approach to estimating MI/PMI, as well as residuals using GRNNs, for data with different degrees of normality and non-linearity.

Objective 3: This objective builds on Objectives 1 and 2 by using the guidelines developed in Objectives 1 and 2 for the most appropriate bandwidth estimators for MI/PMI and RE to investigate the effectiveness of alternative approaches to dealing with boundary issues associated with bandwidth selection in improving PMI IVS performance for data with different distributions. The specific objectives are: (1) 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) (**I3 in Fig. 1.1**). (2) 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 (**Chapter 4**). In the context of PMI IVS, this develops and tests guidelines for the best approach to estimating MI/PMI, as well as residuals, for data with different degrees of normality and non-linearity, considering both bandwidth estimation and boundary issues. Consequently, the guidelines presented under this objective represent best practice guidelines for PMI IVS and are therefore able to meet the ultimate objective of this thesis.

It is important to note the relationship between the issues addressed in Objectives 1 to 3, as this has an influence on the order of the objectives presented above. In Fig 1.1, it can be seen that I1 is not influenced by any other parts of the system, while the rest of the system is affected by this issue. Consequently, I1 is investigated primarily and its outcome is able to benefit both I2 and I3. Similarly, I2 (only affected by I1) has strong impacts on I3, therefore it is studied next and the results of I2, in conjunction with those of I1, contribute to I3. Finally, I3, the performance of which is influenced by both I1 and I2, is addressed by consideration of the previous studies. In this way, the analytic procedure becomes rigorous and reliable, with clear logic and minimal side-effects and overlaps.

1.3 Thesis overview

The present thesis is organised into five chapters. In addition to the Introduction (**Chapter 1**), the main body (**Chapters 2 to 4**) is formed by three journal papers. The critical findings, contributions and suggested future research are then summarised in **Chapter 5**. Supplementary materials for Chapters 2 to 4 (three journal papers) are presented in **APPDIX A to C**, which summarise additional supporting analytic figures and tables and mathematical explanations and derivations (i.e. Gaussian reference rule, 2-stage direct plug-in, smoothed cross validation, bivariate reflection correction, and local linear/quadratic polynomial regression). The synopsis, including the content and linkage to the objectives, of each chapter in the main body is outlined in the following sections.

Chapter 2 (Journal paper 1) (Li et al., 2014b) is focused on the development of a systematic way of determining the optimal bandwidth (also known as the smoothing parameter) for the application of GRNN based RE. This is because the performance of GRNNs is essentially controlled by values of one or more bandwidths and insufficient attention has been given to the best way to estimate the bandwidths of GRNNs within environmental and water resource applications, particularly, with data that have varying degrees

of normality, linearity and distinct modelling objectives (**I1 in Fig. 1.1**). In order to overcome such issue, nine different bandwidth estimation methods that have different assumptions on normality, linearity and modelling objectives, as well as that of a benchmark MLPANN model, are assessed in terms of accuracy and computational efficiency for a number of synthetic data sets with distinct data properties [**Objective 1** (1)]. Of these methods, five are based on bandwidth estimators used in kernel density estimation, and four are based on single and multivariable calibration strategies (details can be found in Section 2.3). Preliminary guidelines for the bandwidth selection of GRNNs based RE are developed in accordance with the critical findings of the synthetic tests and then validated on one water quality (forecasting river salinity in the River Murray in South Australia one, five and 14 days in advance) and one water quantity problem (prediction of runoff in the Kentucky River basin in the USA one day in advance) [**Objective 1** (2)].

As discussed in Section 1.2, the bandwidth selection issue for GRNN based RE (**I1 in Fig. 1.1**) has a pronounced influence on the performance of PMI estimation, affected by both the bandwidth selection issue (**I2 in Fig. 1.1**) and the boundary issue (**I3 in Fig. 1.1**). Consequently it is studied as the first priority in Chapter 2.

Chapter 3 (Journal paper 2) (Li et al., 2015) focuses on the performance of PMI IVS under the impact of the bandwidth selection issue, as the currently applied PMI IVS methods in environmental and water resources depend predominately on the Gaussian reference rule (GRR), while the distribution of most water resources data is generally far from normal, which leads to inaccurate IVS for data that are highly non-Gaussian (**I2 in Fig. 1.1**). This issue is taken into account through the investigation of the performance of PMI IVS using six different kernel bandwidth techniques with varying Gaussian dependence [**Objective 2** (1)]. Of these methods, five are kernel based approaches, and one depends on a single variable calibration strategy (details can be found in Section 3.3). The preliminary guidelines for the selection of the most appropriate methods for obtaining the accurate and efficient PMI IVS are determined based on the results of the synthetic case studies with data having various degrees of non-normality and are then

validated for two semi-real case studies developed based on the forecasting of river salinity in the River Murray, South Australia and predicting of flow in the Kentucky River basin, USA [**Objective 2** (2)].

As mentioned in Section 1.2, the preliminary guidelines developed to select the optimal bandwidth for RE in PMI IVS (**I1 in Fig. 1.1**) developed in Chapter 2 are merged into the ones established in Chapter 3. This results in the complete exploration of the performance of PMI IVS influenced by the bandwidth selection issue (**I2 in Fig. 1.1**).

Chapter 4 (Journal paper 3) (Li et al., 2014a) addresses the boundary issue, which is caused by the adoption of the symmetrical kernel at the unsymmetrical boundary during kernel based MI and RE within PMI IVS, which has not been considered or investigated thus far in the environmental and water resources fields (**I3 in Fig. 1.1**). Systematic studies are conducted by investigating the effectiveness of sixteen approaches. Of these approaches, three are benchmark approaches without consideration of the boundary issue, two aim to improve the boundary issue in MI, seven aim to minimise the effect of the boundary issue in RE, and four take into account the boundary issue in both MI and RE (details can be found in Section 4.3). In addition, the effect of the bandwidth issue is effectively addressed in all sixteen approaches based on the guidelines developed for **Objectives 1 and 2** [**Objective 3** (1)]. The preliminary guidelines that are developed based on the results of the above studies, which attenuate the boundary issue associated with the selection of the most appropriate bandwidth estimator for data with different degrees of normality, are validated for two semi-real case studies used in journal papers 1 and 2 [**Objective 3** (2)].

By recalling Section 1.2, the boundary issue is not the only driving force on the performance of PMI IVS, since selection of the bandwidth also affects the accuracy of PMI IVS. Therefore, the boundary issue (**I3 in Fig. 1.1**) is studied after the bandwidth issue in PMI IVS has been addressed explicitly (**I1 and I2 in Fig. 1.1**). By resolving I3 in conjunction with the outcomes of I1 and I2, the ultimate objective within this thesis, mentioned in Section 1.2, is achieved in Chapter 4.

CHAPTER 2 JOURNAL PAPER 1 -

Selection of Smoothing Parameter Estimators for General Regression Neural Networks - Applications to Hydrological and Water Resources Modelling

Statement of Authorship

Title of Paper	Selection of Smoothing Parameter Estimators for General Regression Neural Networks - Applications to Hydrological and Water Resources Modelling
Publication Status	◉ Published, ○ Accepted for Publication, ○ Submitted for Publication, ○ Publication style
Publication Details	Li, X., Zecchin, A.C., Maier, H.R., 2014b. 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.

Author Contributions

By signing the Statement of Authorship, each author certifies that their stated contribution to the publication is accurate and that permission is granted for the publication to be included in the candidate's thesis.

Name of Principal Author (Candidate)	Xuyuan Li		
Contribution to the Paper	Undertook literature review, developed analytic procedure and numerical models, developed software, and prepared manuscript		
Signature		Date	

Name of Co-Author	Dr. Aaron C. Zecchin		
Contribution to the Paper	Supervised manuscript preparation and reviewed draft		
Signature		Date	

Name of Co-Author	Professor Holger R. Maier		
Contribution to the Paper	Supervised manuscript preparation and reviewed draft		
Signature		Date	

Name of Co-Author			
Contribution to the Paper			
Signature		Date	

Abstract

Multi-layer perceptron artificial neural networks are used extensively in hydrological and water resources modelling. However, a significant limitation with their application is that it is difficult to determine the optimal model structure. General regression neural networks (GRNNs) overcome this limitation, as their model structure is fixed. However, there has been limited investigation into the best way to estimate the parameters of GRNNs within water resources applications. In order to address this shortcoming, the performance of nine different estimation methods for the GRNN smoothing parameter is assessed in terms of accuracy and computational efficiency for a number of synthetic and measured data sets with distinct properties. Of these methods, five are based on bandwidth estimators used in kernel density estimation, and four are based on single and multivariable calibration strategies. In total, 5674 GRNN models are developed and preliminary guidelines for the selection of GRNN parameter estimation methods are provided and tested.

Software availability

Software name: GRNNs

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

Email: xli@civeng.adelaide.edu.au;

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 ; all versions of Visual Studio 2012, 2010 and 2008 are supported except Visual Studio Express; 256 MB RAM

Software requirements: PGI Visual Fortran 2003 or later version

Language: English

Size: 4.74 MB

Availability: Free to download for research purposes from the following website:

http://www.ecms.adelaide.edu.au/civeng/research/water/software/generalised-regression-neural-network/

https://github.com/xuyuanli/GRNNs

2.1 Introduction

Over the last two decades, artificial neural networks (ANNs) have been used extensively in the field of hydrological and water resources modelling, and their popularity is still increasing (Abrahart et al., 2012; Maier et al., 2010; Wu et al., 2014b). In the vast majority of these applications, multi-layer perceptrons (MLPs) have been used as the most common model architecture (Maier et al., 2010; Wu et al., 2014b). While the use of MLPs has generally resulted in good model performance, their development is complicated by the fact that there are no rigorous methods for determining an appropriate model structure. Determination of the optimal number of hidden nodes is especially difficult, unless sophisticated Bayesian approaches are used (Kingston et al., 2008; Zhang et al., 2011), which are computationally demanding and require substantial technical expertise to implement. Therefore, the optimal model structure is generally determined by trial and error (Maier et al., 2010; Wu et al., 2014). This process usually involves a number of steps, including (i) selection of a trial model structure, (ii) calibration of the model with the selected structure, and (iii) evaluation of the predictive performance of the calibrated model. These steps are repeated for models with different trial structures and the model structure that results in the best predictive performance is selected. Consequently, the model structure that is found to be optimal is a function of a number of factors, including:

(i) *The trial model structures selected for evaluation:* As the potential number of different model structures is generally large, the performance of a subset of all possible structures is usually evaluated. This can be achieved using different approaches, including ad-hoc, stepwise (e.g. constructive, pruning) or global approaches (Maier et al., 2010). Consequently, as different approaches generally result in the evaluation of different model structures, the structure obtained is a function of the adopted approach.

(ii) *The calibration method used:* The predictive performance of a model with a particular structure is a function of the quality of the calibration (training) process. Finding the combination of model parameters

(connection weights) that gives the best predictive performance for a given network structure is complicated by the presence of a large number of local optima in the error surface (Kingston et al., 2005b). This is particularly the case if gradient-based calibration (training) methods are used (Maier and Dandy, 1999), such as the most commonly used back-propagation algorithm (Maier et al., 2010; Wu et al., 2014). In addition to the choice of calibration (training) methods, the parameters that control the searching behaviour of these methods (e.g. learning rate and momentum when the back-propagation algorithm is used) can also have a significant impact on the best predictive model performance obtained for a particular model structure (Maier and Dandy, 1998a, b). Consequently, unless the predictive performance that corresponds to the global optimum in the error surface can be identified for all models with different structures, it is not possible to identify which model structure results in the best predictive performance with certainty. As a result, the optimal model structure obtained is a function of the quality of the model calibration process.

(iii) *The calibration data used:* The available data are generally split into different subsets for calibration (training) and validation, which can be done using a number of different methods (see Maier et al., 2010). Consequently, which data points are included in the different subsets can vary, depending on which data division method is used (Bowden et al., 2002; May et al., 2010; Wu et al., 2012; Wu et al., 2013). This can also have an impact on which model structure is found to result in the best predictive performance. This is because different data points will result in different error surfaces during calibration, thereby potentially affecting calibration difficulty (see (ii)) and producing different global and local optima, which is likely to change which model structure results in the lowest error.

Given the factors described above, it is generally not possible to isolate the impact of model structure on the predictive performance of MLPs, making it difficult to know which model structure should be used. In addition, the trial-

and-error process generally used to determine the optimal structure of MLPs is computationally expensive, as it necessitates the development of a potentially large number of models.

Although there are other alternative ANN based approaches, including Radial Basis Functions (RBFs) (Buhmann, 2003), Recurrent Neural Networks (RNNs) (Williams and Zipser, 1989) and Probabilistic Neural Networks (PNNs) (Specht, 1990), General regression neural networks (GRNNs) (Specht, 1991) provide an alternative ANN model structure that has been shown to perform well in a number of studies in water resources applications (Bowden et al., 2005b; Bowden et al., 2006; Cigizoglu and Alp, 2006; Gibbs et al., 2006) and overcomes the shortcomings associated with MLPs discussed above, as the structure of GRNNs is fixed (Bowden et al., 2005a). This removes the ambiguity associated with determining which model structure is optimal. In addition, it increases the computational efficiency of the model development process, as there is no need to develop a number of models with different structures in order to determine which is optimal.

However, a potential issue with the application of GRNNs to hydrological and water resources problems is that there has been limited work on determining which smoothing parameter estimation methods should be adopted. As GRNNs are essentially a Nadaraya-Watson kernel regression method (Cai, 2001), parameter estimation only involves the determination of optimal values of one or more smoothing parameters, also known as kernel bandwidths. However, this is not a trivial issue, as illustrated by the vast amount of literature on kernel bandwidth estimation as applied to density estimation (Bowman, 1984; Hall et al., 1992; Park and Marron, 1990; Rudemo, 1982; Scott and Terrell, 1987; Wand and Jones, 1995). Overestimating the smoothing parameter can result in over-smoothing of the estimated density (i.e. kernel based probability density function (PDF)). In this case, the detailed local information (for instance the variation of daily rainfall in hydrological applications) will not be captured in the estimated density. In contrast, if values of the smoothing parameter are underestimated, the general trend of the

estimated density (for instance the overall rainfall trend within a given time period) can be disturbed by localised features or noise.

Among the extensive literature on smoothing parameter (or kernel bandwidth) estimation in other areas of research, such as mathematics and statistics, there are a number of different approaches to obtaining optimal estimates of kernel density, which are based on assumptions about the form of the PDF and different fitness function types (i.e. the objective function on which the estimator is based). Consequently, their relative merits for determining the optimal values of the smoothing parameters for water resources GRNN models are likely to vary from case study to case study, depending on the distribution of the data and the modelling objective function used. However, the relationship between the performance of GRNNs with smoothing parameters obtained using different kernel density estimation methods and the properties of the water resources data used to develop them has not been considered previously, making it difficult to know which methods to use for particular case studies.

Therefore, the objectives of the current study are: (i) to compare the performance, in terms of both predictive accuracy and computational cost, of GRNN models for which smoothing parameters have been estimated using a range of methods, as well as that of a benchmark MLP model, for case studies with data that have varying degrees of normality, linearity and different modelling objectives (e.g. matching average or extreme events); and (ii) to develop and test empirical guidelines for the selection of the most appropriate methods for GRNN smoothing parameter estimation based on the properties of the available data (i.e. degree of normality and non-linearity) and the modelling objective.

The remainder of this paper is organised as follows. A brief introduction to GRNNs is provided in Section 2.2, followed by the Methodology in Section 2.3. Results and discussion are given in Section 2.4, and conclusion and recommendations are provided in Section 2.5.

2.2 GRNNs

According to Bowden et al. (2005a), GRNNs can be treated as supervised feedforward ANNs with a fixed model architecture. The general architecture of GRNNs is illustrated in Fig. 2.1.

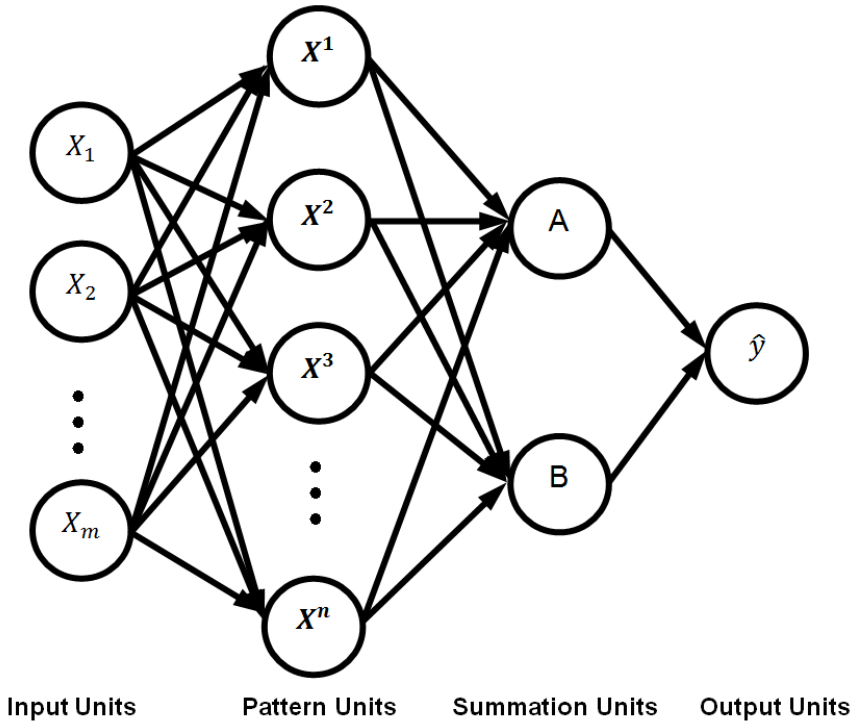


Figure 2.1 General architecture of a GRNN

(based upon Gibbs et al. (2006))

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 (the patterns) for $j = 1, \ldots, n$, where n is the number of observations, $\boldsymbol{X}^j = \left[X_1^j \ldots X_m^j\right]^T$ are the observed input data and y^j are the observed output data; and $\hat{y}$ be the GRNN estimate of the actual output y. If the joint density $f(\boldsymbol{X}, y)$ is known, the conditional expectation of output y given input $\boldsymbol{X}$ is given as

$$E[y|\boldsymbol{X}] = \frac{\int_{-\infty}^{\infty} y f(\boldsymbol{X},y) dy}{\int_{-\infty}^{\infty} f(\boldsymbol{X},y) dy} \tag{2.1}$$

The joint density $f(\boldsymbol{X}, y)$ in Eq. (2.1) is generally unknown, however, the empirical joint density of the observed input/output pairs $(\boldsymbol{X}^j, y^j)$, j = 1, …, n can be estimated by the Gaussian kernel-based estimator as

$$\hat{f}(\boldsymbol{X}, y) = \frac{1}{2\pi^{\frac{m+1}{2}} h^{(m+1)}} \frac{1}{n} \sum_{j=1}^{n} exp\left[-\frac{(\boldsymbol{X}-\boldsymbol{X}^j)^T(\boldsymbol{X}-\boldsymbol{X}^j)}{2h^2}\right] exp[-\frac{(y-y^j)^2}{2h^2}] \qquad (2.2)$$

where h is the kernel smoothing parameter (Cacoullos, 1966; Parzen, 1962). Note that this approximation is commonly known as Parzen window density estimation. It is valid, however, only if the underlying density is continuous and the first partial derivative at any $\boldsymbol{X}$ is small. Specht (1991) combined the conditional expectation of y (Eq. (2.1)) with the Parzen window density estimation $\hat{f}(\boldsymbol{X}, y)$ (Eq. (2.2)) to obtain the following estimator for y

$$\hat{y}(\boldsymbol{X}, h) = \frac{\sum_{j=1}^{n} y^j \exp(-\frac{D_j^2(\boldsymbol{X})}{2h^2})}{\sum_{j=1}^{n} \exp(-\frac{D_j^2(\boldsymbol{X})}{2h})} \qquad (2.3)$$

Where D_j^2 is the scalar function

$$D_j^2 = (\boldsymbol{X} - \boldsymbol{X}^j)^T(\boldsymbol{X} - \boldsymbol{X}^j) \qquad (2.4)$$

which measures the Euclidian distance between the input $\boldsymbol{X}$ and the observed data points $\boldsymbol{X}^j$. Within this equation, the smoothing parameter h is the only unknown parameter that needs to be obtained by training (calibration).

With respect the GRNN formulation, the expression in Eq. (2.3) can be implemented by the four-unit (or layer) parallel network shown in Fig. 2.1. The GRNN consists of input, pattern, summation and output units that are fully connected. According to Specht (1991), the input units are formed by the elements of the input vector $\boldsymbol{X}$, and these then feed into each of the pattern units. The pattern units record D_j^2 , the sum of squared (or absolute) difference between an input vector $\boldsymbol{X}$ and the observed data $\boldsymbol{X}^j$, and then feed into a nonlinear activation function (e.g. the exponential function as in Eq. (2.3)) before passing into the summation units. The summation units contain two parts, A and B, which correspond to the numerator and denominator in Eq.

(2.3), respectively. Part A (the numerator) contains a dot product between the observed output records y^j and the weights $exp\left(-\frac{D_j^2(\boldsymbol{X})}{2h^2}\right)$ from the pattern units, while part B (the denominator) only includes the weights from the pattern units. The quotient of parts A and B is the predicted output $\hat{y}$.

In Fig. 2.1, the model architecture of GRNNs is fixed by the fact that the number of input nodes is determined by the number of inputs *m*; the number of pattern nodes depends on the size of the observed input data *n*; and the nodes in the summation units always consist of a denominator node and a numerator node.

Within this study, a slightly generalised version of the GRNN estimator in Eq. (2.3) is considered, namely

$$\hat{y}(\boldsymbol{X},h)=\frac{\sum_{j=1}^{n} y^j exp\left(-\frac{1}{2}\sum_{i=1}^{m}\frac{\left(x_i-x_i^j\right)^2}{h_i^2}\right)}{\sum_{j=1}^{n} exp\left(-\frac{1}{2}\sum_{i=1}^{m}\frac{\left(x_i-x_i^j\right)^2}{h_i^2}\right)} \tag{2.5}$$

where the primary difference between Eq. (2.3) and Eq. (2.5) is the adoption of a unique smoothing parameter h_i for each dimension of the input space i = 1, …, n. The advantage of this form of the GRNN is that it enables an independent scaling of the kernel smoothing, as opposed to a common smoothing, along each dimension of the input space.

2.3 Methodology

The approach to the systematic assessment of the performance of GRNNs with different bandwidth estimators is illustrated in Fig. 2.2. As can be seen, there are four main steps: (i) procurement of input and output data with different degrees of normality and non-linearity; (ii) estimation of the optimal GRNN smoothing parameter (bandwidth) for these different input or output data using a number of different smoothing parameter estimators; (iii)

development of benchmark MLP models; and (iv) assessment of model performance. Details of each of these steps are given in the subsequent sections.

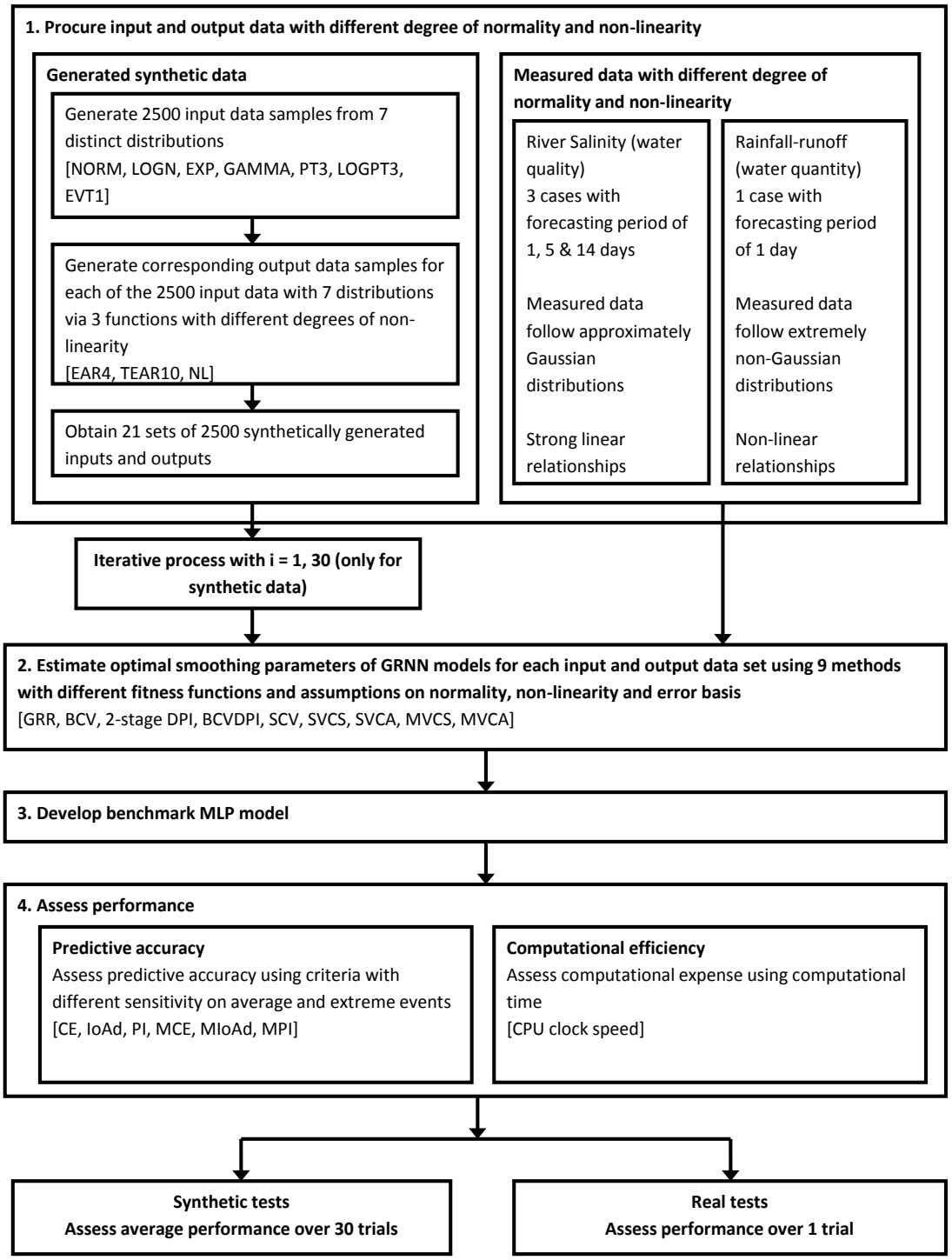


Figure 2.2 Overview of proposed assessment approach

2.3.1 Procurement of input/output data with different degrees of normality and non-linearity

As can be seen from Fig. 2.2, two different approaches to procuring input and output data with different degrees of normality and non-linearity were used, including the generation of synthetic data and the use of measured data, as outlined below.

Synthetically generated data

Procurement of the synthetic data involved the generation of input data from distributions with differing degrees of normality, and the subsequent generation of the corresponding output data using synthetic models with different degrees of non-linearity. Data were generated from seven distinct distributions, 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) (see Fig. 2.2). These distributions were used because they are the most commonly adopted distributions in hydrological problems (Chow et al., 1988), and have the ability to generate data with a large range of skewness and kurtosis, which are measures of the degree of non-normality (Bennett et al., 2013). The properties of each distribution are given in Tables 2.1 and 2.2. For each distribution, an additional 25 data points were generated for each of the exogenous inputs in the time series models, as the first 25 points were rejected in order to prevent initialisation effects (May et al., 2008b). All data sets were split into training (60%), testing (20%) and validating sets (20%) using the DUPLEX method (see May et al., 2010), in accordance with the guidelines suggested by Wu et al. (2013).

Table 2.1 Details of the simulated input distributions for the time series models (EAR4, TEAR10)

Distribution	Key Parameters	s	k	Normality
NORM	Mean=3.0; sd =1.0	0.000	-0.013	High
GAMMA	Shape=2.0; Scale=1.0	1.370	2.638	High
LOGN	Mean=0.5; sd=1.0	5.326	53.694	Low
EXP	Rate=1.0	2.132	7.219	Moderate
PT3	Shape=2.5; Scale=3.0; Location=2.0	1.251	2.381	High
LOGPT3	Shape=0.5; Scale=0.2; Location=2.0	4.792	43.265	Low
EVT1	Shape=0.0; Scale=0.5; Location=10.0	1.198	2.880	High

(Key parameters in the table are used to simulate the exogenous input variable; the skewness and kurtosis shown in the table are the averaged values of all input and output data)

Table 2.2 Details of the simulated input distributions for the nonlinear model (NL)

Distribution	Key Parameters	s	k	Normality
NORM	Mean=3.0; sd =1.0	1.826	5.158	High
GAMMA	Shape=2.0; Scale=1.0	10.520	192.091	Low
LOGN	Mean=0.5; sd=0.4	5.389	47.767	Low
EXP	Rate=1.0	14.029	334.408	Low
PT3	Shape=0.5; Scale=1.0; Location=0.5	16.271	514.270	Low
LOGPT3	Shape=0.5; Scale=0.2; Location=0.5	14.261	390.522	Low
EVT1	Shape=0.1; Scale=0.0; Location=10.0	1.788	9.807	Moderate

(Key parameters in the table are used to simulate each of the input variables; the skewness and kurtosis shown in the table are the averaged values of all input and output data)

The synthetic models used to produce the output data included a linear exogenous auto-regressive time series model (EAR4), a threshold exogenous auto-regressive time series model (TEAR10), and a nonlinear input-output function (NL) (see Fig. 2.2), as they represent relationships with increasing degrees of non-linearity and are based on synthetic models used in previous studies (Bowden et al., 2005a; Galelli and Castelletti, 2013; May et al., 2008b). The equation for the linear exogenous auto-regressive time series of order four (EAR4) is given by

$$x_t = 0.6x_{t-1} - 0.4x_{t-4} + p_{t-1} + 0.1\varepsilon_t \quad (2.6)$$

where x_t is the output time series; x_{t-n} is the input time series with lag n; p_{t-n} is the exogenous input with lag n; and $0.1\varepsilon_t$ is the introduced error term.

The equation for the nonlinear exogenous auto-regressive time series model of order ten (TEAR10) is given by

$$x_t = \begin{cases} -0.5x_{t-6} + 0.5x_{t-10} - 0.3p_{t-1} + 0.1\varepsilon_t; & x_{t-6} \leq 0 \\ 0.8x_{t-10} - 0.3p_{t-1} + 0.1\varepsilon_t; & otherwise \end{cases} \quad (2.7)$$

and the equation for the nonlinear input-output function (NL) is given by

$$y = (x_2)^3 + x_6 + 5\sin(x_9) + 0.1\varepsilon_t \quad (2.8)$$

The first two synthetic models (Eqs. (2.6) and (2.7)) were modified versions of the synthetic models used in May et al. (2008b) and the third synthetic model (Eq. (2.8)) was modified from the one used in Bowden et al. (2005a). For the first two synthetic models, the modifications include the introduction of an independent lagged input p_{t-1} into all exogenous AR models, and the p_{t-1} were sampled from the distributions outlined in Table 2.1. For the third synthetic model, the significance (coefficient) of each input was slightly modified and each input was sampled based on the distributions outlined in Table 2.2. In addition, the error term $0.1\varepsilon_t$ was added to all models to introduce noise into the models without obscuring the influence of the actual independent variables. The noise term ε_t followed the standard normal distribution $N(0,1)$.

Real case studies

In order to further test the impact of the degree of normality and non-linearity of the data on the predictive performance and computational efficiency of the different GRNN parameter estimation methods investigated, as well as the performance of the empirical guidelines for the selection of the most appropriate methods for GRNN smoothing parameter estimation developed based on the results from the synthetic data, two case studies with data with different degrees of normality and non-linearity were selected. The first case study was concerned with forecasting salinity in the River Murray in South Australia one, five and 14 days in advance and the second with the prediction of runoff in the Kentucky River basin in the USA one day in advance. The

data division procedure used for both real case studies was identical to the one used for the synthetic case studies (see Section 2.1.1).

The salinity case has been studied extensively in the context of ANN modelling (Bowden et al., 2005b; Fernando et al., 2009; Kingston et al., 2005a; Maier and Dandy, 1996; Maier and Dandy, 2000a). According to Maier and Dandy (1996), salinity in the River Murray is a function of upstream inflows of salinity, flow, river level and groundwater level. Maier and Dandy (2000) also found that different combinations of inputs contribute to the output during different forcasting periods. In line with this finding, different GRNNs were developed in this study to predict salinity at Murray Bridge one, five and 14 days in advance (Table 2.3). Different input variables with different lags (Table 2.3) were associated with each output in a given forecasting period, where the inputs were selected from previous studies (e.g. Maier and Dandy, 1996; Maier and Dandy, 2000; Kingston et al., 2005b). All data covered the period 1987 to 1990, and were the same as the data used by Maier and Dandy (1996; 2000).

Analysis of the input data shows that the salinity based inputs are approximately normally distributed (average $s = -1.11$ & $k = 0.319$), although distributions of some lagged inputs have multiple peaks and the distribution of the water level based input is mildly non-Gaussian (average $s = 5.96$ & $k = 2.57$). According to Bowden (2003), the input and output data contain strongly linear components. Consequently, the data for this case study are close to mildly non-normal and the relationship to be modelled is close to linear.

Table 2.3 Inputs and outputs used to forecast salinity at Murray Bridge 1, 5, & 14 days in advance

	Inputs				Output			
Case No.	Location	Variable	Abbreviation	Lags	Location	Variable	Abbreviation	Forecasting Period
1	Murray Bridge	Salinity	MBS	1	Murray Bridge	Salinity	MBS	1
	Mannum	Salinity	MAS	1				
2	Murray Bridge	Salinity	MBS	1	Murray Bridge	Salinity	MBS	5
	Mannum	Salinity	MAS	1				
3	Mannum	Salinity	MAS	1	Murray Bridge	Salinity	MBS	14
	Morgan	Salinity	MOS	1				
	Waikerie	Salinity	WAS	1, 5				
	Loxton	Salinity	LOS	1				
	Lock 7 Lower	Flow rate	L7F	1				
	Lock 1 Upper	River level	L1UL	1				

Table 2.4 Inputs and output used to model rainfall-runoff from the Kentucky River basin

Inputs				Output			
Location	**Variable**	**Abbreviation**	**Lags**	**Location**	**Variable**	**Abbreviation**	**Forecasting Period**
Manchester	Mean daily effective rainfall	P	0,1,2	Lock & Dam 10	Mean daily runoff	Q	1
Hyden							
Jackson							
Heidelberg							
Lexington Airport							
Lock & Dam 10	Mean daily runoff	Q	1,2				

Table 2.5 Selected smoothing parameter estimators with different fitness functions and assumptions of normality and error basis

Applied method	Fitness function	Dependence on		Sensitive to event	No. of smoothing parameters	Optimizer
		Normality	Error basis			
GRR	AMISE	High	Mean	Average	Single	None
BCV	AMISE	High	Mean	Average	Multiple	GSS
2-stage DPI	AMISE	Low	Mean	Average	Multiple	None
BCVDPI	AMISE	Low	Mean	Average	Multiple	GSS
SCV	EMISE	Low	Mean	Average	Multiple	GSS
SVC	MAE/RMSE	None	Mean/squared	Average/Extreme	Single	GSS
MVC	MAE/RMSE	None	Mean/squared	Average/Extreme	Multiple	PSO

(GSS refers to the golden section search algorithm (Press et al., 1992); PSO stands for the particle swarm optimisation algorithm (Poli et al., 2007); MAE is the mean absolute error; RMSE denotes the root mean squared error)

The rainfall-runoff problem from the Kentucky River basin has also been extensively studied in the ANN literature (Bowden et al., 2012; Jain and Srinivasulu, 2004; Srinivasulu and Jain, 2006; Wu et al., 2013). The catchment area is approximately 10240 km^2 and the average daily total rainfall measurements come from five rain gauges located at Manchester, Hyden, Jackson, Heidelberg, and Lexington Airport. The average daily streamflow at Lock and Dam 10 are used as the output. Jain and Srinivasulu (2004) suggested five significant inputs (i.e. lagged effective rainfall $P(t), P(t-1), P(t-2)$ and lagged runoff $Q(t-1), Q(t-2)$). Therefore, the effective rainfall, with lags from the present day to two days prior, and the flow with lags of the first two days, were adopted as inputs (Table 2.4). The data used in this paper were identical to the 13 years of training data (1960-1972) utilised by Jain and Srinivasulu (2004).

Analysis of the input and output data shows that the distributions of lagged effective rainfall and flow are extremely non-Gaussian (averaged $s = 5.11$ & $k = 34.8$). Although the linearity of the rainfall- runoff problem in the Kentucky River basin has not previously been analysed, the general rainfall-runoff problem is well recognised as being highly nonlinear (Coulibaly et al., 2001; Dawson et al., 2002; Hu et al., 2001; Jain and Indurthy, 2003), and therefore the data are likely to contain a strong nonlinear structure. Consequently, the data for this case study are considered to be highly non-normal and the relationship to be modelled is likely to be highly non-linear.

2.3.2 Estimation of GRNN smoothing parameters using different estimation methods

The parameters for all of the GRNN models for the synthetic tests and real case studies were estimated using nine methods. Of these methods, five are adopted from the literature on kernel bandwidth selection for kernel density estimation, and four are based on single and multivariable calibration optimisation strategies. The methods adopted from the kernel density estimation literature are: the Gaussian reference rule (GRR); biased cross validation (BCV); 2-stage direct plug-in (DPI); a combination of BCV and

DPI (BCVDPI); smoothed cross validation (SCV). The methods based on calibration optimisation strategies are as follows: single variable calibration with squared error as the objective function (SVCS); single variable calibration with mean absolute error as the objective function (SVCA); multi-variable calibration with squared error as the objective function (MVCS); and multi-variable calibration with mean absolute error as the objective function (MVCA) (Fig. 2.2). These methods were selected as they are based on different fitness functions and assumptions of normality and error basis, as shown in Table 2.5. Details of these smoothing parameter estimators are given in the following subsections.

Gaussian reference rule (GRR)

The GRR based smoothing parameter estimator is the most commonly used estimator. It is based on minimising the asymptotic mean integrated squared error (AMISE) under the integrability assumption of an unknown probability function f of the given data (Scott, 1992; Wand and Jones, 1995). Under these assumptions, the derived AMISE has the expression

$$AMISE\{\hat{f}(.;h)\} = (nh)^{-1}R(K) + \frac{1}{4}h^4\mu_2(K)^2\widetilde{R(f'')} \tag{2.9}$$

where K is the kernel function; $R(K) = \int [K(x)]^2 dx$ is the integrated square of the kernel function; $\mu_2(K) = \int x^2 K(x) dx$ is the second moment of K; and $\widetilde{R(f'')}$ represents the approximation of the integrated squared second derivative of f. By assuming that the data follow a Gaussian distribution, and adopting a Gaussian kernel, the GRR based smoothing parameter estimator that minimises the AMISE is derived as

$$\hat{h}_{GRR,i} = (\frac{4}{m+2})^{1/(m+4)}\sigma_i n^{-1/(m+4)} \tag{2.10}$$

where σ_i is the sample standard deviation of the X_i^j (usually standardised first). As outlined in Table 2.5, this approach depends heavily on the Gaussian assumption.

Biased cross validation (BCV)

As with the GRR, the BCV (Scott and Terrell, 1987) based smoothing parameter estimation method aims to minimise the AMISE, and is based on the assumption that the data are normally distributed. However, as the BCV is a combination of cross-validation and 'plug-in' bandwidth selection described by Wand and Jones (1995), it is potentially more robust than the GRR based approach through optimisation. The AMISE is expressed as follows by substituting the estimated $\widetilde{R(f'')}$ into Eq. (2.9)

$$AMISE_{BCV,i}(h) = (nh)^{-1}R(K) + \frac{1}{4}h^4\mu_2(K)^2 n^{-2} \sum\sum_{p\neq q}(K'' * K'')(X_i^{\boldsymbol{p}} - X_i^{\boldsymbol{q}}) \quad (2.11)$$

where $*$ indicates the convolution operation. The BCV smoothing parameter is then given as

$$\hat{h}_{BCV,i} = argmin_h\{AIMSE_{BCV,i}(h)\} \quad (2.12)$$

As illustrated in Table 2.5, the underlying assumptions for the estimator $\hat{h}_{BCV,i}$ are similar to $\hat{h}_{GRR,i}$ (Eq. (2.10)), however $\hat{h}_{BCV,i}$ is determined by minimising the $AIMSE_{BCV,i}(h)$ through an optimisation process (in the current study, the golden section search (GSS) (Press et al., 1992) was used).

Two-stage direct plug-in (DPI)

The motivating idea behind the DPI (Park and Marron, 1992) is to approximate the unknown term $\widetilde{R(f'')}$ with $\hat{\varphi}_r(g)$ (which is a pilot kernel estimation of the *r-th* order integrated squared density derivative); g is the pilot kernel bandwidth; L is the pilot kernel; and r is the stage number into Eq. (2.9) to obtain a computable form for the asymptotically optimal bandwidth. By minimising AMISE (Eq. (2.9)) and replacing $\widetilde{R(f'')}$ with a pilot kernel bandwidth estimation $\hat{\varphi}_4(g)$, the DPI based smoothing parameter expression, for each input dimension *i*, becomes

$$\hat{h}_{DPI,i} = \left[\frac{R(K)}{[\mu_2(K)]^2\hat{\varphi}_4(g)n}\right]^{1/5} \quad (2.13)$$

where $\hat{\varphi}_4(g) = n^{-1}\sum_{i=1}^{n}\hat{L}^{(4)}(X^i; g)$ represents the fourth order integrated squared density derivative, which is approximated by the pilot kernel L, with the corresponding pilot bandwidth as g (Hall and Marron, 1987; Jones and Sheather, 1991). The asymptotic mean squared error (AMSE) based optimal overall pilot bandwidth g is

$$g = \left[\frac{k!L^{(r)}(0)}{-\mu_k(L)\hat{\varphi}_{r+k}n}\right]^{1/(r+k+1)} \tag{2.14}$$

where k is the order of the pilot kernel L; r is the stage number of L; $\mu_k(L) = \int u^k L(u)du$ is the *k-th* moment of L. The stage number r determines how many kernel estimations are required to approximate $\hat{\varphi}_4(g)$ based upon the higher order integrated squared density derivative. Although it has been found that more stages can result in a better estimation when using the DPI, the improvement comes at a significant cost in terms of computational efficiency (Wand and Jones, 1995). The commonly suggested number of stages is $r = 2$ (Park and Marron, 1992), which was adopted in this study. For a 2-stage DPI based estimator, the corresponding fitness function and assumptions on linearity and error basis are identical to those for the GRR and BCV based approaches, while the dependence on the Gaussian assumption is effectively reduced by the pilot kernel based fourth order integrated squared density derivative, as shown in Table 2.5.

Combination of biased cross validation and two-stage direct plug-in (BCVDPI)

The BCVDPI estimator is a combination of the BCV and 2-stage DPI, and is achieved by replacing the estimated term $\widetilde{R(f'')}$ in Eq. (2.8) with the 2-stage DPI based $\hat{\varphi}_4(g)$ as follows

$$AMISE_{BCVDPI,i}(h) = (nh)^{-1}R(K) + \frac{1}{4}h^4\mu_2(K)^2\hat{\varphi}_4(g)_{DPI} \tag{2.15}$$

Although the BCVDPI has no closed form (it requires the solution of an optimisation problem), it inherits the positive attributes of a reduced dependence on the Gaussian assumption in comparison to the DPI. The

optimal smoothing parameter by minimising $AIMSE_{BCVDPI,i}(h)$ can be expressed, for each input dimension i, as

$$\hat{h}_{BCVDPI,i} = argmin_h\{AIMSE_{BCVDPI,i}(h)\} \tag{2.16}$$

The fitness function and assumptions of the BCVDPI based approach are identical to those of the 2-stage DPI approach. The main difference between these two approaches is that the former uses GSS based optimisation due to the biased cross-validation procedure, while the latter does not.

Smoothed cross validation (SCV)

The concept behind SCV is very similar to that underpinning the DPI approach, except that SCV attempts to minimise the exact MISE, rather than the AMISE (Eq. (2.9)) used in the DPI method. The MISE can also be approximated as

$$MISE\{\hat{f}(.;h)\} \approx (nh)^{-1}R(K) + \int (K_h * f - f)(x)^2 dx \tag{2.17}$$

By replacing $\int (K_h * f - f)(x)^2 dx$ with $\widehat{ISB}(h)$, where $\widehat{ISB}(h)$ is an estimation of the integrated squared bias, Eq. (2.16) can be re-written as

$$EMISE_{SCV,i}(h) = (nh)^{-1}R(K) + \widehat{ISB}(h) \tag{2.18}$$

where $\widehat{ISB}(h)$ is given by

$$\widehat{ISB}(h) = n^{-2}\sum_{p=1}^{n}\sum_{q=1}^{n}(K_h * K_h * L_g * L_g - 2 * K_h * L_g * L_g + L_g * L_g)(X_i^{\boldsymbol{p}} - X_i^{\boldsymbol{q}}) \tag{2.19}$$

where K_h and L_g are Gaussian kernels with kernel bandwidth h and pilot kernel bandwidth g, respectively (Hall et al., 1992; Wand and Jones, 1995). The pilot kernel bandwidth g is a function of a series of pilot kernel bandwidths, each estimated based upon sequentially higher order integrated squared density derivatives (Wand and Jones, 1995). The optimal smoothing parameter is determined by finding the parameter $\hat{h}_{SCV,i}$, which minimises

$EMISE_{SCV,i}(h)$ through optimisation (GSS), as shown in Eq. (2.20) for the *i*-th input

$$\hat{h}_{SCV,i} = argmin_h\{EIMSE_{SCV,i}(h)\} \tag{2.20}$$

Although the assumptions with regard to normality, linearity, and error basis of the SCV based method are very similar to those of the 2-stage DPI based approach (Table 2.5), the fitness function of the SCV method is based upon an exact, rather than asymptotic, estimation of MISE. Therefore, the predictive accuracy of SCV is expected to be the same as or better than that of the DPI approach (Wand and Jones, 1995).

Single variable calibration (SVC) and multi-variable calibration (MVC)

The most commonly applied trial and error approaches to bandwidth estimation can be classified as single variable calibration (SVC) and multi-variable calibration (MVC). The SVC estimator assumes that a common smoothing parameter is applicable to all input vectors, which increases computational efficiency compared with the MVC estimator, for which smoothing parameter estimates have to be obtained for each input vector, but at the cost of potential reductions in modelling accuracy and flexibility (Gibbs et al., 2006). The fitness function used to define the SVC and MVC estimators can be either extreme event oriented (e.g. squared error) or average event oriented (e.g. mean absolute error) (Dawson et al., 2007). The combination of different optimisation algorithms and modelling objectives results in four smoothing parameter estimators, namely SVCS, SVCA, MVCS, and MVCA. The mathematical formulations of these four estimators can be written as

$$\hat{h}_{SVCS} = argmin_h\{\sum_{i=1}^{n}\left[y^i - \hat{y}(\boldsymbol{X}^j, h)\right]^2\} \tag{2.21}$$

$$\hat{h}_{SVCA} = argmin_h\{\sum_{i=1}^{n}\left|y^i - \hat{y}(\boldsymbol{X}^j, h)\right|\} \tag{2.22}$$

$$\hat{\boldsymbol{h}}_{MVCS} = argmin_{\boldsymbol{h}}\{\sum_{i=1}^{n}\left[y^i - \hat{y}(\boldsymbol{X}^j, \boldsymbol{h})\right]^2\} \tag{2.23}$$

$$\hat{\boldsymbol{h}}_{MVCA} = argmin_{\boldsymbol{h}}\{\sum_{i=1}^{n}\left|y^i - \hat{y}(\boldsymbol{X}^j, \boldsymbol{h})\right|\} \tag{2.24}$$

where $\hat{y}(\boldsymbol{X}^j, \boldsymbol{h})$ is the GRNN prediction based upon the bandwidth vector $\boldsymbol{h} = [h_1 \quad \cdots \quad h_m]^T$. The optimal single smoothing parameter in Eqs. (2.21) and (2.22) is achieved by minimising the errors (either squared errors or mean absolute errors) between the observed data y^i and the predictions $\hat{y}(\boldsymbol{X}^j, h)$. In contrast, the optimal bandwidth matrix in Eqs. (2.23) and (2.24) is obtained by minimising the errors (either squared errors or mean absolute errors) between the observed records y^i and the predictions $\hat{y}(\boldsymbol{X}^j, \boldsymbol{h})$. Unlike the previous methods, the fitness functions of the SVC and MVC based approaches depend only upon the calibration error between observed and predicted output data. Consequently, these approaches are independent of Gaussian assumptions (Table 2.5). In this research, GSS was used to obtain the bandwidths of the SVC estimators, while the evolutionary strategy particle swarm optimisation (PSO) algorithm (Poli et al., 2007), which was written in Fortran, was used for this purpose for the MVC approaches.

2.3.3 Development of benchmark MLP model

In order to assess the performance of the different GRNN models in absolute terms, standard MLPs were developed as benchmarks using the systematic approach outlined in Wu et al. (2014). The model inputs/outputs and training, testing and validation data were identical to those used in the development of the GRNN models. A single hidden layer was used and the optimal number of hidden nodes was determined by trial and error, considering a range of 0-5. The optimal number of hidden nodes for the different models was as follows: 2 (EAR4), 2 (TEAR10), 3 (NL), 3 (river salinity 1 day), 3 (river salinity 5 day), 4 (river salinity 14 day), and 4 (flow 1 day), respectively. The back-propagation (BP) algorithm (with learning rate of 0.1 and momentum of 0.1) was used for calibration.

2.3.4 Model performance assessment

As mentioned in the Introduction and shown in Fig. 2.2, model performance criteria included predictive accuracy and computational efficiency. The

specific measures adopted to assess these two aspects of performance are outlined in the subsequent sections.

Predictive accuracy

As discussed in Bennett et al. (2013), careful selection of appropriate predictive performance measures is extremely important. In this study, predictive accuracy was characterised by six dimensionless criteria (listed in Fig. 2.2), commonly used as evaluation metrics for hydrological prediction problems (Bennett et al., 2013; Dawson et al., 2007; Krause et al., 2005). These criteria include the coefficient of efficiency (CE), the index of agreement (IoAd), the persistence index (PI), and modified forms of CE, IoAd, and PI. These measures were chosen because: they are commonly used in hydrology; they have clear cut-off points to distinguish different extents of accuracy (good, satisfactory, or poor); and they are sensitive to different types of events, which assists performance characterisation with respect to the modelling objective. Particularly, CE compares the performance of the model to a model that only contains the mean of the observations; IoAd compares the sum of squared error to the potential error; and PI compares the sum of squared error to the error based on the predictions of previous observations (Bennett et al., 2013). In order to be able to assess the impact of the modelling objective on model performance, modified versions of these metrics were also used, in which squared error terms are replaced with absolute error terms (see Krause et al., 2005).

Although predictive accuracy was assessed using all of the six performance metrics mentioned above, only the performance based on the averaged IoAd and modified IoAd (MIoAd) is presented in the body of the paper, while the performance based on the other metrics can be found in the APPENDIX-A (Figs. A.1, A.3, & A.5). IoAd is a measure of the overall agreement between the observed and modelled records, and is expressed as

$$IoAd = 1 - \frac{\sum_{i=1}^{n}(y^j - \hat{y}^j)^2}{\sum_{i=1}^{n}(|\hat{y}^j - \bar{y}| + |y^j - \bar{y}|)^2} \quad (2.25)$$

where $\boldsymbol{y^j}$ is the individual observation, $\boldsymbol{\hat{y}^j}$ is the corresponding approximation and $\bar{y}$ is the sample mean of the observations. IoAd is sensitive to the mean and variance differences between the observed and modelled records; however, it is insensitive to systematic positive or negative errors. Good performance corresponds to IoAd values greater than or equal to 0.9, and model performance with an IoAd less than 0.8 is considered to be poor (Dawson et al., 2007).

The adopted MIoAd is very similar to Eq. (2.25), except that the squared error terms are replaced by the absolute value in both the numerator and denominator, so that performance becomes average event, rather than extreme event, sensitive. Details of the derivations and applications of the MIoAd can be found in Krause et al. (2005).

The reason for detailing the sensitivity of the performance criteria to the average trends and extreme events is so that an assessment of the impact of the error basis of the fitness functions used by the different smoothing parameter estimators on the performance of the GRNN models with different modelling objectives can be made.

Computational efficiency

Computational efficiency was measured by computational time (CT) (measured by a dual processor 2.6 GHz Intel Machine), which was based on the average CPU clock speed (in seconds), as shown in Fig. 2.2.

2.3.5 Test regime

The test regime was implemented in accordance with Fig. 2.2. Overall, 630 synthetic data sets with 1,575,000 data points were generated, which consisted of 30 replicates of time series generated using 3 different models, for each of which input data were generated from 7 different distributions. Each of the 630 data sets was then divided into training, testing and validation sets and used to calibrate and validate 9 GRNN models, each using 1 of 9 different smoothing parameter estimation techniques, resulting in a total of 5670

GRNN models for the synthetic data. In addition to the experiments with the synthetic data, 4 experiments were conducted with the real data, 3 for the salinity data with different forecasting periods and 1 for the rainfall runoff data. MLPANNs were also developed for each of the 30 replicates of the synthetic data sets and for the 4 experiments with real data. As part of the model development process, the residuals of the training data of all GRNNs and MLPs were checked for replicative validity (see APPENDIX-A Figs. A.2, A.4, and A.6) in accordance with the recommendations of Wu et al. (2014). The residuals were generally 'white noise', indicating that all models can be considered replicatively valid.] The performance of all 5674 models was assessed using the 6 selected predictive accuracy criteria, as well as computational time. Because of the large computational requirements, all tests were coded in PGI Visual Fortran 2008 and run on a Linux 2.6.32.2 operating system. The software used for conducting the numerical experiments is available for others to use, as per the details in the Software Availability at the beginning of this paper.

2.4 Results and discussion

2.4.1 Synthetic case studies

The predictive accuracy for the validation data and computational efficiency of all GRNN models for the synthetic data are summarised in Fig. 2.3 and Fig. 2.4, respectively. The key findings in relation to the impact of the degree of normality, the degree of non-linearity and the modelling objective on GRNN performance (predictive accuracy and computational efficiency) for the different smoothing parameter estimators are presented in *Performance of different smoothing parameter estimation methods*, with the results of the comparison with the MLP benchmark models summarised in *Comparison with MLP*. Preliminary empirical guidelines for the selection of the most appropriate GRNN smoothing parameter estimator based on the properties of the data and the modelling objective derived from the results of the experiments on the synthetic data sets are presented in *Suggested rules and guidelines for use.*

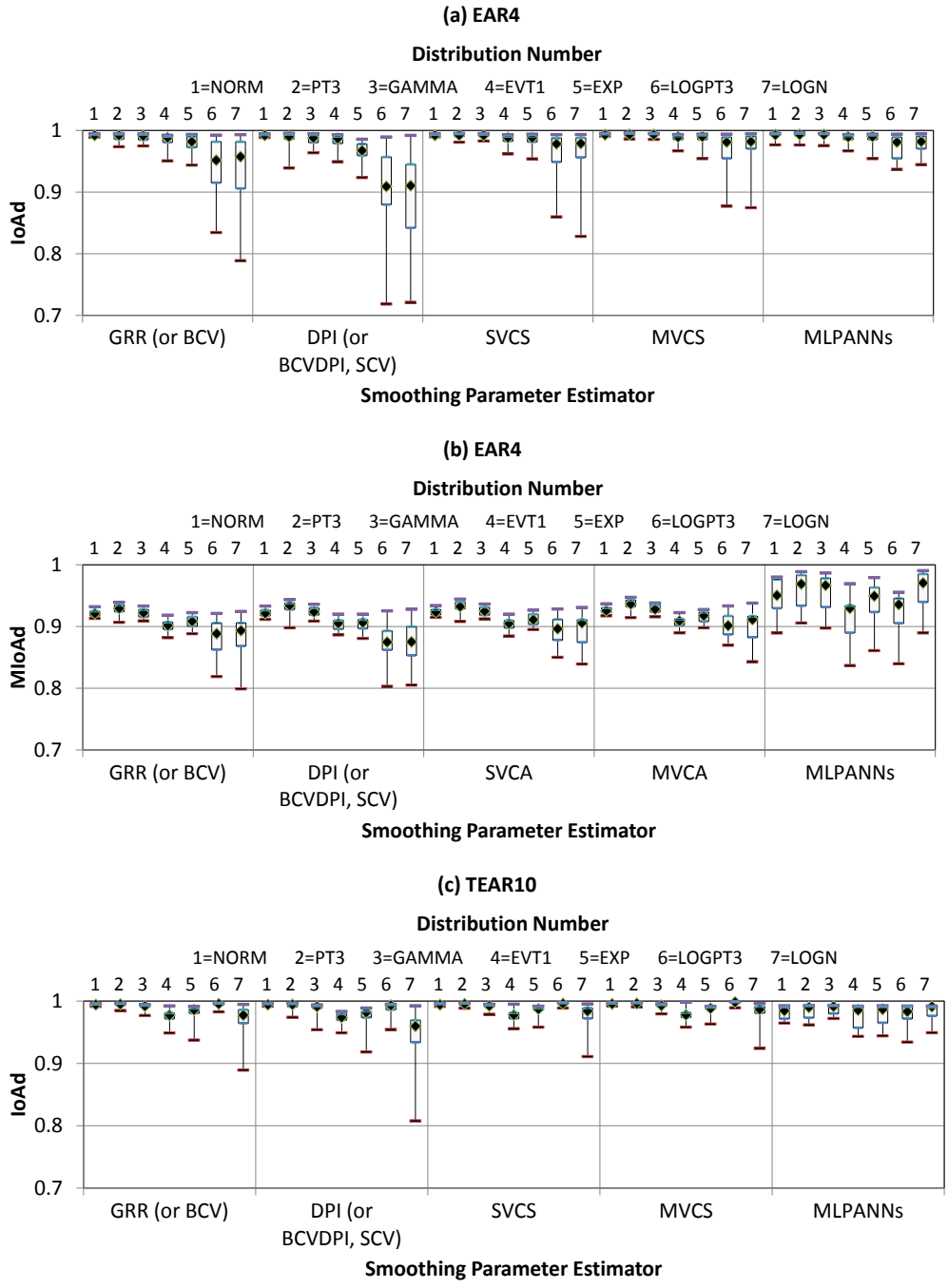
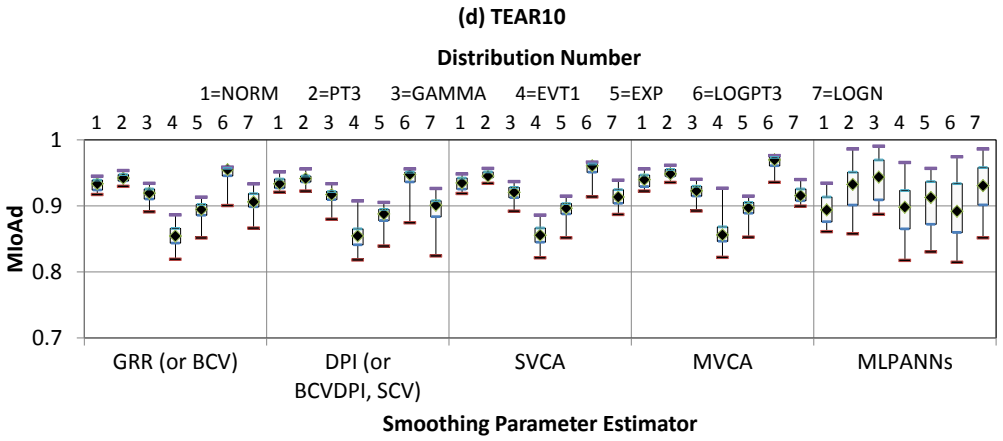
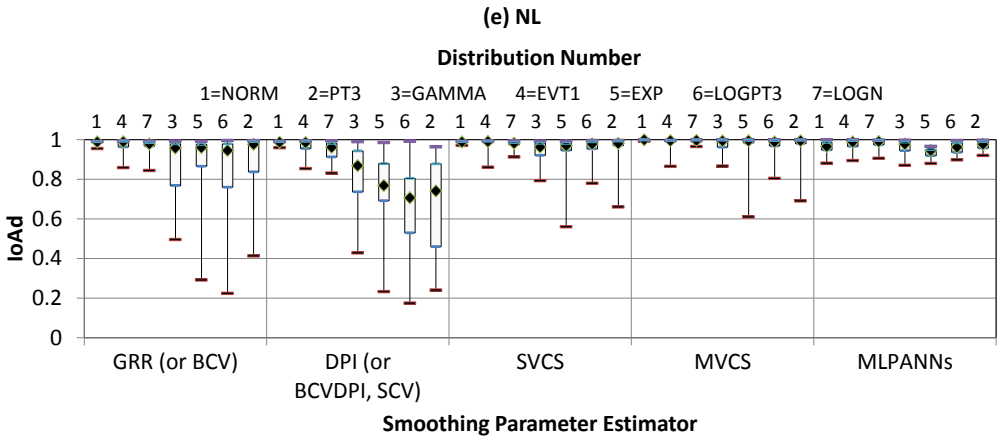


Figure 2.3 Predictive accuracy for the validation data of MLPs and GRNNs for different synthetic data-generating models and distributions for which optimal parameters have been obtained using different methods





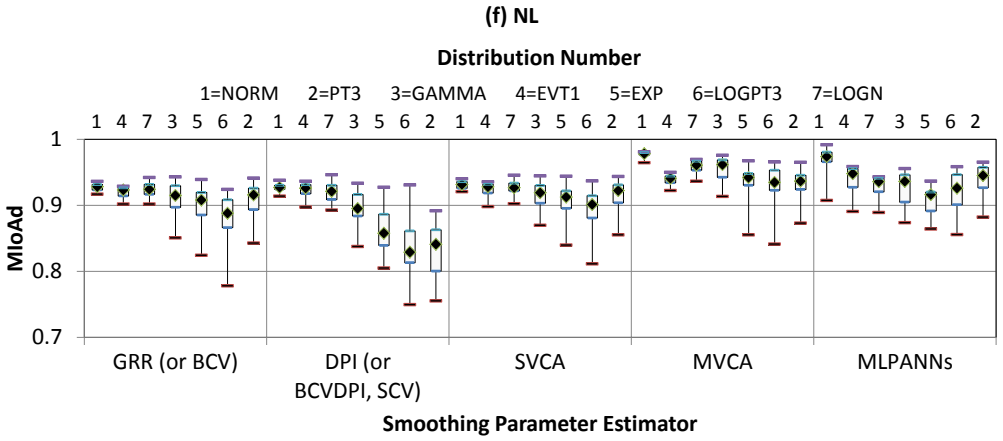


Figure 2.3 *(Continued)*

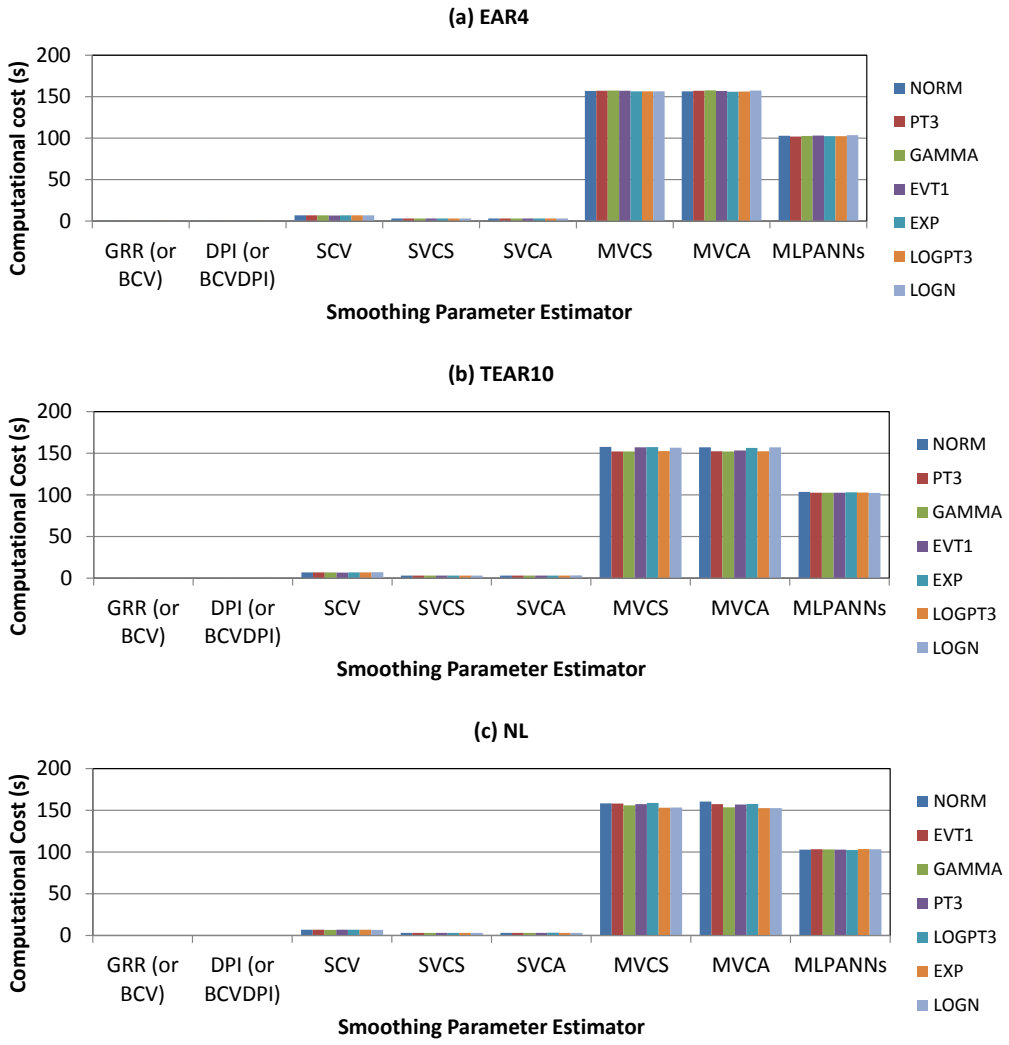


Figure 2.4 Computational efficiency of MLPs and GRNNs for different synthetic data-generating models and distributions for which optimal parameters have been obtained using different methods

Performance of different smoothing parameter estimation methods

Overall, the results indicate that the predictive performance of the GRNN models reduces as the degree of non-Gaussianity in the data increases, especially when the GRR, BCV, DPI, BCDPI and SCV methods were used for smoothing parameter estimation. This suggests that the DPI (or BCVDPI) and SCV methods are not consistently effective in improving the predictive performance of GRNN models for non-Gaussian data compared with using the GRR, despite their reduced reliance on the normality assumption and their increased computational cost. In fact, in many instances, use of these parameter estimation methods resulted in a decrease in predictive performance

compared with that obtained using the GRR, particularly for the more extreme distributions (i.e. LOGPT3, EXP, LOGN in Fig. 2.3).

In contrast, use of the SVCS/SVCA and MVCS/MVCA methods was generally successful in terms of improving the predictive performance of the GRNN models for data with high degrees of non-normality compared with the models for which the GRR was used for smoothing parameter estimation. In fact, when the SVCS/SVCA and MVCS/MVCA methods are used, there is very little degradation in predictive performance with an increase in the non-normality of the data. This is most likely because these smoothing parameter estimation techniques do not rely on any Gaussian assumptions. This makes use of the SVCS/SVCA approaches a particularly attractive option for highly non-Gaussian data, on account of their much smaller computational cost compared with the MVCS/MVCA methods.

While the trends described above apply to all three synthetic data sets, they manifest themselves more strongly for the non-linear (NL) case. This suggests that the combination of non-linear and non-Gaussian data has the potential to result in a marked degradation in the predictive performance of GRNNs, unless the SVCS/SVCA or MVCS/MVCA methods are used. It should also be noted that for the NL case, there was a noticeable improvement in predictive performance when the MVCS/MVCA approach was used instead of the SVCS/SVCA method. However, this improvement was achieved at a significantly increased computational cost.

Comparison with MLP

In the vast majority of cases, the predictive performance of the MLP models was similar to that of the GRNN models for which the SVCS/SVCA and MVCS/MVCA methods were used for smoothing parameter estimation, although the MLPs performed slightly better than the best-performing GRNNs in some instances. In addition, for Gaussian or nearly Gaussian data, the predictive performance of the GRNNs for which the GRR was used for smoothing parameter estimation was very similar to that of the MLPs. Consequently, the results suggest that if a bandwidth estimation technique is

used that is appropriate for the distribution of the data, the predictive performance of GRNNs is very similar to that of MLPs. In addition, this can generally be achieved at a much reduced computational cost, unless the MVCS/MVCA bandwidth estimation technique is used. Furthermore, use of GRNNs eliminates the uncertainty associated with the selection of an appropriate MLP model geometry.

Suggested rules and guidelines for use

Based on the findings of the 5670 computational experiments with the synthetically generated data, a set of preliminary empirical guidelines has been developed for selecting the most appropriate smoothing parameter estimation technique based on the degree of normality and degree of non-linearity of the data, as well as the modelling objective (Fig. 2.5). It should be noted that the smoothing parameter estimation techniques included in the suggested guidelines represent reasonable trade-offs between predictive accuracy and computational efficiency, although it is acknowledged that which trade-offs are optimal is also a function of case-study dependent circumstances and / or user preferences.

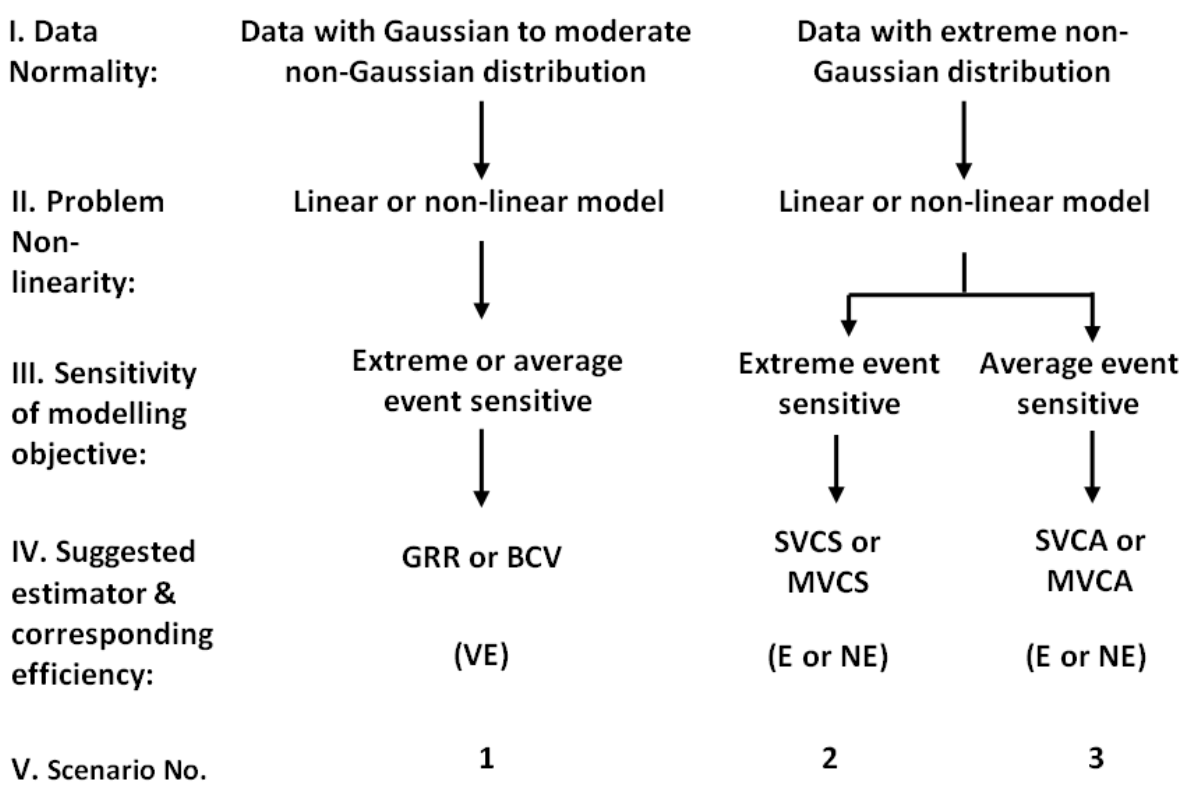


Figure 2.5 Suggested smoothing parameter estimators under different problem situations

(VE = comparatively very computationally efficient, E = comparatively moderately computationally efficient, and NE = comparatively not computationally efficient)

Based on Fig. 2.5, the preliminary empirical guidelines for selecting an appropriate method for estimating the parameter(s) of GRNNs can be grouped into a number of scenarios, as explained below:

Scenario 1: If the problem has input/output data that are mainly mildly non-Gaussian (average $s < 5$ & $k < 30$), the GRR (or BCV) smoothing parameter estimator is recommended, irrespective of linearity and model objective, as these methods are observed to provide good accuracy for these cases at a comparatively high computational efficiency.

Scenario 2: If (i) inputs and outputs are extremely non-Gaussian (average $s > 5$ & $k > 30$) and (ii) the modelling objective is to capture extreme events for a linear or non-linear problem, the use of SVCS or MVCS is suggested. However, this observed increase in predictive accuracy comes at the cost of significantly decreased computational efficiency (particularly for the MVCS).

Scenario 3: If the problem is as in Scenario 2 (extremely non-Gaussian data & linear or non-linear problem), but with a modelling objective that is average magnitude event sensitive, SVCA or MVCA should be adopted.

2.4.2 Real case studies

The results for the two real case studies are given in Figs. 2.6 and 2.7. Fig. 2.6 (a), (b), and (c) show the predictive accuracy for the validation data of river salinity at Murray Bridge 1, 5, and 14 days in advance and the corresponding computational efficiency is illustrated in Fig. 2.7 (a), (b), (c). Fig. 2.6 (d) displays the predictive accuracy for the validation data of runoff at Lock and Dam 10 in the Kentucky River basin 1 day in advance and the corresponding computational efficiency is given in Fig. 2.7 (d).

River salinity at Murray Bridge

By considering the properties of the data for the salinity case study (Table 2.3), and the modelling objective of capturing the averaged salinity trends, this case study corresponds to Scenario 1 in Fig. 2.5. Given this, the predictive performance of the GRNNs developed using the GRR or BCV based methods

was expected to be superior in terms of an appropriate trade-off between predictive accuracy and computational efficiency. This is confirmed by the results, which indicate that predictive performance was not affected significantly by using the different smoothing parameter estimation methods. Although the methods that have reduced reliance on the Gaussian assumption result in a slight increase in predictive performance, this is probably not outweighed by the additional computational costs incurred. However, as mentioned previously, the method that is considered most appropriate is case study and user dependent. For example, if high predictive accuracy was critical in this case and computational efficiency was not an issue, the MVCA based approach would be preferable. As was the case for the synthetic case studies, the predictive performance of the GRNNs is very similar to that of the MLPs, but at a significantly reduced computational cost.

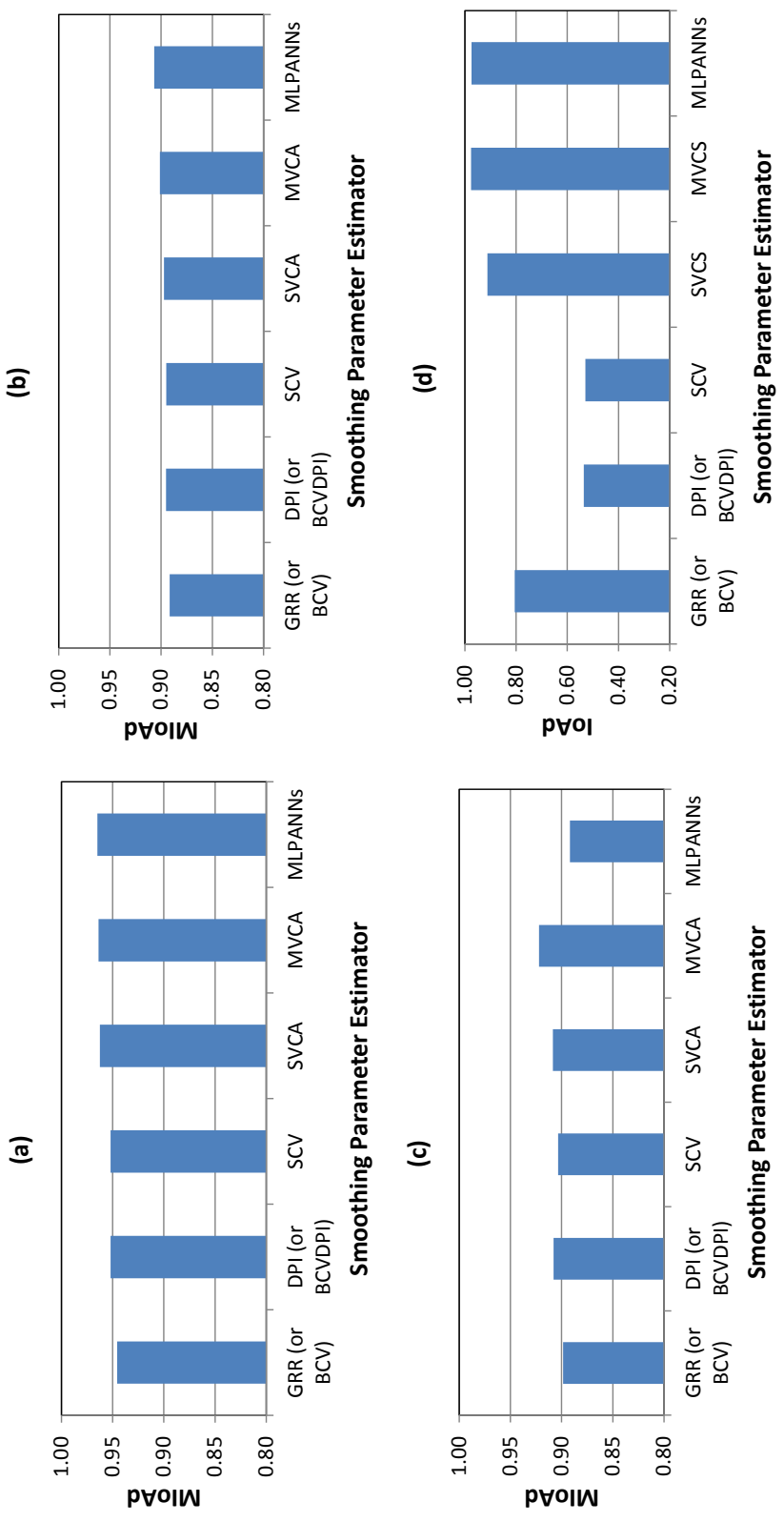


Figure 2.6 Predictive accuracy of MLPs and GRNNs with different smoothing parameter estimators for the validation data for the real case studies ((a), (b), and (c): river salinity at Murray Bridge 1, 5, and 14 days in advance; (d): runoff at Lock and Dam 10 in the Kentucky River basin 1 day in advance)

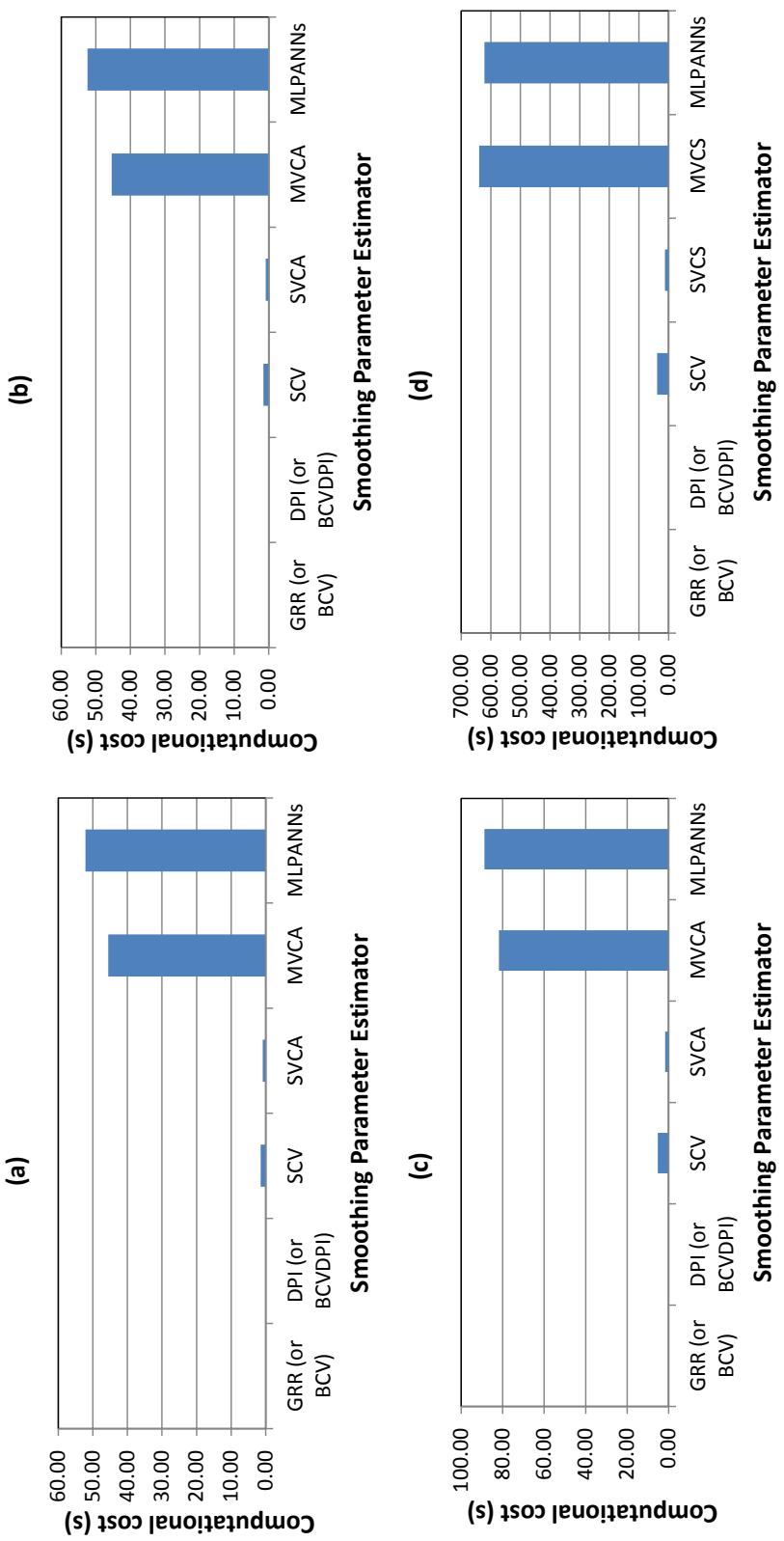


Figure 2.7 Predictive efficiency of MLPs and GRNNs with different smoothing parameters for the validation data for the real case studies

((a), (b), and (c): river salinity at Murray Bridge 1, 5, and 14 days in advance; (d): runoff at Lock and Dam 10 in the Kentucky River basin 1 day in advance)

Rainfall-runoff in Kentucky River basin

By considering the properties of the data for the rainfall-runoff case study (Table 2.4), and the modelling objective of capturing extreme events, this case study corresponds to Scenario 2 in Fig. 2.5. Given this, the predictive performance of the GRNNs developed using the SVCS and MVCS based methods was expected to be superior.

As shown in Fig. 2.6(d), the predictive performance of the GRNNs developed using the SVCS and MVCS based methods was indeed significantly better than that of the GRNNs developed using the other parameter estimation methods and was as good as that of the MLPs. In this case, the SVCS method provided the best trade-off between predictive accuracy and computational efficiency. However, if predictive accuracy was critical, the large increase in computational cost incurred (Fig. 2.7 (d)) for a small increase in predictive accuracy (Fig. 2.6 (d)) when using the MVCS method might be warranted.

2.5 Summary and conclusions

Artificial neural networks (ANNs) have been used extensively for hydrological and water resources modelling over the last two decades. In the vast majority of studies, multi-layer perceptrons (MLPs) have been used as the ANN model architecture. However, obtaining the optimal structure of such models is not an easy task. By using general regression neural networks (GRNNs) as the ANN model architecture, this problem can be overcome, as GRNNs have a fixed model structure. However, there has been limited investigation into the best way to estimate the parameters of GRNNs. In order to address this shortcoming, the performance of nine different GRNN parameter estimation methods was assessed in terms of accuracy and computational efficiency for data with distributions of varying degrees of normality and non-linearity on both synthetic and measured data. In addition, the impact of the objective function on model performance was assessed. In total, 5674 GRNN models were developed as part of the computational

experiments conducted. As a way of benchmarking, the predictive performance and computational efficiency of the GRNN models was also compared with that of MLP models.

The main results from the synthetic case studies show that:

1. The predictive performance of GRNNs developed using the GRR, BCV, DPI, BCVDPI, and SCV based methods was generally influenced by the distribution of the input/output data because of their dependence on the Gaussian assumption (assuming the underlying density follows a normal distribution).
2. Compared to the GRNNs developed using the GRR, use of the DPI, BCVDPI, and SCV based methods did not effectively improve predictive performance, despite their decreased dependence on the Gaussian assumption and increased computational cost.
3. The predictive accuracy of GRNNs developed using the SVCA/SVCS and MVCA/MVCS based methods was relatively insensitive to the distribution of the input/output data because of their independence of the Gaussian assumption.
4. There is a distinct trade-off between predictive accuracy and computational efficiency for the methods investigated, with a reduction in computational efficiency for the methods that are least affected by the Gaussian assumption (i.e. SVCA/SVCS and MVCA/MVCS) by several orders of magnitude.
5. If an appropriate smoothing parameter estimation technique is used, the predictive performance of the GRNN models is very similar to that of the MLPANN models, although slightly worse in some instances. However, the computational cost of developing the GRNN models is generally significantly less. In addition, there is no uncertainty in relation to the selection of the most appropriate model structure.

Based on the general observations of the relationship between the performance of the different GRNN parameter estimation methods and the properties of the data and modelling objectives, preliminary empirical guidelines for selecting the GRNN parameter estimation method that

represents good trade-offs between predictive accuracy and computational efficiency were developed.

The validity of the guidelines was tested and confirmed for two case studies with real data, including the forecasting of salinity in the River Murray in South Australia and a rainfall-runoff study in the Kentucky River basin in the USA.

While the results of this study provide useful insights and guidance on the selection of appropriate parameter estimation methods for GRNNs, further research into the possibility of improving the predictive performance of some of the methods that rely on the Gaussian assumption to some degree is warranted, as these methods are much more computationally efficient than the methods that are found to perform well with extremely non-Gaussian data in this study. In particular, the stage number used in the DPI, BCVDPI, and SCV methods may not be sufficient to describe extreme distributions with data accumulated at the boundary and a long tail. The boundary issue (Karunamuni and Alberts, 2005b; Scott, 1992), as another critical issue with the same importance as the bandwidth, needs to be studied further for problems that contain extreme data distributions.

2.6 Acknowledgments

This research was aided by the suggestions and the original code of GRNN from Dr. Rob May and Dr. Greer Humphrey. The authors would also like to thank the three anonymous reviewers, whose input has improved the quality of this paper significantly.

CHAPTER 3 JOURNAL PAPER 2 -

Improved PMI-Based Input Variable Selection Approach for Artificial Neural Network and Other Data Driven Environmental and Water Resource Models

Statement of Authorship

Title of Paper	Improved PMI-based input variable selection approach for artificial neural network and other data driven environmental and water resource models
Publication Status	◉ Published, ○ Accepted for Publication, ○ Submitted for Publication, ○ Publication style
Publication Details	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 Software 65 15-29 DOI: 10.1016/j.envsoft.2014.11.028

Author Contributions

By signing the Statement of Authorship, each author certifies that their stated contribution to the publication is accurate and that permission is granted for the publication to be included in the candidate's thesis.

Name of Principal Author (Candidate)	Xuyuan Li		
Contribution to the Paper	Undertook literature review, developed analytic procedure and numerical models, developed software, and prepared manuscript		
Signature		Date	

Name of Co-Author	Professor Holger R. Maier		
Contribution to the Paper	Supervised manuscript preparation and reviewed draft		
Signature		Date	

Name of Co-Author	Dr. Aaron C. Zecchin		
Contribution to the Paper	Supervised manuscript preparation and reviewed draft		
Signature		Date	

Name of Co-Author			
Contribution to the Paper			
Signature		Date	

Abstract

Input variable selection (IVS) is one of the most important steps in the development of artificial neural network and other data driven environmental and water resources models. Partial mutual information (PMI) is one of the most promising approaches to IVS, but has the disadvantage of requiring kernel density estimates (KDEs) of the data to be obtained, which can become problematic when the data are non-normally distributed, as is often the case for environmental and water resources problems. In order to overcome this issue, preliminary guidelines for the selection of the most appropriate methods for obtaining the required KDEs are determined based on the results of 3,780 trials using synthetic data with distributions of varying degrees of non-normality and six different KDE techniques. The validity of the guidelines is confirmed for two semi-real case studies developed based on the forecasting of river salinity and rainfall-runoff modelling problems.

3.1 Introduction

Artificial neural networks (ANNs) have been applied successfully and extensively to environmental (Adeloye et al., 2012; Ibarra-Berastegi et al., 2008; Luccarini et al., 2010; Maier and Dandy, 1997b; 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., 2012; ASCE, 2000a, b; Dawson and Wilby, 2001; Maier and Dandy, 2000b; Maier et al., 2010; Wolfs and Willems, 2014; Wu et al., 2014b) problems over the last two decades. One of the most important steps in the ANN model development process is the selection of appropriate inputs (Galelli et al., 2014; Humphrey et al., 2014; Maier et al., 2010; May et al., 2011; May et al., 2008b; Wu et al., 2014b). According to Bowden et al. (2005a), if potential inputs that have a pronounced relationship with the modelled output are not included in the model, the performance of the resulting model will be compromised. Conversely, if redundant or superfluous inputs are included, computational efficiency is decreased, calibration becomes more difficult and model parameters are less well defined, potentially making model validation in terms of physical plausibility, as well as knowledge extraction, problematic (Dawson et al., 2014; Galelli et al., 2014; Haimi et al., 2013; Humphrey et al., 2014; Maier et al., 2010; May et al., 2011; Mount et al., 2013).

Given the importance and likely impact of input variable selection (IVS), it is somewhat surprising that in most studies, ad-hoc approaches are used (Maier et al., 2010; Wu et al., 2014b). However, a number of quantitative approaches to IVS for ANN water resources models have already been developed and utilized, such as sensitivity analysis (Jain et al., 1999; Maier and Dandy, 1997a), the Gamma test (Agalbjörn et al., 1997; Noori et al., 2011), partial mutual information (PMI) (Bowden et al., 2005a), hybrid independent component analysis and input variable selection filter (Trappenberg et al., 2006), principal component analysis (Hu et al., 2007), use of the Box-Jenkins method (Box et al., 2013), cross-correlation analysis (Chua and Wong, 2010), distributed evaluation of local sensitivity analysis (Rakovec et al., 2014), recursive variable selection (RVS) embedded in dynamic emulation models

(Castelletti et al., 2012a; Castelletti et al., 2012b), and tree-based iterative input variable selection (Galelli and Castelletti, 2013). Among these, PMI IVS is one of the most promising approaches, as it has a number of desirable properties, such as the ability to account for input relevance, the ability to cater to both linear and non-linear input-output relationships and the ability to determine the relative contribution (significance) of selected inputs (May, 2010). In addition, it has already been applied successfully to a number of studies (Bowden et al., 2005a; Bowden et al., 2005b; Fernando et al., 2009; He et al., 2011; May et al., 2008a; May et al., 2008b; Wu et al., 2013).

However, current implementations of PMI IVS approaches are not without their limitations. Generally, kernel density estimation (KDE) is used to approximate the probability density function (PDF) needed for the calculation of MI (Bowden et al., 2005a; Bowden et al., 2005b; He et al., 2011; May et al., 2008a; May et al., 2008b; Sharma, 2000a, b). One of the reasons for this is that simple methods exist for KDE that are a function of only a single parameter, the kernel bandwidth, otherwise termed the smoothing parameter (Scott, 1992; Wand and Jones, 1995). While many methods exist for estimating the bandwidth, in almost all existing PMI IVS studies dealing with environmental and water resources problems (e.g. Bowden et al., 2005a,b; May et al., 2008a,b; He et al., 2011) the Gaussian reference rule (GRR) is used for this purpose. The inherent limitation of this implementation of the PMI algorithm is that the input/output data are assumed to follow a Gaussian distribution. However, this is unlikely to be the case, as the distribution of most environmental and water resources data is generally far from normal. As a result, use of the GRR for determining the bandwidth for the KDE needed for MI estimation is likely to result in inaccurate IVS for data that are highly non-Gaussian (Galelli et al., 2014; Humphrey et al., 2014), and over-smoothed bandwidths have been found to result in more accurate MI estimates for such data (Harrold et al., 2001). Consequently, there is a need to investigate the effectiveness of alternative approaches to estimating the bandwidth in PMI IVS so that the performance of this commonly-used algorithm can be improved for data that follow non-Gaussian distributions.

In order to overcome the limitations of existing PMI IVS implementations outlined above, the objectives of the current study are: 1) to assess if, and to what degree, the performance of PMI IVS can be improved for data with different degrees of normality by using alternative bandwidth estimators with reduced reliance on the assumption that the data are normally distributed; and 2) to develop and test a set of preliminary guidelines for selecting the most appropriate bandwidth estimator for data with different degrees of normality. Consequently this paper makes a specific contribution in terms of improving the performance of the PMI algorithm for data that are encountered most commonly in practice.

The remainder of this paper is organised as follows. A detailed explanation of PMI IVS is provided in Section 2, followed by the methodology for meeting the objectives in Section 3. The results are presented and discussed in Section 4. The developed guidelines are validated on the semi-real studies in Section 5, before a summary and conclusions are given in Section 6.

3.2 PMI IVS

Although PMI IVS has been described in Sharma (2000a), Bowden et al. (2005a), May et al. (2008b), He et al. (2011), and May et al. (2011), the implementation of the KDE in 2-D used in this paper has not been explained clearly thus far in this field of research. Consequently, the overall procedure, mathematical details, and relevant assumptions of the PMI IVS algorithm implemented in this paper are discussed in detail below for the sake of completeness. As illustrated in Fig. 3.1, the **first step** is to procure candidate inputs $\boldsymbol{X}$ and output(s) y from the available data in accordance with an understanding of the system. Let: $\boldsymbol{X} = [X_1 \dots X_m]^T$ 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 $j = 1, \dots, n$, where n is the number of observations, $\boldsymbol{X}^j = \left[X_1^j \dots X_m^j\right]^T$ are the observed input data and y^j are the observed output data.

The **second step** is to estimate the marginal PDF of each individual input $f(X_i)$ and the output $f(y)$. The PDF is approximated by kernel density estimation (KDE) in accordance with

$$\hat{f}(X_i) = \frac{1}{n}\sum_{j=1}^{n} K_h(X_i - X_i^j) \tag{3.1}$$

The kernel type K_h used in Eq. (3.1) is the most commonly used Gaussian kernel since the selection of kernel type has negligible impact on the accuracy of KDE (May et al., 2008b; Scott, 1992; Wand and Jones, 1995). The expression of the 1D Gaussian kernel is

$$K_h(\boldsymbol{X}) = \frac{1}{(\sqrt{2\pi}|h|)} exp\left(-\frac{\boldsymbol{X}^2}{2h^2}\right) \tag{3.2}$$

In Eq. (3.2), h is the univariate kernel bandwidth, which determines the accuracy of the KDE (Duong and Hazelton, 2003; Scott, 1992; Wand and Jones, 1995). This single dimensional bandwidth, used for the marginal PDF estimation, directly contributes to the bandwidth matrix used for the joint PDF estimation (as explained later). As mentioned previously, in most studies, the Gaussian reference rule (GRR) has been used for the estimation of the kernel bandwidth in PMI IVS due to its high computational efficiency, ease of implementation, and reasonable stability (Bowden et al., 2005a; He et al., 2011; Huang and Chow, 2005; May et al., 2008b).

The **third step** is to calculate the joint PDF $f(X_i, y)$ between the *i*-th input and the output, which requires the development of a 2-D bandwidth matrix for the joint KDE. The currently used bivariate bandwidth matrix for standardised data is

$$\boldsymbol{H} = h_i^2 \begin{bmatrix} S_{x,i}^2 & S_{xy,i} \\ S_{xy,i} & S_y^2 \end{bmatrix} \tag{3.3}$$

where $S_{x,i}^2$ is the sample variance of the input X_i; $S_{xy,i}$ is the covariance between input X_i and output y, S_y^2 is the sample variance of the output y, and $h_i (h_i = h_{x,i} = h_y)$ is the estimated 1-D kernel bandwidth if the data are standardised, or for non-standardised data

$$\boldsymbol{H} = \begin{bmatrix} h_{x,i}^2 & \rho_{xy,i}h_{x,i}h_y \\ \rho_{xy,i}h_{x,i}h_y & h_y^2 \end{bmatrix} \tag{3.4}$$

(known as a hybrid class of bandwidth matrix), where $\rho_{xy,i}$ is the correlation coefficient between input X_i and output y. According to Wand and Jones (1993), the diagonal terms of the bandwidth matrix adjust the shape of the joint PDF, while the off-diagonal terms control the orientation. The empirical joint density of the i-th X_i input and the output y can be estimated by the Gaussian kernel-based estimator as

$$\hat{f}(X_i, y) = \frac{1}{n}\sum_{j=1}^{n} K_{\boldsymbol{H}}\left(\begin{bmatrix} X_i \\ y \end{bmatrix} - \begin{bmatrix} X_i^j \\ y^j \end{bmatrix}\right) \tag{3.5}$$

where the multivariate kernel is given by

$$K_H(\boldsymbol{X}) = \frac{1}{(\sqrt{(2\pi)^m|\boldsymbol{H}|})}\exp\left[-\frac{1}{2}\boldsymbol{X}^T\boldsymbol{H}^{-1}\boldsymbol{X}\right] \tag{3.6}$$

It should be noted that this approximation is commonly known as the Parzen window density estimation (Cacoullos, 1966; Parzen, 1962). This is valid, however, only if the underlying density is continuous and the first partial derivative at any $\boldsymbol{X}$ is small.

According to Shannon (1948), MI is then approximated as

$$I_{X_i,y} \approx \frac{1}{n}\sum_{j=1}^{n} log[\frac{f\left(X_i^j, y^j\right)}{f\left(X_i^j\right)f(y^j)}] \tag{3.7}$$

(marginal PDFs $f\left(X_i^j\right)$ and $f\left(y^j\right)$ are as defined in Eq. (3.1)) in the **fourth step**. The input with the greatest MI value is the most significant input among the candidate inputs. The significant inputs are selected by means of these four steps during the first run of the algorithm and added to the significant input set X_s, that is, the set is updated to include $X_{i^*} \in X_s$ where $i^* = argmax\{I_{v_i,u}\}$.

In order to remove any redundant information, RE is required in the **fifth step**. RE is at the core of the 'partial' aspect of PMI IVS and the mutual information shared between the residual inputs and output is called PMI (the

term used after the 1st iteration of the PMI IVS). Typically, a general regression neural network (GRNN) (Specht, 1991) is used as the residual estimator in PMI IVS (e.g. May et al., 2008b; He et al., 2011). The residual estimator is used to update the inputs and output by removing the influence of the selected input variables. The updated input is defined as the difference between the current value of the unselected inputs v_i and the estimation of v_i based on the selected input X_{i^*} and is given by

$$v_i^j \leftarrow v_i^j - \hat{m}_{v_i}(X_{i^*}^j) \tag{3.8}$$

where $\hat{m}_{v_i}(X_{i^*}^j)$ is the residual estimate of v_i based on X_{i^*}which removes the shared information between the selected input $X_{i^*}^j$ and the remaining inputs v_i. Similarly, the updated output is

$$u^j \leftarrow u^j - \hat{m}_u(X_{i^*}^j) \tag{3.9}$$

where $\hat{m}_u(X_{i^*}^j)$ is the residual estimate of u based on X_{i^*}, which again eliminates the shared information between the selected inputs X_{i^*} and the output u.

The **sixth step** is to judge the selected input against the chosen stopping criterion. 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). After updating the input and output variables based on the selected input variable, the corresponding PMI is estimated as

$$I_{v_i,u} \approx \frac{1}{n}\sum_{j=1}^{n} \log[\frac{f\left(v_i^j,u_i^j\right)}{f\left(v_i^j\right)f\left(u_i^j\right)}] \tag{3.10}$$

based on Eqs. (3.7), (3.8), and (3.9). If the PMI value of the selected input is still significant according to the applied termination criterion, the above steps are repeated, as shown in Fig. 3.1, until all significant inputs X_s have been determined. In this way, the algorithm can accommodate a large number of potential input variables, as demonstrated in Fernando et al. (2009).

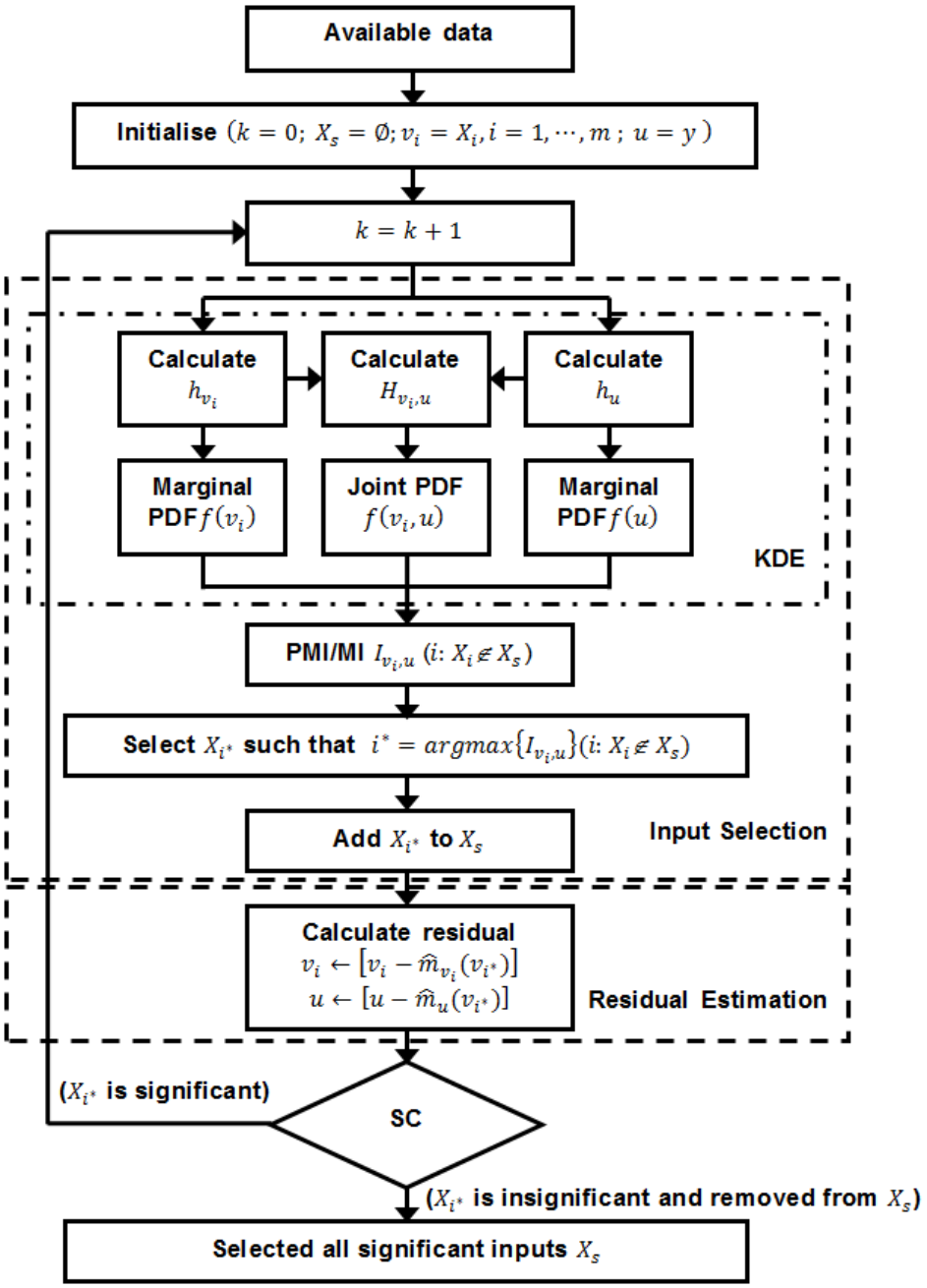


Figure 3.1 Procedure of PMI IVS adopted in this study

(The superscript is omitted, as all operations are performed over the input data $j = 1, \cdots, n$)

3.3 Methodology

The adopted procedure for assessing if, and to what degree, the performance of PMI IVS can be improved for data with different degrees of normality by using alternative bandwidth estimators is outlined in Fig. 3.2. This proposed approach contains three main steps: (i) generation of input/output data for a

range of distributions (with different degrees of normality); (ii) estimation of the kernel PDF and MI for these data using a number of different kernel bandwidth estimators; (iii) assessment of the performance of the IVS process.

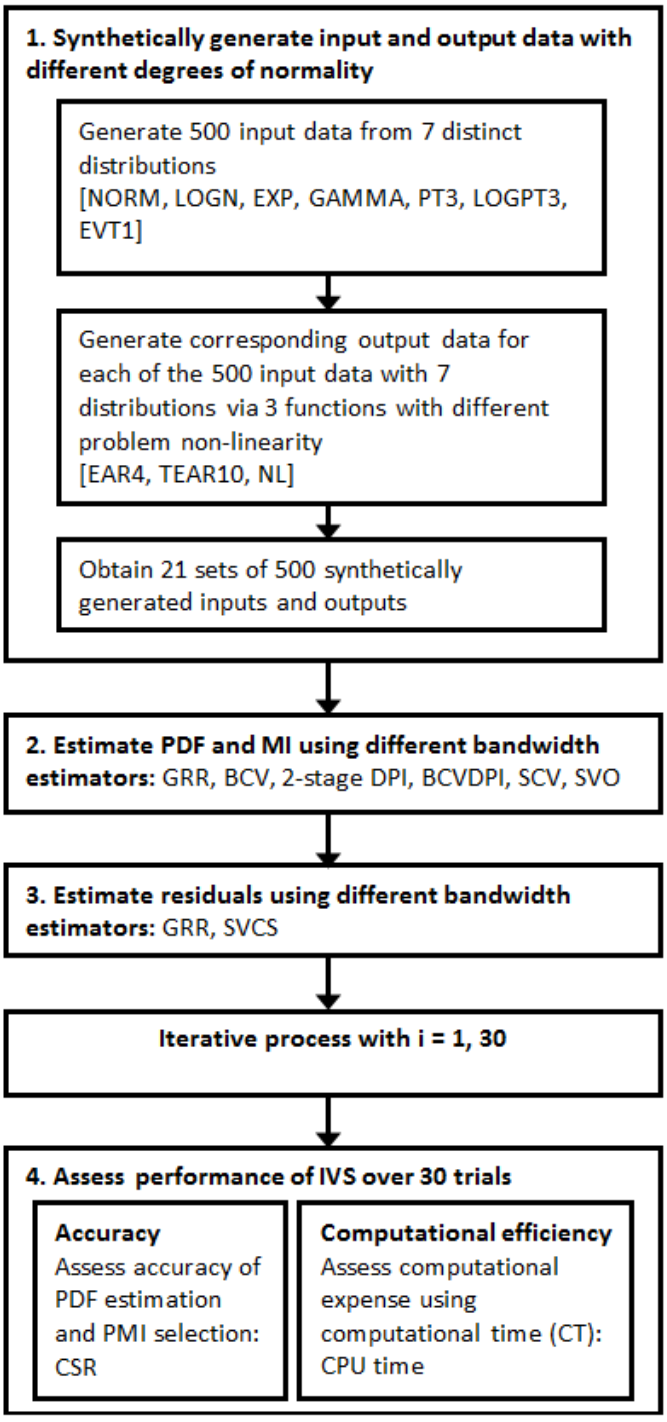


Figure 3.2 Outline of the proposed experimental approach

3.3.1 Generation of 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 with different degrees of normality were generated from distributions with differing degrees of normality, and the corresponding output data were obtained by substituting the generated inputs into synthetic models. Seven distinct distributions were used for input data generation, 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). The degree of normality of the input/output data was measured using skewness and kurtosis in accordance with Bennett et al. (2013). The properties of each distribution are listed in Table 3.1 and 3.2. Although time series of different lengths (i.e. 500, 1,000, and 2,000) were considered in preliminary tests, their impact on the results was found to be insignificant. Therefore 500 data points were generated and the first additional 25 points were rejected in order to prevent initialisation effects (May et al., 2008b).

Table 3.1 Details 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
NORM	Mean=3.0; sd =1.0	0.000	-0.013	High
GAMMA	Shape=2.0; Scale=1.0	1.370	2.638	High
LOGN	Mean=0.5; sd=1.0	5.326	53.694	Low
EXP	Rate=1.0	2.132	7.219	Moderate
PT3	Shape=2.5; Scale=3.0; Location=2.0	1.251	2.381	High
LOGPT3	Shape=0.5; Scale=0.2; Location=2.0	4.792	43.265	Low
EVT1	Shape=0.0; Scale=0.5; Location=10.0	1.198	2.880	High

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

Table 3.2 Details 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
NORM	Mean=3.0; sd =1.0	1.826	5.158	High
GAMMA	Shape=2.0; Scale=1.0	10.520	192.091	Low
LOGN	Mean=0.5; sd=0.4	5.389	47.767	Low
EXP	Rate=1.0	14.029	334.408	Low
PT3	Shape=0.5; Scale=1.0; Location=0.5	16.271	514.270	Low
LOGPT3	Shape=0.5; Scale=0.2; Location=0.5	14.261	390.522	Low
EVT1	Shape=0.1; Scale=0.0; Location=10.0	1.788	9.807	Moderate

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

The three synthetic models used for generating the known outputs, given a set of inputs, included a linear exogenous auto-regressive time series model (EAR4), a threshold exogenous auto-regressive time series model (TEAR10), and a non-linear input-output function (NL), as they are representative of general water engineering problem scenarios with increasing degrees of problem non-linearity and are based on those used for this purpose in previous studies (Bowden et al., 2005b; Galelli and Castelletti, 2013; Li et al., 2014b; May et al., 2008b). The equation of the EAR4 model is given by

$$x_t = 0.6x_{t-1} - 0.4x_{t-4} + p_{t-1} + 0.1\varepsilon_t \tag{3.11}$$

where x_t stands for the output time series; x_{t-n} represents the input time series with lag n; p_{t-n} is the exogenous input with lag n; and $0.1\varepsilon_t$ is the introduced error term (as explained later). The equation for the TEAR10 model is given by

$$x_t = \begin{cases} -0.5x_{t-6} + 0.5x_{t-10} - 0.3p_{t-1} + 0.1\varepsilon_t; & x_{t-6} \le 0 \\ 0.8x_{t-10} - 0.3p_{t-1} + 0.1\varepsilon_t; & \text{otherwise} \end{cases} \tag{3.12}$$

and the equation for NL is given by

$$y = (x_2)^3 + x_6 + 5\sin(x_9) + 0.1\varepsilon_t \tag{3.13}$$

The first two synthetic models (Eqs. (3.11) and (3.12)) were modified from those used May et al. (2008b) through the introduction of an independent lagged input p_{t-1} into all exogenous AR models, and the p_{t-1} were sampled

from the distributions outlined in Table 3.1. The third synthetic model (Eq. (3.13)) was modified from the one used in Bowden et al. (2005a) through a slight adjustment of the significance (coefficient) of each input, and each input was sampled based on the distributions outlined in Table 3.2. For all three synthetic models, the error term $0.1\varepsilon_t$ was added to introduce noise without obscuring the influence of the actual independent variables. The noise term ε_t followed a standard normal distribution $N(0,1)$. In addition, for each synthetic model, 22 redundant or irrelevant input variables of different lags were included, so that the effectiveness of PMI IVS could be tested.

3.3.2 Estimation of PDF and MI using different bandwidth estimators

The kernel bandwidths used to estimate the PDF and MI for the synthetic and semi-real data sets were approximated by six different bandwidth estimators, including the Gaussian reference rule (GRR), biased cross validation (BCV), 2-stage direct plug-in (DPI), a combination of BCV and DPI (BCVDPI), smoothed cross validation (SCV) and single variable optimisation (SVO) (Fig. 3.2). These bandwidth estimators were selected because they have distinct dependence on the Gaussian assumption. The mathematical details of each method are given in the following sections.

Gaussian reference rule (GRR) As the most commonly used bandwidth estimator, the GRR is applied as the benchmark approach in this study. It approximates the bandwidth by minimising the asymptotic mean integrated squared error (AMISE) between the unknown probability function f of the given data and the KDE $\hat{f}(\cdot;h)$ under the integrability assumption of f, in accordance with Scott (1992) and Wand and Jones (1995). The expression of AMISE is given as

$$\mathrm{AMISE}\{\hat{f}(\cdot;h)\} = (nh)^{-1}R(K) + \frac{1}{4}h^4\mu_2(K)^2\widetilde{R(f'')} \tag{3.14}$$

where K is the kernel function; $R(K)$ is the integrated square of the kernel function; $\mu_2(K)$ is the second moment of K; and $\widetilde{R(f'')}$ is the integrated squared second derivative of f. According to Wand and Jones (1995),

although it is ideal to determine the bandwidth by directly investigating the mean squared error (MSE) (summation of bias and variance), its expression depends on the bandwidth in a complicated way, which makes it difficult to interpret the impact of the bandwidth on the performance of the KDE. Consequently, AMISE was developed with consideration of the bias and the variance of the approximated kernel density function $\{\hat{f}(\cdot;h)\}$ (assuming that the bandwidth approaches 0 at a rate slower than n^{-1} and K has a finite 4th moment and symmetry about origin) to overcome such issues and the optimal univariate bandwidth with respect to the AMISE can be derived as

$$\hat{h}_{GRR,i} = \left(\frac{3}{4}\right)^{\frac{1}{5}} \sigma n^{\frac{-1}{5}} \tag{3.15}$$

by assuming that the data follow a Gaussian distribution and by adopting a Gaussian kernel. A detailed derivation of Eq. (3.15) is given in Wand and Jones (1995) and Scott (1992). The detailed derivation is also given in APPENDIX-B B.1.

Biased cross validation (BCV) Although the BCV based bandwidth estimator also minimises the AMISE, and depends on the Gaussian assumption through minimising the AMISE under the assumption of normally distributed data, it is a combination of a cross-validation and 'plug-in' approach, which is potentially more stable than the GRR (Scott and Terrell, 1987) as its asymptotic variance is considerably lower. The BCV is achieved via replacing the unknown $\widetilde{R(f'')}$ in Eq. (3.14) by a cross-validation kernel estimator $\widetilde{R(f'')} = n^{-2}\sum\sum_{p\neq q}(K'' * K'')(X_i^p - X_i^q)$ and the optimal bandwidth is then determined by minimising the approximation of the AMISE with the cross-validation term. Therefore its expression is given as

$$\hat{h}_{BCV,i} = argmin_h \left\{(nh)^{-1}R(K) + \frac{1}{4}h^4\mu_2(K)^2 n^{-2}\sum\sum_{p\neq q}(K'' * K'')(X_i^p - X_i^q)\right\} \tag{3.16}$$

where K'' denotes the second derivative of kernel K and $*$ is the convolution operation and the golden section search (GSS) method (Press et al., 1992) was

applied for the purpose of univariate optimisation in the current study. A detailed derivation of Eq. (3.16) is given in Wand and Jones (1995).

2-stage direct plug-in (DPI) As with the GRR and BCV based approaches, the DPI estimates the optimal bandwidth by minimising the AMISE. For univariate KDE, the optimal bandwidth for Eq. (3.14) can be derived as $\left[\frac{R(K)}{\mu_2(K)^2\widehat{R(f'')}n}\right]^{\frac{1}{5}}$ in accordance with Wand and Jones (1995). The DPI is then established through replacing the unknown $\widehat{R(f'')}$ in $\left[\frac{R(K)}{\mu_2(K)^2\widehat{R(f'')}n}\right]^{\frac{1}{5}}$ by a pilot kernel estimation of the r-th order integrated squared density derivative $\hat{\varphi}_r(g)$ (where g is the pilot kernel bandwidth; L is the pilot kernel; and r is the stage number), according to Park and Marron (1992). Hence the univariate bandwidth estimator of DPI becomes

$$\hat{h}_{DPI,i} = \left[\frac{R(K)}{\mu_2(K)^2\hat{\varphi}_4(g)n}\right]^{\frac{1}{5}} \tag{3.17}$$

where $\hat{\varphi}_4(g)$ is the fourth order integrated squared density derivative, which is approximated by the pilot kernel L with a pilot bandwidth g (Hall and Marron, 1987; Jones and Sheather, 1991). Although the pilot kernel L can be identical to the Gaussian kernel K, the pilot bandwidth g is estimated by minimising the asymptotic mean squared error (AMSE), resulting in

$$g = \left[\frac{K!L^{(r)}(0)}{-\mu_k(L)\hat{\varphi}_{r+k}(g)n}\right]^{\frac{1}{r+k+1}} \tag{3.18}$$

where k represents the order of the pilot kernel L (normally $k = 2$); r is the stage number of L; and $\mu_k(L)$ is the k-th moment of L. Although the stage number r determines how many kernel estimations are required to approximate $\hat{\varphi}_4(g)$ based upon the higher order integrated squared density derivative and more stages can result in a better estimation, determination of the optimal stage number is not trivial and there is a trade-off between an increase in accuracy and computational efficiency (Wand and Jones, 1995). Consequently, the stage number used for the current study was two, as suggested by Aldershof (1991) and Park and Marron (1992), which results in

a desirable balance between the effectiveness and computational cost of the pilot kernel. The motivation behind the DPI is that the dependence of the Gaussian assumption is attenuated by introducing the pilot kernel estimation with $r > 0$, which makes the estimation more sensitive to the actual distribution. In fact, the GRR can be treated as a special case of the DPI with $r = 0$ (see APPENDIX-B B.1). A detailed derivation of Eqs. (3.17) and (3.18) can be found in Wand and Jones (1995), which can also be illustrated in APPENDIX-B B.1.

Combination of BCV and DPI (BCVDPI) The BCVDPI is simply a combination of the BCV and the DPI based approaches. The motivation behind this method is to maintain the advantage of low asymptotic variance in BCV, while adding the feature of reduced Gaussian dependence from the pilot kernel estimator used in DPI. Hence, the BCVDPI is implemented by replacing the cross-validation kernel estimator $n^{-2}\sum\sum_{p\neq q}(K'' * K'')(X_i^p - X_i^q)$ in $\hat{h}_{BCV,i}$ (Eq. (3.16)) with the $\hat{\varphi}_4(g)$ used in $\hat{h}_{DPI,i}$ (Eq. (3.17)), resulting in the following expression

$$\hat{h}_{BCVDPI,i} = argmin_h \left\{(nh)^{-1}R(K) + \frac{1}{4}h^4\mu_2(K)^2\hat{\varphi}_4(g)_{DPI}\right\} \tag{3.19}$$

As such, the BCVDPI inherits the reduced dependence on the Gaussian assumption from the 'plug-in' term $\hat{\varphi}_4(g)$ and the optimal bandwidth is approximated by minimising the AMISE, which was obtained for the BCVDPI in this study by optimisation with the GSS.

Smoothed cross validation (SCV) Although the concept behind the SCV based bandwidth estimator is similar to that underpinning the aforementioned four approaches, SCV aims to minimise the exact MISE (EMISE), rather than the AMISE used in the other four methods. The main difference between the EMISE and AMISE is that the former estimates MISE as a summation of the exact integrated squared bias and the approximation of the integrated variance of $\hat{f}(\cdot; h)$, while the later approximates MISE by integrating MSE (summation of bias and variance) with the integrability assumption and the asymptotic feature of the integrated squared bias. The EMISE derived for SCV is given as

$$\text{EMISE}_{SCV,i}(h) = (nh)^{-1}R(K) + \widehat{ISB}(h) \tag{3.20}$$

where the integrated squared bias $\widehat{ISB}(h)$ is estimated by

$$\widehat{ISB}(h) = n^{-2}\sum_{p=1}^{n}\sum_{q=1}^{n}\left(K_h * K_h * L_g * L_g - 2 * K_h * L_g * L_g + L_g * L_g\right)(X_i^p - X_i^q) \tag{3.21}$$

where K_h and L_g are the Gaussian kernels with kernel bandwidth h and pilot kernel bandwidth g, respectively (Hall et al., 1992; Wand and Jones, 1995). g is a function of a series of pilot kernel bandwidths, each estimated based upon sequentially higher order integrated squared density derivatives, and up to the 10th order was applied in this study based on Wand and Jones (1995). The SCV based optimal univariate bandwidth is then determined as

$$\hat{\text{h}}_{\text{SCV,i}} = \text{argmin}_{\text{h}}\{\text{EMISE}_{\text{SCV,i}}(\text{h})\} \tag{3.22}$$

A detailed derivation of Eq. (3.22) can be found in Wand and Jones (1995), which is also given in APPENDIX-B B.1. Although the dependence on the Gaussian assumption of SCV is also reduced by introducing the pilot kernel estimation, which is similar to that of the DPI, the predictive accuracy of the former is expected to be the same as or better than that of the latter due to minimising EMISE, rather than AMISE.

Single variable optimisation (SVO) Unlike the five estimators mentioned above, SVO, developed in this paper, determines the best bandwidth by minimising the Kolmogorov-Smirnov (K-S) statistic (Parsons and Wirsching, 1982) between the empirical and estimated CDFs. This method does not depend on the Gaussian assumption, nor the approximation of the MISE. The optimal univariate kernel bandwidth is determined as

$$\hat{h}_{SVO,i} = argmin_h\{sup_{j=1\ldots n}\left|F_{emp}\left(X_i^j\right) - F_{est}\left(X_i^j\right)\right|\} \tag{3.23}$$

where $F_{emp}\left(X_i^j\right)$ is the empirical CDF of the input variable estimated by a histogram; $F_{est}\left(X_i^j\right)$ is the estimated kernel-based CDF of the input variable; and *sup* represents the supremum function. The adopted optimiser was the

GSS. The performance of the empirical histogram is a function of the histogram bin width, therefore a number of bin widths (from 0.001 to 1.0) were tested via sensitivity analysis. Although alternative ways can be used to estimate the histogram bin width for each case, the results of the sensitivity analysis (as shown in APPENDIX-B Figs. B.2.4 to B.2.6) suggest that a bin width of 0.01 was adequate for the purposes of this study.

It should be noted that the introduced kernel bandwidth estimators were implemented directly for the estimation of the univariate marginal PDF, which then extended to the bivariate joint PDF in conjunction with the bandwidth matrix, as mentioned in Section 2 (as in Eqs. (3.3) to (3.6)).

3.3.3 Performance assessment

As mentioned in the Introduction and described in Fig. 3.2, PMI performance was assessed based on selection accuracy and computational efficiency. Selection accuracy was characterised by the correct selection rate (CSR), which corresponds to the percentage of times the correct inputs are selected in the 30 independent trials with different instances of a particular data set, as was done in May et al. (2008b) and Galelli and Castelletti (2013). In addition, the degree of over- and under-estimation of the correct inputs was also assessed, in order to provide additional information on selection accuracy (see Galelli et al., 2014).

Computational efficiency was measured using the average CPU time (measured by a dual processor 2.6 GHz Intel Machine).

3.3.4 Test regime

The software used for conducting the numerical experiments was coded in Fortran 90/95 and run on a Linux 2.6.32.2 operating system. As outlined in Fig. 3.2, 630 synthetic data sets were generated, which consisted of a combination of 30 replicates, for each of the three synthetic models with input data generated from the seven distributions. For the 630 data sets, each of the

6 different kernel bandwidth estimators was used for KDE, resulting in a total of 3,780 tests for the synthetic case studies.

The residual estimation required for PMI estimation (see Section 2) was carried out using a GRNN, as was the case in previous studies (e.g. Bowden et al., 2005a; May et al., 2008b; Fernando et al., 2009). The empirical guidelines proposed by Li et al. (2014b) for identifying the most appropriate bandwidth estimation approach based on the distributional properties of the data were used in order to isolate the impact of different bandwidth estimators for residual estimation on IVS accuracy as much as possible. Details of the GRNN bandwidth estimators used for the different datasets resulting from the application of these empirical guidelines are given in Table 3.3.

Table 3.3 GRNN bandwidth estimation techniques used for residual estimation during the PMI IVS process (based on the guidelines from Li et al. (2014b))

Synthetic data set 1				**EAR4**			
Data distribution	NORM	EVT1	PT3	GAMMA	EXP	LOGN	LOGPT3
Bandwidth estimator	GRR	GRR	GRR	GRR	GRR	SVCS	SVCS
Synthetic data set 2				**TEAR10**			
Data distribution	NORM	EVT1	PT3	GAMMA	EXP	LOGN	LOGPT3
Bandwidth estimator	GRR	GRR	GRR	GRR	GRR	SVCS	SVCS
Synthetic data set 3				**NL**			
Data distribution	NORM	EVT1	LOGN	PT3	EXP	LOGPT3	GAMMA
Bandwidth estimator	GRR	GRR	SVCS	SVCS	SVCS	SVCS	SVCS

(GRR denotes for Gaussian reference rule; SVCS stands for single variable calibration with squared error based fitness function)

The Akaike Information Criterion (AIC) (Akaike, 1974) was used as the stopping criterion (i.e. to decide when to stop adding inputs to the selected set) because it offers a trade-off between model accuracy and generalisation ability (Akaike, 1974; Bennett et al., 2013; Dawson et al., 2007; May et al., 2008b), has been found to perform well compared with alternative stopping criteria (May et al., 2008b) and has been successfully applied to a number of previous studies using PMI IVS (e.g. May et al., 2008a,b; He et al., 2011; Wu et al., 2013). The AIC stopping criterion for PMI IVS is computed as

$$AIC = n \times ln[\frac{1}{n}\sum_{j=1}^{n}(y^j - \hat{y}^j)^2] + 2e \quad (3.24)$$

where $\hat{y}^j$ denotes the estimated output and e is the number of effective inputs, measured by the trace of the $n \times n$ hat-matrix in KDE (May et al., 2008b). The performance of all 3,780 synthetic tests was assessed against the performance criteria detailed in Section 3.3.

3.4 Results and discussion

Within the following, Section 4.1 focuses on assessing the selection accuracy of the PMI- IVS methods with different bandwidth estimators applied to the synthetic data sets, and Section 4.2 focusses on computational efficiency. The empirical guidelines for the selection of the most appropriate bandwidth estimators for PMI IVS are presented in Section 4.3.

3.4.1 Selection accuracy

The accuracy of the PMI algorithm with alternative bandwidth estimators for the three synthetic models is summarised in Figs. 3.3, 3.4 and 3.5. As can be seen from Fig. 3.3, for the EAR4 model, the use of alternative bandwidth estimators did not result in any significant improvement in CSR when the input/output data followed Gaussian or nearly Gaussian distributions (average $s < 1.3$ and $k < 3$; i.e. NORM, EVT1, and PT3). For instance, the CSRs when the GRR was used were all above 96.7% for the NORM, EVT1, and PT3 distributions, indicating very high selection accuracy. This result can be explained by the fact that the alternative bandwidth estimators did not provide a significant improvement in KDE accuracy compared with the GRR, as assessed using the Kolmogorov-Smirnov (K-S) statistic (Parsons and Wirsching, 1982), as shown in Figs. 3.6(a), 3.6(b) and 3.6(c). This is not surprising, as the Gaussian assumption used in the KDE is consistent with the actual input/output data distributions, which resulted in an insignificant difference between the empirical and estimated CDFs (Figs. 3.6(a), 3.6(b) and 3.6(c)). To better understand the causes for these findings, the predictive

accuracy of the GRNN models used for residual estimation at each step of the PMI process was assessed using the coefficient of efficiency (CE) (Fig. 3.7), which measures the difference in predictive performance of the model and a model that only contains the mean of the observations (Bennett et al., 2013). As can be seen, the predictive accuracy of the GRNN models was very high, as indicated by CE values close to 1. Consequently, errors in residual estimation were unlikely to contribute to any inaccuracies in PMI IVS.

For data that were moderately non-Gaussian (average $1.3 < s < 5$ and $3 < k < 30$; i.e. GAMMA and EXP), the alternative bandwidth estimators (DPI, BCVDPI, SCV, and SVO) increased the CSR (Fig. 3.3). For example, for data following the EXP distribution, use of the GRR resulted in a CSR of 86.7%, whereas the CSRs for the alternative bandwidth estimators were much higher at 96.7% (SVO), 93.3% (SCV and DPI) and 90.0% (BCVDPI). As can be seen from Figs. 3.3, 3.6(e), and 3.6(f), the trend in improvement in CSR for the different bandwidth estimation techniques is matched by a similar trend in KDE accuracy, suggesting that the improved KDE has a direct impact on CSR. This is because the DPI, BCVDPI, SCV, and SVO based estimators have a reduced dependence on the assumption that the data follow a Gaussian distribution compared with the GRR. As was the case for the data that followed mildly non-Gaussian distributions, the accuracy of the GRNNs used for residual estimation was very high (Fig. 3.7), suggesting that the residual estimation step in the PMI process was unlikely to have any negative impact on CSR.

When the average distributions of the input/output data were extremely non-Gaussian (average $s > 5$ and $k > 30$; i.e. LOGN and LOGPT3), use of the alternate bandwidth estimators still resulted in a noticeable improvement in CSR (Fig. 3.3). However, this improvement was less pronounced for the most extreme distribution (LOGPT3), increasing CSR from 43.3% when the GRR was used to just over 60% when the DPI, BCVDPI, SCV and SVO were used. This is significantly lower than the CSR (over 90%) obtained for all other distributions. The reason for this is likely to be a combination of inaccuracy in KDE, as well as residual estimation. As can be seen in Fig. 3.6(g),

although the use of SVO resulted in improved KDE, the K-S statistic is still outside the 95% confidence limits. In addition, there are significant errors in residual estimation, as shown in Fig. 3.7, even though the bandwidth estimator was based on the empirical guidelines suggested by Li et al. (2014b). As seen in the LOGN and LOGPT3 boxplots in Fig. 3.7, despite the relatively high median, very low CE values were obtained for some of the 30 trials, which is likely to have a negative impact on CSR. These residual estimation inaccuracies are most likely caused by boundary issues (Scott, 1992; Karunamuni and Alberts, 2005), as discussed in Li et al. (2014b), which occur when a symmetrical kernel is applied at a bounded and unsymmetrical boundary, resulting in an under-estimated density near the boundary.

It should also be noted that while the results suggest that improved accuracy in KDE results in improved PMI selection accuracy, consideration of the average ratio of the bandwidths of the 30 replicates used in the MI calculation (see Eq. (3.25)) is also informative.

$$Ratio\ of\ the\ bandwidths = \frac{\hat{h}_{pro,i}}{\hat{h}_{GRR,i}} \tag{3.25}$$

where $\hat{h}_{pro,i}$ stands for the estimated bandwidth based on the proposed bandwidth estimators and $\hat{h}_{GRR,i}$ is the estimated bandwidth based on the GRR (Eq. (3.15)). As part of an empirical study on the effect of different bandwidth ratios on the accuracy of MI estimation, Harrold et al. (2001) found that for highly non-Gaussian data, an over-smoothed bandwidth performs best, with an optimal bandwidth ratio of 1.5. This general finding is confirmed by the results of this study (Table 3.4), which show that bandwidth ratios increase with the degree of non-Gaussianity for the bandwidth estimators that result in more accurate KDE. In addition, the GRR based PMI IVS is found to mainly underestimate the correct number of significant inputs (shown in APPENDIX-B, Fig. B.2.1) for the non-Gaussian cases (e.g. LOGN and LOGPT3), which is consistent with the results (i.e. NL and Bank cases) in Galelli et al. (2014). This can be ascribed to the underestimated bandwidth, as the severity of underestimating the correct number of significant inputs is proportional to the bandwidth ratio outline in Table 3.4. However, alternative bandwidth estimators (i.e. DPI, BCVDPI, SCV, and SVO) tend to correct such

underestimation with increased bandwidths, which sometimes even result in slight overestimation.

Table 3.4 Average ratio of different kernel bandwidths under different distribution scenarios for EAR4 model

	NORM	EVT1	PT3	GAMMA	EXP	LOGN	LOGPT3
GRR	-	-	-	-	-	-	-
BCV	0.964	0.954	0.997	0.984	1.033	1.007	0.997
DPI	0.958	0.886	1.039	0.971	1.265	1.716	1.804
BCVDPI	0.958	0.886	1.039	0.971	1.265	1.716	1.804
SCV	0.971	0.856	1.046	0.967	1.268	1.737	1.804
SVO	0.493	0.418	0.810	0.791	1.190	1.399	1.497

(The average ratio is between each of the alternative kernel bandwidth estimators and the GRR)

The general trends observed for the EAR4 model were confirmed by those obtained for the TEAR10 and NL models, except for the comparatively low accuracy when SVO was used for the NORM and LOGPT3 distributions for the data generated from the TEAR10 model and the overall reduction in CSR for the data generated from the NL model. Even the alternative bandwidth estimators (i.e. DPI, BCVDPI, SCV, and SVO) were found to tend to underestimate the correct number of significant inputs, as shown in APPENDIX-B Fig. B.2.3. This observation is likely to be the result of the combined effect of the reduced KDE and residual estimation accuracy due to boundary issues, particularly influenced by increased problem non-linearity, as discussed below. For example, the non-Gaussianity of the NL model, as measured by skewness and kurtosis, is much more severe than that of the EAR4 and TEAR10 models (as shown in Tables 3.1 and 3.2), suggesting increased potential impact of boundary issues on KDE and residual estimation. For kernel based PDF and MI estimation, the corresponding accuracy of the KDE of the NL model is generally slightly worse than that of the EAR4 and TEAR10 models, as indicated by the K-S values in Figs. 3.6 and 3.8. For residual estimation, the overall accuracy of the NL model was found to be significantly less than that of the EAR4 model, as shown in Figs. 3.7 and 3.9. This can be explained by the fact that the univariate GRNN used for residual estimation is essentially a Nadaraya-Watson regression and therefore the corresponding bias is a function of the regression function $m(X_i)$ and the probability density function $f(X_i)$ with respect to input X_i. According to Fan

(1992), Ruppert and Wand (1994), and Masry (1996), this bias increases as the boundary issue becomes severe. Consequently, the accuracy of residual and PMI estimation is likely to be compromised as the influence of boundary issues increases with increasing problem non-linearity and non-normality.

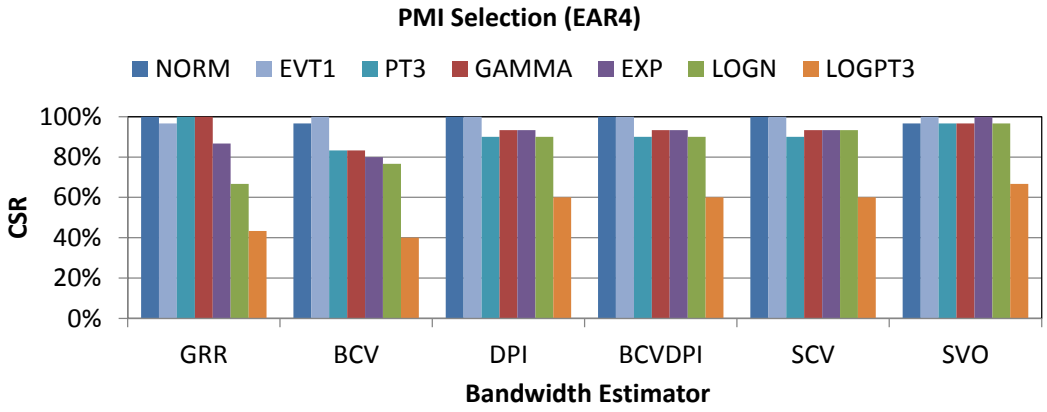


Figure 3.3 Correct selection rate of EAR4 model with alternative bandwidth estimators

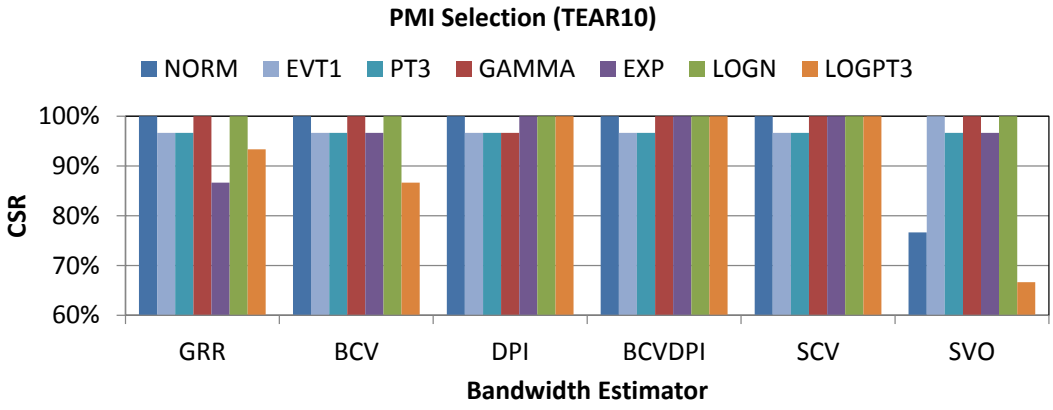


Figure 3.4 Correct selection rate of TEAR10 model with alternative bandwidth estimators

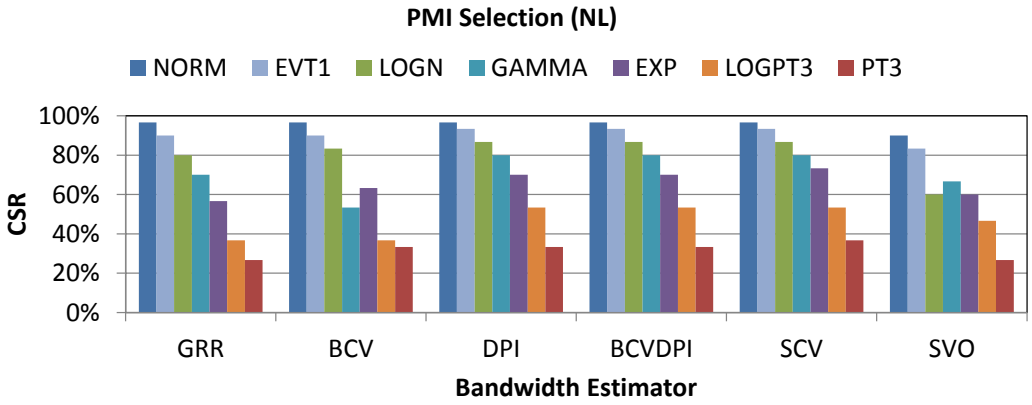


Figure 3.5 Correct selection rate of NL model with alternative bandwidth estimators

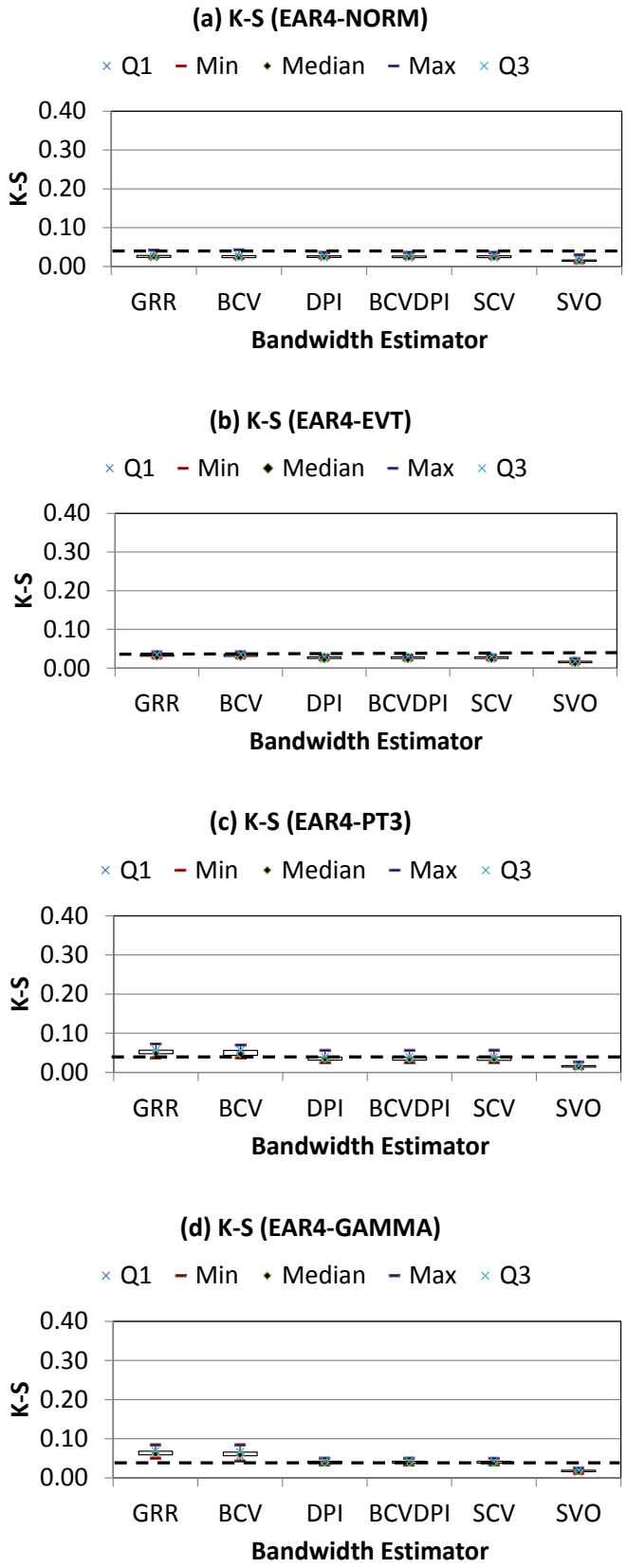
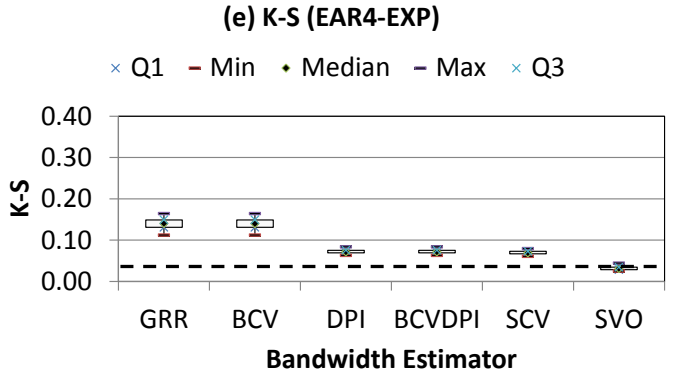
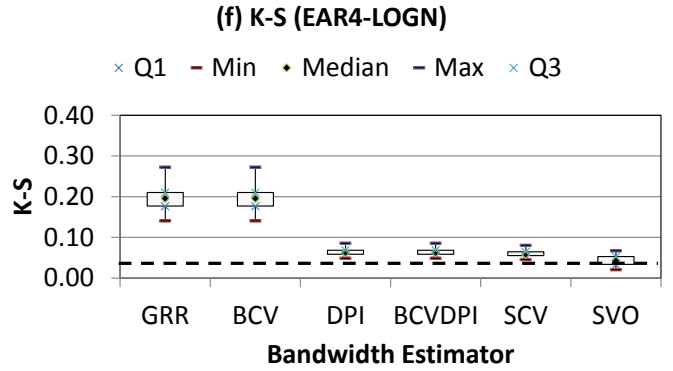


Figure 3.6 KDE accuracy measured by K-S statistics for EAR4 & TEAR10 models

(The dashed line indicates the 95% confidence interval for kernel density estimation based on the Kolmogorov-Smirnov (K-S) statistic (Parsons and Wirsching, 1982))





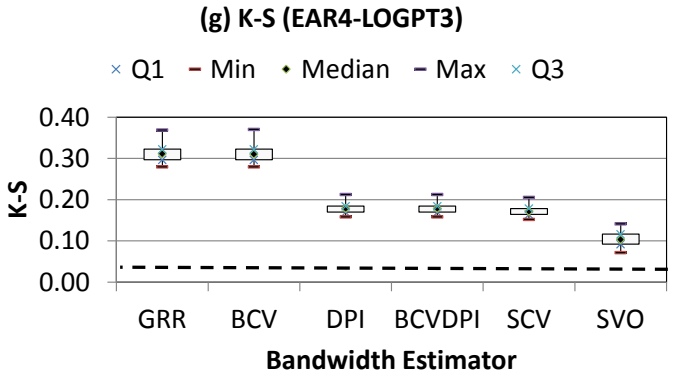


Figure 3.6 *(Continued)*

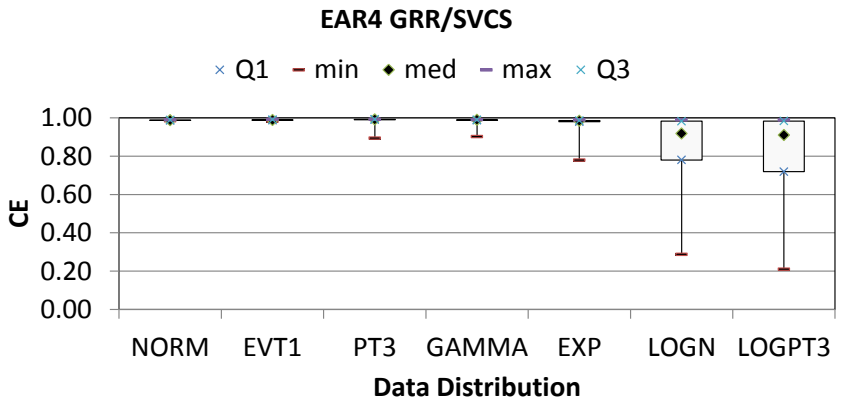


Figure 3.7 Residual accuracy measured by CE for EAR4 model

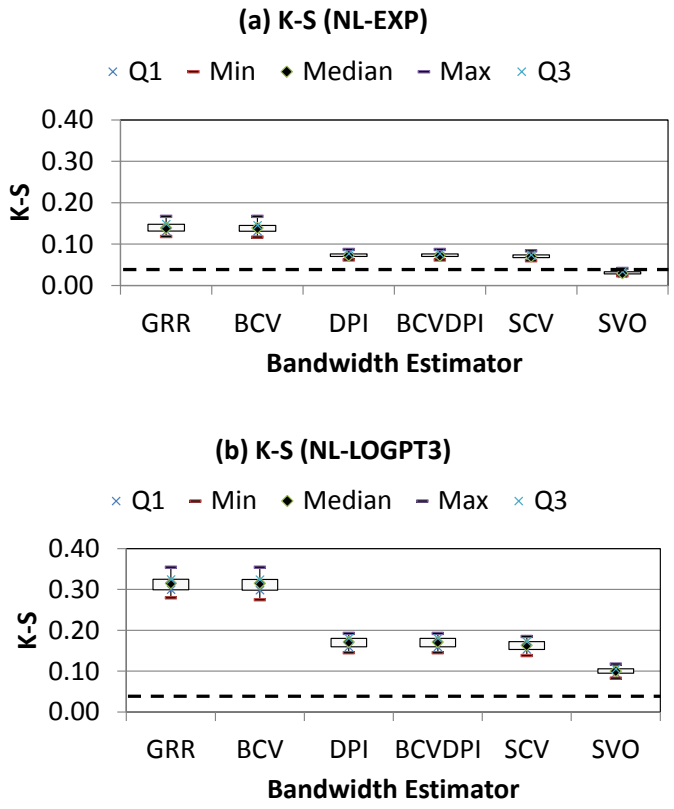


Figure 3.8 KDE accuracy measured by K-S statistics for NL model

(The dashed line indicates the 95% confidence interval for kernel density estimation based on the Kolmogorov-Smirnov (K-S) statistic (Parsons and Wirsching, 1982))

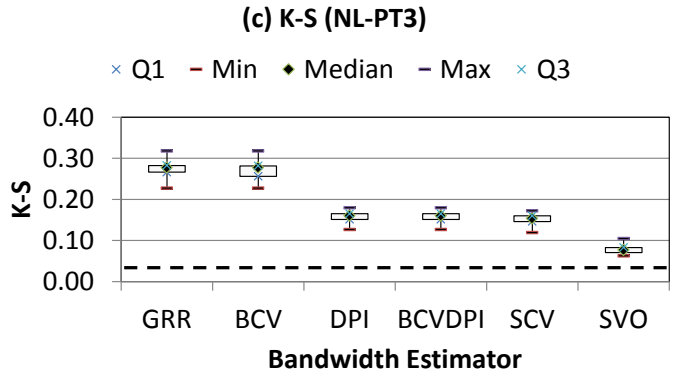


Figure 3.8 *(Continued)*

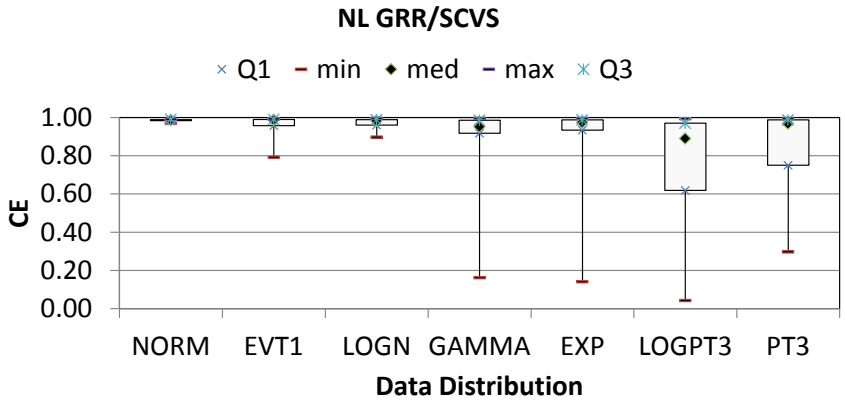


Figure 3.9 Residual accuracy measured by CE for NL model

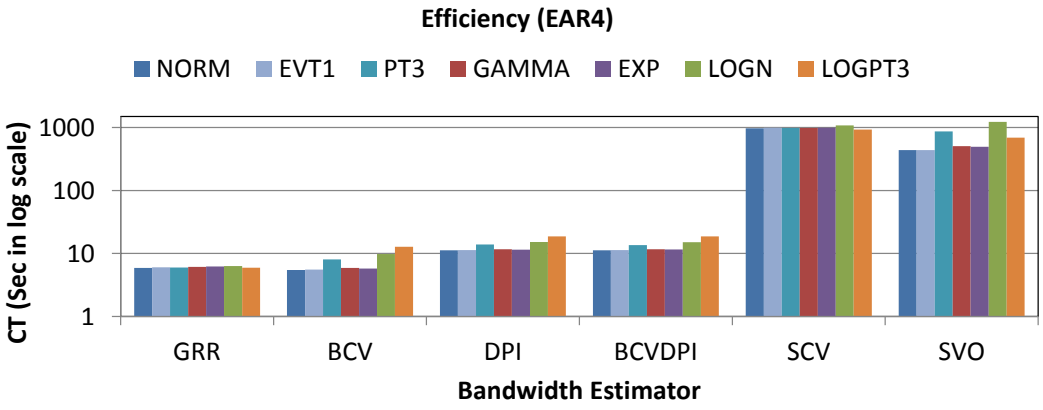


Figure 3.10 Computational efficiency of EAR4 model with different bandwidth estimators

3.4.2 Computational efficiency

The computational efficiency of different bandwidth estimators used for the EAR4 model is given in Fig. 3.10. The GRR based method was found to be the most efficient overall. This can be explained by the fact that the only unknown parameter is the size of the applied data after standardisation (May et al., 2008b). The computational expense of the BCV approach was close to that of the GRR because the fitness functions used are identical, although the BCV requires an additional iterative optimisation process. The average runtimes for both DPI and BCVDPI were double that required by the GRR. This is because of the additional time required for the estimation of the pilot bandwidths during each iteration of the MI estimation (Wand and Jones, 1995). The efficiency of using SVO for bandwidth estimation is significantly less than that of the methods discussed thus far, with an average runtime of 667s, which is over 110 times greater than that associated with the GRR. The increased computational requirements of SVO are a result of the need to estimate the fitness function for each trial bandwidth during the optimisation process. Use of the SCV method was most inefficient, with an average runtime of over 160 times greater than that for the GRR. The inefficiency of SCV can be ascribed to the need to approximate a high order integrated squared density derivative during each iteration of the MI estimation (Wand and Jones, 1995), as well as the optimisation searching process. These findings were supported by the results for the TEAR10 and NL models (See Figs. B.2.7 and B.2.8 in APPENDIX-B).

3.4.3 Suggested rules and guidelines

The preliminary empirical guidelines for selecting the most appropriate kernel bandwidth estimation technique based on the degree of normality of the data (according to the findings of the 3,780 computational experiments with the synthetically generated data) are given in Fig. 3.11. It should be noted that the proposed guidelines represent reasonable trade-offs between selection accuracy and computational efficiency, although it is acknowledged that the

best trade-off is also a function of case-study dependent features and user preferences.

As can be seen in Fig. 3.11, the preliminary empirical guidelines can be categorised into three scenarios, as described below:

Scenario 1: If most of the input/output data follow Gaussian or nearly Gaussian distributions (average $s<1.3$ and $k<3$), the GRR is suggested for residual estimation and the GRR (or BCV) is recommended for MI estimation, as these methods are able to provide good selection accuracy at a comparatively greater computational efficiency.

Scenario 2: If the input/output data are mainly moderately non-Gaussian (average $1.3<s<5$ and $3<k<30$), the GRR is suggested for residual estimation and the DPI (or BCVDPI) is recommended for MI estimation, so that selection accuracy can be improved with only a small reduction in computational efficiency, in comparison with using the GRR and BCV.

Scenario 3: If the input/output data are mainly extremely non-Gaussian (average $s>5$ and $k>30$), the SVC is suggested for residual estimation and the DPI (or BCVDPI) is recommended for MI estimation. While these methods will decrease computational efficiency significantly, they are also likely to result in a marked increase in selection accuracy.

I. Data Distribution:	Gaussian or nearly Gaussian distribution ($s<1.5$ & $k<3$)	Slight to moderate non-Gaussian distribution ($1.5<s<5$ & $3<k<30$)	Extreme non-Gaussian distribution ($s>5$ & $k>30$)
II. *Suggested bandwidth for residual & corresponding efficiency:	GRR based GRNN (VE)	GRR based GRNN (VE)	SVCS based GRNN (NE)
III. **Suggested bandwidth for KDE & MI & corresponding efficiency:	GRR or BCV (VE)	DPI or BCVDPI (E)	DPI or BCVDPI (E)
IV. Scenario No.:	1	2	3

Figure 3.11 Suggested bandwidth estimators under different distribution scenarios

(VE = comparatively very computationally efficient, E = comparatively moderately computationally efficient, and NE = comparatively not computationally efficient; *recommendation based on Li et al. (2014b); ** recommendation based on present study)

3.5 Testing of proposed rules and guidelines

The rules and guidelines proposed in Section 4.3 were 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., 2005a; Li et al., 2014b; 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., 2014b; Srinivasulu and Jain, 2006; Wu et al., 2013). The 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 outputs, 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). Details of each semi-real case study are given in the subsequent sections.

River salinity at Murray Bridge

The study area of the first semi-real case is illustrated in Fig. 3.12. According to Maier and Dandy (1996), 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 at locations specified in Table 3.5). Consequently, these six inputs were used to generate the corresponding outputs ($MBS + 13$). Other redundant or irrelevant candidate inputs listed in Table 3.5 were also introduced for the purpose of testing the effectiveness of PMI IVS.

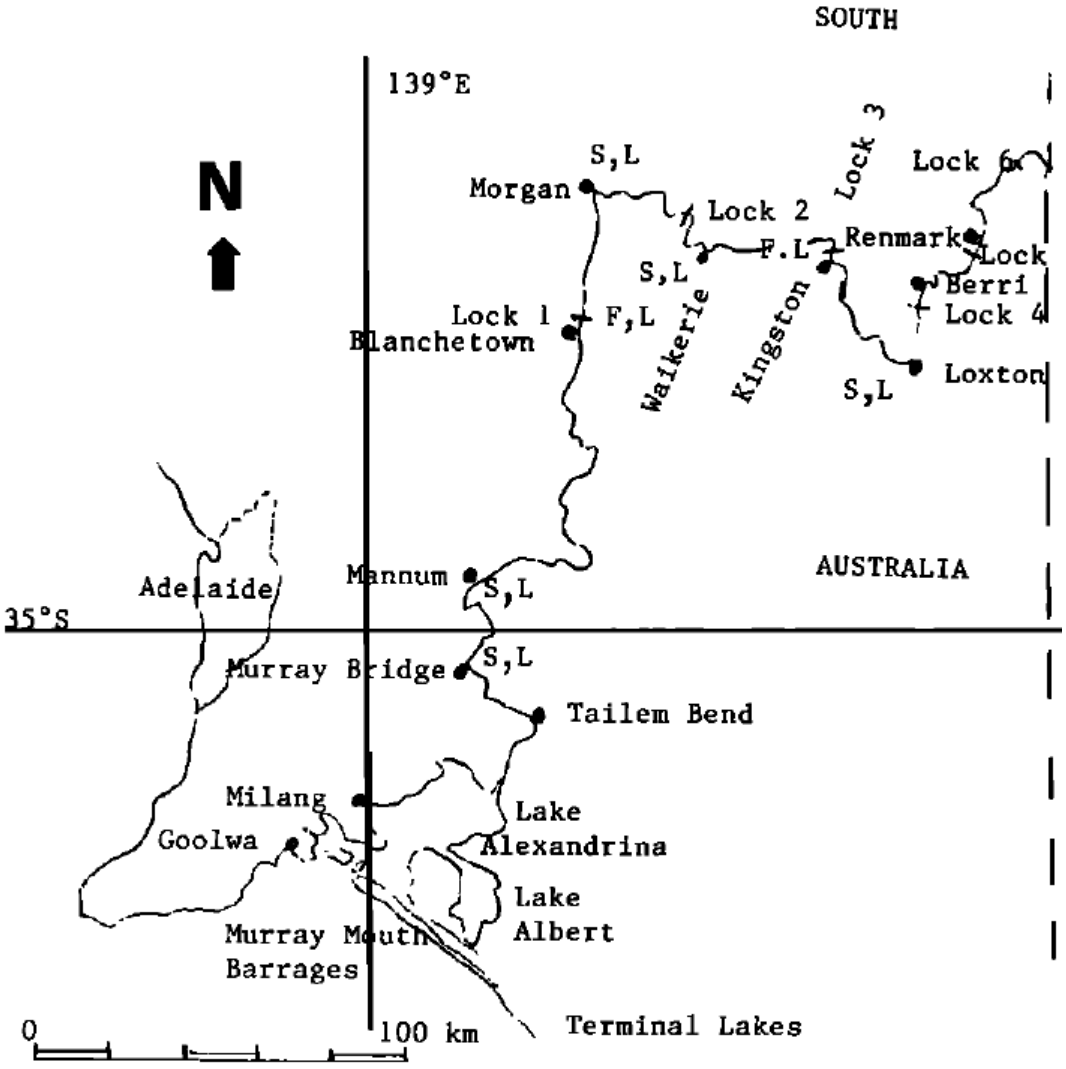


Figure 3.12 The River Murray in South Australia (Maier and Dandy, 1996)

Table 3.5 Candidate inputs and output for the salinity case study

Candidate Inputs				Output			
Location	Variable	Abbreviation	Lags	Location	Variable	Abbreviation	Forecasting Period
Mannum	Salinity	MAS	1,3,5,7,9	Murray Bridge	Salinity	MBS	14
Morgan	Salinity	MOS	1,3,5,7,9				
Waikerie	Salinity	WAS	1,2,3,4,5				
Loxton	Salinity	LOS	1,2,3,4,5				
Murray Bridge	Salinity	MBS	1,3,5,7,9				
Lock 1 Upper	River level	L1UL	-3,-1,1,3,5				

In order to generate the known outputs from the real inputs, standard multilayer perceptron (MLP) artificial neural networks (ANNs) were developed using the approach outlined in Wu et al. (2014). The historical records from 1987 to 1990 were split into training (60%), testing (20%) and validating sets (20%) using the DUPLEX method (May et al., 2010), in accordance with the guidelines suggested by Wu et al. (2013). A single hidden layer was used and the optimal number of hidden nodes was determined by trial and error, considering a range of 0 to 6. The optimal model structure was found to be 6-4-1. The back-propagation algorithm (with learning rate of 0.1 and momentum of 0.1) was used for model calibration. The test inputs were then re-simulated 30 times based on the real observations in order to obtain data sets that contained a certain degree of variation, while still maintaining the major time patterns and data distributions. This enabled IVS performance to be evaluated over 30 independent trials. The corresponding output was obtained by substituting the simulated inputs into the trained ANN model. The input/output data contain strongly linear components and follow a mildly non-Gaussian distribution, according to Bowden (2003), Wu et al. (2013) and Li et al. (2014b). Consequently, this study corresponds to Scenario 2 in Fig. 3.11. Given this, the selection performance of the PMI using the DPI (and BCVDPI) for KDE and the GRR for residual estimation was expected to be superior in terms of an appropriate trade-off between selection accuracy and computational efficiency.

Based on the results in Fig. 3.13, this was observed to be the case. The CSR resulting from the use of the proposed approach was 96.7%, compared with 83.3% when the GRR and BCV approaches were used for KDE. Although use of the SCV and SVO methods also resulted in a CSR of 96.7%, the associated computational cost was significantly greater. Consequently, the DPI/BCVDPI based method provided a good trade-off between selection accuracy and computational efficiency for this study, as suggested by the proposed guidelines (Fig. 3.11).

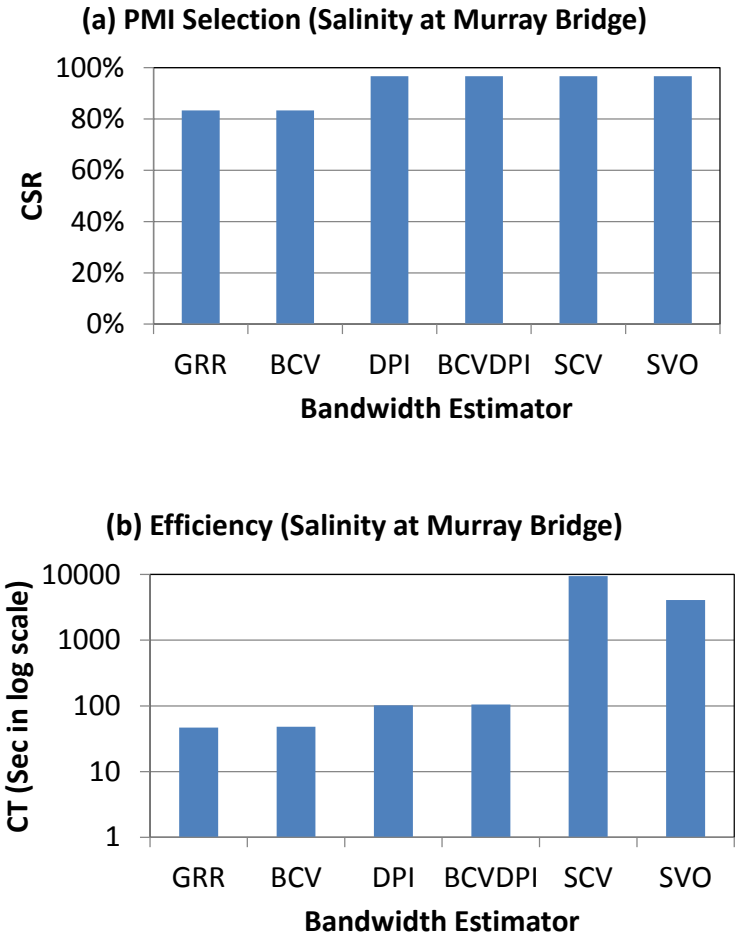


Figure 3.13 Correct selection rate and efficiency of salinity forecast at Murray Bridge with proposed and alternative bandwidth estimators

Rainfall-runoff in Kentucky River Basin

The second semi-real data set is concerned with rainfall-runoff modelling in the Kentucky River Basin in the USA (Fig. 3.14). The output variable for this case study is the forecast flow at Lock and Dam 10 one day in advance (Jain and Srinivasulu, 2004). The corresponding inputs, including average daily effective rainfall and runoff with specific lag time (i.e. average daily effective rainfall: P(t), P(t-1) and average daily runoff: Q(t-1), Q(t-2) at locations specified in Table 3.6), together with other redundant or irrelevant candidate inputs, are summarized in Table 3.6, which are the same as those used by Bowden (2003), Wu et al. (2013) and Li et al. (2014b).

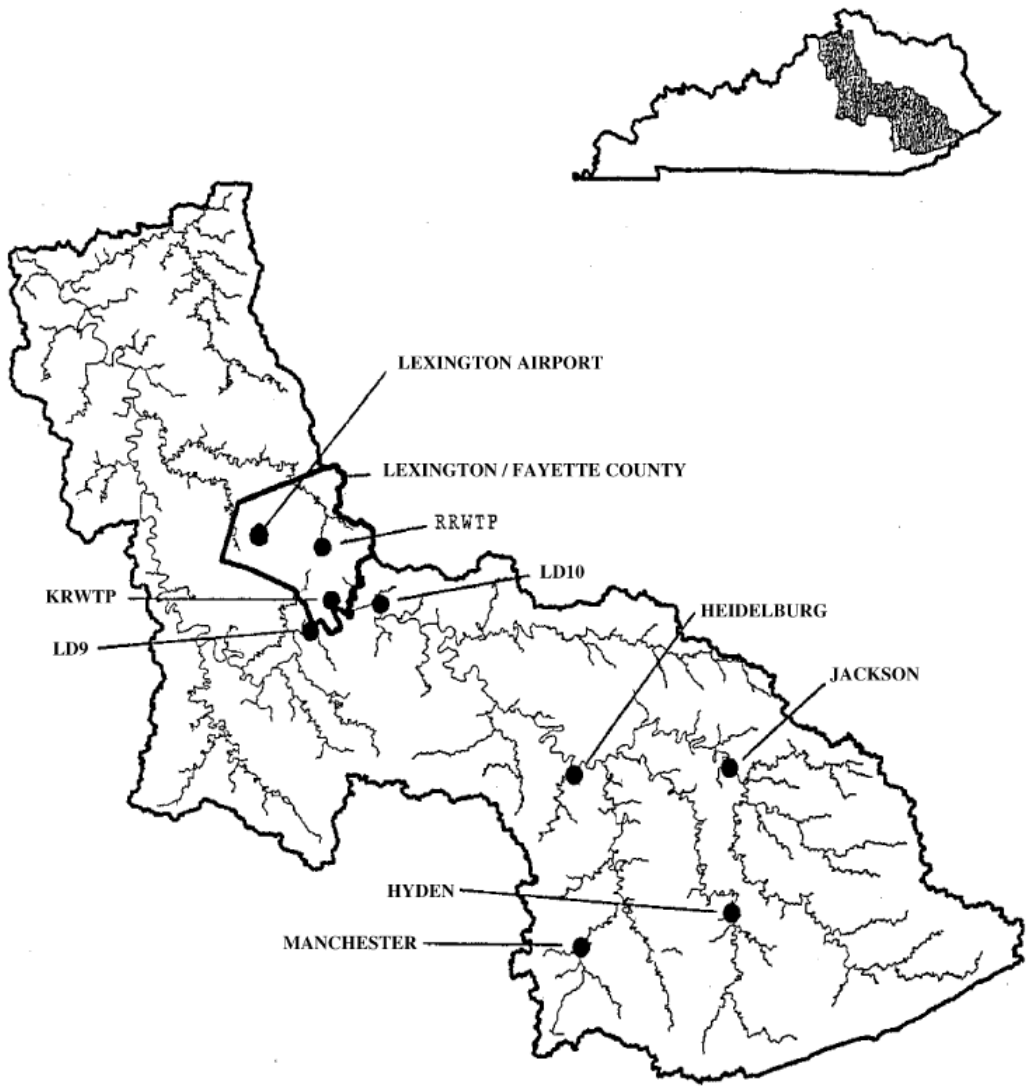


Figure 3.14 The Kentucky River Basin in USA (Jain et al., 2004)

Table 3.6 Candidate inputs and output used for the rainfall-runoff case study

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

The historical rainfall-runoff records from 1960 to 1972 were used for developing the MLP-ANNs using the approach described for the salinity case study. The optimal model structure was determined as 4-4-1. Thirty sets of inputs and outputs were generated using the procedure described for the salinity case study. It should be noted that the input/output data contain non-linear components and follow extremely non-Gaussian distributions, as discussed by Wu et al. (2013), Li et al. (2014b) , and Galelli et al. (2014). Consequently, this study corresponds to Scenario 3 in Fig. 3.11. Given this, the selection performance of the PMI using the DPI (and BCVDPI) for KDE was expected to be superior in terms of an appropriate trade-off between selection accuracy and computational efficiency.

As indicated in Fig. 3.15(a), use of the approach suggested in the proposed guidelines derived from the synthetic data (i.e. DPI with SVC) clearly results in the best CSR, with an accuracy of 96.7%. This is much higher than the CSR of 77.8% when the 'standard' approach (i.e. GRR with GRR) is used. While this increased selection accuracy comes at a significant increase in computational cost (i.e. 68 times more computationally expensive), as shown in Fig. 3.15(b), this still seems to provide the best trade-off between selection accuracy and computational efficiency, as suggested by the proposed guidelines (Fig. 3.11).

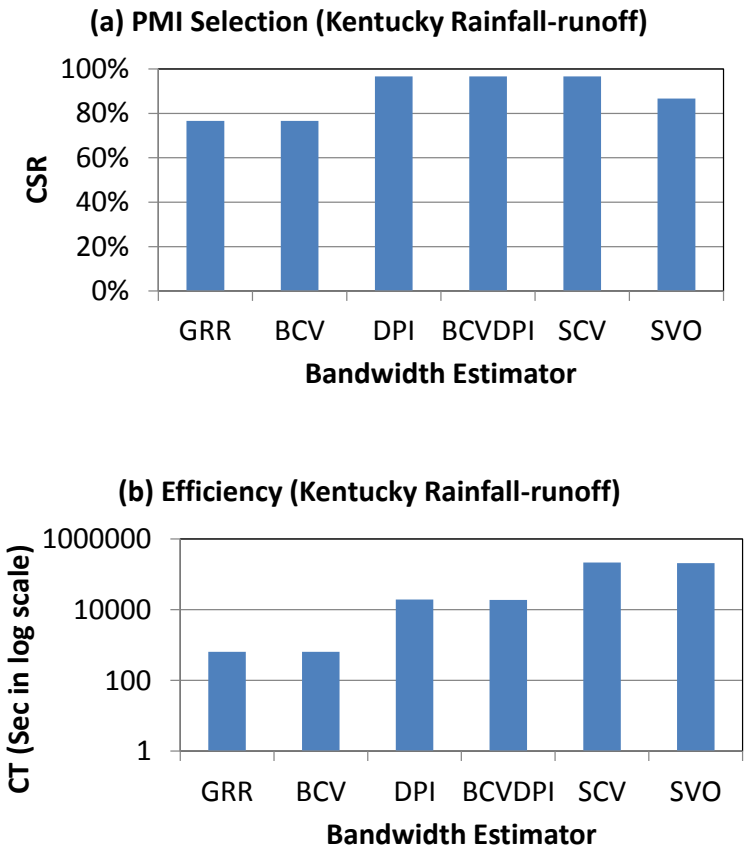


Figure 3.15 Correct selection rate and efficiency of flow forecast at Kentucky River Basin with proposed and alternative bandwidth estimators

3.6 Summary and conclusions

Input variable selection (IVS), as one of the most important steps in the development of ANN and other data driven environmental and water resources models, determines the quality and quantity of information used in the modelling process. Partial mutual information (PMI) is one of the most promising approaches to IVS, as it is able to account for the relevance and redundancy of all candidate inputs and can be used for both linear and non-linear problems. However, one disadvantage of using PMI is that it requires kernel density estimates (KDEs) of the data to be obtained, which can become problematic when the data are non-normally distributed, as is often the case for environmental and water resources problems. However, this is an issue that has been ignored in previous studies on the application of PMI IVS, in which the Gaussian reference rule (GRR) has generally been used to obtain

the required KDEs. This is likely to result in a reduced CSR for data that are non-Gaussian, as shown by Galelli et al. (2014) and Humphrey et al. (2014).

In order to develop an improved approach to PMI IVS for data that are non-normally distributed, the selection performances of PMI with six different kernel bandwidth estimators for KDE were assessed in terms of selection accuracy and computational efficiency for input/output data with distinct degrees of normality on three synthetic data sets. The results from the 3,780 trials with the synthetic data were used to develop empirical guidelines for the choice of the most appropriate bandwidth estimation techniques for data with different degrees of non-normality. The validity of these guidelines was then tested on the two semi-real data sets.

The results of the synthetic case studies suggest that the use of GRR-based bandwidth estimators only results in good input selection accuracy if the input/output data follow Gaussian or nearly Gaussian distributions, which is in line with the results obtained by Galelli et al. (2014) and Humphrey et al. (2014). As a result of their reduced dependence on the Gaussian assumption, DPI, BCVDPI, SCV, and SVO based bandwidth estimators generally result in marked improvements in CSR for problems with data that follow non-Gaussian distributions. However, there is a distinct trade-off between selection accuracy and computational efficiency.

One of the major outcomes of this paper is the development of the empirical guidelines based on the synthetic tests. As shown in Fig. 3.11, the suggested bandwidth estimators for KDE used in the MI calculation should be used in conjunction with the bandwidth estimators for residual estimation suggested by Li et al. (2014b). The results for the two semi-real data sets, which follow mildly and extremely non-Gaussian distributions, support the validity of the proposed guidelines for the selection of appropriate bandwidth estimation methods for data with different degrees of non-normality. It should be noted that the proposed guidelines are valid for environmental and water resource applications with data that have distributional properties similar to those provided in the guidelines, and that the implementation of the guidelines is

also likely to benefit other data-driven environmental and water resources models, even though they were only tested on MLPs.

Although the results of this study indicate that the use of alternate bandwidth estimators can result in significant improvements in PMI input selection accuracy for data that are non-normally distributed, these improvements were not as pronounced for extremely non-Gaussian data and the non-linear synthetic case study. This is likely due to boundary issues associated with KDE for highly non-Gaussian data (Karunamuni and Alberts, 2005b; Scott, 1992). Consequently, future research should focus on potential improvements to input variable selection accuracy as a result of the consideration of such boundary issues. In addition, alternative methods for dealing with non-Gaussian data in the context of PMI IVS, such as transforming the input data to normality (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. Finally, there is a need to assess the performance of the proposed modifications to the implementation of the PMI algorithm on a broader set of data and against that of other IVS algorithms (see Galelli et al., 2014).

3.7 Acknowledgments

This research was aided by the suggestions from Prof. A. Sharma and the original code from Dr. R.J. May (GRR based PMI) and Dr. G.B. Humphrey (GRR based GRNN). The authors would also like to thank the three anonymous reviewers, whose input has improved the quality of this paper significantly.

CHAPTER 4 JOURNAL PAPER 3 -

Improved Partial Mutual Information-Based Input Variable Selection by Consideration of Boundary Issues Associated With Bandwidth Estimation

Statement of Authorship

Title of Paper	Improving partial mutual information-based input variable selection by consideration of boundary issues associated with bandwidth estimation
Publication Status	○ Published, ○ Accepted for Publication, ◉ Submitted for Publication, ○ Publication style
Publication Details	Li, X., Zecchin, A.C., Maier, H.R., 2014a. Improving partial mutual information-based input variable selection by consideration of boundary issues associated with bandwidth estimation. Environmental Modelling and Software, submitted on 04/12/2014.

Author Contributions

By signing the Statement of Authorship, each author certifies that their stated contribution to the publication is accurate and that permission is granted for the publication to be included in the candidate's thesis.

Name of Principal Author (Candidate)	Xuyuan Li		
Contribution to the Paper	Undertook literature review, developed analytic procedure and numerical models, developed software, and prepared manuscript		
Signature		Date	

Name of Co-Author	Dr. Aaron C. Zecchin		
Contribution to the Paper	Supervised manuscript preparation and reviewed draft		
Signature		Date	

Name of Co-Author	Professor Holger R. Maier		
Contribution to the Paper	Supervised manuscript preparation and reviewed draft		
Signature		Date	

Name of Co-Author			
Contribution to the Paper			
Signature		Date	

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, which plays an important role during two steps of the PMI algorithm. In total, the effectiveness of 16 different approaches is tested on synthetically generated data, and the results used to develop preliminary guidelines for the selection of the most appropriate PMI variants based on the degree of non-linearity and normality of the data. These guidelines are validated on two semi-real case studies, showing that by using the proposed guidelines, the correct inputs can be identified in 100% of trials, even if the data are highly non-linear or extremely non-Gaussian.

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; all versions of Visual Studio 2012, 2010 and 2008 are supported except Visual Studio Express; 256 MB RAM

Software requirements: PGI Visual Fortran 2003 or later version

Language: English

Size: 4.55MB

Availability: Free to download for research purposes from the following website:

https://github.com/xuyuanli/IVS_PMI_2014

4.1 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 (e.g. Galelli et al., 2014; Maier et al., 2010; Wu et al., 2014a; Wu et al., 2014b). 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 challenge of the IVS problem, a large number of approaches, categorised as either model free (on the basis of a statistical measures of significance between the candidate inputs and the output) or model based (depending on the adoption of an optimization algorithm that is used to determine 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, 2000a), aiming to determine the smallest number of inputs that best charaterise the input-output relationship with the least amount of variable irrelvance or redundancy (Galelli et al., 2014; Guyon and Elisseeff, 2003). Among the various 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; Galelli et al., 2014; May et al., 2008b; Wu et al., 2014b; Wu et al., 2013).

The PMI IVS approach was introduced by Sharma (2000a) and is based on Shannon's principle (Shannon, 1948), otherwise termed Shannon's entropy, which measures the MI between a random input variable and a random output

variable. As part of the PMI algorithm, inputs are chosen using a forward selection approach, during which one input variable is selected at each iteration, 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), which quantifies the reduction in uncertainty with respect to the output due to observation of an input variable, 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 partial mutual information (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), which is 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 estimates have been obtained using kernel density based methods in order to approximate marginal and joint PDFs and residual estimates have been obtained using kernel based regression methods for the estimation of kernel based weights (e.g. Bowden et al., 2005a; Bowden et al., 2005b; Gibbs et al., 2006; He et al., 2011; Li et al., 2015; May et al., 2008a; May et al., 2008b). 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 (otherwise termed 'smoothing parameter') selection and how well any boundary issues are addressed, as pointed by Santhosh and Srinivas (2013), Scott (1992), and Wand and Jones (1995), as discussed below.

The bandwidth selection issue is caused by the fact that 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)), there is no clear consensus as to which bandwidth estimator performs best for general cases and in almost all existing

PMI IVS studies in environmental modelling the Gaussian reference rule (GRR) has been used for bandwidth estimation due to its simplicity (e.g. Bowden et al., 2005a; Bowden et al., 2005b; He et al., 2011; May et al., 2008a; May et al., 2008b). 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.

Another potential problem with kernel based methods in environmental and water resources modelling 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 due to the fact that data can be bounded in accordance with their physical feasibility (e.g. rainfall-runoff data are bounded at 0mm).

While the impact of different bandwidth estimators for MI and RE on the performance of PMI IVS has been assessed recently, and empirical guidelines proposed for the selection of the optimal bandwidth for MI and residual estimation for data following different distributions (Li et al., 2015), the impact of boundary issues associated with MI and residual estimation on the performance of PMI IVS has not yet been considered, 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). 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 are given in Section 6.

4.2 Background on PMI IVS and Boundary Issues

4.2.1 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, 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, $\boldsymbol{X}^j = \left[X_1^j \ldots X_m^j\right]^T$ are the observed input data and y^j are the observed output data

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 KDE and the marginal PDF (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;

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 with respect to y due to 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) $\hat{m}_y(X_{i^*})$ between the selected input(s) X_{i^*} and the output y and (ii) obtaining the residuals ($y - \hat{m}_y(X_{i^*})$) 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, 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 $v_i = X_i - \hat{m}_{X_i}(X_{i^*})$ and output residual $u = y - \hat{m}_y(X_{i^*})$ 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 residual estimation (Step 5). 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 residual estimation by using modelling approaches that are not reliant on kernel regression methods. Background information of the boundary issue and of its relevance to PMI IVS are given in the following subsection.

4.2.2 Boundary issues in PMI IVS

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

$$\hat{f}(X_i; \boldsymbol{H}) = \frac{1}{n}\sum_{j=1}^{n} K_H\left(X_i - X_i^j\right) \tag{4.1}$$

where X_i represents the ith input vector and K_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

$$K_H(\boldsymbol{X}) = \frac{1}{(\sqrt{2\pi}|\boldsymbol{H}|)^m} \exp\left[-\frac{1}{2}\boldsymbol{X}^T\boldsymbol{H}^{-1}\boldsymbol{X}\right] \quad (4.2)$$

In Eq. (4.2), $\boldsymbol{H}$ is the kernel bandwidth matrix (or kernel bandwidth for univariate problems). 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 of $\hat{f}$ is bounded, and in the absence of exponentially falling tails (e.g. support $[0, a]$), strong under-estimation occurs for all data points in the boundary region (which are within a distance of the bandwidth h from the boundary) because of the nonzero kernel density estimation 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 (2005a), and Wand and Jones (1995). These are the so-called 'boundary issues' associated with non-parametric kernel-based estimations.

A graphical representation of boundary issue in 2D is also provided in Fig. 4.1 in accordance with Hazelton and Marshall (2009). In Fig. 4.1, the kernel density estimates are an approximation of data on the location of childhood leukaemia and lymphoma in North Humberside, England. It can be seen that the left-hand estimate without boundary correction has a smoothed edge, while the right hand estimate with boundary correction has a sharper and significantly higher edge at the same point. This indicates strong under-estimates for all data points in the boundary region, as mentioned above.

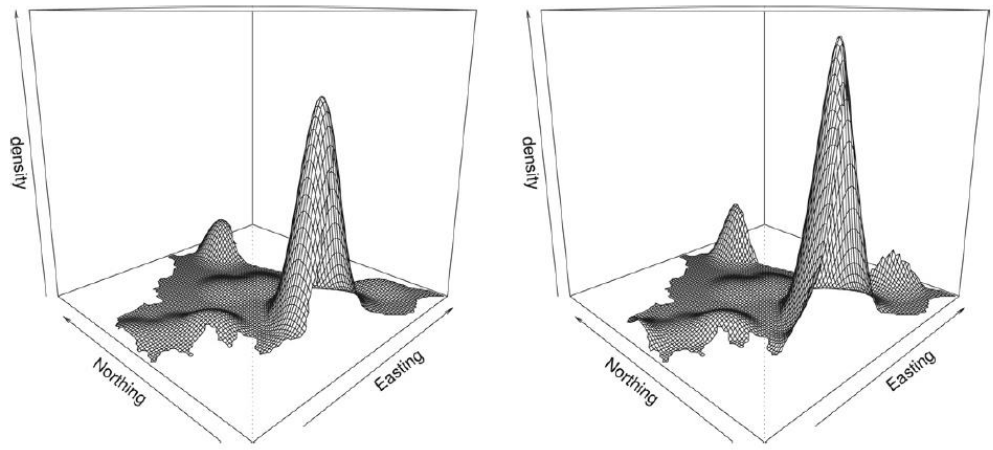


Figure 4.1 Graphical representation of the boundary issue in 2D (Hazelton and Marshall, 2009)

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 kernel regression estimation, in step 6) when the observations are bounded and/or follow non-Gaussian distributions (e.g. with high skewness and kurtosis).

4.2.3 Potential solutions to solve 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. 4.2. 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. 4.2:

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 as 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 $\hat{f}$;
- 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 (Hall and Park, 2002; Jakeman et al., 2006), which removes boundary issues by introducing an additional empirical data perturbation term $\hat{\alpha}$, constructed specifically to adjust the bias of density estimate 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. 4.2, 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 to 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, 1997b; 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, 2000b; Maier et al., 2010; Wolfs and Willems, 2014; Wu et al., 2014a; Wu et al., 2014b) problems. In addition, they are independent from boundary issues due to their kernel free features (Maier et al., 2010; Wu et al., 2014b), although a major drawback of MLPANNs is their generally high computational requirements. In this paper, only selected and appropriate approaches from the aforementioned methods in Fig. 4.2 are implemented to fulfil the required objectives. Details of the analytical processes associated with the different approaches are described in the subsequent section.

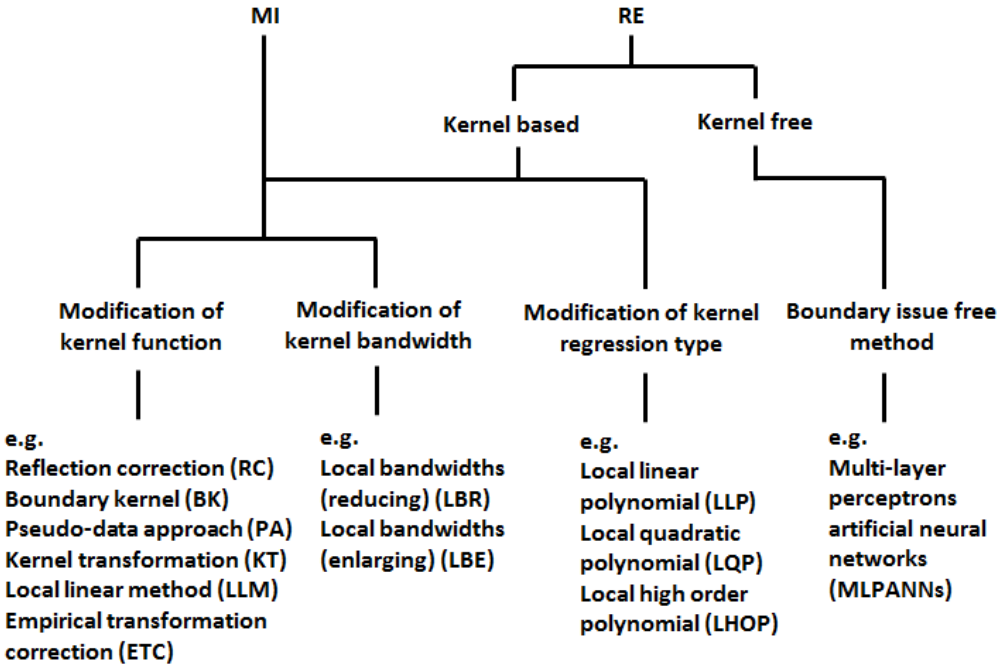


Figure 4.2 Taxonomy of methods for dealing with boundary issues in mutual information and residual estimation

4.3 Methodology

The approach adopted for the systematic assessment of methods for addressing boundary issues on the performance of PMI IVS is outlined in Fig. 4.3. 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 used to indicate different severities of boundary issues); (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 residual estimation. Details of each of these steps are given in the subsequent sections.

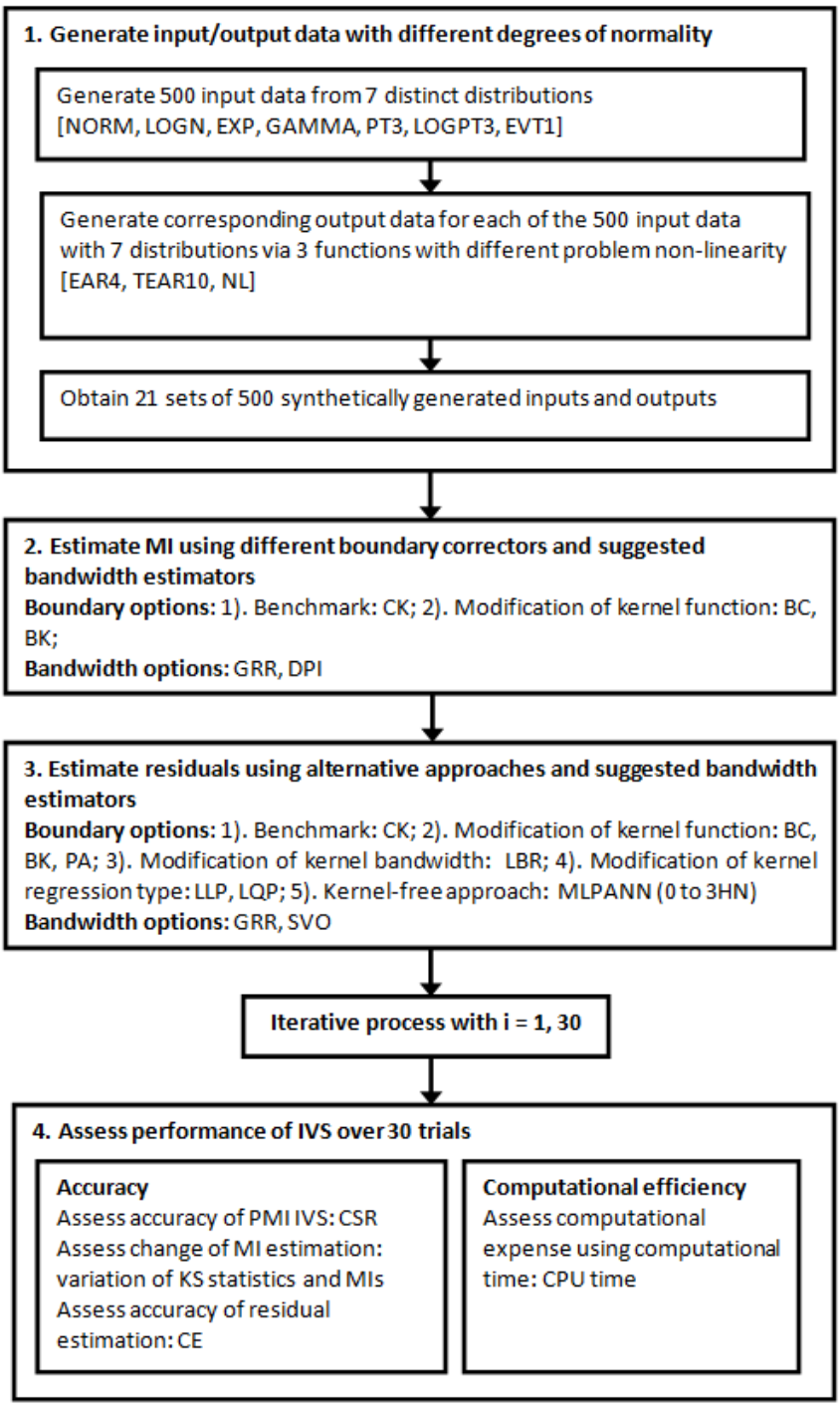


Figure 4.3 Overview of the proposed analysis for the PMI IVS influenced by bandwidth and boundary issues

4.3.1 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 4.1 and 4.2. In total, 525 data 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 4.1 Details 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
NORM	Mean=3.0; sd =1.0	0.000	-0.013	High
GAMMA	Shape=2.0; Scale=1.0	1.370	2.638	High
LOGN	Mean=0.5; sd=1.0	5.326	53.694	Low
EXP	Rate=1.0	2.132	7.219	Moderate
PT3	Shape=2.5; Scale=3.0; Location=2.0	1.251	2.381	High
LOGPT3	Shape=0.5; Scale=0.2; Location=2.0	4.792	43.265	Low
EVT1	Shape=0.0; Scale=0.5; Location=10.0	1.198	2.880	High

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

Table 4.2 Details 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
NORM	Mean=3.0; sd =1.0	1.826	5.158	High
GAMMA	Shape=2.0; Scale=1.0	10.520	192.091	Low
LOGN	Mean=0.5; sd=0.4	5.389	47.767	Low
EXP	Rate=1.0	14.029	334.408	Low
PT3	Shape=0.5; Scale=1.0; Location=0.5	16.271	514.270	Low
LOGPT3	Shape=0.5; Scale=0.2; Location=0.5	14.261	390.522	Low
EVT1	Shape=0.1; Scale=0.0; Location=10.0	1.788	9.807	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., 2014b; May et al., 2008b).

The equation of the EAR4 model is given by

$$x_t = 0.6x_{t-1} - 0.4x_{t-4} + p_{t-1} + 0.1\varepsilon_t \tag{4.3}$$

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 $0.1\varepsilon_t$ is the introduced error term (explained shortly).

The equation for the TEAR10 model is given by

$$x_t = \begin{cases} -0.5x_{t-6} + 0.5x_{t-10} - 0.3p_{t-1} + 0.1\varepsilon_t; & x_{t-6} \le 0 \\ 0.8x_{t-10} - 0.3p_{t-1} + 0.1\varepsilon_t; & otherwise \end{cases} \tag{4.4}$$

The equation for NL is given by

$$y = (x_2)^3 + x_6 + 5\sin(x_9) + 0.1\varepsilon_t \tag{4.5}$$

The first two time series models are modified from May et al. (2008b) by introducing an additional independent lagged input p_{t-1} into 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. All three synthetic models have also been studied in Li et al. (2014b, 2015). The error term ε_t follows a standard normal distribution $N(0,1)$, which introduces noise without obscuring the influence of the actual independent variables. In the present study, all data are scaled between 0 and 1.

4.3.2 Estimate MI using different boundary correctors and suggested bandwidth estimators

Although a number of potential methods aiming to ameliorate boundary issues by means of modification of the kernel function have been introduced in Section 2.2, 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, 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. (4.1) and (4.2). As mentioned in Section 2, this method does not provide any boundary correction, and is therefore used as a benchmark approach.

Refection 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

$$\hat{f}(X_i; h_x) = \begin{cases} \frac{1}{n}\sum_{j=1}^{n}\left[K_{h_x}\left(X_i - X_i^j\right) + K_{h_x}(X_i + X_i^j)\right]; \; h_x \geq X_i \geq 0 \\ \frac{1}{n}\sum_{j=1}^{n}\left[K_{h_x}\left(X_i - X_i^j\right)\right]; \; X_i > h_x \\ 0; X_i < 0 \end{cases} \tag{4.6}$$

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

$$\hat{f}(X_i, y; \boldsymbol{H}) = \begin{cases} \frac{1}{n}\sum_{j=1}^{n}\left[K_H\left(\begin{bmatrix} X_i \\ y \end{bmatrix} - \begin{bmatrix} X_i^j \\ y^j \end{bmatrix}\right) + K_H\left(\begin{bmatrix} X_i \\ y \end{bmatrix} - \begin{bmatrix} -X_i^j \\ -y^j \end{bmatrix}\right)\right]; h_x \geq X_i \geq 0, \; h_y \geq y \geq 0 \\ \frac{1}{n}\sum_{j=1}^{n}\left[K_H\left(\begin{bmatrix} X_i \\ y \end{bmatrix} - \begin{bmatrix} X_i^j \\ y^j \end{bmatrix}\right) + K_H\left(\begin{bmatrix} X_i \\ y \end{bmatrix} - \begin{bmatrix} -X_i^j \\ y^j \end{bmatrix}\right)\right]; h_x \geq X_i \geq 0, \; y > h_y \\ \frac{1}{n}\sum_{j=1}^{n}\left[K_H\left(\begin{bmatrix} X_i \\ y \end{bmatrix} - \begin{bmatrix} X_i^j \\ y^j \end{bmatrix}\right) + K_H\left(\begin{bmatrix} X_i \\ y \end{bmatrix} - \begin{bmatrix} X_i^j \\ -y^j \end{bmatrix}\right)\right]; X_i > h_x, \; h_y \geq y \geq 0 \\ \frac{1}{n}\sum_{j=1}^{n}\left[K_H\left(\begin{bmatrix} X_i \\ y \end{bmatrix} - \begin{bmatrix} X_i^j \\ y^j \end{bmatrix}\right)\right]; X_i > h_x, \; y > h_y \\ 0; X_i < 0 \, , y < 0 \end{cases} \tag{4.7}$$

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

$$\boldsymbol{H} = \begin{bmatrix} h_x^2 & \rho_{xy} h_x h_y \\ \rho_{xy} h_x h_y & h_y^2 \end{bmatrix} \tag{4.8}$$

(known as a hybrid class of bandwidth matrix), where h_y is the bandwidth for output y and ρ_{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-C C.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 to 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

$$B(u; h_x) = \frac{\left[\left(a_3^{(1)} + 4a_2\right) - \left(a_2^{(1)} + 3a_1\right)u\right] K_{h_x}(u)}{\left(a_3^{(1)} + 4a_2\right)a_0 - \left(a_2^{(1)} + 3a_1\right)a_1} \tag{4.9}$$

where $a_\alpha^{(\gamma)} = \int u^\alpha D^\gamma K_h(u)\, du$; $D^\gamma K_h(u) = \left(\partial^{\int u K_h(u) du} / \partial\, u^{\int u K_h(u) du}\right)$.

$K_h(u)$; and $u = (X_i - X_i^j)/h_x$. This 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 the density 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 $B(u; h_x)$ has an adaptive form for the scaled data within $(u_{min}, 1)$ and a fixed form for the scaled data within $(1, u_{max})$, thereby being able to add the underestimated density back within the boundary region while keeping the density unchanged in the free region.

By extending this concept into two dimensions, the expression of the bivariate BK is given as

$$B(u,v;\boldsymbol{H}) = \frac{b_0 K_H(u,v) + b_1 u K_H(u,v) + b_2 v K_H(u,v)}{b_0 a_{00} + b_1 a_{10} + b_2 a_{01}} \tag{4.10}$$

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 $v = (y - y^j)/h_y$. This results in a linear combination of three kernels, which is able to eliminate the $O(h)$ extra bias term that is present in the bivariate case when compared with the univariate one (Eq. 4.9). Similar to the univariate BK $B(u; h_x)$, the bivariate BK $B(u, v; \boldsymbol{H})$ is again 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 $(1, v_{max})$). The detailed mathematic derivations and explanations of Eqs. (4.9) and (4.10) can be found in Marshall and Hazelton (2010).

4.3.3 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. 4.3 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. 4.3, but are also theoretically applicable to univariate approaches to residual estimation. 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. (2014a), as outlined in Table 4.3.

Table 4.3 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., 2014a)

$$\hat{y}_{GRNN}(X_i, h) = \frac{\sum_{j=1}^{n} y^j exp\left[-\frac{\left(X_i - X_i^j\right)^2}{2h_x^2}\right]}{\sum_{j=1}^{n} exp\left[-\frac{\left(X_i - X_i^j\right)^2}{2h_x^2}\right]} \tag{4.11}$$

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. (4.11) with the RC in Eq. (4.6). The expression for the estimator then becomes

$$\hat{y}_{RC}(X_i,h)=\begin{cases}\dfrac{\sum_{j=1}^{n}y^j\left[exp\left(-\frac{\left(X_i-X_i^j\right)^2}{2h_x^2}\right)+exp\left(-\frac{\left(X_i+X_i^j\right)^2}{2h_x^2}\right)\right]}{\sum_{j=1}^{n}\left[exp\left(-\frac{\left(X_i-X_i^j\right)^2}{2h_x^2}\right)+exp\left(-\frac{\left(X_i+X_i^j\right)^2}{2h_x^2}\right)\right]}; h_x \geq X_i \geq 0\\ \dfrac{\sum_{j=1}^{n}y^j\left[exp\left(-\frac{\left(X_i-X_i^j\right)^2}{2h_x^2}\right)\right]}{\sum_{j=1}^{n}\left[exp\left(-\frac{\left(X_i-X_i^j\right)^2}{2h_x^2}\right)\right]}; X_i > h_x\\ 0;\ X_i < 0\end{cases} \tag{4.12}$$

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. (4.9)) is plugged into Eq. (4.11), resulting in the following expression

$$\hat{y}_{BK}(X_i,h)=\frac{\sum_{j=1}^{n}y^j\left\{\frac{\left[\left(a_3^{(1)}+4a_2\right)-\left(a_2^{(1)}+3a_1\right)u\right]K_h(u)}{\left(a_3^{(1)}+4a_2\right)a_0-\left(a_2^{(1)}+3a_1\right)a_1}\right\}}{\sum_{j=1}^{n}\left\{\frac{\left[\left(a_3^{(1)}+4a_2\right)-\left(a_2^{(1)}+3a_1\right)u\right]K_h(u)}{\left(a_3^{(1)}+4a_2\right)a_0-\left(a_2^{(1)}+3a_1\right)a_1}\right\}} \tag{4.13}$$

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

$$X^{(-j)}=-5X^{\left(\frac{j}{3}\right)}-4X^{\left(\frac{2j}{3}\right)}+\frac{10}{3}X^{(j)}, j=1,\cdots,n \tag{4.14}$$

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

$$\hat{f}(X_i) = \frac{1}{nh}\left\{\sum_{j=1}^{n} K_h\left[(X_i - X_i^j)/h\right] + \sum_{j=1}^{l} K_h\left[(X_i - X_i^{(-j)})/h\right]\right\} \tag{4.15}$$

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. 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 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. (4.11) and Eq. (4.15), the expression for GRNN(PA) is given by

$$\hat{y}_{PA}(X_i, h) = \frac{\sum_{j=1}^{n} y^j\left\{\sum_{j=1}^{n} K_h\left[(X_i - X_i^j)/h\right] + \sum_{j=1}^{l} K_h\left[(X_i - X_i^{(-j)})/h\right]\right\}}{\sum_{j=1}^{n} K_h\left[(X_i - X_i^j)/h\right] + \sum_{j=1}^{l} K_h\left[(X_i - X_i^{(-j)})/h\right]} \tag{4.16}$$

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 $a \le X_i^j \le c$, where a and c are left and right boundaries, is defined by

$$h_{X_i^j} = \begin{cases} \max\left(X_i^j - a, \varepsilon\right); if\ a \le X_i^j < (h + a) \\ \max\left(c - X_i^j, \varepsilon\right); if\ (c - h) < X_i^j \le c \\ h; otherwise \end{cases} \tag{4.17}$$

and $\varepsilon = 0.001$ is added to avoid zero bandwidth values and the regression model used is identical to Eq. (4.11).

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

$$\hat{y}_{LP}(X_i; p, h) = \boldsymbol{e}_1^T \begin{bmatrix} \hat{s}_0 & \cdots & \hat{s}_p \\ \vdots & \ddots & \vdots \\ \hat{s}_p & \cdots & \hat{s}_{2p} \end{bmatrix}^{-1} \begin{bmatrix} \hat{t}_0 \\ \vdots \\ \hat{t}_p \end{bmatrix} \tag{4.18}$$

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. (4.18), giving

$$\hat{y}_{LLP}(X_i; 1, h) = n^{-1}\sum_{j=1}^{n} \frac{\left\{\hat{s}_2 - \hat{s}_1(X_i^j - X_i)\right\}K_h\left(X_i^j - X_i\right)y^j}{\hat{s}_2\hat{s}_0 - \hat{s}_1\hat{s}_1} \tag{4.19}$$

Local quadratic polynomial regression (LQP) Although the general expression for the LQP and LLP is identical (Eq. (4.18)), 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. (4.18), the univariate equation for the LQP is given as

$$\hat{y}_{LQP}(X_i; 2, h) =$$
$$n^{-1}\sum_{j=1}^{n} \frac{\left[(\hat{s}_2\hat{s}_4 - \hat{s}_3\hat{s}_3) - (\hat{s}_1\hat{s}_4 - \hat{s}_2\hat{s}_3)\left(X_i^j - X_i\right) + (\hat{s}_1\hat{s}_3 - \hat{s}_2\hat{s}_2)\left(X^i - X\right)^2\right]K_h\left(X_i^j - X_i\right)y^i}{\left[\hat{s}_0(\hat{s}_2\hat{s}_4 - \hat{s}_3\hat{s}_3) - \hat{s}_1(\hat{s}_4\hat{s}_1 - \hat{s}_3\hat{s}_2) + \hat{s}_2(\hat{s}_1\hat{s}_3 - \hat{s}_2\hat{s}_2)\right]} \tag{4.20}$$

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 optimal number of hidden nodes for the different models is 2 (EAR4), 2 (TEAR10), and 3 (NL). The back-propagation (BP) algorithm (with learning rate of 0.1 and momentum of 0.1) is used for calibration. This is consistent with the procedure implemented by Li et al. (2015).

4.3.4 Test regime

As outlined in Fig. 4.3, 630 synthetic data sets are simulated, which include 30 replicates for each of the three synthetic models, 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 residual estimation (as shown in Table 4.4), 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 residual estimation (R1 to R7), and four take into account the boundary issue in both MI and residual estimations (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. (2014b, 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 residual estimation 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 residual estimations 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 of methods that are likely to result in similar outcomes. In addition, the influence of the bandwidth selection issue in both MI and residual estimations is minimised by following the guidelines proposed by Li et al. (2014b, 2015), as specified in Sections 3.2 and 3.3, respectively.

Table 4.4 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 successfully applied 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 open 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.

4.3.5 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. 4.3, 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 residual estimation 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 (KS) 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 variation of the KS is expressed as follows

$$KS\ variation\ (\%) = \frac{KS_{A1} - KS_{A2}}{KS_{A1}} \times 100\% \quad (4.21)$$

where the KS 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 4.4). A positive KS variation indicates improvement of accuracy, and vice versa. It should be noted that the performance of the empirical kernel based CDF is a function of the bin width, therefore a number of bin widths (from 0.001 to 1.0) have been tested through a sensitivity analysis. Bin widths of 0.01 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

$$MI\ variation\ (\%) = \frac{MI_{A1} - MI_{A2}}{MI_{A1}} \times 100\% \quad (4.22)$$

and indicates to what extent the improvement or deterioration in kernel density estimation can be propagated to the estimation of MI. When considering Eqs. (4.21) and (4.22), high KS and MI variations indicate effective improvement of boundary issue in MI estimates as a result of boundary correction in the estimation of marginal PDFs. High MI variation but low KS variation corresponds to effective improvement of the boundary issue in MI estimates due to boundary correction in the estimation of joint PDFs, while low MI variation suggests insignificant impact of boundary issue in MI estimates, 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).

4.4 Results and Discussion

Within this section, the selection accuracy of the PMI IVS method with different approaches to addressing the boundary issue (see Table 4.4) 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.

4.4.1 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. 4.4. 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.

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 residual estimation (B3), compared with the most commonly used approach (B1), but generally CSRs do not exceed 90% (Fig. 4.4). 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 4.5(a)). 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. 4.5(a) and 4.5(b)). This can be explained by considering the expression of MI (Shannon, 1948), which is given as

$$I_{X_i,y} \approx \frac{1}{n}\sum_{j=1}^{n} log[\frac{f\left(x_i^j,y^j\right)}{f\left(x_i^j\right)f(y^j)}] \tag{4.23}$$

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

$$I_{X_i,y} \approx \frac{1}{n}\sum_{j=1}^{n} log\left\{\frac{f\left(x_i^j,y^j\right)\Delta x_i^j y^j}{\left[f\left(x_i^j\right)\Delta x_i^j\right]\left[f(y^j)\Delta y^j\right]}\right\} \tag{4.24}$$

where $\Delta x_i^j y^j$, Δx_i^j, and Δy^j indicate variations in the marginal and joint densities due to the boundary correction. This equation is equivalent to

$$I_{X_i,y} \approx \frac{1}{n}\sum_{j=1}^{n} log[\frac{f\left(x_i^j,y^j\right)}{f\left(x_i^j\right)f(y^j)}] + \left\{log\left(\Delta x_i^j y^j\right) - log\left(\Delta x_i^j\right) - log\left(\Delta y^j\right)\right\} \tag{4.25}$$

In Eq. (4.25), 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. 4.4) correspond to higher values of the Coefficients of Efficiency of these models compared with that for method B3 (Fig. 4.6). In contrast, there reverse applies for method R2. Similar results can also be found in APPENDIX-C Fig. C.2.3. The effectiveness of 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.

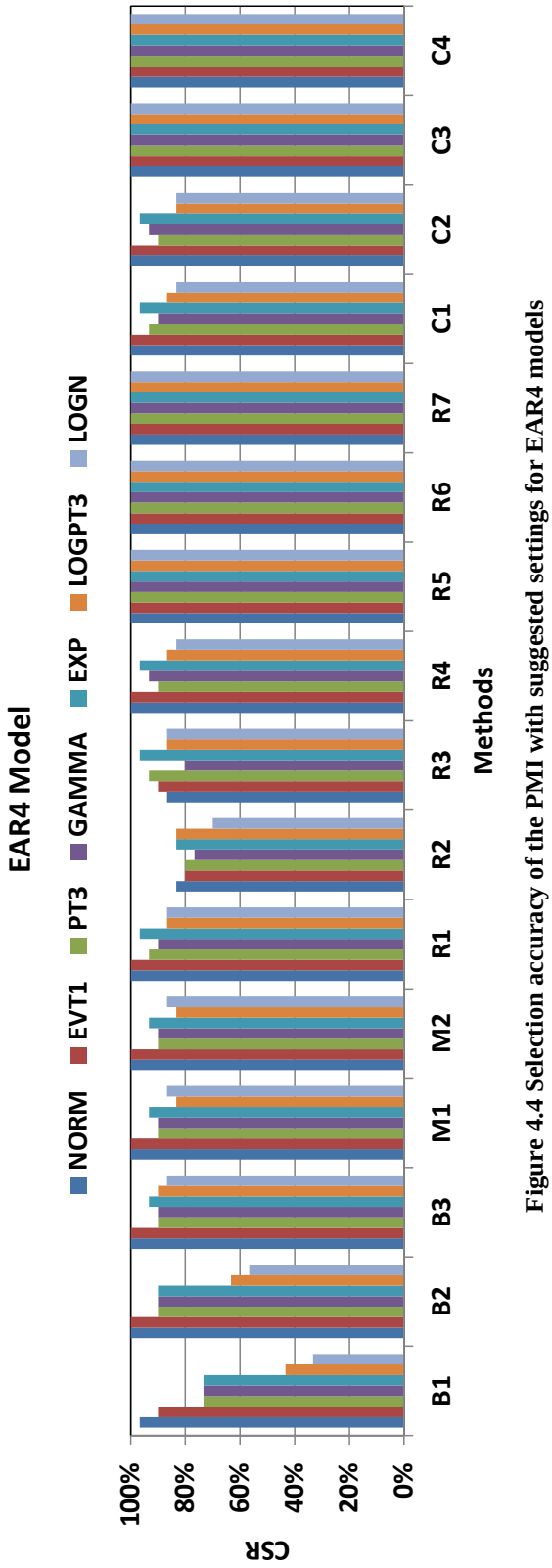


Figure 4.4 Selection accuracy of the PMI with suggested settings for EAR4 models

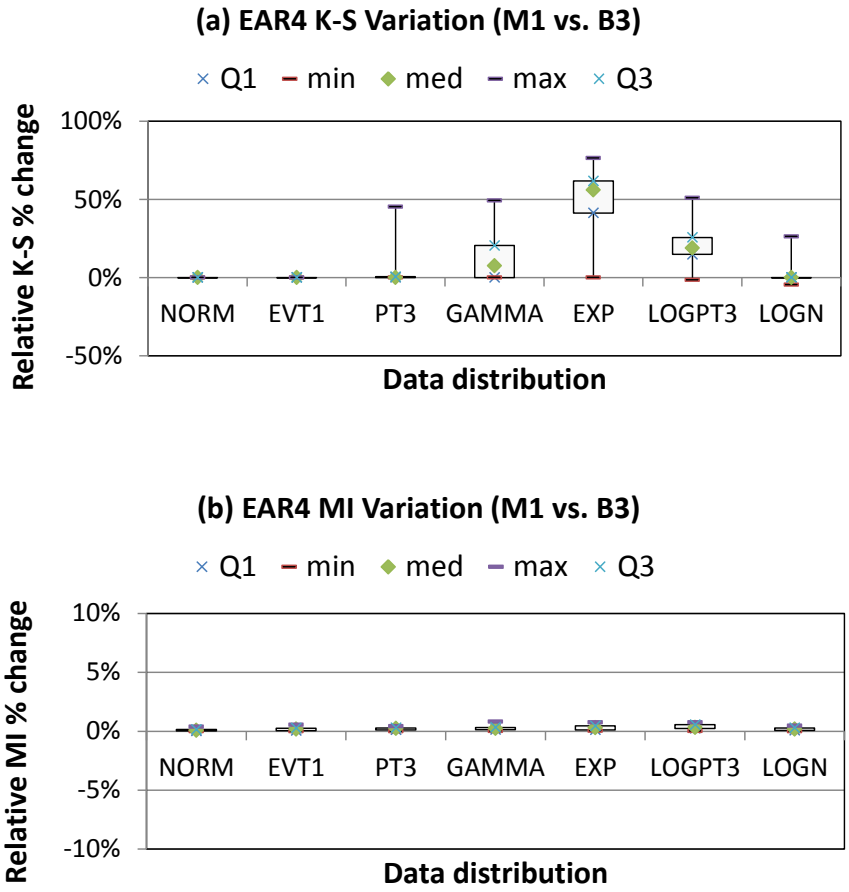


Figure 4.5 Relative change of K-S and MI in-between M1 and B3 for EAR4 model

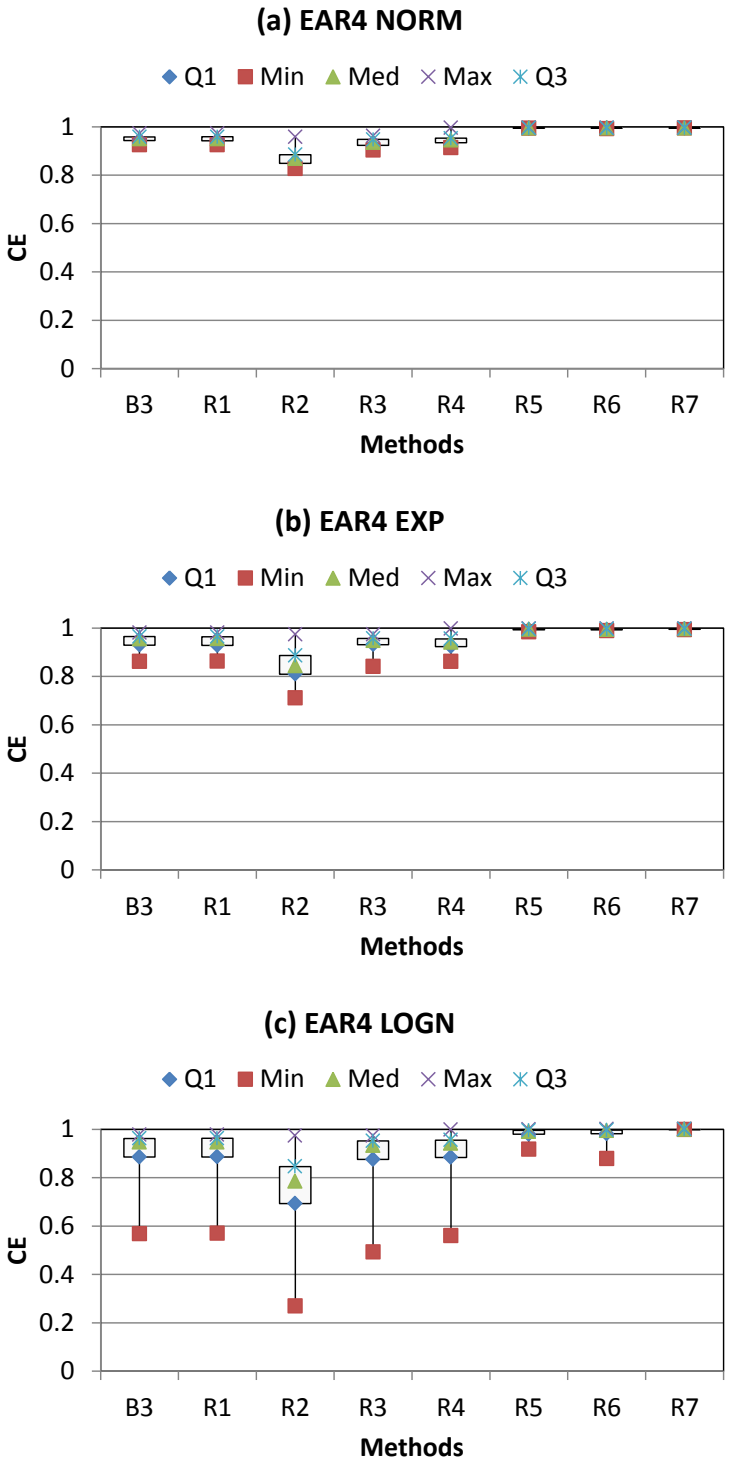


Figure 4.6 Accuracy of residual estimation with alternative estimators for EAR4 model (3 cases)

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.

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. 4.7) and NL (Fig. 4.8) models, with additional supporting information provided in APPENDIX-C Figs. C.2.1 to C.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. 4.9) than the NL model (Fig. 4.10). This is most likely due the different degrees of non-linearity of the datasets. 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 error with increased bandwidths, which is consistent with the finding in Harrold et al. (2001) and Li et al. (2015).

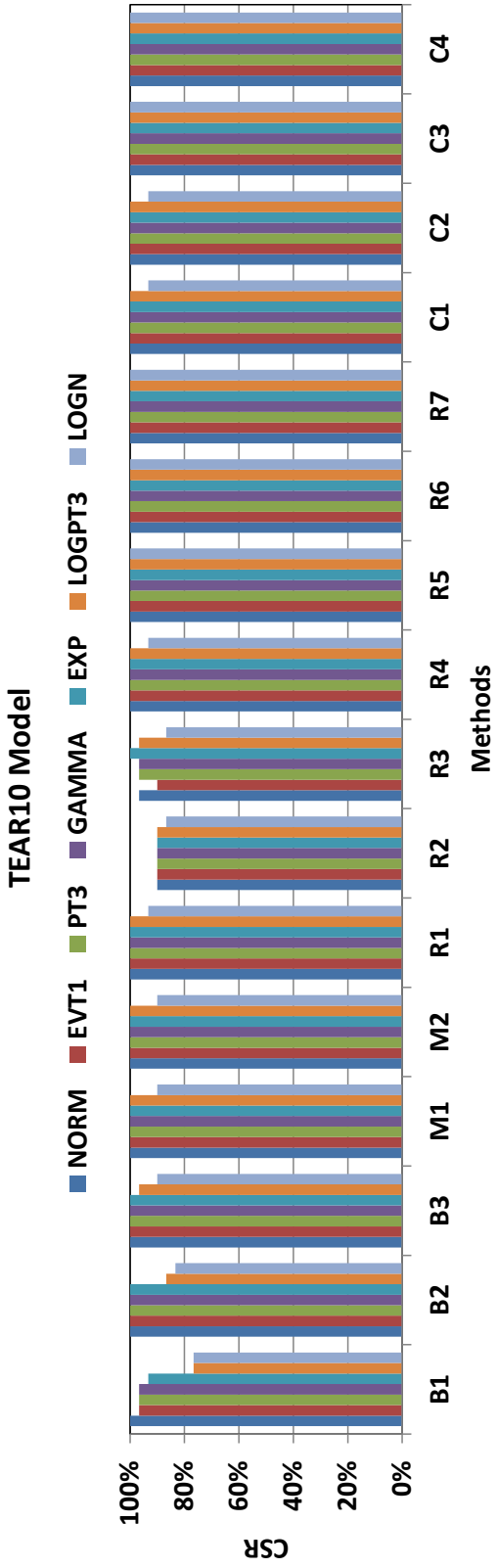


Figure 4.7 Selection accuracy of the PMI with suggested settings for TEAR10 models

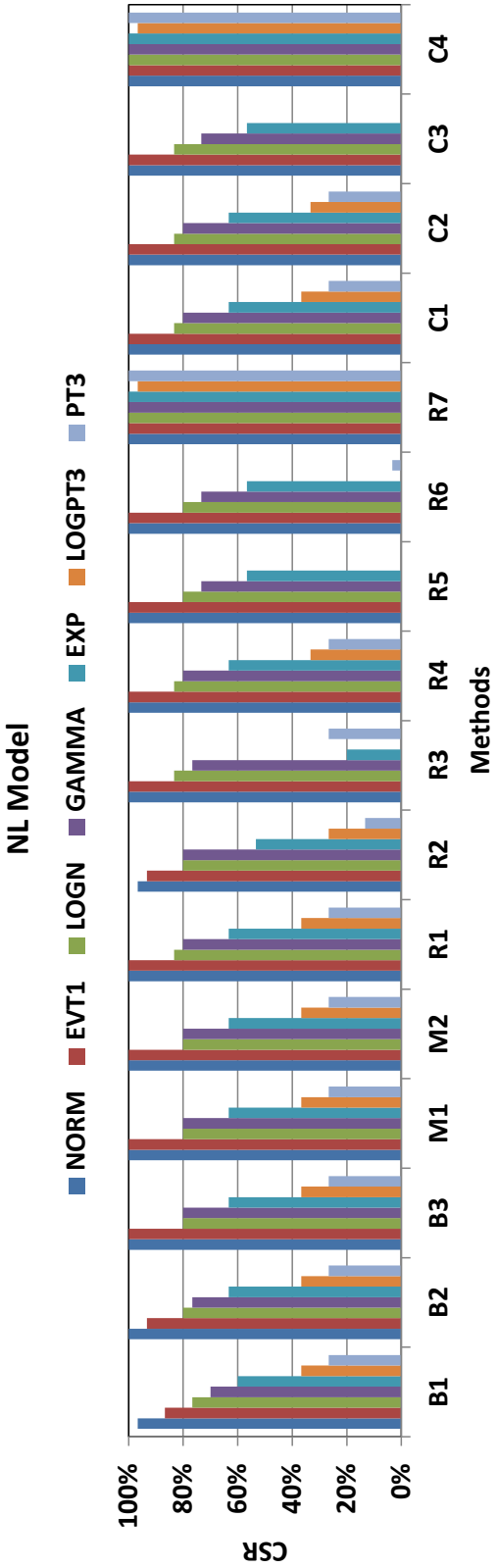


Figure 4.8 Selection accuracy of the PMI with suggested settings for NL models

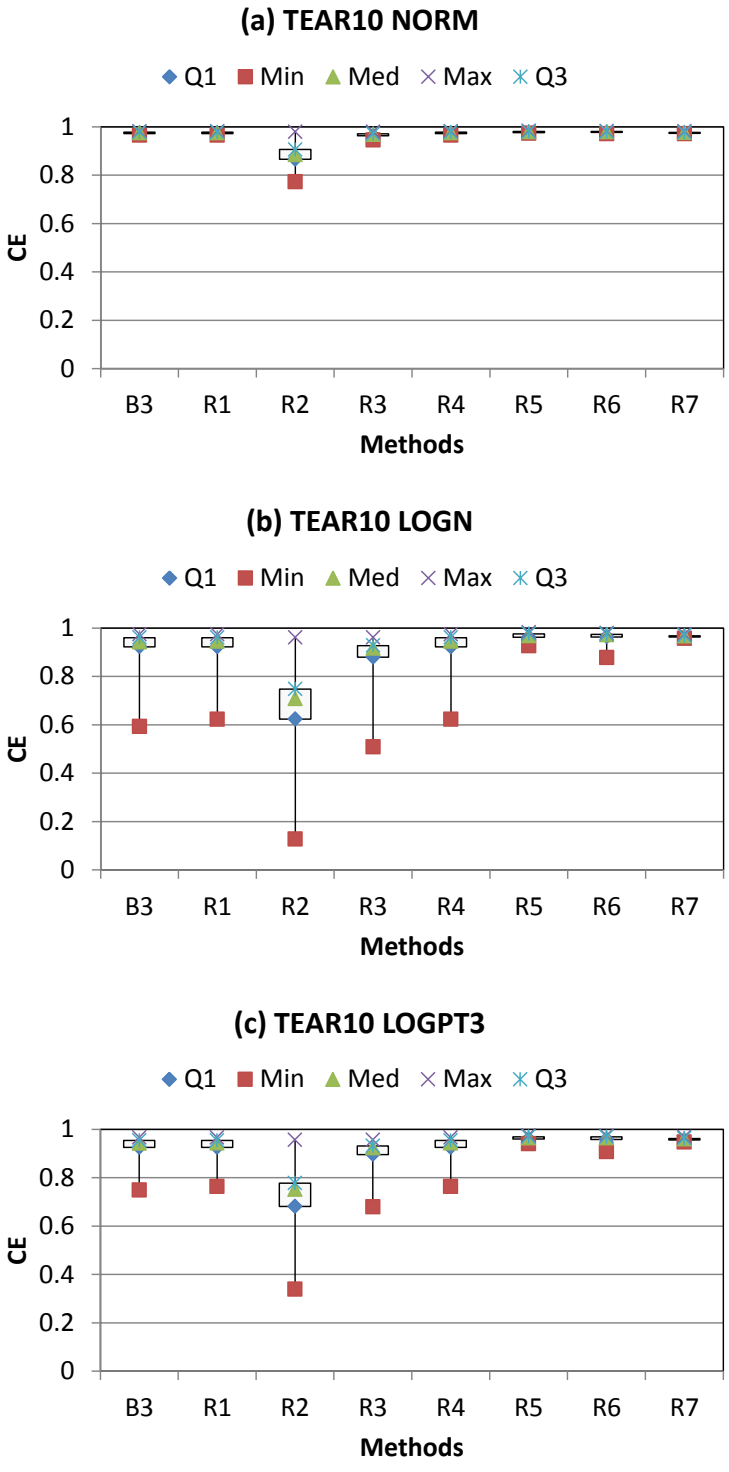


Figure 4.9 Accuracy of residual estimation with alternative estimators for TEAR10 model (3 cases)

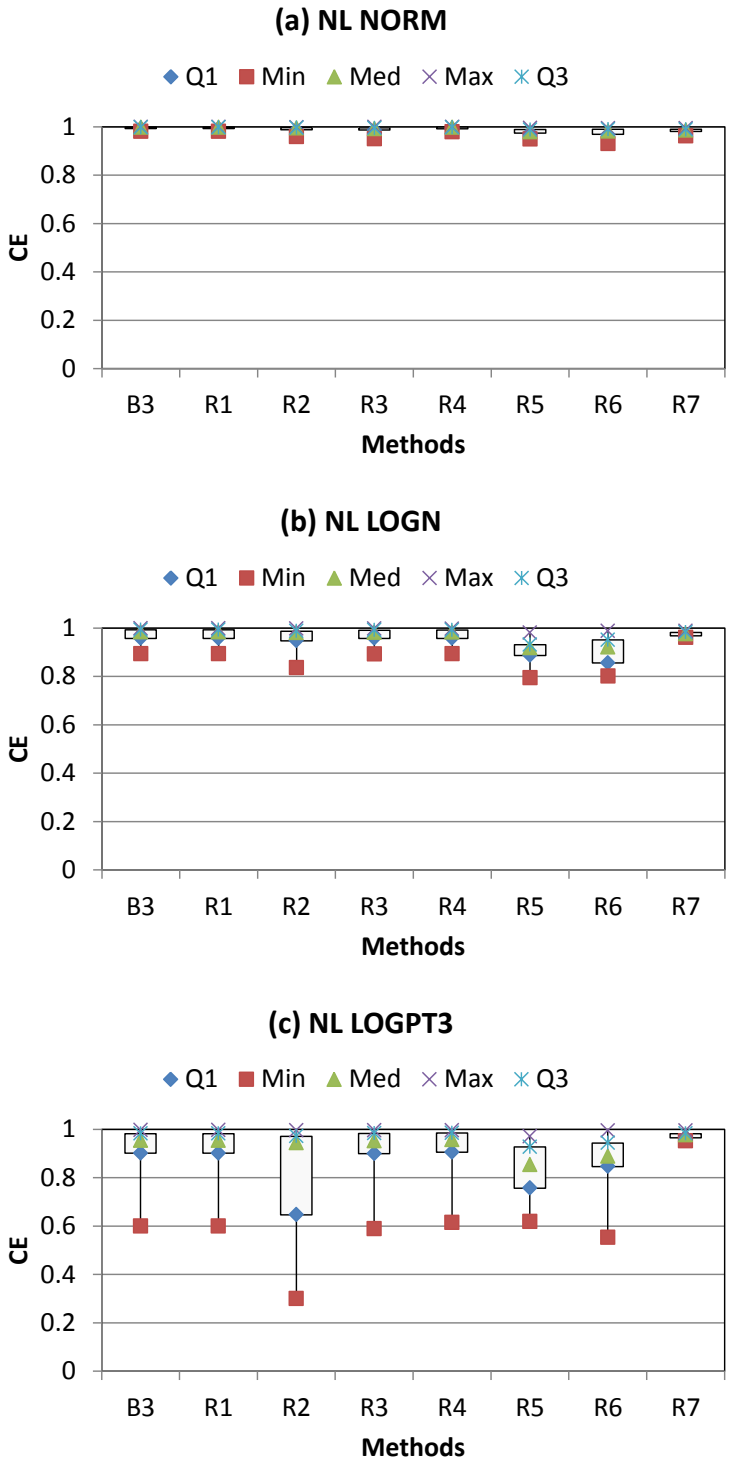


Figure 4.10 Accuracy of residual estimation with alternative estimators for NL model (3 cases)

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. 4.4 and 4.7), 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 residual estimation of non-linear problems, as influenced by both the boundary issue and problem nonlinearity, cannot be effectively improved by using local linear (1^{st} order) or quadratic (2^{nd} 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. 4.8), 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*.

4.4.2 Computational efficiency

The computational efficiency of the different PMI IVS approaches investigated is displayed in Fig. 4.11. As can be seen, the conventional benchmark approach (B1) is 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 residual

estimation 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 residual estimation (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. 4.11), 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.

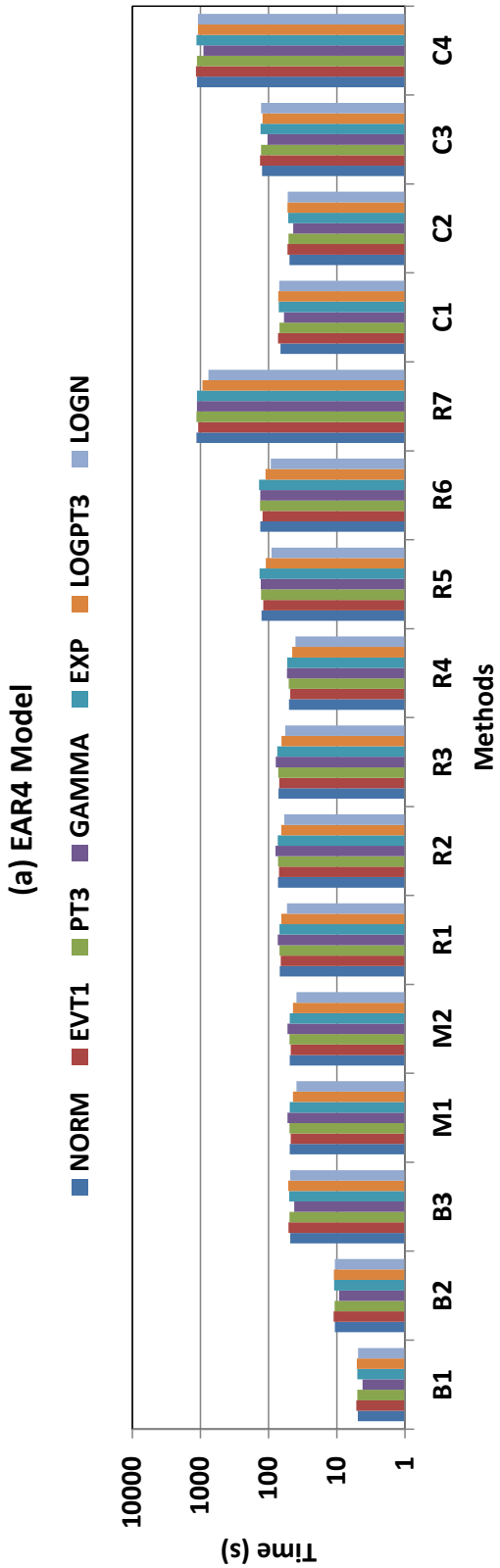


Figure 4.11 Selection efficiency of the PMI IVS with tested methods for EAR4 models

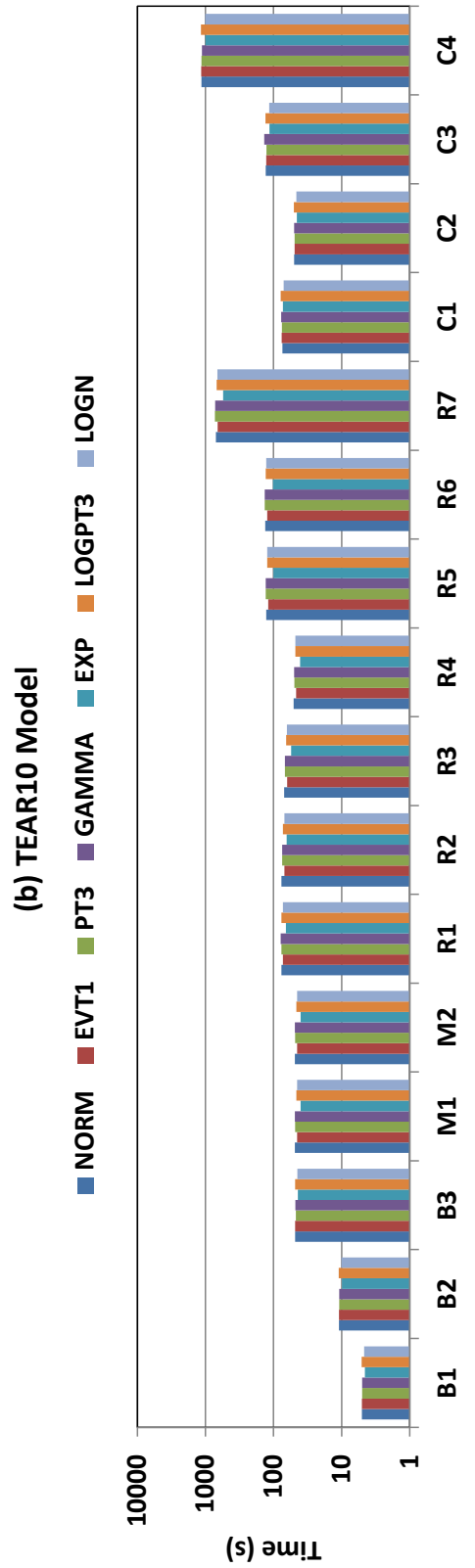


Figure 4.11 ***(Continued)***

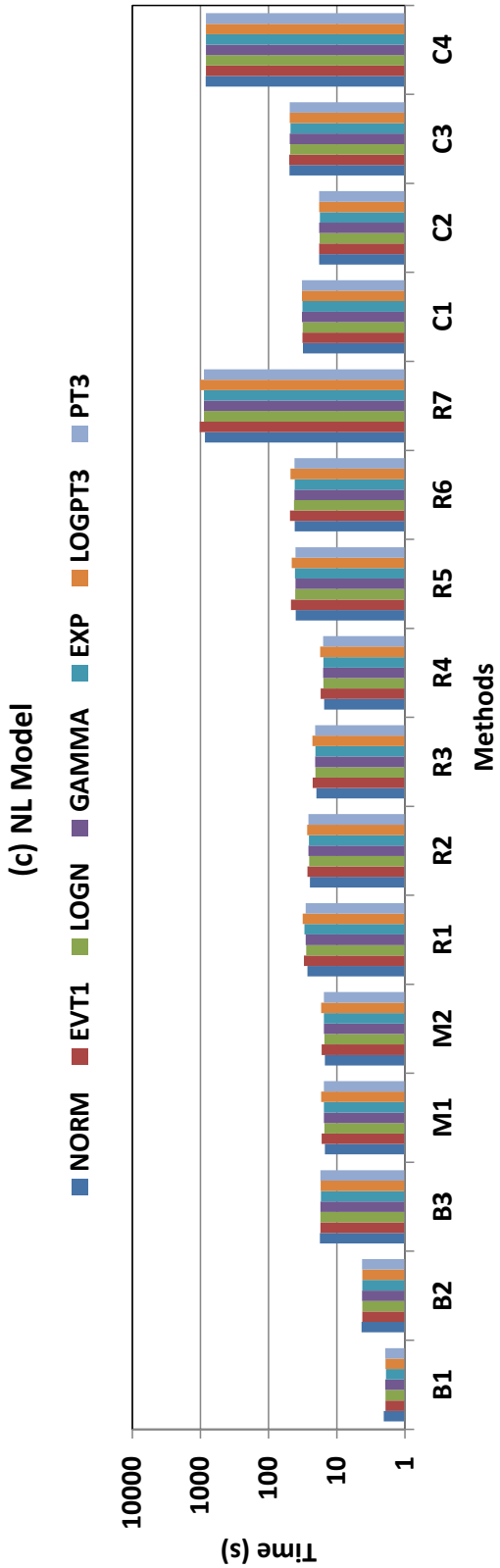


Figure 4.11 *(Continued)*

4.4.3 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. (2014b, 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. 4.12. 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.

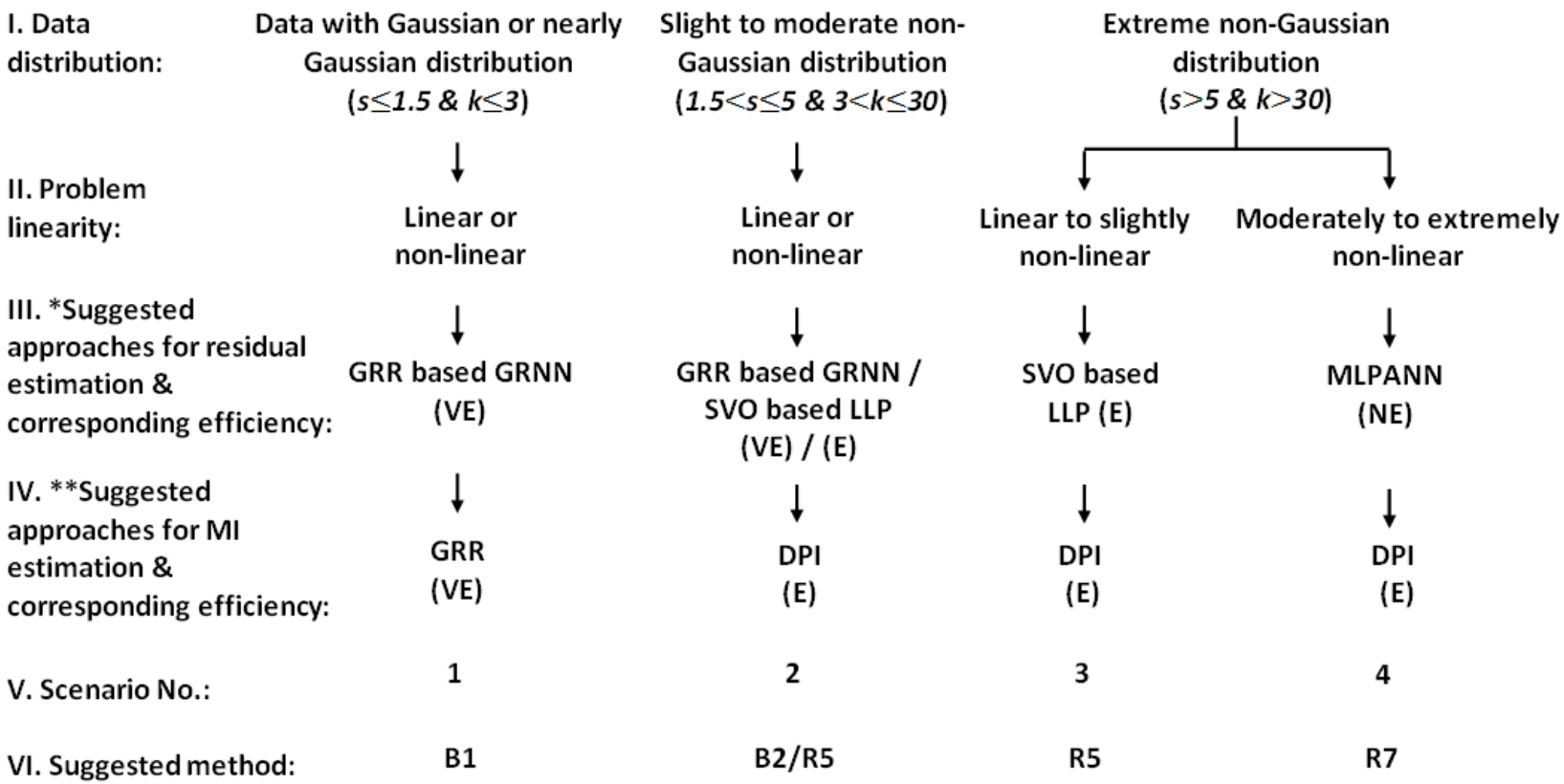


Figure 4.12 Suggested PMI IVS approaches under distinct scenarios

(VE = comparatively very computationally efficient, E = comparatively computationally efficient, and NE = comparatively not computationally efficient; *recommendation based on Li et al. (2014b) 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 residual estimation 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 residual estimation 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 residual estimation 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 residual estimation 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 residual estimation 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.

4.5 Validation on Murray Bridge and Kentucky River Basin case studies

4.5.1 Background

The rules and guidelines proposed in Section 4.4.3 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., 2005a; Li et al., 2014b; Li et al., 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., 2014b; Li et al., 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 4.5.

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 4.6.

Table 4.5 Candidate inputs and output used to forecast salinity at Murray Bridge 14 days in advance

Candidate Inputs				Output			
Location	Variable	Abbreviation	Lags	Location	Variable	Abbreviation	Forecasting Period
Mannum	Salinity	MAS	1,3,5,7,9	Murray Bridge	Salinity	MBS	14
Morgan	Salinity	MOS	1,3,5,7,9				
Waikerie	Salinity	WAS	1,2,3,4,5				
Loxton	Salinity	LOS	1,2,3,4,5				
Murray Bridge	Salinity	MBS	1,3,5,7,9				
Lock 1 Upper	River level	L1UL	-3,-1,1,3,5				

Table 4.6 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				

4.5.2 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 outputs, 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).

4.5.3 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., 2014b; Li et al., 2015; Wu et al., 2013). Consequently, these data correspond to Scenario 2 in Fig. 4.12. 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. 4.13 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. 4.12).

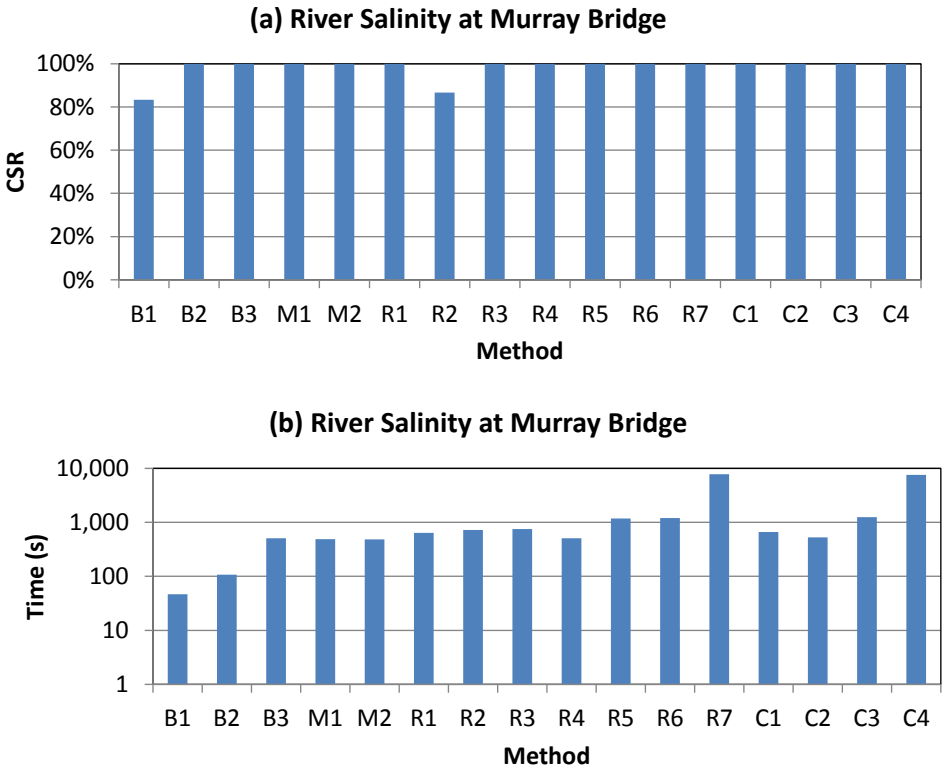


Figure 4.13 Selection accuracy and efficiency of the PMI IVS with suggested settings for Murray Bridge case

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., 2014b; Li et al., 2015; Wu et al., 2013), it corresponds to Scenario 4 in Fig. 4.12. 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. 4.14 (a) and 4.14 (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. 4.14 (b), this provide the most robust selection accuracy, as suggested by the proposed guidelines (Fig. 4.12).

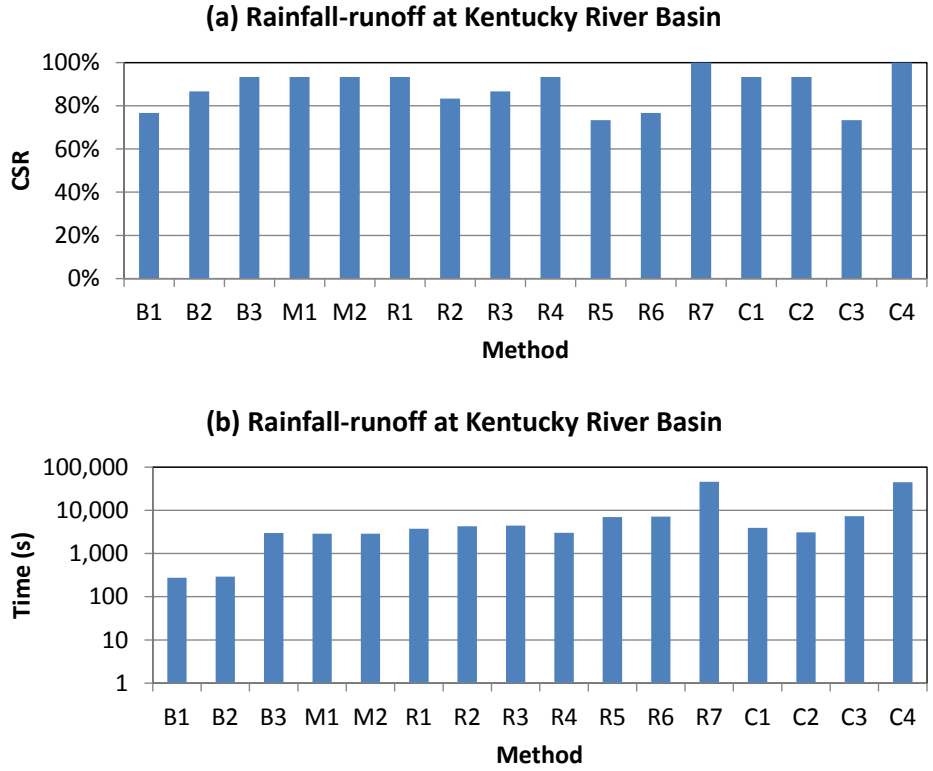


Figure 4.14 Selection accuracy and efficiency of the PMI IVS with suggested settings for Kentucky River basin case

4.6 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 residual estimation (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 for 16 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. 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 (Fig. 4.12), 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.

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 4.4), 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.

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. 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).

4.7 Acknowledgments

This research was aided by the code and suggestions from Dr. R.J. May (GRR based MI/PMI), Dr. G.B. Humphrey (GRR based GRNN) and Dr. J.C. Marshall and Dr. M.L. Hazelton (bivariate boundary kernel). The authors would also like to thank the three anonymous reviewers, whose input has improved the quality of this paper.

CHAPTER 5 CONCLUSIONS

5.1 Thesis summary

Artificial neural networks (ANNs) are one of the most commonly used data driven models for addressing environmental and water resources problems and they have been applied successfully and extensively over the last two decades. The performance of ANNs is essentially determined by the quality of the methods used in the various steps of their development, which consist of data collection, data processing, input variable selection (IVS), data division, calibration, validation, and application to real problems. IVS, as one of the most important steps in the development of ANNs and other data driven environmental and water resources models, as it determines the quality and quantity of information used in the modelling process.

Despite the existence of a large number of IVS techniques, partial mutual information (PMI) is one of the most promising approaches to IVS, as it is able to account for the relevance and redundancy of all candidate inputs and can be used for both linear and non-linear problems. However, current implementations of PMI IVS are not without their limitations. To the best of the author's knowledge, on one hand, the Gaussian reference rule (GRR), which assumes that the input/output data follow a Gaussian distribution, is still predominately used for the estimation of the kernel bandwidth within PMI IVS, even though the distribution of most water resources data is generally far from normal (this is known as bandwidth selection issue). On the other hand, the impact of the boundary issue, which is a result of the use of a symmetrical kernel at boundary, has not been addressed in environmental and water resources applications, although this contributes to an under-estimation of the kernel density near the boundary. As a result, the performance of current implementations of PMI IVS is compromised by both bandwidth selection and boundary issues. Consequently, the corresponding ultimate objective of this thesis is to improve the performance of PMI IVS by

investigating the impact of bandwidth selection and boundary issues for ANNs and other data driven environmental and water resources models.

In order to achieve the ultimate objective of this research, three detailed objectives and papers are established. Firstly, the performance of GRNN based residual estimation (RE), as part of PMI IVS, is assessed through the investigation of nine bandwidth estimators with various Gaussian dependence. Secondly, the performance of PMI IVS is studied through five bandwidth estimators with varying Gaussian dependence, as well as the proposed suggestions of bandwidth estimation of GRNN based RE. Thirdly, the performance of PMI IVS is further investigated by introducing sixteen methods that attenuate boundary issues associated with the guidelines of bandwidth selection with distinct data properties, obtained through the studies of the first two objectives. All the methods are assessed on synthetic models with distinct problem non-linearity. 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 with different degrees of normality are generated from distributions with differing degrees of normality, and the corresponding output data are obtained by substituting the generated inputs into synthetic models.

Based on the findings of the research presented in this thesis, it is suggested that:

1. The performance of PMI IVS is influenced by both bandwidth selection and boundary issues.

2. Currently implemented PMI IVS methods (i.e. depending on the Gaussian assumption without boundary correction) only result in reliable IVS if the input/output data follow Gaussian or nearly Gaussian distributions.

3. Bandwidths with reduced dependence on the Gaussian assumption can effectively improve selection accuracy for data that are non-Gaussian.

4. The proposed methods are very effective in addressing the boundary issue

5. It is vital to consider both bandwidth selection and boundary issues simultaneously.

The guidelines for selecting appropriate methods for MI/PMI and RE based on the properties of the available data appear to be very effective when tested on the semi-real validation data. The 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 enable the benefits of utilising measured input data (i.e. not generated from a known distribution) to be combined with those of having known outputs, 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). Although the developed guidelines are applied to datasets in which variables have similar distributions in Chapters 2, 3, and 4, this does not limit the methodological contribution of this research. As such, it is expected that this research is able to provide more robust and rigorous applications of PMI IVS for ANNs and other data driven environmental and water resources models.

5.2 Research contributions

The overall contribution of the present research is the effective improvement of PMI IVS, by considering a balance between accuracy and efficiency, through the investigation of both bandwidth selection and boundary issues. Based on the research presented in Chapters 2 to 4 of this thesis, details of critical contributions are summarised as follows:

1. The first contribution of this research is that it proposes rigorous and novel analytical procedures for assessing if, and to what degree, the performance of residuals and MI/PMI is affected by bandwidth selection and boundary issues.

For each study, a rigorous and novel analytical framework, including simulation of synthetic cases, adoption of investigated methods, application to semi-real problem based cases, and examination of modelling performance, is designed and implemented.

2. The second contribution of this research is that it provides an explanation for the suboptimal performance of conventional PMI IVS under the influence of the bandwidth selection and boundary issues. It is confirmed that use of GRR based bandwidth estimator only results in good input selection accuracy if the input/output data follow Gaussian or nearly Gaussian distributions. In contrast, 2-stage direct plug-in (DPI), combination of biased cross validation and DPI (BCVDPI), smoothed cross validation (SCV), and single variable optimisation (SVO) based bandwidth estimators, as a result of their reduced dependence on the Gaussian assumption, generally result in pronounced improvements in selection accuracy. The use of local linear polynomial (LLP) regression and multi-layer perceptron artificial neural network (MLPANN) models for RE is found to result in marked improvement when dealing with boundary issues, as a result of their increased resistance to the boundary issue for problems with data bounded at certain point(s).

3. The third contribution of this research is the development of effective preliminary guidelines based on the results of extensive controllable synthetic studies to deal with bandwidth selection and boundary issues under different scenarios categorised by data normality and problem linearity. By consolidating the established preliminary guidelines within all three papers and recalling Fig. 4.12, it is suggested that

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

(2) If the input/output data follow moderately non-Gaussian (average $1.3 < s \leq 5$ and $3 < k \leq 30$) distributions, a PMI approach with

the GRR based GRNN for RE and the DPI for MI estimation (B2) is suggested, so that correct selection rate (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), a PMI approach with the SVO based LLP for RE and the DPI for MI estimation (R5) is proposed for IVS. It should be noted that increasing computational challenges are expected when introducing the DPI and the SVO based LLP.

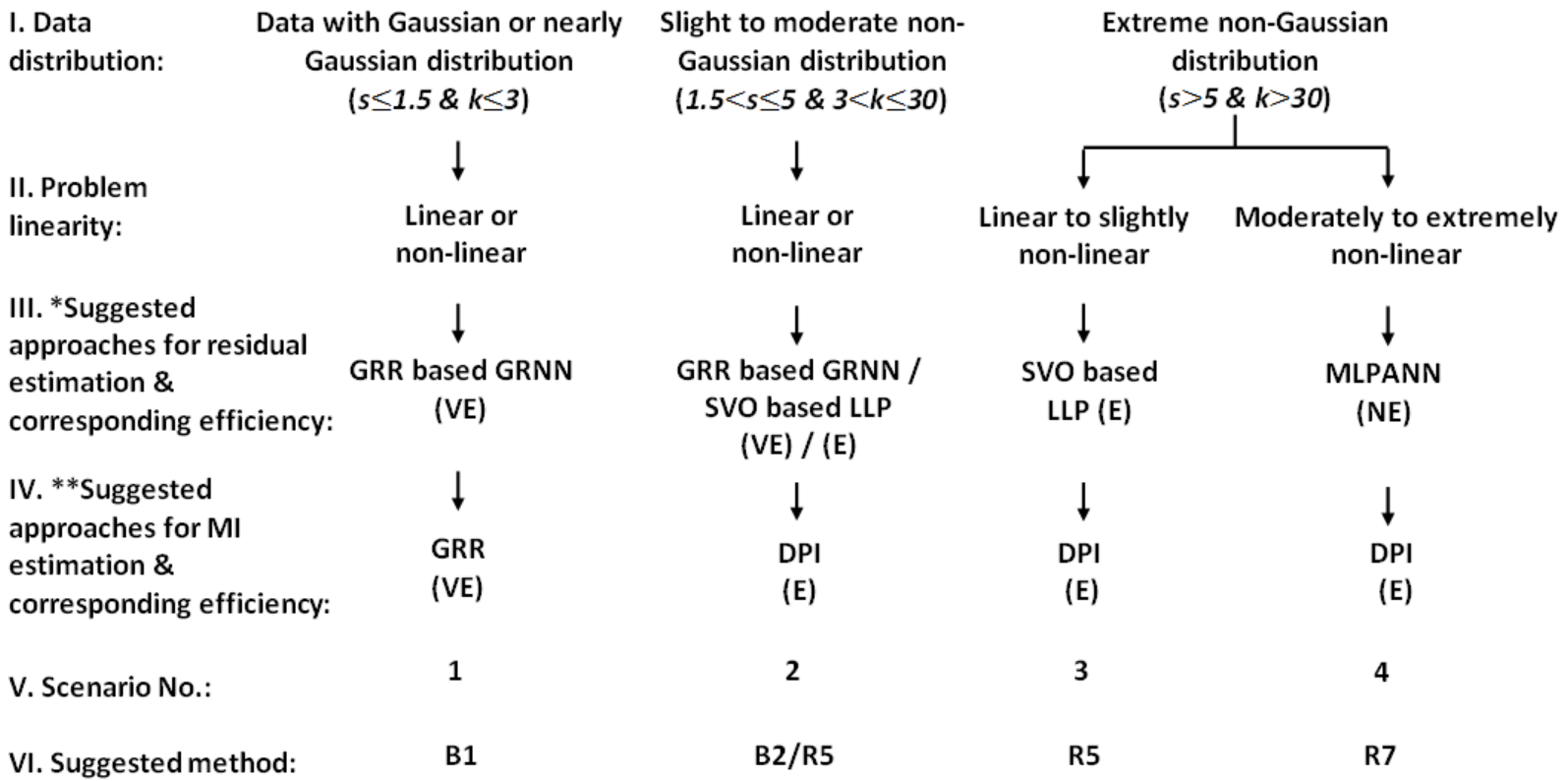
(3) If most of the input/output data follow extremely non-Gaussian (average $s > 5$ and $k > 30$) distributions and the problem is linear or slightly non-linear, a PMI approach with the SVO based LLP for RE and the DPI for MI estimation (R5) should be implemented, as the combined impact of bandwidth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

(4) If the same conditions as in Scenario 3 apply, except that the problem becomes moderately to extremely non-linear, a PMI approach with the MLPANN for RE and the DPI for MI estimation (R7) 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.

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.

When applying the proposed guidelines to different water resources and environmental modelling problems, it is recommended to check the average distributions (skewness and kurtosis) of input and output variables first and to

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 bandwidth and boundary issue should be considered in accordance with scenarios 3 and 4 in Fig. 4.12. While most water quality models mainly include input and output variables that follow Gaussian or nearly Gaussian distributions (e.g. concentration of dissolved oxygen in the river), therefore scenarios 1 and 2 in Fig. 4.12 should be implemented for the sake of good selection accuracy at the best possible computational efficiency. However, it is also acknowledged that the application of the proposed guidelines is also a function of case-study dependent features and user preferences.



Recall Figure 4.11 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. (2014b) and present study; **recommendation based on Li et al. (2015))

4. The fourth contribution of this research is more robust and reliable software based applications of the proposed PMI IVS for realistic environmental and water resources problems. A number of programs have been developed in accordance with the preliminary guidelines discussed in each of the journal papers and they are free to download for research purposes from the following website:

http://www.ecms.adelaide.edu.au/civeng/research/water/software/generalised-regression-neural-network/
https://github.com/xuyuanli/GRNNs
https://github.com/xuyuanli/IVS_PMI_2014

5.3 Publications

List of works contained within this thesis:

Paper 1 presented in Chapter 2 (Li et al., 2014b): Li, X., Zecchin, A.C., Maier, H.R., 2014b. 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.

Paper 2 presented in Chapter 3 (Li et al., 2015): 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 Software 65 15-29 DOI: 10.1016/j.envsoft.2014.11.028

Paper 3 presented in Chapter 4 (Li et al., 2014a): Li, X., Zecchin, A.C., Maier, H.R., 2014a. Improving partial mutual information-based input variable selection by consideration of boundary issues associated with bandwidth estimation. Environmental Modelling and Software, submitted on 04/12/2014.

List of works resulting from research associated with thesis but not contained within:

Li, X., Maier, H.R., Zecchin, A.C., 2013. Improving PMI based input selection by using different kernel bandwidths for artificial neural network models (extended abstract), *20th International Congress on Modelling and Simulation (MODSIM2013)*: Adelaide, Australia.

5.4 Recommendations for future research

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., 2014a; Galelli et al., 2014). However, the computational expense of some methods described in the proposed guidelines is of concern and the development of alternative methods with equivalent selection accuracy but better computational efficiency is suggested for future research. Alternative methods for dealing with non-Gaussian data in the context of PMI IVS that deserve consideration in this context include:

1) **transforming the input data to normality** (e.g. Bowden et al., 2003), which requires a combination of normalising the data and transforming the kernel. As such, the computational efficiency can be improved by applying the Gaussian assumption to the normalised data, while the bandwidth and boundary issues are addressed simultaneously. The major challenges of this approach are to determine the most effective data transformation method(s) and to derive the corresponding transformation kernel(s) in 1D and 2D.
2) **estimating the required densities using histogram-based methods** (e.g. Fernando et al., 2009), which could potentially perform as well as the

proposed guidelines, but with better efficiency due to the fact that such methods are not affected by the boundary issue, which is only associated with kernel based approaches. However, the major challenge of this histogram-based method is to approximate the optimal histogram bin width, so that it is neither too large nor too small for general cases. This challenge is, in fact, technically similar to the bandwidth selection issue.

As part of the PMI IVS approach, the stopping criterion can also affect the stability of selection accuracy, which has been mentioned before in this research (Section 3.2 and 4.2.1 PMI IVS). However, in this research only AIC, suggested by May et al. (2008b), is used. As a consequence, the impact of the stopping criterion also requires further investigation to secure the robustness of the proposed guidelines. Alternative stopping criteria that could be considered for this purpose include bootstrapping, tabulated critical values, and the Hampel test, as discussed and tested in May et al. (2008b).

In addition, the data used for the synthetic tests are pre-determined and controllable with low degree of noise, collinearity and interdependency. In contrast, the data for realistic water resources and environmental problems can be far more complicated. Consequently, 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). All future analysis and tests of IVS are also strongly recommended to follow the systematic approach proposed by Galelli et al. (2014).

REFERENCES

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.

Abrahart, R.J., 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.

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.

Agalbjörn, S., Končar, N., Jones, A., 1997. A note on the gamma test. Neural Computing and Applications 5(3) 131-133.

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

Aldershof, B.K., 1991. Estimation of integrated squared density derivatives. The University of North Carolina, Chapel Hill.

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

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

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., 2003. Data transformation for neural network models in water resources applications. Journal of Hydroinformatics 5 245-258.

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., 2002. Optimal division of data for neural network models in water resources applications. Water Resources Research 38(2) 2.1-2.11.

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.

Bowden, G.J., Nixon, J.B., Dandy, G.C., Maier, H.R., Holmes, M., 2006. Forecasting chlorine residuals in a water distribution system using a general regression neural network. Mathematical and Computer Modelling 44(5) 469-484.

Bowman, A.W., 1984. An alternative method of cross-validation for the smoothing of density estimates. Biometrika 71(2) 353-360.

Box, G.E., Jenkins, G.M., Reinsel, G.C., 2013. Time series analysis: forecasting and control. John Wiley and Sons, Hoboken, New Jersey.

Buhmann, M.D., 2003. Radial basis functions: theory and implementations, vol. 12. Cambridge university press.

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

Cai, Z., 2001. Weighted nadaraya–watson regression estimation. Statistics and Probability Letters 51(3) 307-318.

Castellano, G., Fanelli, A.M., 2000. Variable selection using neural-network models. Neurocomputing 31(1) 1-13.

Castelletti, A., Galelli, S., Ratto, M., Soncini-Sessa, R., Young, P., 2012a. A general framework for dynamic emulation modelling in environmental problems. Environmental Modelling and Software 34 5-18.

Castelletti, A., Galelli, S., Restelli, M., Soncini-Sessa, R., 2012b. Data-driven dynamic emulation modelling for the optimal management of environmental systems. Environmental Modelling and Software 34 30-43.

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

Chua, L.H.C., Wong, T.S.W., 2010. Improving event-based rainfall-runoff modeling using a combined artificial neural network-kinematic wave approach. Journal of Hydrology 390(1-2) 92-107.

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

Coulibaly, P., Bobée, B., Anctil, F., 2001. Improving extreme hydrologic events forecasting using a new criterion for artificial neural network selection. Hydrological Processes 15(8) 1533-1536.

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., 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.

Dawson, C.W., Harpham, C., Wilby, R.L., Chen, Y., 2002. Evaluation of artificial neural network techniques for flow forecasting in the River Yangtze, China. Hydrology and Earth System Sciences 6(4) 619-626.

Dawson, C.W., Mount, N., Abrahart, R., Louis, J., 2014. Sensitivity analysis for comparison, validation and physical-legitimacy of neural network-based hydrological models. Journal of Hydroinformatics 16(2) 407-424.

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

De Vos, N., Rientjes, T., 2007. Multi-objective performance comparison of an artificial neural network and a conceptual rainfall—runoff model. Hydrological Sciences Journal 52(3) 397-413.

Duong, T., Hazelton, M., 2003. Plug-in bandwidth matrices for bivariate kernel density estimation. Journal of Nonparametric Statistics 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.G., 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.

Haimi, H., Mulas, M., Corona, F., Vahala, R., 2013. Data-derived soft-sensors for biological wastewater treatment plants: An overview. Environmental Modelling and Software 47 88-107.

Hall, P., Marron, J.S., 1987. Estimation of integrated squared density derivatives. Statistics and Probability Letters 6(2) 109-115.

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.

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.

Hu, T., Lam, K., Ng, S., 2001. River flow time series prediction with a range-dependent neural network. Hydrological Sciences Journal 46(5) 729-745.

Hu, T., Wu, F., Zhang, X., 2007. Rainfall-runoff modeling using principal component analysis and neural network. Nordic Hydrology 38(3) 235-248.

Huang, D., Chow, T.W.S., 2005. Effective feature selection scheme using mutual information. Neurocomputing 63 325-343.

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., Indurthy, S.K.V.P., 2003. Comparative analysis of event-based rainfall-runoff modeling techniques—deterministic, statistical, and artificial neural networks. Journal of Hydrologic Engineering 8(2) 93-98.

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.

Jain, S., Das, A., Srivastava, D., 1999. Application of ANN for reservoir inflow prediction and operation. Journal of Water Resources Planning and Management 125(5) 263-271.

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.

Jones, M., Sheather, S., 1991. Using non-stochastic terms to advantage in kernel-based estimation of integrated squared density derivatives. Statistics and Probability Letters 11(6) 511-514.

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

Karunamuni, R.J., Alberts, T., 2005b. On boundary correction in kernel density estimation. Statistical Methodology 2(3) 191-212.

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

Kingston, G.B., Maier, H.R., Lambert, M.F., 2005b. Calibration and validation of neural networks to ensure physically plausible hydrological modeling. Journal of Hydrology 314(1) 158-176.

Kingston, G.B., Maier, H.R., Lambert, M.F., 2008. Bayesian model selection applied to artificial neural networks used for water resources modeling. Water Resources Research 44(4) W04419.

Krause, P., Boyle, D., Bäse, F., 2005. Comparison of different efficiency criteria for hydrological model assessment. Advances in Geosciences 5(5) 89-97.

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 Software 65 15-29 DOI: 10.1016/j.envsoft.2014.11.028.

Li, X., Zecchin, A.C., Maier, H.R., 2014a. Improving Partial Mutual Information-based input variable selection by consideration of boundary issues associated with bandwidth estimation. Environmental Modelling and Software, submitted.

Li, X., Zecchin, A.C., Maier, H.R., 2014b. 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., 1996. The use of artificial neural networks for the prediction of water quality parameters. Water Resources Research 32(4) 1013-1022.

Maier, H.R., Dandy, G.C., 1997a. Determining inputs for neural network models of multivariate time series. Computer-Aided Civil and Infrastructure Engineering 12(5) 353-368.

Maier, H.R., Dandy, G.C., 1997b. 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., 1998a. The effect of internal parameters and geometry on the performance of back-propagation neural networks: an empirical study. Environmental Modelling and Software 13(2) 193-209.

Maier, H.R., Dandy, G.C., 1998b. Understanding the behaviour and optimising the performance of back-propagation neural networks: an empirical study. Environmental Modelling and Software 13(2) 179-191.

Maier, H.R., Dandy, G.C., 1999. Empirical comparison of various methods for training feed-Forward neural networks for salinity forecasting. Water Resources Research 35(8) 2591-2596.

Maier, H.R., Dandy, G.C., 2000a. Application of artificial neural networks to forecasting of surface water quality variables: issues, applications and challenges, In: Govindaraju, R.S., Rao, A.R. (Ed.), Artificial Neural Networks in Hydrology. Springer: Kluwer Academic Publishers, The Netherlands, pp. 595-605.

Maier, H.R., Dandy, G.C., 2000b. 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., Burch, M.D., 1998. Use of artificial neural networks for modelling cyanobacteria Anabaena spp. in the River Murray, South Australia. Ecological Modelling 105(2) 257-272.

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.J., Dandy, G.C., Maier, H.R., 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. Modeling microalgal abundance with artificial neural networks: Demonstration of a heuristic 'Grey-Box'to deconvolve and quantify environmental influences. Environmental Modelling and Software 38 27-39.

Mount, N., Dawson, C., Abrahart, R., 2013. Legitimising data-driven models: exemplification of a new data-driven mechanistic modelling framework. Hydrology and Earth System Sciences 17(7) 2827-2843.

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 (Salmo trutta L.) in Iberian rivers. Environmental Modelling and Software 59 30-43.

Noori, R., Karbassi, A., Moghaddamnia, A., Han, D., Zokaei-Ashtiani, M., Farokhnia, A., Gousheh, M.G., 2011. Assessment of input variables determination on the SVM model performance using PCA, Gamma test, and forward selection techniques for monthly stream flow prediction. Journal of Hydrology 401(3) 177-189.

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.

Park, B.U., Marron, J.S., 1992. On the use of pilot estimators in bandwidth selection. Journal of Nonparametric Statistics 1(3) 231-240.

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.

Poli, R., Kennedy, J., Blackwell, T., 2007. Particle swarm optimization. Swarm Intelligence 1(1) 33-57.

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.

Rakovec, O., Hill, M., Clark, M., Weerts, A., Teuling, A., Uijlenhoet, R., 2014. Distributed Evaluation of Local Sensitivity Analysis (DELSA), with application to hydrologic models. Water Resources Research 50(1) 409-426.

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., 2000a. 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.

Sharma, A., 2000b. Seasonal to interannual rainfall probabilistic forecasts for improved water supply management: Part 3 - A nonparametric probabilistic forecast model. Journal of Hydrology 239(1-4) 249-258.

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

Specht, D.F., 1990. Probabilistic neural networks. Neural Networks 3(1) 109-118.

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.

Trappenberg, T., Ouyang, J., Back, A., 2006. Input variable selection: mutual information and linear mixing measures. Knowledge and Data Engineering, IEEE Transactions on 18(1) 37-46.

Wand, M.P., Jones, M.C, 1993. Comparison of smoothing parameterizations in bivariate kernel density estimation. Journal of the American Statistical Association 88(422) 520-528.

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

Wei, Q., Lu, Z., Chen, K., Ma, Y., 2010. Channel Selection for Optimizing Feature Extraction in an Electrocorticogram-Based Brain-Computer Interface. Journal of Clinical Neurophysiology 27(5) 321.

Williams, R.J., Zipser, D., 1989. A learning algorithm for continually running fully recurrent neural networks. Neural Computation 1(2) 270-280.

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.C., Maier, H.R., 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., Dandy, G.C., Maier, H.R., 2012. A method for comparing data splitting approaches for developing hydrological ANN models, International Congress on Environmental Modelling and Software (6th: 2012: Leipzig, Germany) iEMSs 2012.

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.

Yang, J., Li, L., Wang, A., 2011. A partial correlation-based Bayesian network structure learning algorithm under linear SEM. Knowledge-Based Systems 24(7) 963-976.

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.

Zhang, X., Liang, F., Yu, B., Zong, Z., 2011. Explicitly integrating parameter, input, and structure uncertainties into Bayesian Neural Networks for probabilistic hydrologic forecasting. Journal of Hydrology 409(3) 696-709.

APPENDICES

APPENDIX-A Supplementary Material from Paper 1 (Chapter 2)

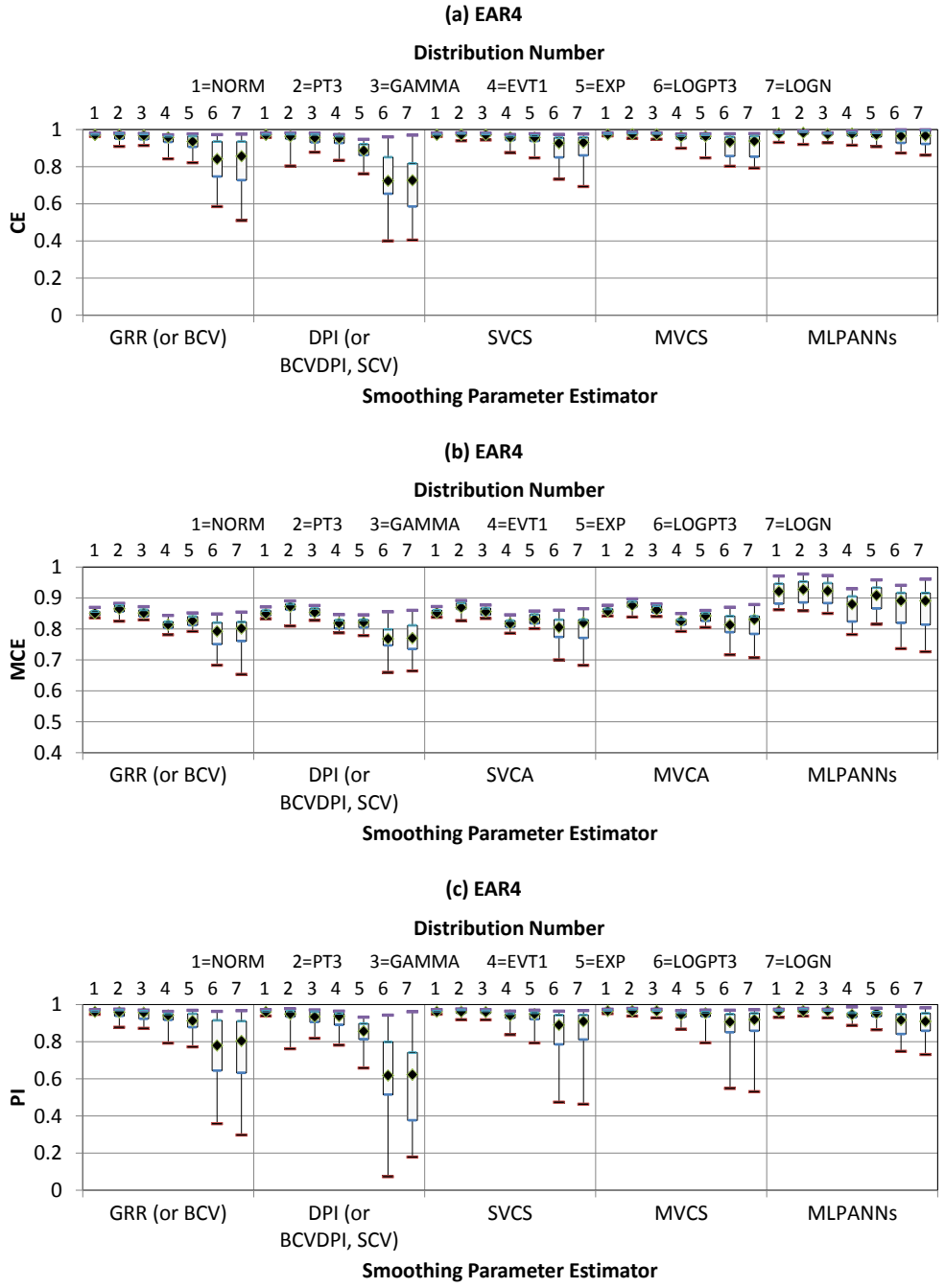
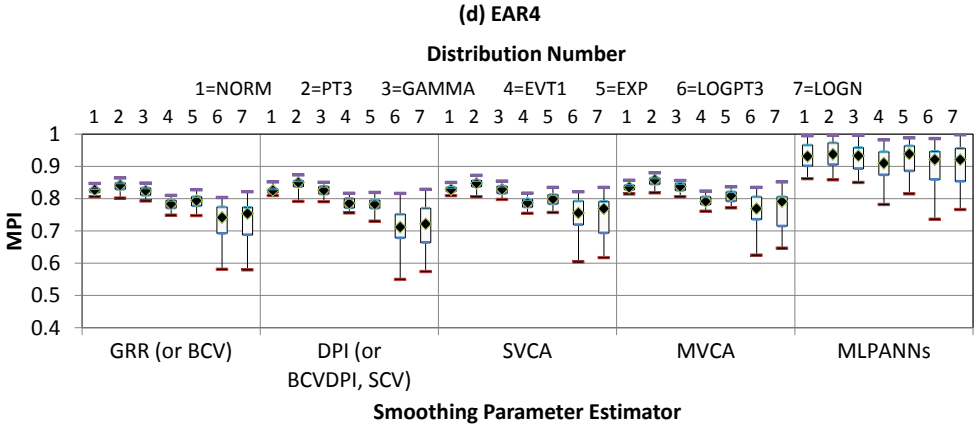
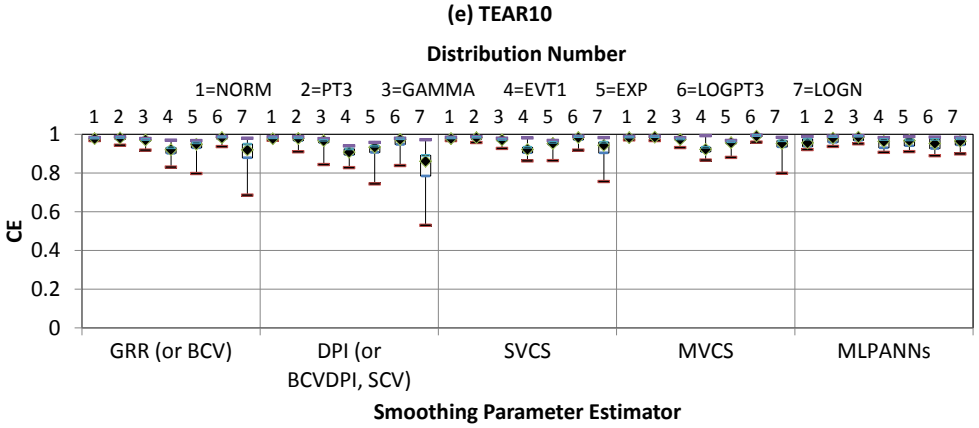


Figure A.1 Predictive accuracy for the validation data of MLPs and GRNNs, measured by CE, MCE, PI & MPI, for different synthetic data-generating models and distributions for which optimal parameters have been obtained using different methods





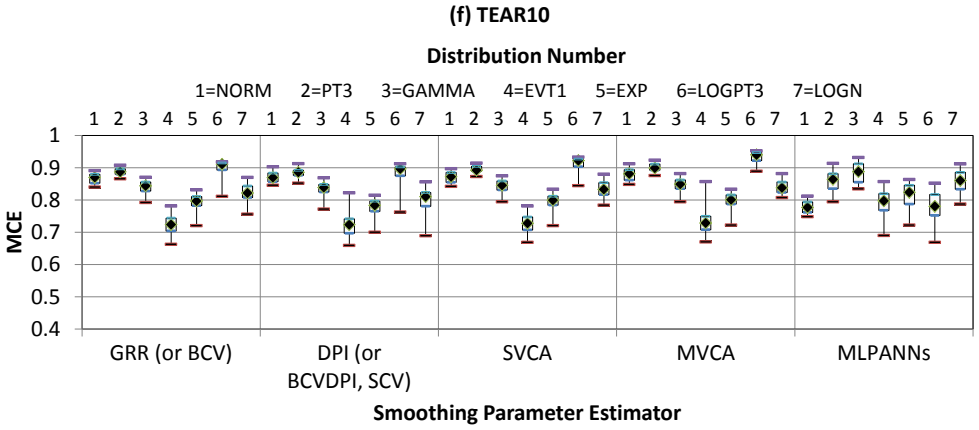
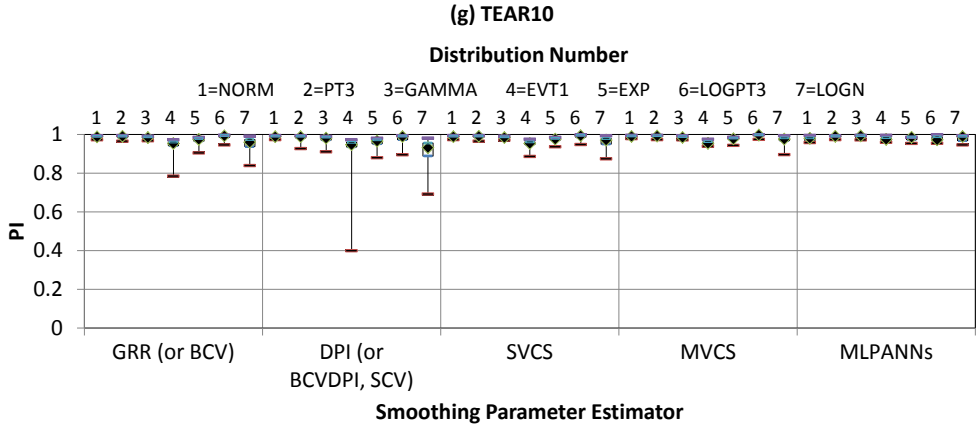
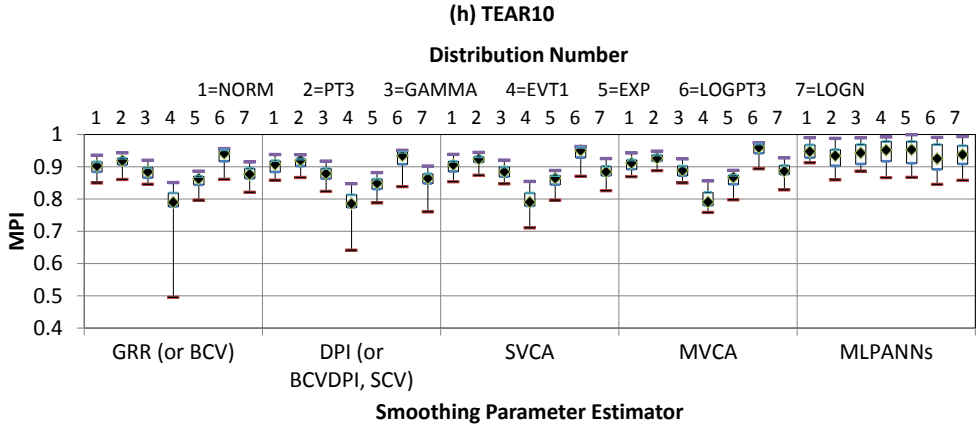


Figure A.1 *(Continued)*





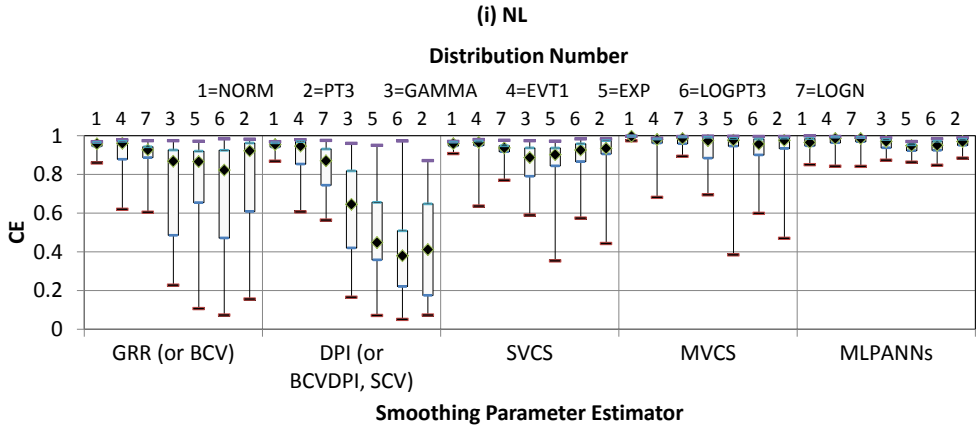
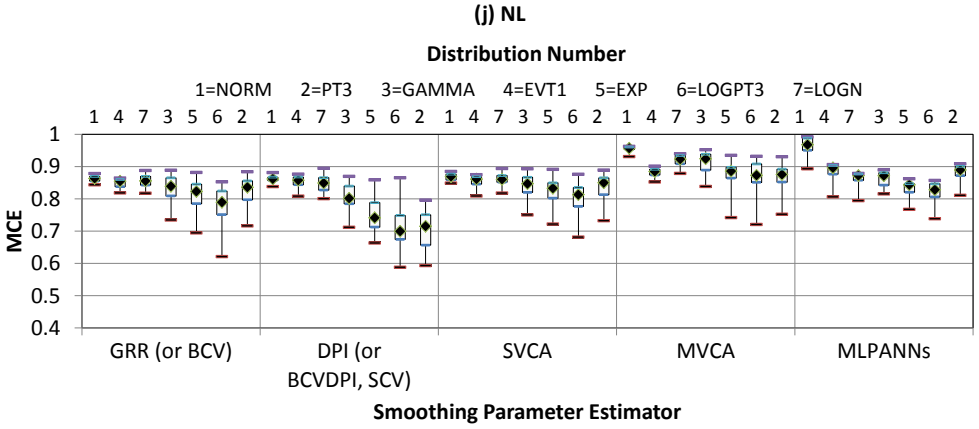
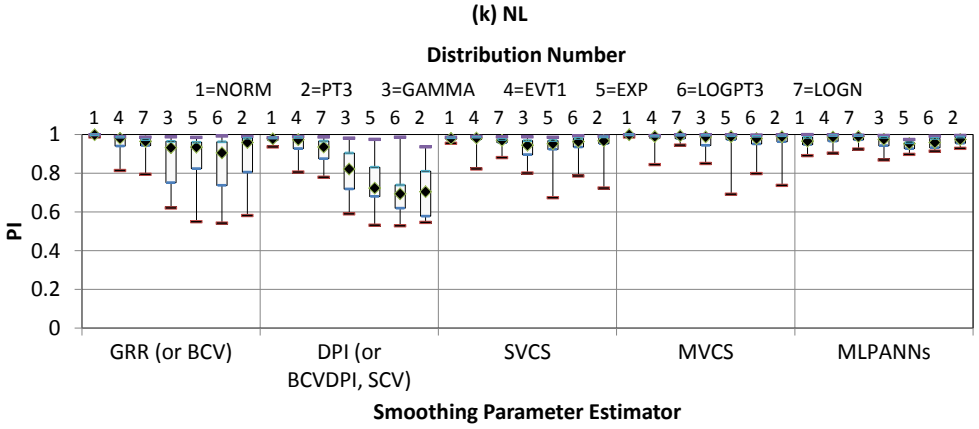


Figure A.1 *(Continued)*





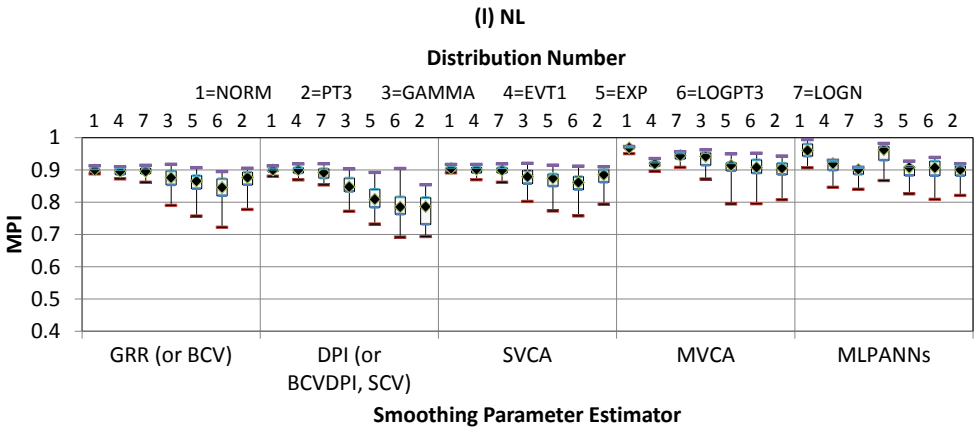


Figure A.1 *(Continued)*

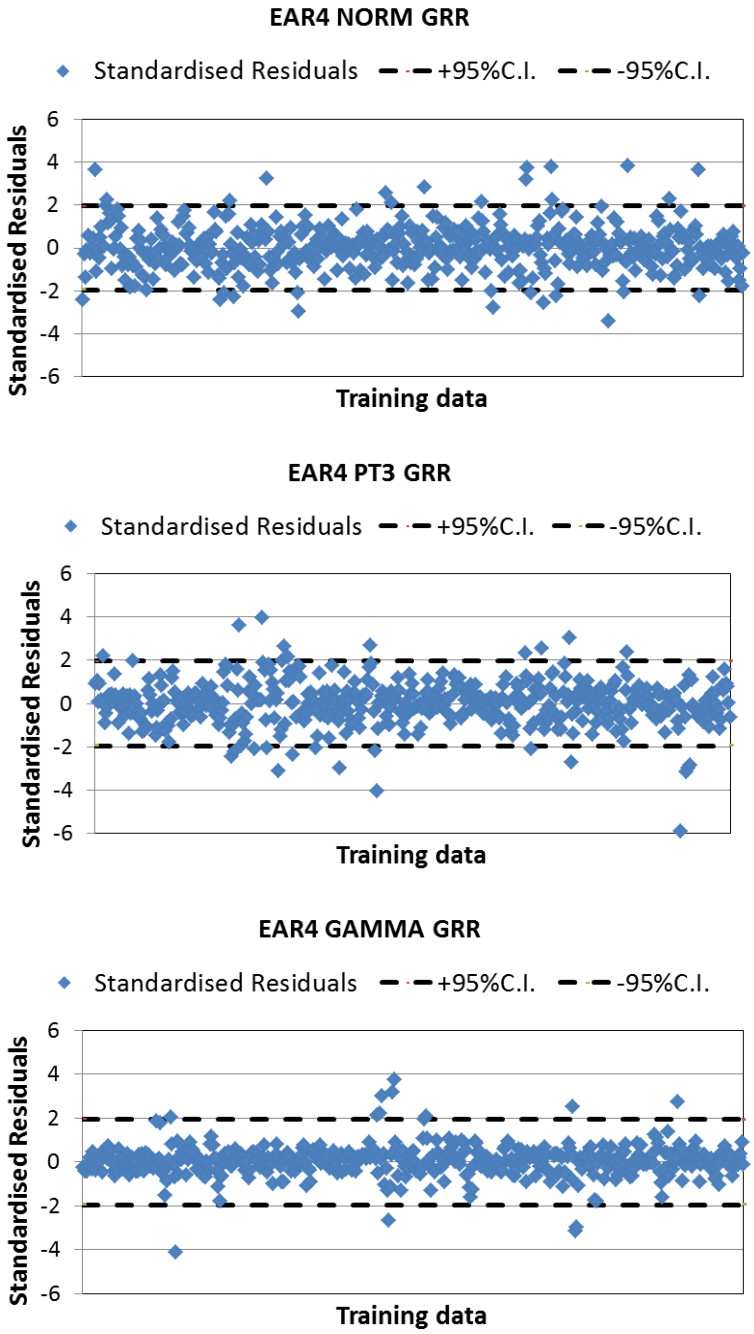


Figure A.2 Standardised residuals for the training data of MLPs and GRNNs with different smoothing parameters for EAR4 model with different distributions

(performance of the BCV was similar to that of the GRR; performance of the BCVDPI and SCV was similar to that of the DPI; similar plots were also observed for TEAR10 & NL models**)**

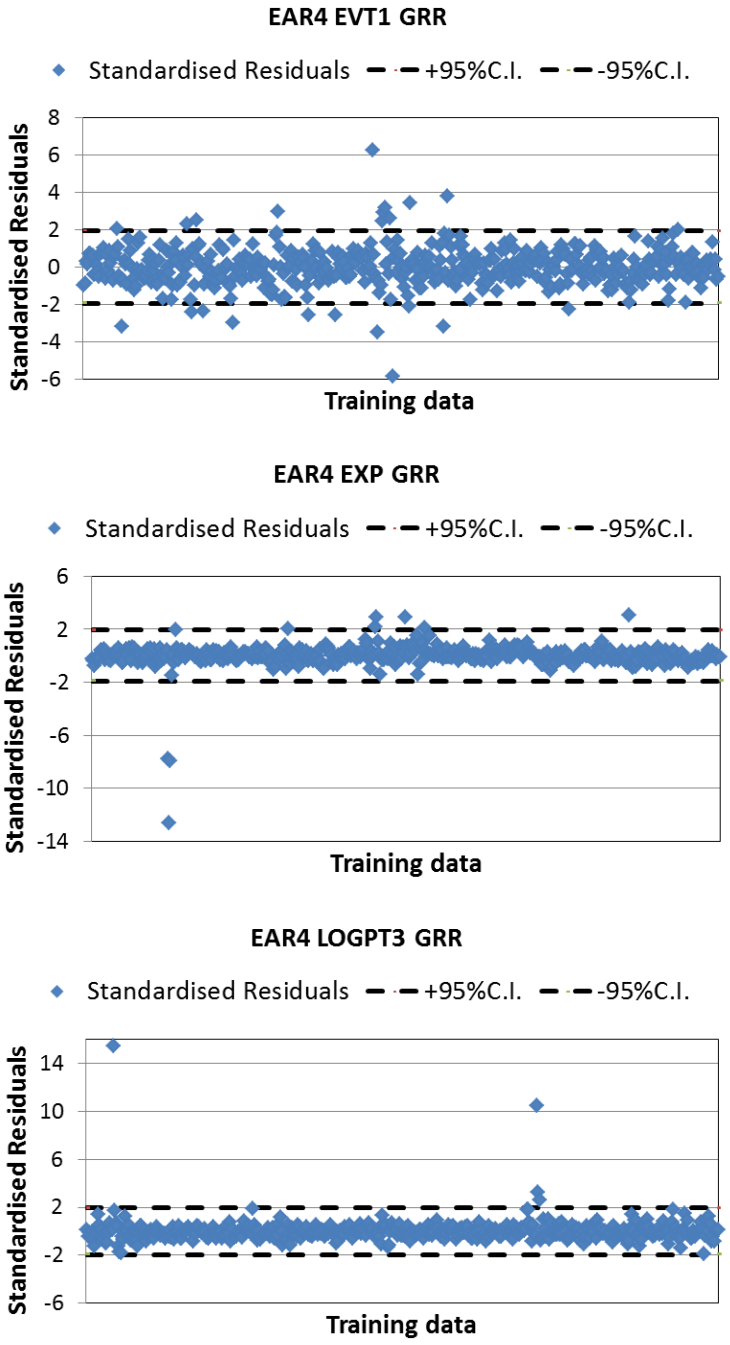


Figure A.2 *(Continued)*

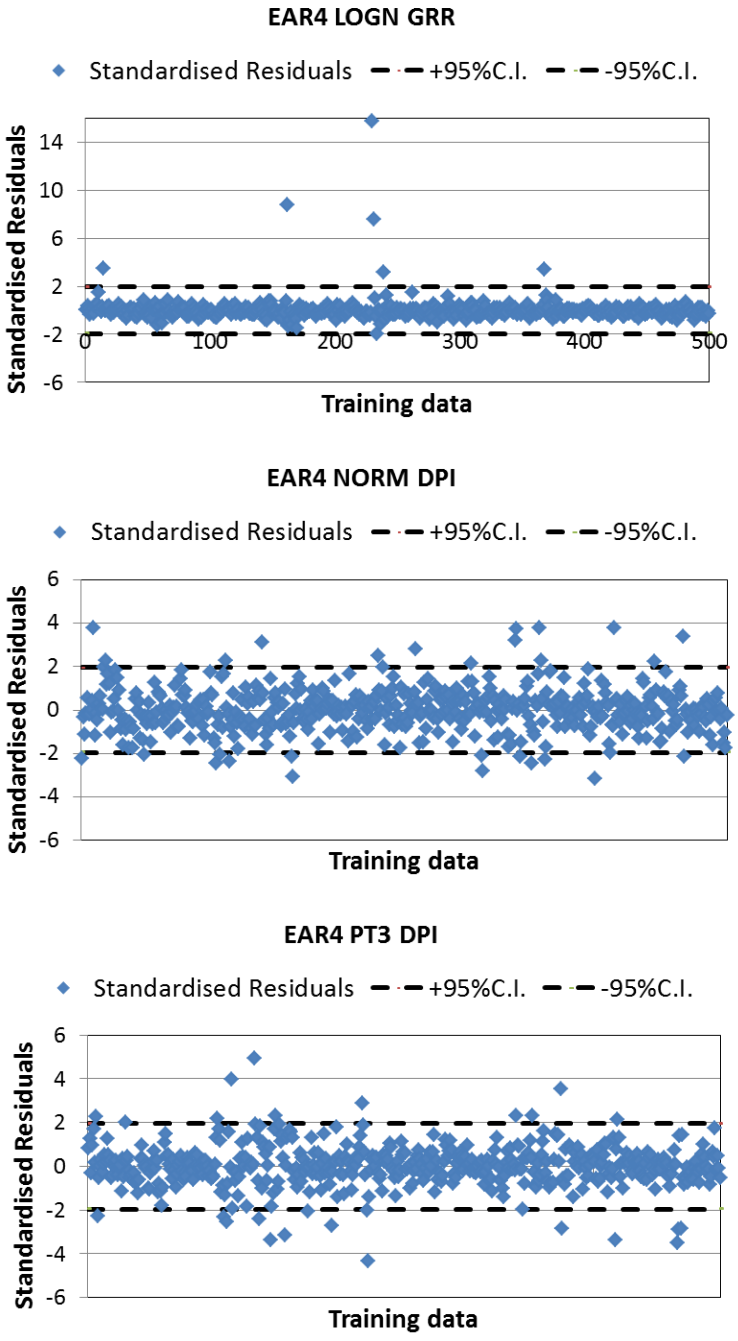


Figure A.2 *(Continued)*

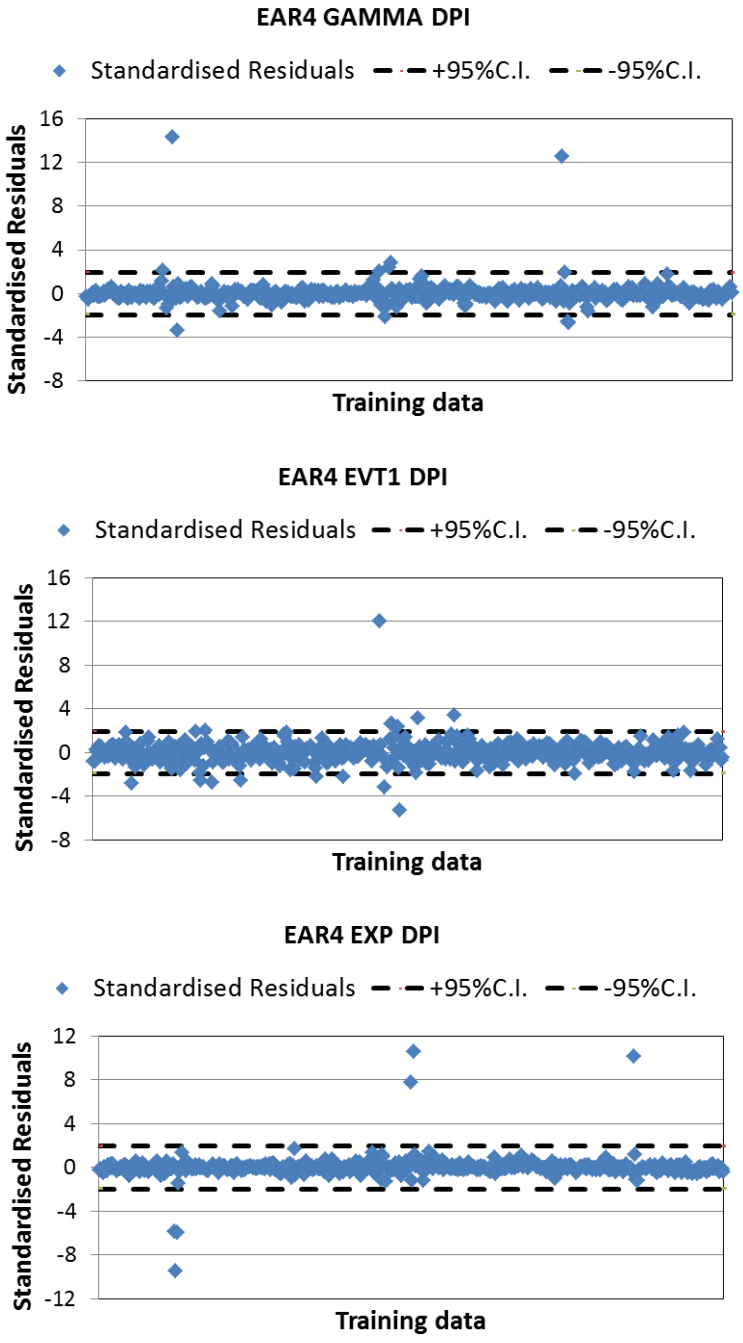


Figure A.2 *(Continued)*

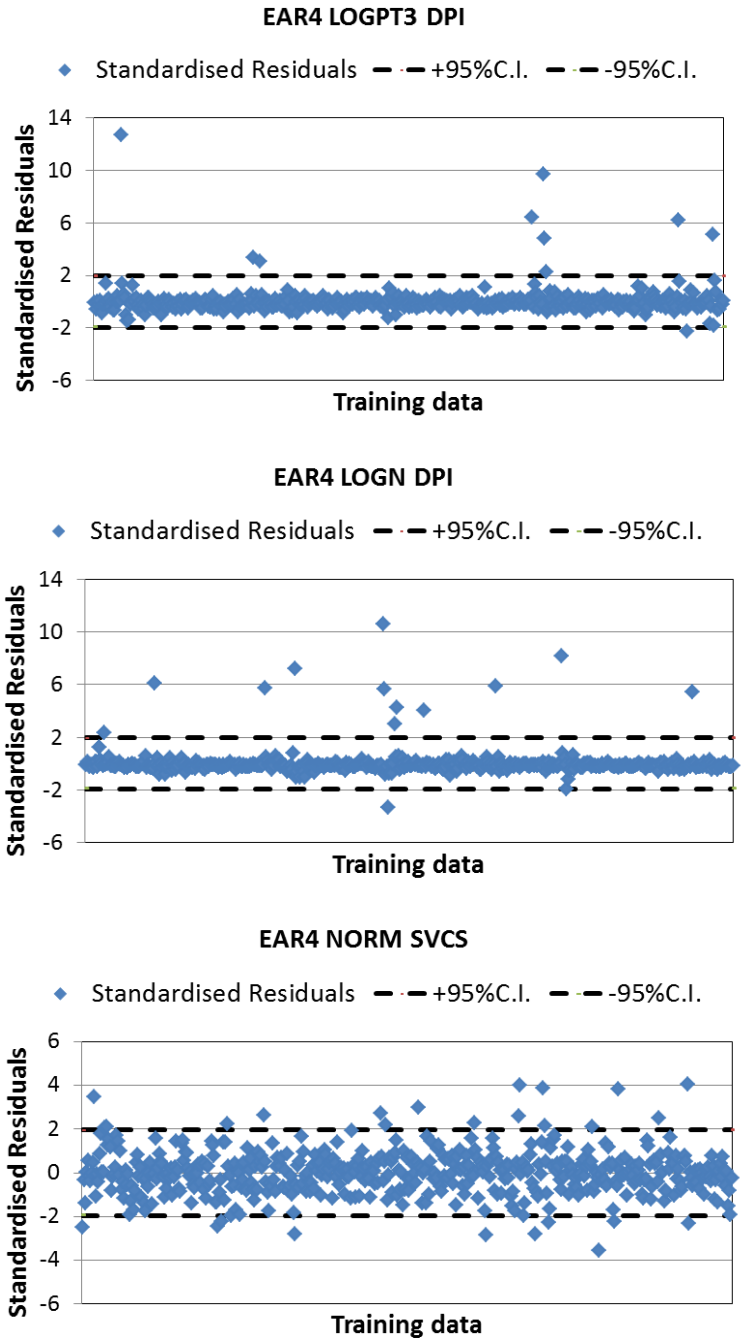


Figure A.2 *(Continued)*

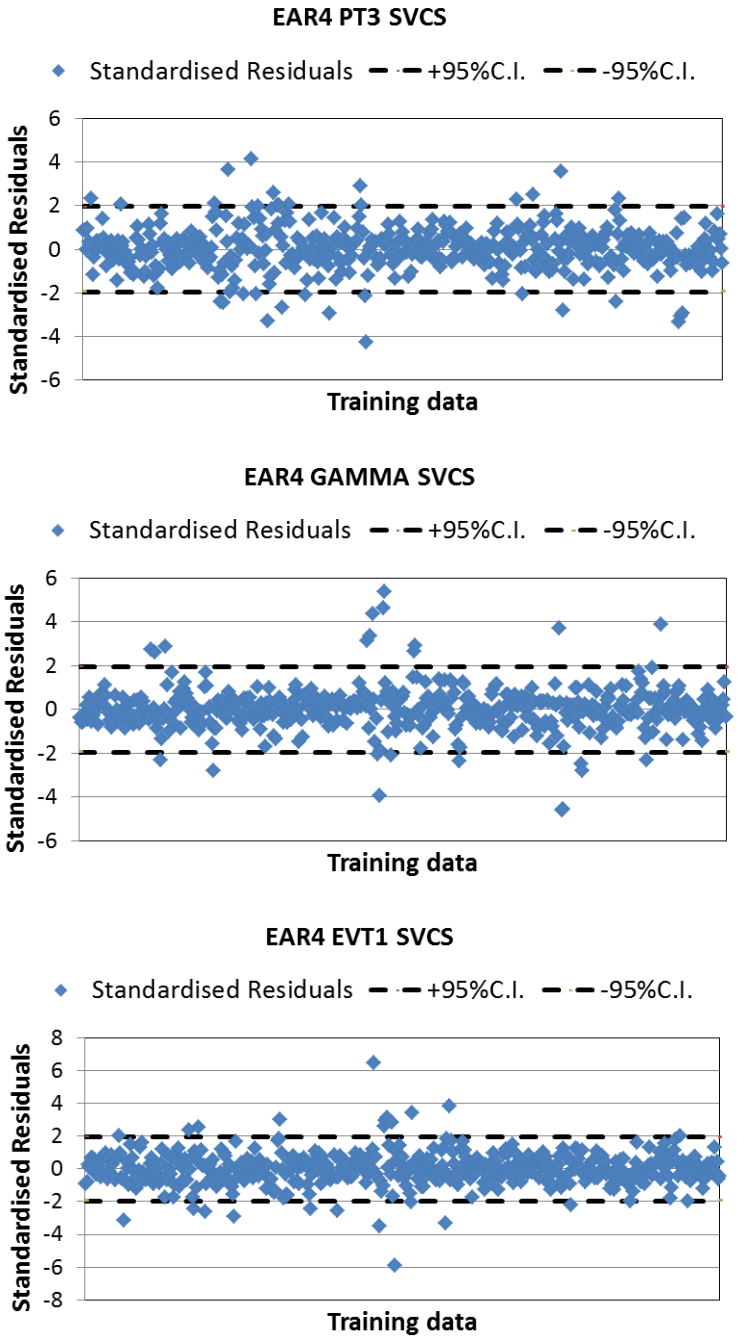


Figure A.2 *(Continued)*

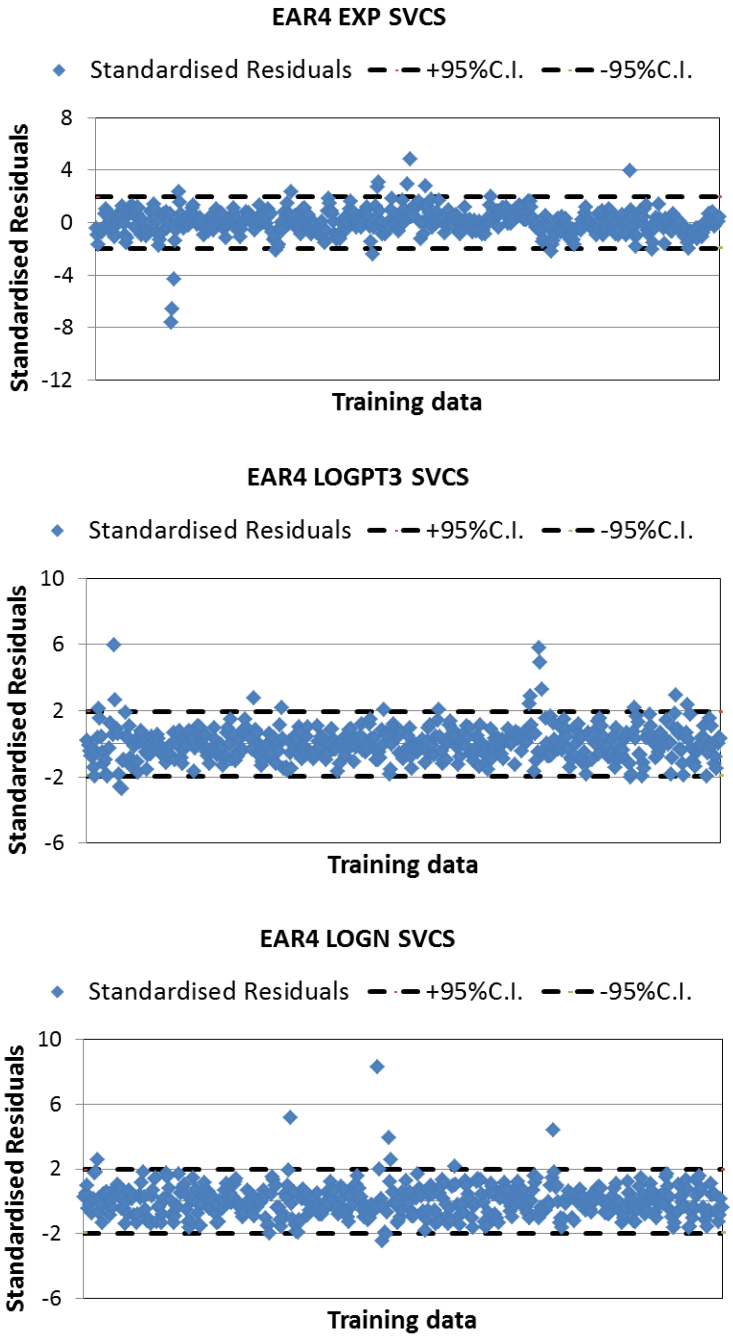


Figure A.2 *(Continued)*

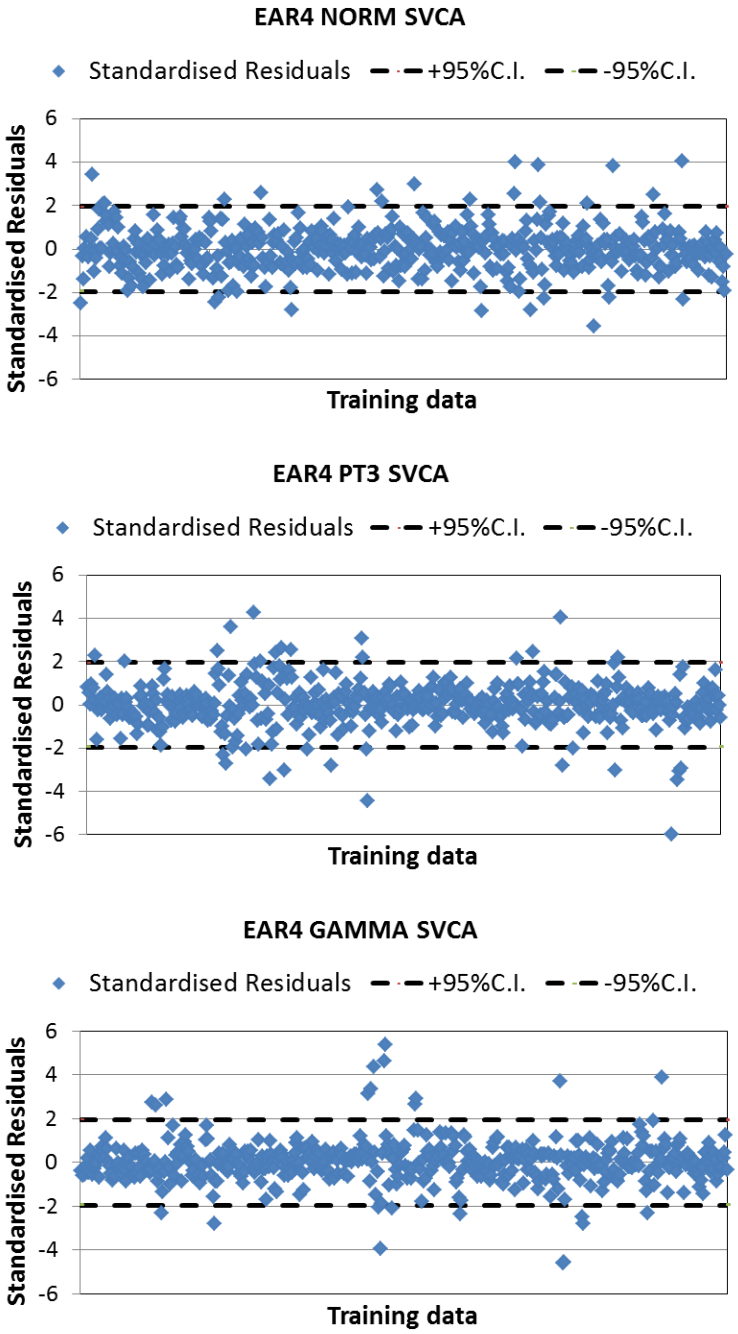


Figure A.2 *(Continued)*

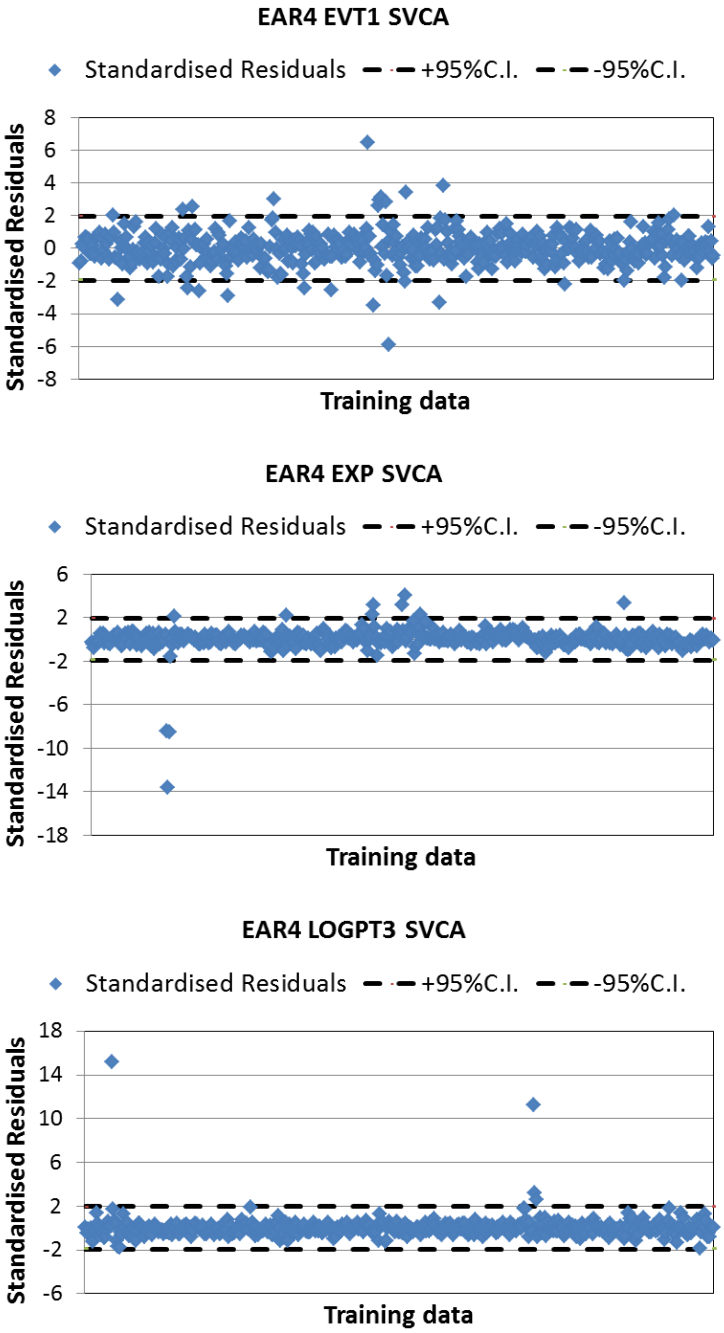


Figure A.2 *(Continued)*

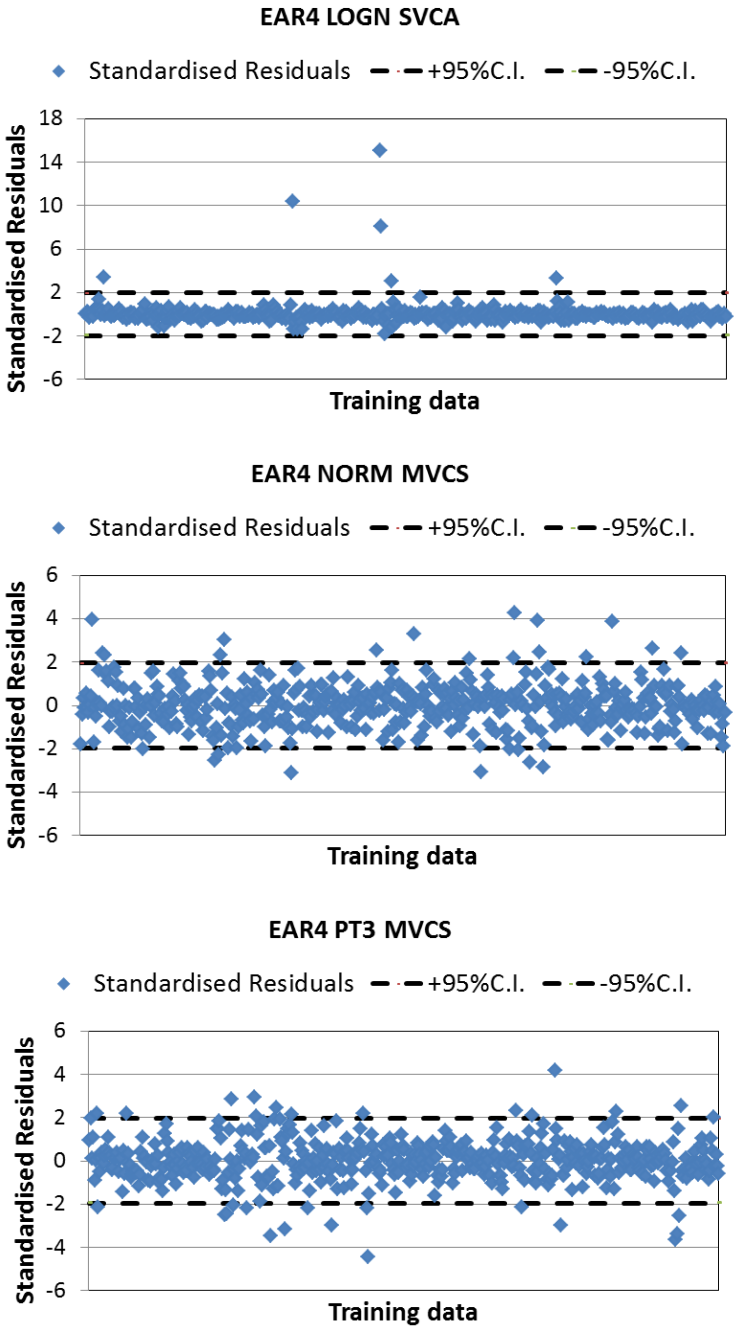


Figure A.2 *(Continued)*

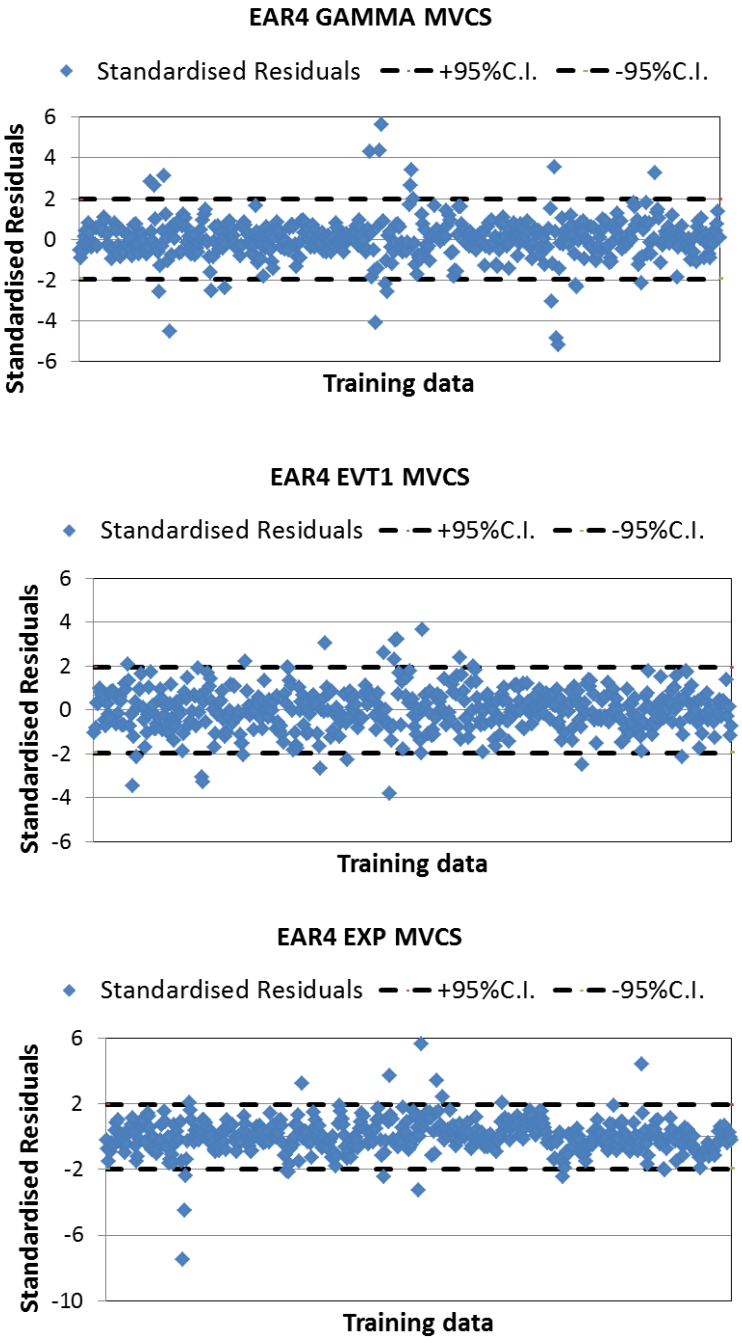
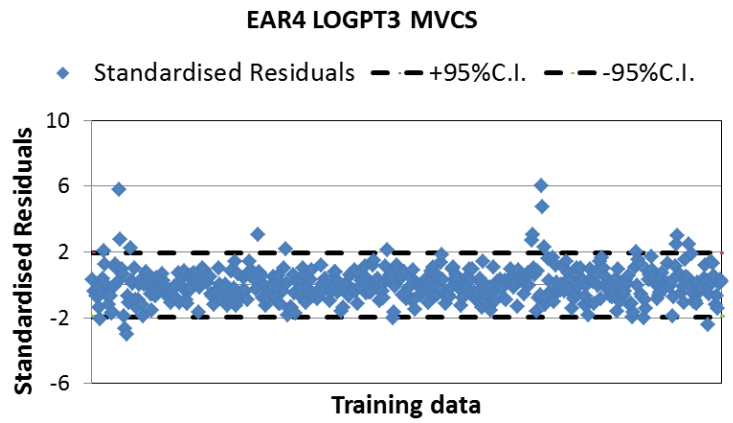
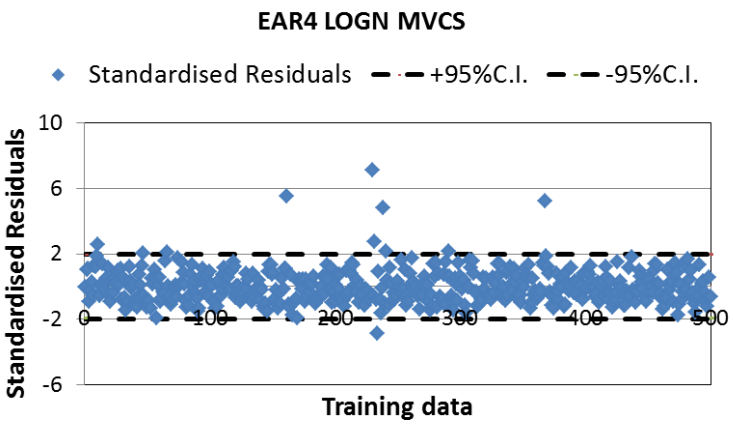


Figure A.2 *(Continued)*





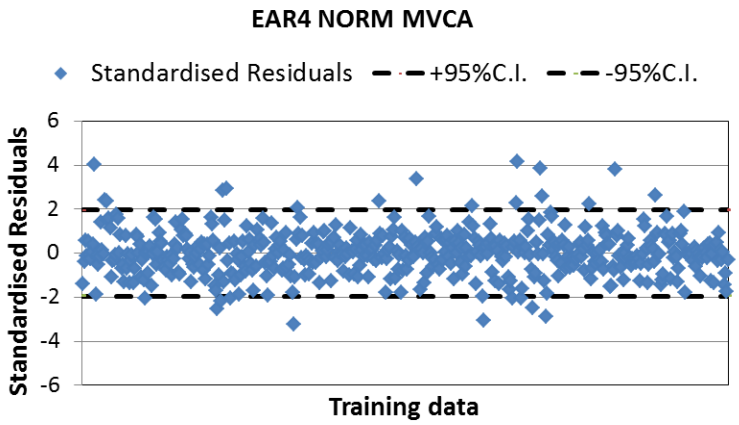


Figure A.2 *(Continued)*

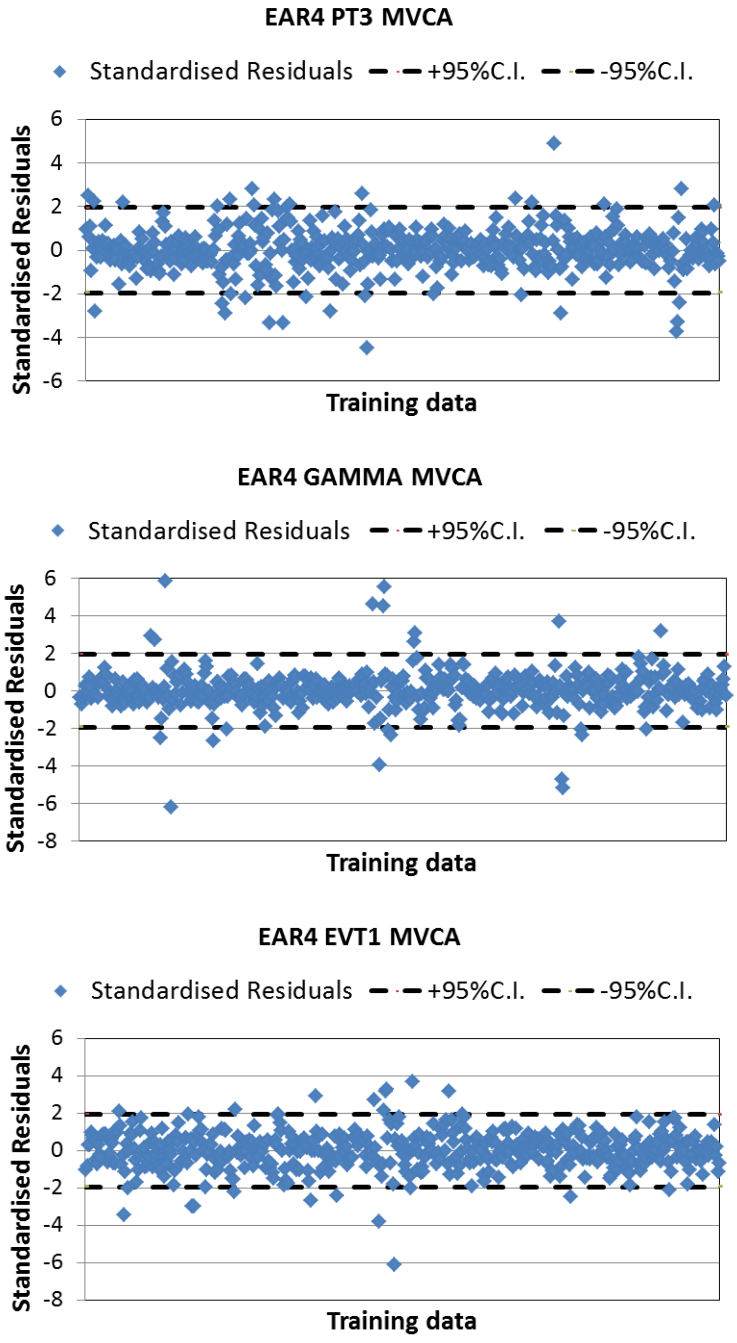


Figure A.2 *(Continued)*

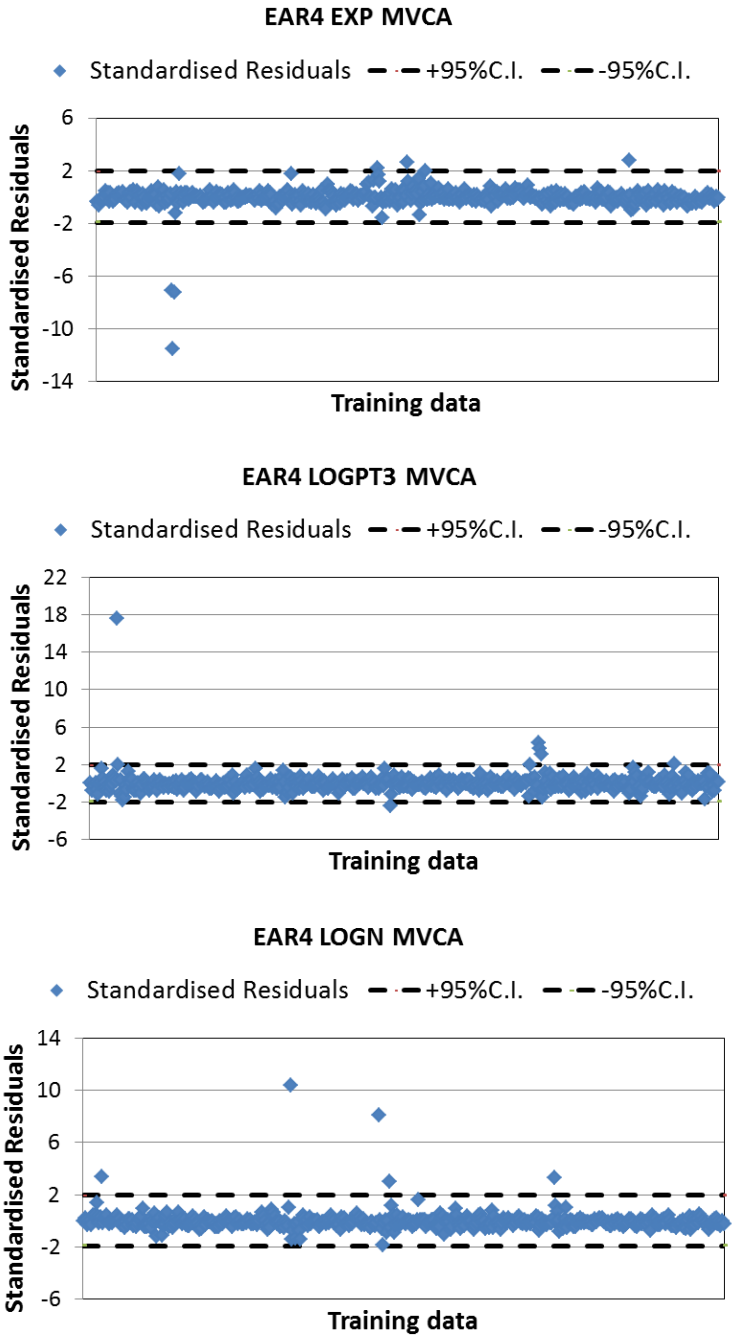


Figure A.2 *(Continued)*

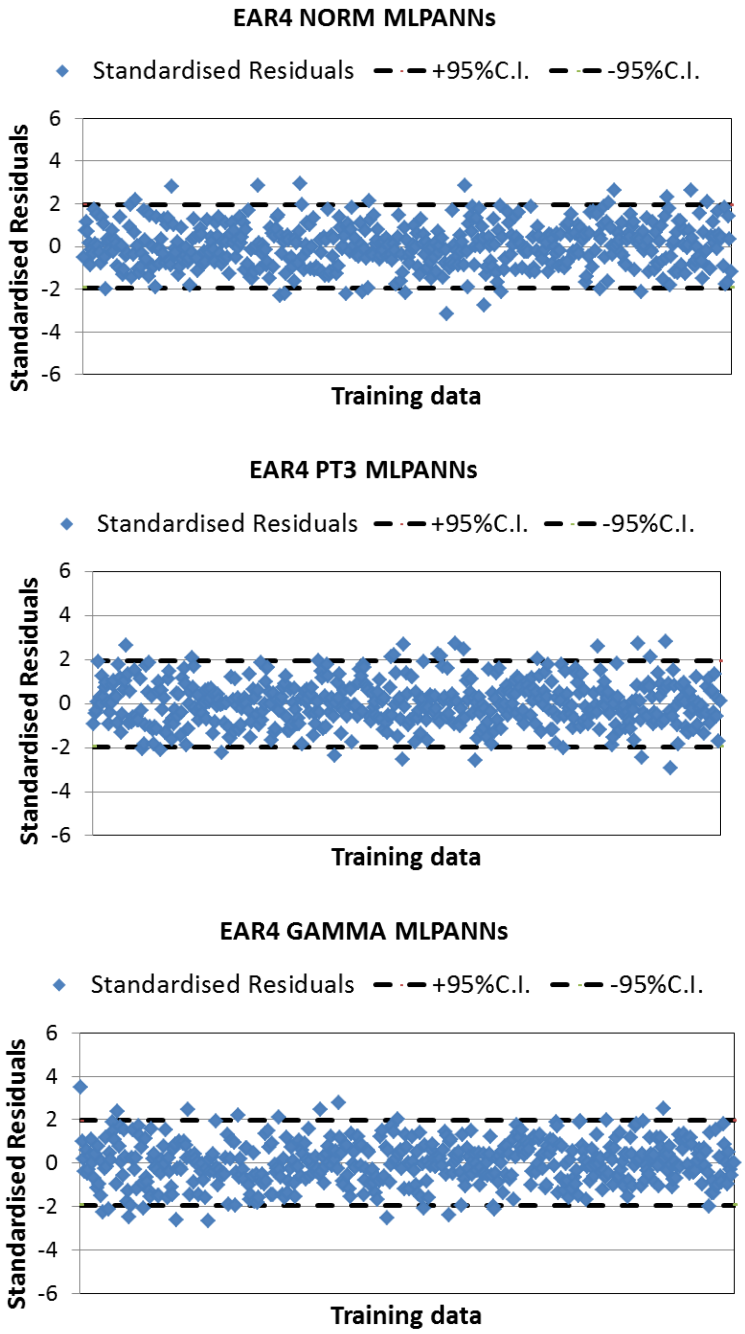


Figure A.2 *(Continued)*

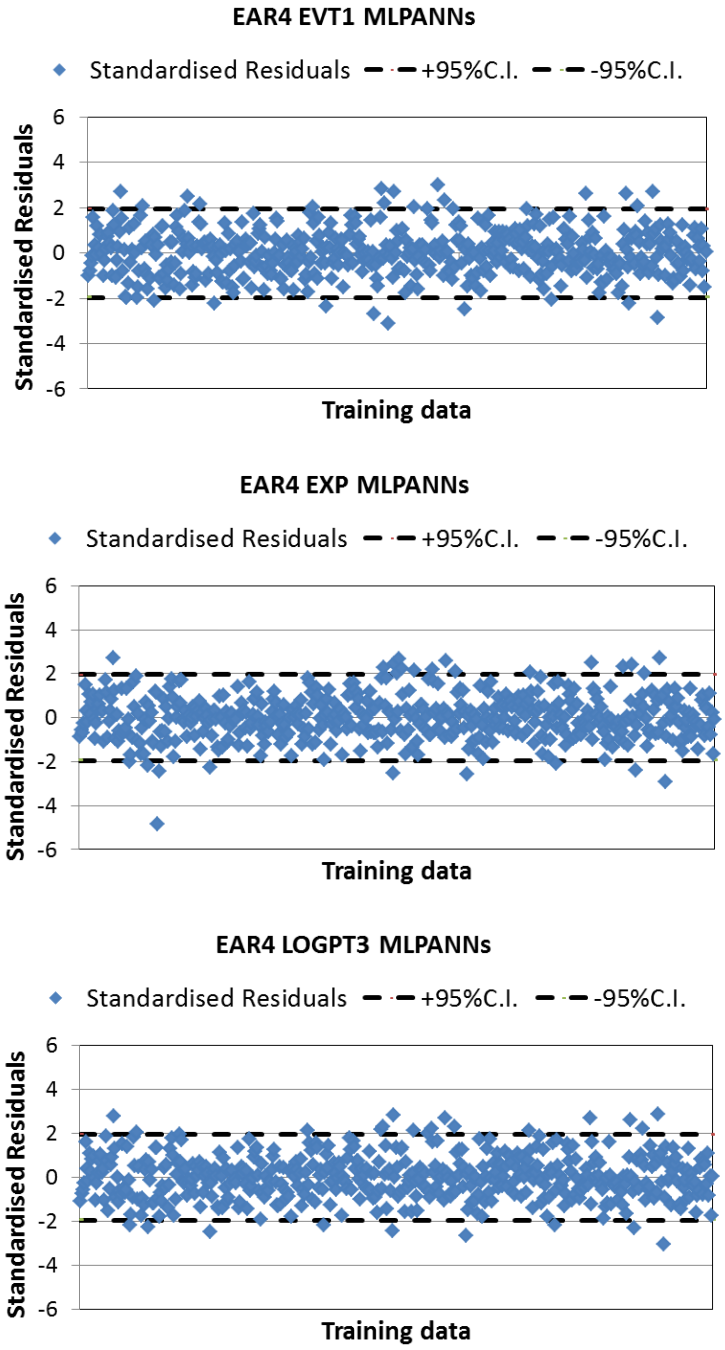


Figure A.2 *(Continued)*

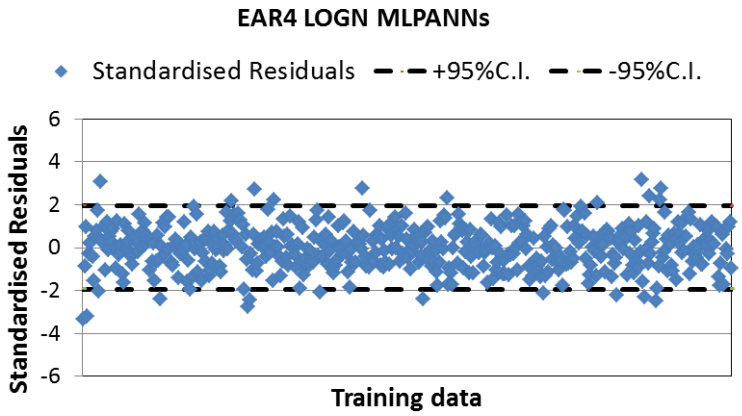


Figure A.2 *(Continued)*

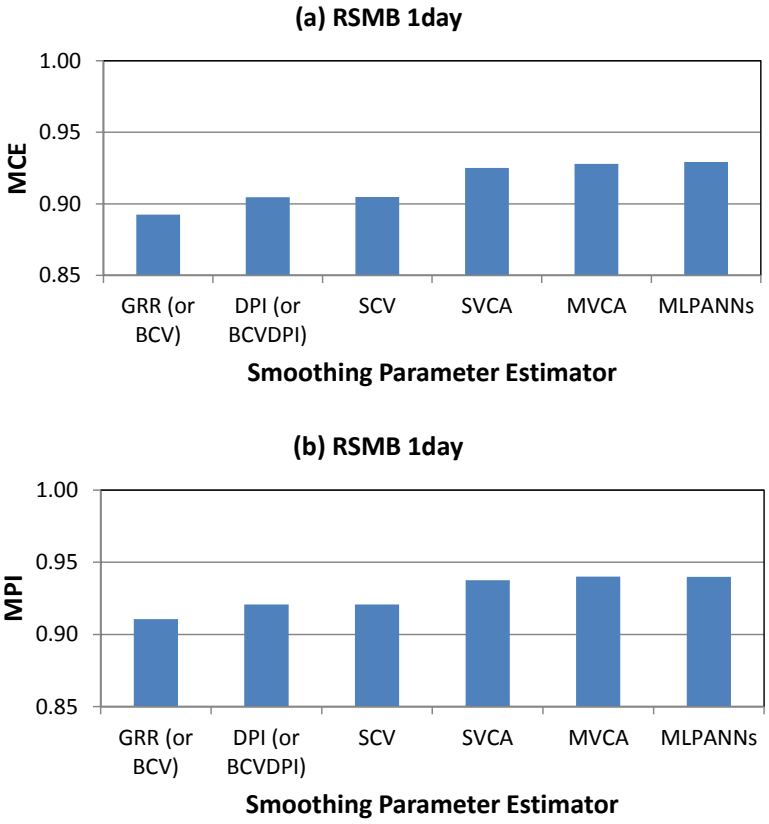


Figure A.3. Predictive accuracy for the validation data of MLPs and GRNNs with different smoothing parameters for river salinity at Murray Bridge 1 day in advance

(similar plots were also observed for 5 days & 14 days in advance)

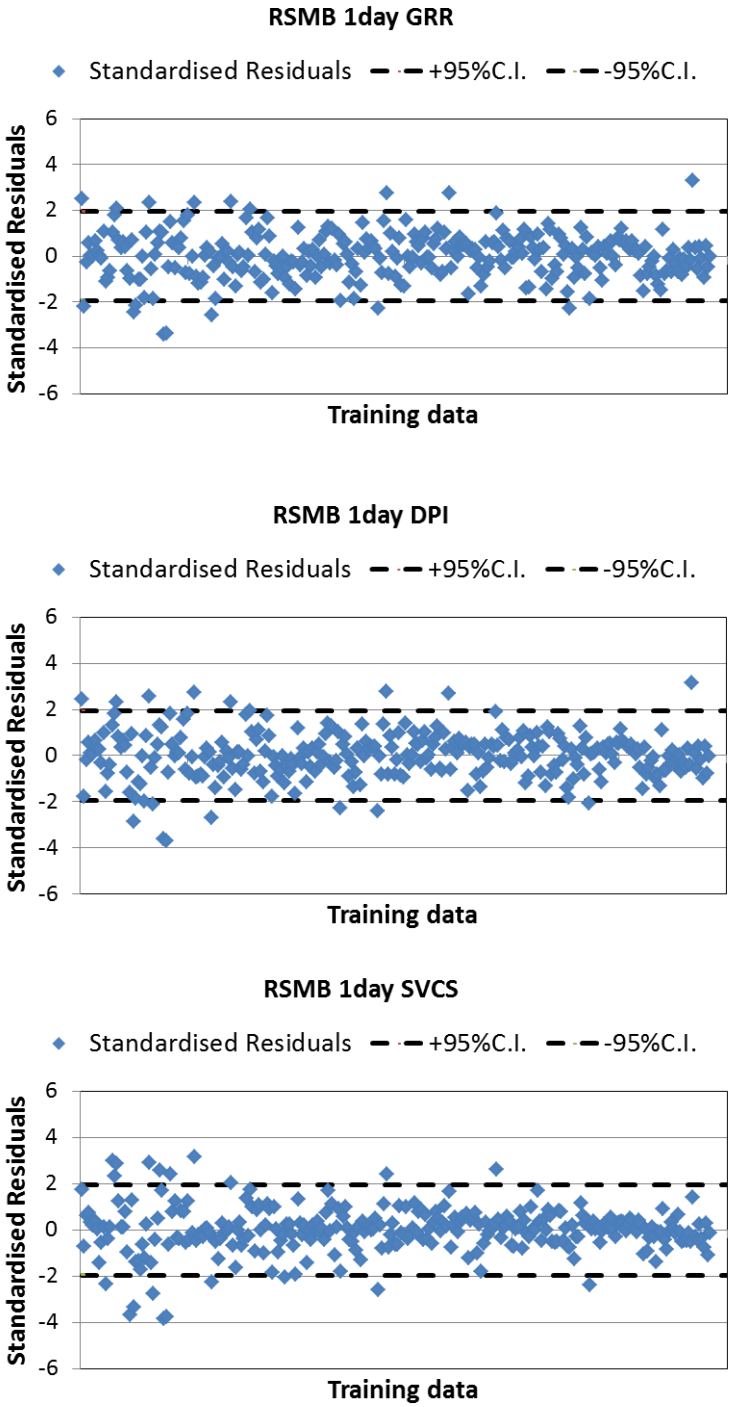
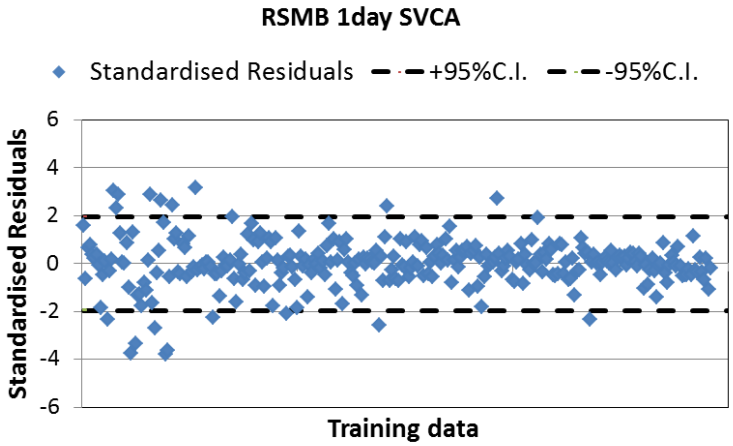
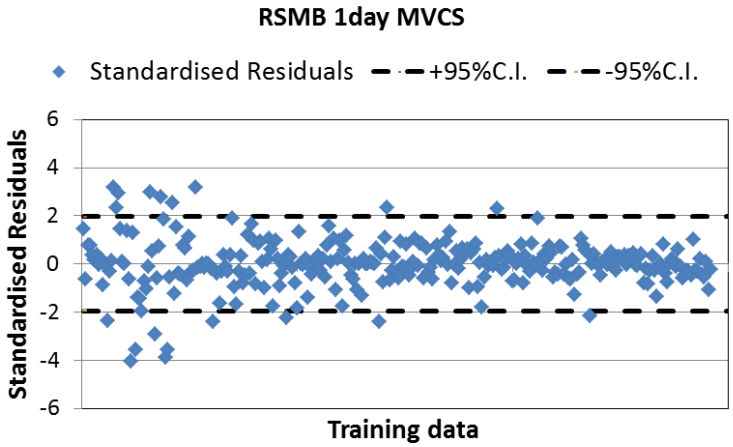


Figure A.4 Standardised residuals for the training data of MLPs and GRNNs with different smoothing parameters for river salinity at Murray Bridge 1 day in advance (plots of the BCV were similar to those of the GRR; plots of the BCVDPI and SCV were similar to those of the DPI; similar plots were also observed for 5 days & 14 days in advance**)**





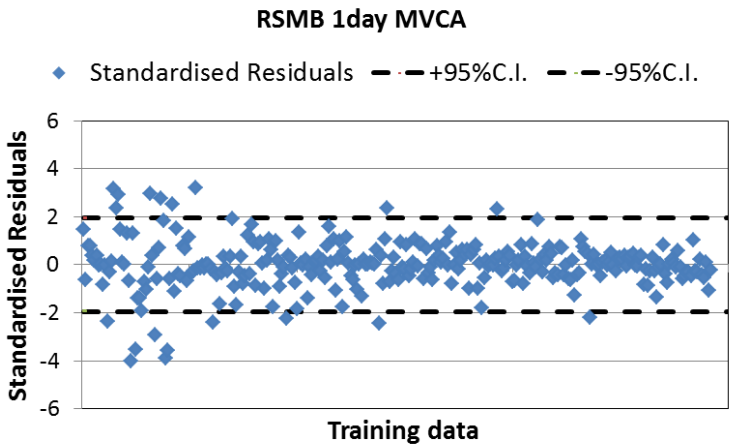


Figure A.4 *(Continued)*

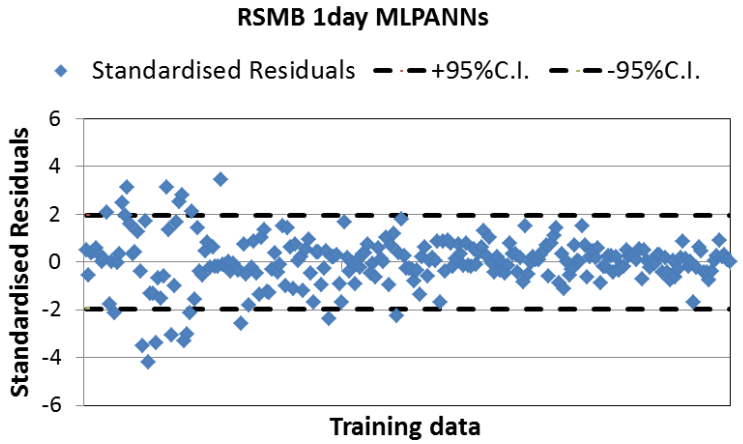
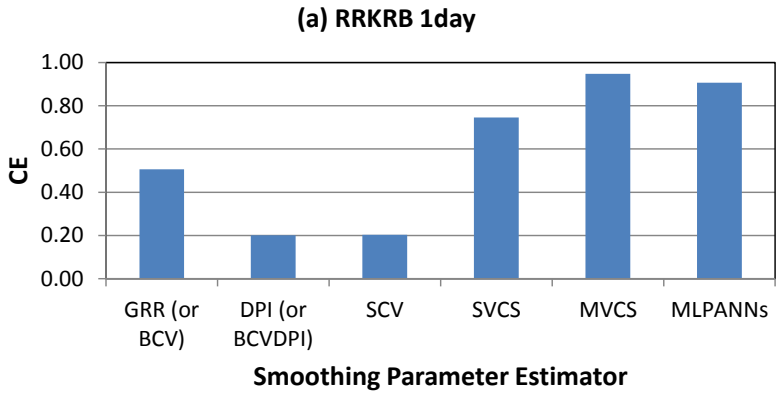


Figure A.4 *(Continued)*



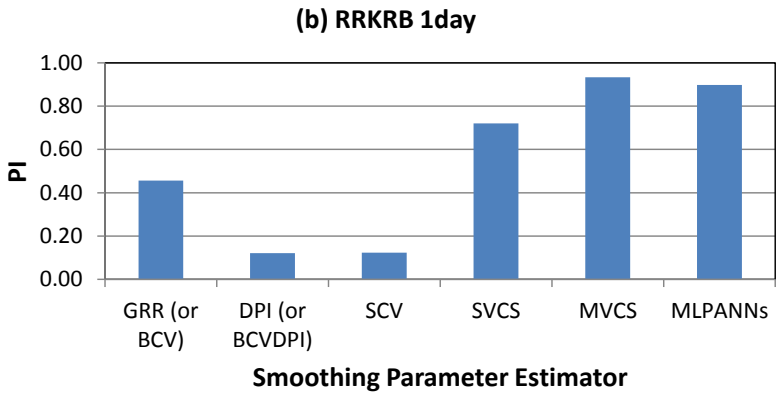


Figure A.5. Predictive accuracy for the validation data of MLPs and GRNNs with different smoothing parameters for runoff at Lock and Dam 10 in the Kentucky River basin1 day in advance

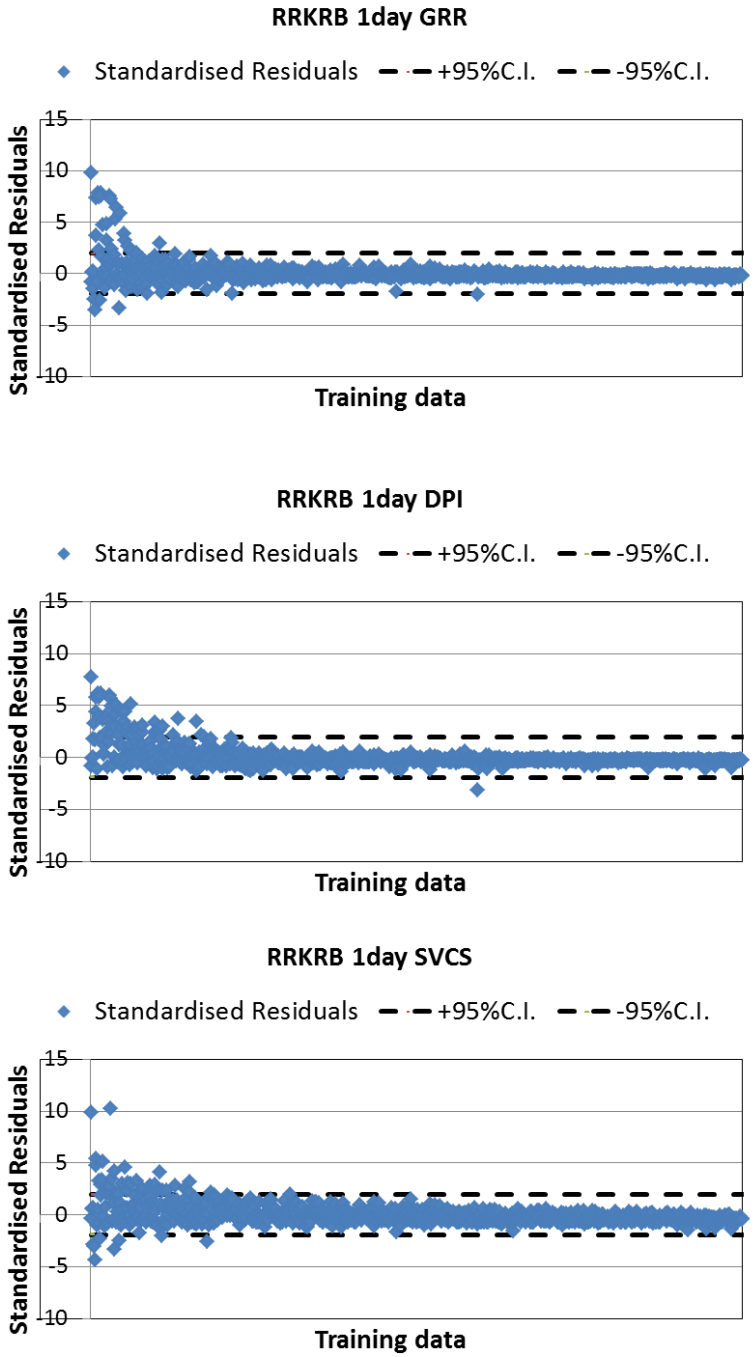


Fig. A.6 Standardised residuals for the training data of MLPs and GRNNs with different smoothing parameters for runoff at Lock and Dam 10 in the Kentucky River basin1 day in advance(plots of the BCV were similar to those of the GRR; plots of the BCVDPI and SCV were similar to those of the DPI**)**

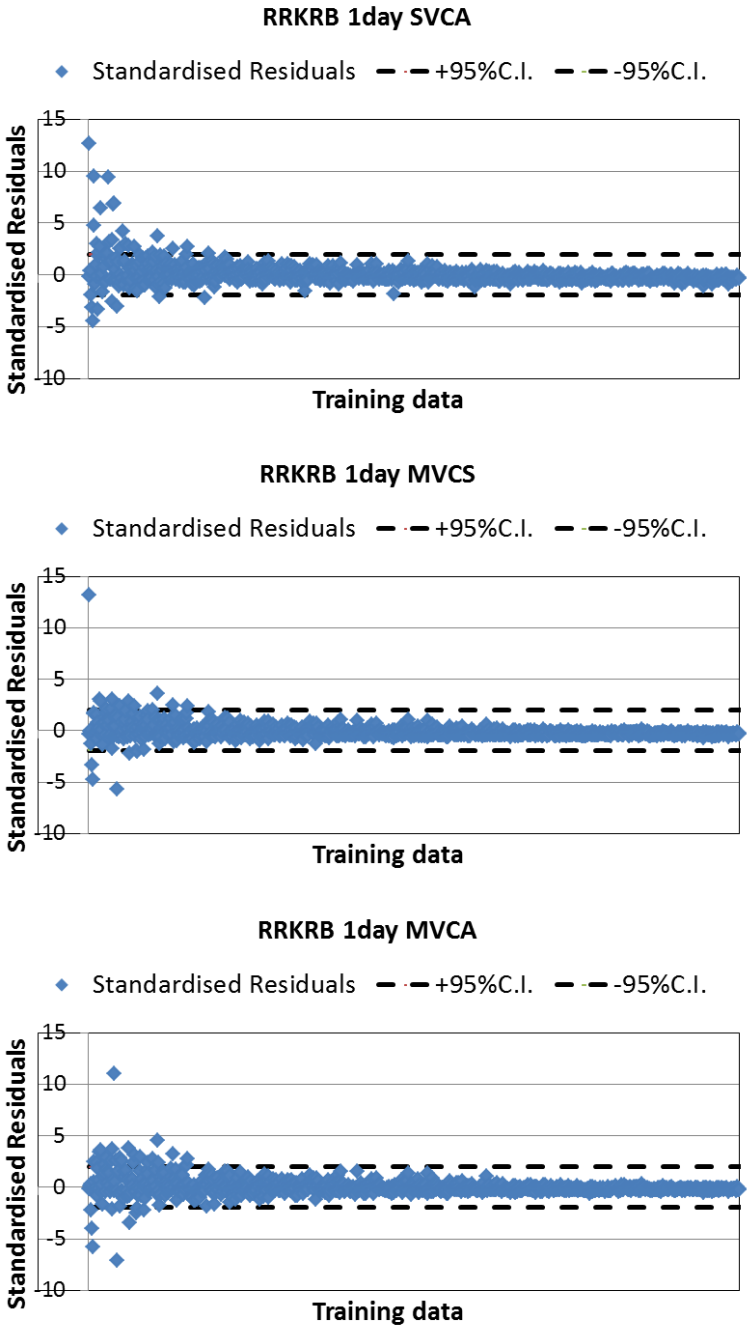


Figure A.6 *(Continued)*

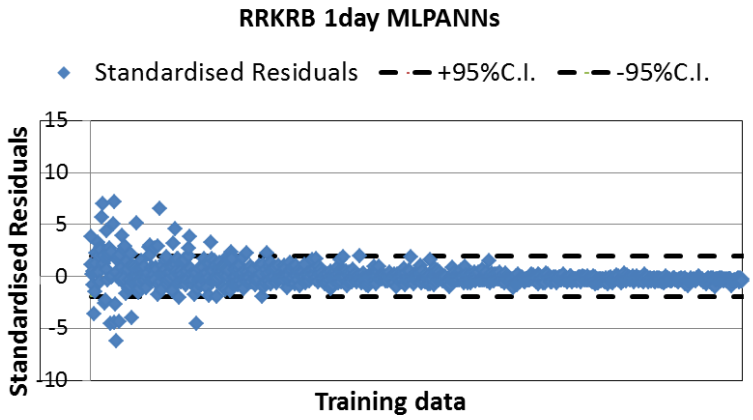


Figure A.6 *(Continued)*

APPENDIX-B Supplementary Material from Paper 2 (Chapter 3)

B.1 Mathematical derivations

Derivation of Gaussian reference rule

Let$f$ be the Gaussian density function $N(\mu, \sigma)$, K be the Gaussian kernel, and $h = \frac{R(K)}{\mu(K)^2 R(f'')}^{\frac{1}{5}} n^{\frac{-1}{5}}$ be the optimal bandwidth with respect to asymptotic mean integrated squared error (AMISE), then

$$f = \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-(x-\mu)^2}{2\sigma^2}}$$

$$f' = \frac{1}{\sigma\sqrt{2\pi}} \times \frac{-2(x-\mu)}{2\sigma^2} \times e^{\frac{-(x-\mu)^2}{2\sigma_K^2}}$$

$$= \frac{-(x-\mu)}{\sigma^3\sqrt{2\pi}} e^{\frac{-(x-\mu)^2}{2\sigma^2}}$$

$$f'' = \frac{-1}{\sigma^3\sqrt{2\pi}} e^{\frac{-(x-\mu)^2}{2\sigma^2}} + \frac{-2(x-\mu)^2}{2\sigma^5\sqrt{2\pi}} e^{\frac{-(x-\mu)^2}{2\sigma^2}}$$

$$= \frac{-1}{\sigma^3\sqrt{2\pi}} e^{\frac{-(x-\mu)^2}{2\sigma^2}} \times \left[\frac{-(x-\mu)^2}{\sigma^2} + 1\right]$$

$$(f'')^2 = \frac{1}{\sigma^6 2\pi} e^{\frac{-(x-\mu)^2}{\sigma^2}} \times \left[\frac{(x-\mu)^4}{\sigma^4} - \frac{2(x-\mu)^2}{\sigma^2} + 1\right]$$

Let$\frac{y}{\sqrt{2}} = x$, then $\frac{dx}{\sqrt{2}} = dy$

$$\int (f'')^2\, dx = \int \frac{1}{\sigma^6 2\pi} e^{\frac{-(x-\mu)^2}{\sigma^2}} \times \left[\frac{(x-\mu)^4}{\sigma^4} - \frac{2(x-\mu)^2}{\sigma^2} + 1\right] dx$$

$$= \int \frac{1}{\sigma^6 2\pi} e^{\frac{-\left(\frac{y}{\sqrt{2}}-\mu\right)^2}{\sigma^2}} \times \left[\frac{(\frac{y}{\sqrt{2}}-\mu)^4}{\sigma^4} - \frac{2(\frac{y}{\sqrt{2}}-\mu)^2}{\sigma^2} + 1\right] \times \frac{dy}{\sqrt{2}}$$

$$= \int \frac{1}{\sigma^6 2\pi} e^{\frac{-(y-\mu\sqrt{2})^2}{2\sigma^2}} \times \left[\frac{(y-\mu\sqrt{2})^4}{4\sigma^4} - \frac{(y-\mu\sqrt{2})^2}{\sigma^2} + 1\right] \times \frac{dy}{\sqrt{2}}$$

$$= \int \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-(y-\mu\sqrt{2})^2}{2\sigma^2}} \times \left[\frac{(y-\mu\sqrt{2})^4}{4\sigma^4} - \frac{(y-\mu\sqrt{2})^2}{\sigma^2} + 1\right] \times \frac{1}{\sigma^5\sqrt{2\pi}} \times \frac{dy}{\sqrt{2}}$$

for Gaussian distribution, $E(x-\mu)^p \begin{cases} 0, if\ p\ is\ odd \\ \sigma^p(p-1)!!, if\ p\ is\ even \end{cases}$

$$\int (f'')^2\, dx = \frac{1}{2\sigma^5\sqrt{\pi}} \times \left(\frac{\sigma^4 \times 3!!}{4\sigma^4} - \frac{\sigma^2 \times 1!!}{\sigma^2} + 1\right)$$

$$\int (f'')^2\, dx = \frac{1}{2\sigma^5\sqrt{\pi}} \times \left(\frac{3}{4} - 1 + 1\right)$$

$$\int (f'')^2\, dx = \frac{3}{8\sigma^5\sqrt{\pi}}$$

$$K = \frac{1}{\sigma(K)\sqrt{2\pi}} e^{\frac{-[x-\mu(K)]^2}{2\sigma(K)^2}}$$

$$K^2 = \frac{1}{\sigma(K)^2 2\pi} e^{\frac{-[x-\mu(K)]^2}{\sigma(K)^2}}$$

$$\int K^2\, dx = \int \frac{1}{\sigma(K)\sqrt{2\pi}} e^{\frac{-[x-\mu(K)]^2}{\sigma(K)^2}} \times \frac{1}{\sigma(K)\sqrt{2\pi}} dx$$

$$= \int \frac{1}{\sigma(K)\sqrt{2\pi}} e^{\frac{-\left[\frac{y}{\sqrt{2}}-\mu(K)\right]^2}{\sigma(K)^2}} \times \frac{1}{\sigma(K)\sqrt{2\pi}} \frac{dy}{\sqrt{2}}$$

$$= \frac{1}{\sigma(K)2\sqrt{\pi}} \times \frac{\sigma(K)^2 \times 1!!}{\sigma(K)^2}$$

$$= \frac{1}{\sigma(K)2\sqrt{\pi}}$$

for standard normal distribution $\sigma(K) = 1$, $\mu_2(K) = \int x^2 K(x)dx \approx \sigma(K)^2 = 1$, then

$$\hat{h}_{GRR,i} = \frac{R(K)^{\frac{1}{5}}}{R(f'')} n^{\frac{-1}{5}}$$

$$\hat{h}_{GRR,i} = \frac{\int K^2\, dx^{\frac{1}{5}}}{\int (f'')^2\, dx} n^{\frac{-1}{5}}$$

$$\hat{h}_{GRR,i} = \frac{\frac{1}{2\sqrt{\pi}}^{\frac{1}{5}}}{\frac{3}{8\sigma^5\sqrt{\pi}}} n^{\frac{-1}{5}}$$

$$\hat{h}_{GRR,i} = \left(\frac{3}{4}\right)^{\frac{1}{5}} \sigma n^{\frac{-1}{5}}$$

which results in Eqs. (3.15). This also consists with Wand and Jones (1995) and Scott (1992).

Derivation of 2-stage direct plug-in

Let $\hat{\varphi}_r = \frac{(-1)^{r/2} r!}{(2\sigma)^{r+1}(r/2)!\pi^{1/2}}$ be the normal scale (NS), σ be the standard deviation of the sample, then

$$\hat{\varphi}_8^{NS} = \frac{(-1)^4 8!}{(2\sigma)^9 4!\, \pi^{1/2}}$$

$$= \frac{2^7 \times 3^2 \times 5 \times 7}{2^{12} \times 3 \times \sigma^9 \times \pi^{1/2}}$$

$$= \frac{105}{32\sigma^9 \pi^{1/2}}$$

Let K be the Gaussian kernel with $\mu_2(K) = \int x^2 K(x)dx = n^{-1} \sum_{j=1}^{n} (X_i^j - X_i)^2 K_h\left(X_i^j - X_i\right) = 1$ for Gaussian kernel, $K^{(n)}$ be the nth derivative of Kthen,

$$K = \frac{1}{\sigma(K)\sqrt{2\pi}} e^{\frac{-[x-\mu(K)]^2}{2\sigma(K)^2}}$$

$$K^{(1)} = \frac{1}{\sigma(K)^3\sqrt{2\pi}} e^{\frac{-[x-\mu(K)]^2}{2\sigma(K)^2}} \times [\mu(K) - x]$$

$$K^{(2)} = \frac{1}{\sigma(K)^3\sqrt{2\pi}} \left\{ -e^{\frac{-[x-\mu(K)]^2}{2\sigma(K)^2}} + \frac{[x-\mu(K)]^2}{\sigma(K)^2} e^{\frac{-[x-\mu(K)]^2}{2\sigma(K)^2}} \right\}$$

$$K^{(3)} = \frac{1}{\sigma(K)^3\sqrt{2\pi}} \left\{ \frac{3[x-\mu(K)]}{\sigma(K)^2} e^{\frac{-[x-\mu(K)]^2}{2\sigma(K)^2}} - \frac{[x-\mu(K)]^3}{\sigma(K)^4} e^{\frac{-[x-\mu(K)]^2}{2\sigma(K)^2}} \right\}$$

$$K^{(4)} = \frac{1}{\sigma(K)^3\sqrt{2\pi}} \left\{ \frac{3}{\sigma(K)^2} e^{\frac{-[x-\mu(K)]^2}{2\sigma(K)^2}} - \frac{6[x-\mu(K)]^2}{\sigma(K)^4} e^{\frac{-[x-\mu(K)]^2}{2\sigma(K)^2}} + \frac{[x-\mu(K)]^4}{\sigma(K)^6} e^{\frac{-[x-\mu(K)]^2}{2\sigma(K)^2}} \right\}$$

$$K^{(5)} = \frac{1}{\sigma(K)^3\sqrt{2\pi}} \left\{ \frac{-15}{\sigma(K)^4} e^{\frac{-[x-\mu(K)]^2}{2\sigma(K)^2}} + \frac{10[x-\mu(K)]^3}{\sigma(K)^6} e^{\frac{-[x-\mu(K)]^2}{2\sigma(K)^2}} - \frac{[x-\mu(K)]^5}{\sigma(K)^8} e^{\frac{-[x-\mu(K)]^2}{2\sigma(K)^2}} \right\}$$

$$K^{(6)} = \frac{1}{\sigma(K)^3\sqrt{2\pi}} \left\{ \frac{-15}{\sigma(K)^4} e^{\frac{-[x-\mu(K)]^2}{2\sigma(K)^2}} + \frac{45[x-\mu(K)]^2}{\sigma(K)^6} e^{\frac{-[x-\mu(K)]^2}{2\sigma(K)^2}} - \frac{15[x-\mu(K)]^4}{\sigma(K)^8} e^{\frac{-[x-\mu(K)]^2}{2\sigma(K)^2}} + \frac{[x-\mu(K)]^6}{\sigma(K)^{10}} e^{\frac{-[x-\mu(K)]^2}{2\sigma(K)^2}} \right\}$$

For standard normal case, $\sigma(K) = 1$

$$K^{(6)}(0) = \frac{-15}{\sqrt{2\pi}}$$

$$K^{(4)}(0) = \frac{3}{\sqrt{2\pi}}$$

recall Eq. (3.18)

$$g_1 = \left[-2K^{(6)}(0) / \{\mu_2(K)\hat{\varphi}_8^{NS} n\} \right]^{1/9}$$

$$g_1 = \left[-2 \times \frac{-15}{\sqrt{2\pi}} / \left\{\frac{105}{32\sigma^9 \pi^{1/2}} n\right\}\right]^{1/9}$$

Let $\hat{\varphi}_r(g) = n^{-1} \sum_{a=1}^{n} \hat{f}^{(r)}(X_a; g) = n^{-2} \sum_{i=1}^{m} \sum_{j=1}^{n} L_g^{(r)}(X_i^j - X_i)$ be the general case and $L = K$, then

$$\hat{\varphi}_6(g_1) = n^{-2} \sum_{i=1}^{m} \sum_{j=1}^{n} L_{g_1}^{(6)}(X_i^j - X_i)$$

$$= n^{-2} \sum_{i=1}^{m} \sum_{j=1}^{n} \frac{1}{g_1 \sqrt{2\pi}} e^{\frac{-(X_i^j - X_i)^2}{2g_1^2}}$$

$$g_2 = \left[-2K^{(4)}(0) / \{\mu_2(K) \hat{\varphi}_6(g_1) n\}\right]^{1/7}$$

$$g_2 = \left[-2 \times \frac{3}{\sqrt{2\pi}} / \left\{n^{-1} \sum_{i=1}^{m} \sum_{j=1}^{n} \frac{1}{g_1 \sqrt{2\pi}} e^{\frac{-(X_i^j - X_i)^2}{2g_1^2}}\right\}\right]^{1/7}$$

$$\hat{\varphi}_4(g_2) = n^{-2} \sum_{i=1}^{m} \sum_{j=1}^{n} L_{g_2}^{(4)}(X_i^j - X_i)$$

$$= n^{-2} \sum_{i=1}^{m} \sum_{j=1}^{n} \frac{1}{g_2 \sqrt{2\pi}} e^{\frac{-(X_i^j - X_i)^2}{2g_2^2}}$$

recall Eq. (3.17)

$$\hat{h}_{DPI,i} = \left[\frac{R(K)}{[\mu_2(K)]^2 \hat{\varphi}_4(g) n}\right]^{\frac{1}{5}}$$

$$\hat{h}_{DPI,i} = \left[\frac{\frac{1}{2\sqrt{\pi}}}{\left\{n^{-1} \sum_{i=1}^{m} \sum_{j=1}^{n} \frac{1}{g_2 \sqrt{2\pi}} e^{\frac{-(X_i^j - X_i)^2}{2g_2^2}}\right\}}\right]^{\frac{1}{5}}$$

Derivation of the linkage in between GRR and DPI

Recall $\hat{\varphi}_r = \frac{(-1)^{r/2} r!}{(2\sigma)^{r+1}(r/2)!\pi^{1/2}}$, $h = \frac{R(K)}{\mu(K)^2 R(f'')}^{\frac{1}{5}} n^{\frac{-1}{5}}$, $R(K) = \frac{1}{2\sqrt{\pi}}$, $R(f'') = \frac{3}{8\sigma^5\sqrt{\pi}}$ and $\mu_2(K) = 1$ for standard normal case then

$$\hat{\varphi}_4^{NS} \approx R(f'')$$

$$= \frac{(-1)^2 4!}{(2\sigma)^5 2!\, \pi^{1/2}}$$

$$= \frac{2^3 \times 3}{2^6 \sigma^5 \pi^{1/2}}$$

$$= \frac{3}{8\sigma^5 \pi^{1/2}}$$

For $r = 0$,

$$\hat{h}_{DPI,i} = \left[\frac{\frac{1}{2\sqrt{\pi}}}{1^2 \frac{3}{8\sigma^5 \pi^{1/2}} n} \right]^{\frac{1}{5}}$$

$$= \left(\frac{3}{4}\right)^{\frac{1}{5}} \sigma n^{\frac{-1}{5}}$$

$$= \hat{h}_{GRR,i}$$

hence, GRR is equivalent to 0-stage DPI, which is a special case in the DPI family.

Derivation of smoothed cross validation

Let $\hat{\varphi}_r = \frac{(-1)^{r/2} r!}{(2\sigma)^{r+1}(r/2)!\pi^{1/2}}$ be the normal scale (NS), σ be the standard deviation of the sample, then

$$\hat{\varphi}_{12}^{NS} = \frac{(-1)^6 12!}{(2\sigma)^{13} 6!\, \pi^{1/2}}$$

$$= \frac{2^{10} \times 3^5 \times 5^2 \times 7 \times 11}{2^{17} \times 3^2 \times 5 \times \sigma^{13} \times \pi^{1/2}}$$

$$= \frac{945 \times 11}{2^7 \sigma^{13} \pi^{1/2}}$$

$$\hat{\varphi}_8^{NS} = \frac{(-1)^4 8!}{(2\sigma)^9 4!\, \pi^{1/2}}$$

$$= \frac{2^7 \times 3^2 \times 5 \times 7}{2^{12} \times 3 \times \sigma^9 \times \pi^{1/2}}$$

$$= \frac{105}{32\sigma^9 \pi^{1/2}}$$

Let K be the Gaussian kernel and $\mu_2(K) = \int x^2 K(x) dx = n^{-1} \sum_{j=1}^{n} (X_i^j - X_i)^2 K_h\left(X_i^j - X_i\right) = 1$ for Gaussian kernel, then

$$K = \frac{1}{\sigma(K)\sqrt{2\pi}} e^{\frac{-[x-\mu(K)]^2}{2\sigma(K)^2}}$$

$$K^{(1)} = \frac{1}{\sigma(K)^3\sqrt{2\pi}} e^{\frac{-[x-\mu(K)]^2}{2\sigma(K)^2}} \times [\mu(K) - x]$$

$$K^{(2)} = \frac{1}{\sigma(K)^3\sqrt{2\pi}} \left\{ -e^{\frac{-[x-\mu(K)]^2}{2\sigma(K)^2}} + \frac{[x-\mu(K)]^2}{\sigma(K)^2} e^{\frac{-[x-\mu(K)]^2}{2\sigma(K)^2}} \right\}$$

$$K^{(3)} = \frac{1}{\sigma(K)^3\sqrt{2\pi}} \left\{ \frac{3[x-\mu(K)]}{\sigma(K)^2} e^{\frac{-[x-\mu(K)]^2}{2\sigma(K)^2}} - \frac{[x-\mu(K)]^3}{\sigma(K)^4} e^{\frac{-[x-\mu(K)]^2}{2\sigma(K)^2}} \right\}$$

$$K^{(4)} = \frac{1}{\sigma(K)^3\sqrt{2\pi}} \left\{ \frac{3}{\sigma(K)^2} e^{\frac{-[x-\mu(K)]^2}{2\sigma(K)^2}} - \frac{6[x-\mu(K)]^2}{\sigma(K)^4} e^{\frac{-[x-\mu(K)]^2}{2\sigma(K)^2}} + \frac{[x-\mu(K)]^4}{\sigma(K)^6} e^{\frac{-[x-\mu(K)]^2}{2\sigma(K)^2}} \right\}$$

$$K^{(5)} = \frac{1}{\sigma(K)^3\sqrt{2\pi}} \left\{ \frac{-15[x-\mu(K)]}{\sigma(K)^4} e^{\frac{-[x-\mu(K)]^2}{2\sigma(K)^2}} + \frac{10[x-\mu(K)]^3}{\sigma(K)^6} e^{\frac{-[x-\mu(K)]^2}{2\sigma(K)^2}} - \frac{[x-\mu(K)]^5}{\sigma(K)^8} e^{\frac{-[x-\mu(K)]^2}{2\sigma(K)^2}} \right\}$$

$$K^{(6)} = \frac{1}{\sigma(K)^3\sqrt{2\pi}}\left\{\frac{-15}{\sigma(K)^4}e^{\frac{-[x-\mu(K)]^2}{2\sigma(K)^2}} + \frac{45[x-\mu(K)]^2}{\sigma(K)^6}e^{\frac{-[x-\mu(K)]^2}{2\sigma(K)^2}} - \frac{15[x-\mu(K)]^4}{\sigma(K)^8}e^{\frac{-[x-\mu(K)]^2}{2\sigma(K)^2}} + \frac{[x-\mu(K)]^6}{\sigma(K)^{10}}e^{\frac{-[x-\mu(K)]^2}{2\sigma(K)^2}}\right\}$$

$$K^{(7)} = \frac{1}{\sigma(K)^3\sqrt{2\pi}}\left\{\frac{105[x-\mu(K)]}{\sigma(K)^6}e^{\frac{-[x-\mu(K)]^2}{2\sigma(K)^2}} - \frac{105[x-\mu(K)]^3}{\sigma(K)^8}e^{\frac{-[x-\mu(K)]^2}{2\sigma(K)^2}} + \frac{21[x-\mu(K)]^5}{\sigma(K)^{10}}e^{\frac{-[x-\mu(K)]^2}{2\sigma(K)^2}} - \frac{[x-\mu(K)]^7}{\sigma(K)^{10}}e^{\frac{-[x-\mu(K)]^2}{2\sigma(K)^2}}\right\}$$

$$K^{(8)} = \frac{1}{\sigma(K)^3\sqrt{2\pi}}\left\{\frac{105}{\sigma(K)^6}e^{\frac{-[x-\mu(K)]^2}{2\sigma(K)^2}} - \frac{420[x-\mu(K)]^2}{\sigma(K)^8}e^{\frac{-[x-\mu(K)]^2}{2\sigma(K)^2}} + \frac{210[x-\mu(K)]^4}{\sigma(K)^{10}}e^{\frac{-[x-\mu(K)]^2}{2\sigma(K)^2}} - \frac{28[x-\mu(K)]^6}{\sigma(K)^{12}}e^{\frac{-[x-\mu(K)]^2}{2\sigma(K)^2}} + \frac{[x-\mu(K)]^8}{\sigma(K)^{14}}e^{\frac{-[x-\mu(K)]^2}{2\sigma(K)^2}}\right\}$$

$$K^{(9)} = \frac{1}{\sigma(K)^3\sqrt{2\pi}}\left\{\frac{-945[x-\mu(K)]}{\sigma(K)^8}e^{\frac{-[x-\mu(K)]^2}{2\sigma(K)^2}} + \frac{1260[x-\mu(K)]^3}{\sigma(K)^{10}}e^{\frac{-[x-\mu(K)]^2}{2\sigma(K)^2}} - \frac{378[x-\mu(K)]^5}{\sigma(K)^{12}}e^{\frac{-[x-\mu(K)]^2}{2\sigma(K)^2}} + \frac{36[x-\mu(K)]^7}{\sigma(K)^{14}}e^{\frac{-[x-\mu(K)]^2}{2\sigma(K)^2}} - \frac{[x-\mu(K)]^9}{\sigma(K)^{16}}e^{\frac{-[x-\mu(K)]^2}{2\sigma(K)^2}}\right\}$$

$$K^{(10)} = \frac{1}{\sigma(K)^3\sqrt{2\pi}}\left\{\frac{-945}{\sigma(K)^8}e^{\frac{-[x-\mu(K)]^2}{2\sigma(K)^2}} + \frac{4725[x-\mu(K)]^2}{\sigma(K)^{10}}e^{\frac{-[x-\mu(K)]^2}{2\sigma(K)^2}} - \frac{3150[x-\mu(K)]^4}{\sigma(K)^{12}}e^{\frac{-[x-\mu(K)]^2}{2\sigma(K)^2}} + \frac{630[x-\mu(K)]^6}{\sigma(K)^{14}}e^{\frac{-[x-\mu(K)]^2}{2\sigma(K)^2}} - \frac{45[x-\mu(K)]^8}{\sigma(K)^{16}}e^{\frac{-[x-\mu(K)]^2}{2\sigma(K)^2}} + \frac{[x-\mu(K)]^{10}}{\sigma(K)^{18}}e^{\frac{-[x-\mu(K)]^2}{2\sigma(K)^2}}\right\}$$

For standard normal case, $\sigma(K) = 1$

$$K^{(6)}(0) = \frac{-15}{\sqrt{2\pi}}$$

$$K^{(10)}(0) = \frac{-945}{\sqrt{2\pi}}$$

$$g_1 = \left[-2K^{(6)}(0)/\{\mu_2(K)\hat{\varphi}_8^{NS} n\}\right]^{1/9}$$

$$= \left[2 \times \frac{\frac{15}{\sqrt{2\pi}}}{\left\{\frac{105}{32\sigma^9\pi^{1/2}} n\right\}}\right]^{1/9}$$

$$= [2/(7n)]^{1/9} 2^{1/2}\sigma$$

$$g_2 = \left[-2K^{(10)}(0)/\{\mu_2(K)\hat{\varphi}_{12}^{NS} n\}\right]^{1/13}$$

$$= \left[2 \times \frac{\frac{945}{\sqrt{2\pi}}}{\frac{945 \times 11}{2^7\sigma^{13}\pi^{1/2}}}\right]^{1/13}$$

$$= [2/(11n)]^{1/13} 2^{1/2}\sigma$$

Let $\hat{\varphi}_r(g) = n^{-1}\sum_{a=1}^{n} \hat{f}^{(r)}(X_a; g) = n^{-2}\sum_{i=1}^{m}\sum_{j=1}^{n} L_g^{(r)}(X_i^j - X_i)$ be the general case and $L = K$, then

$$\hat{\varphi}_6(g_1) = n^{-2}\sum_{i=1}^{m}\sum_{j=1}^{n} L_{g_1}^{(6)}(X_i^j - X_i)$$

$$= n^{-2}\sum_{i=1}^{m}\sum_{j=1}^{n} \frac{1}{g_1\sqrt{2\pi}} e^{\frac{-(X_i^j - X_i)^2}{2g_1^2}}$$

$$\hat{\varphi}_{10}(g_2) = n^{-2}\sum_{i=1}^{m}\sum_{j=1}^{n} L_{g_2}^{(10)}(X_i^j - X_i)$$

$$= n^{-2}\sum_{i=1}^{m}\sum_{j=1}^{n}\frac{1}{g_2\sqrt{2\pi}}e^{\frac{-(X_i^j-X_i)^2}{2g_2^2}}$$

$$K^{(4)}(0) = \frac{3}{\sqrt{2\pi}}$$

$$K^{(8)}(0) = \frac{105}{\sqrt{2\pi}}$$

$$g_3 = \left[-2K^{(4)}(0)/\{\mu_2(K)\hat{\varphi}_6 n\}\right]^{1/7}$$

$$= \left[-2 \times \frac{\frac{3}{\sqrt{2\pi}}}{\left\{n^{-1}\sum_{i=1}^{m}\sum_{j=1}^{n}\frac{1}{g_1\sqrt{2\pi}}e^{\frac{-(X_i^j-X_i)^2}{2g_1^2}}\right\}}\right]^{1/7}$$

$$g_4 = \left[-2K^{(8)}(0)/\{\mu_2(K)\hat{\varphi}_{10} n\}\right]^{1/11}$$

$$= \left[-2 \times \frac{\frac{105}{\sqrt{2\pi}}}{\left\{n^{-1}\sum_{i=1}^{m}\sum_{j=1}^{n}\frac{1}{g_2\sqrt{2\pi}}e^{\frac{-(X_i^j-X_i)^2}{2g_2^2}}\right\}}\right]^{1/7}$$

recall Eq. (3.22),

$$\hat{h}_{SCV,i} = argmin_h\{EIMSE_{SCV,i}(h)\}$$

$$= (nh)^{-1}\left(2\pi^{1/2}\right)^{-1} + \sum_{i=1}^{m}\sum_{j=1}^{n}\left\{\Phi_{\left(2h^2+2g^2\right)^{1/2}} - 2\Phi_{\left(h^2+2g^2\right)^{1/2}} + \Phi_{\left(2g^2\right)^{1/2}}\right\}(X_i^j - X_i)$$

where

$$g = \hat{C}n^{-23/45}h^{-2}$$

and

$$\hat{C} = \left\{\frac{441}{(64\pi)}\right\}^{\frac{1}{18}} (4\pi)^{-\frac{1}{5}} \hat{\varphi}_4(g_3)^{-\frac{2}{5}} \hat{\varphi}_8(g_4)^{-\frac{1}{9}}$$

$$= \left\{\frac{441}{(64\pi)}\right\}^{\frac{1}{18}} (4\pi)^{-\frac{1}{5}} \left\{ n^{-2} \sum_{i=1}^{m} \sum_{j=1}^{n} L_{g_3}^{(4)} \left(X_i^j - X_i\right) \right\}^{-\frac{2}{5}}$$

$$\left\{ n^{-2} \sum_{i=1}^{m} \sum_{j=1}^{n} L_{g_4}^{(8)} (X_i^j - X_i) \right\}^{-1/9}$$

The derived formulas have been compiled in the Software. Further details can also be referred to Wand and Jones (1995) and Scott (1992).

B.2 Supplementary figures and tables

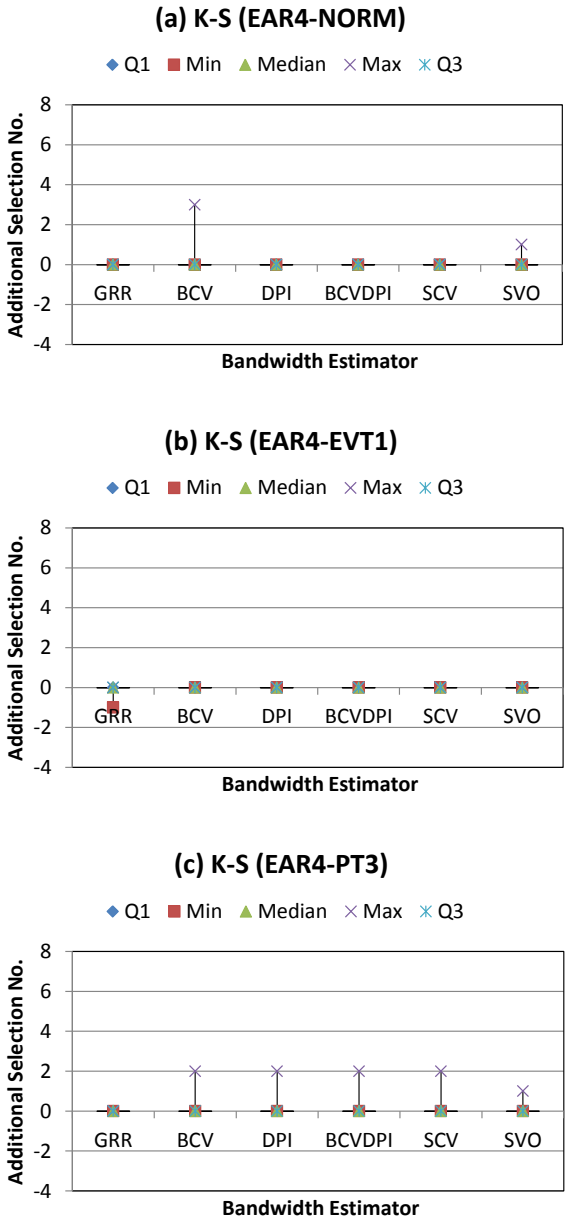
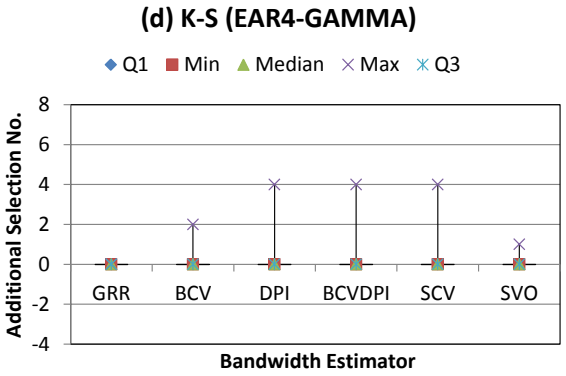
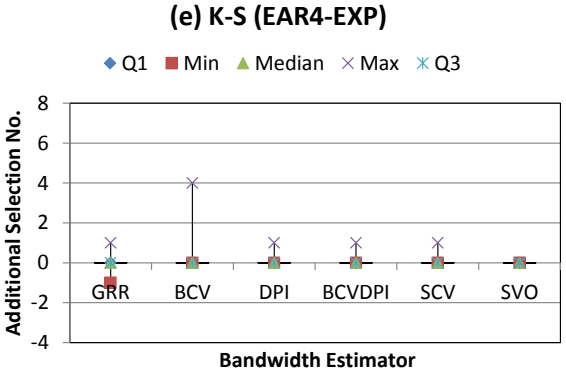
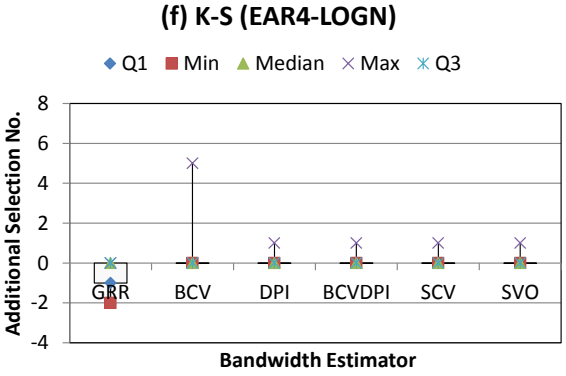


Figure.B.2.1 Number of selected additional inputs of EAR4 model with alternative bandwidth estimators (0 indicates correct number of significant inputs; overestimation occurs if above 0; and underestimation appears if below 0)







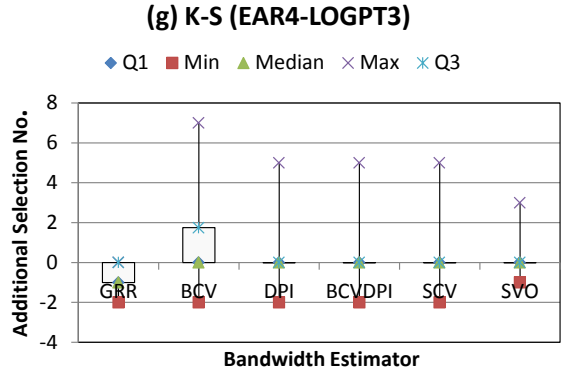


Figure B.2.1 *(Continued)*

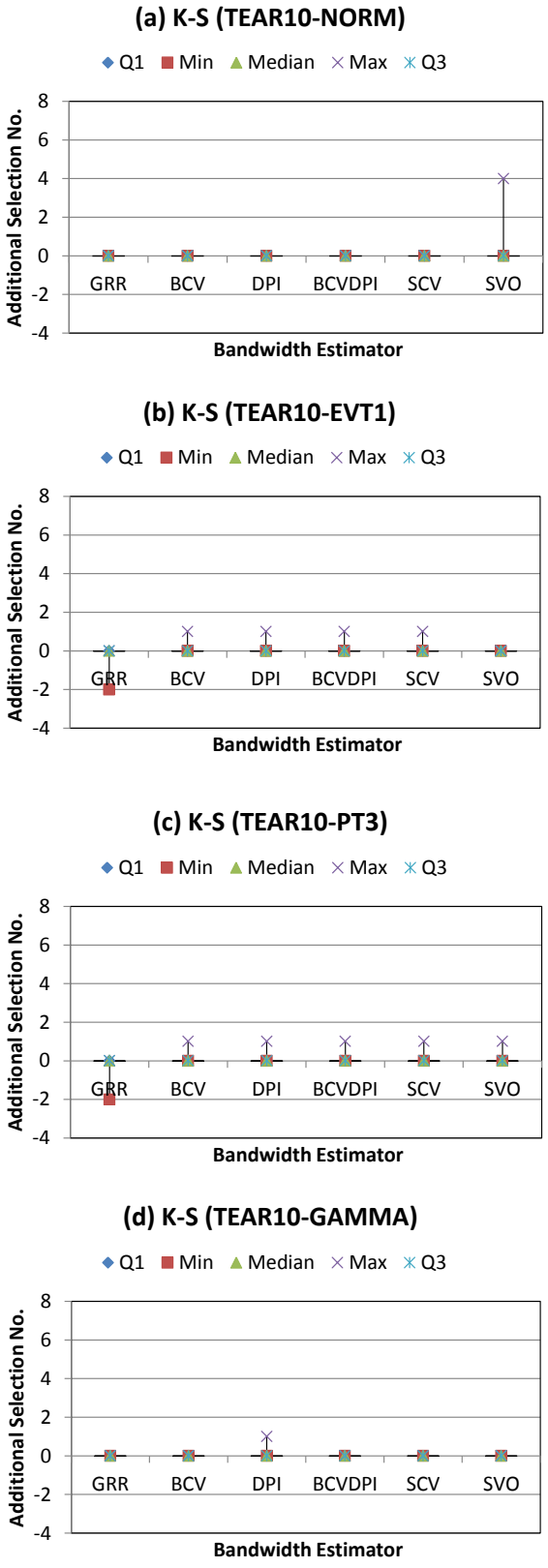
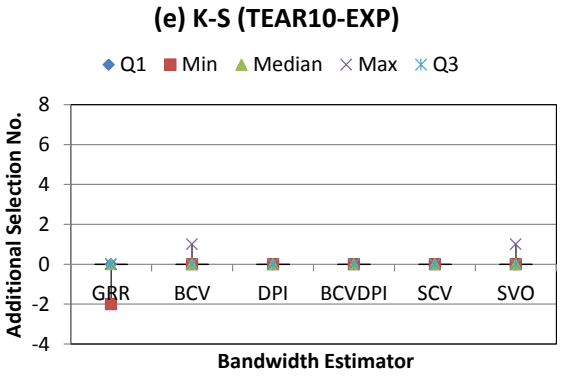
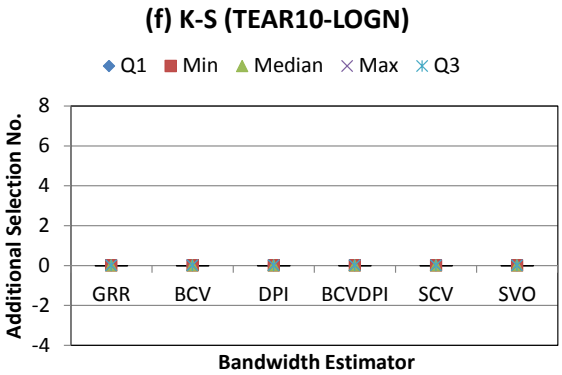


Figure.B.2.2 Number of selected additional inputs of TEAR10 model with alternative bandwidth estimators (0 indicates correct number of significant inputs; overestimation occurs if above 0; and underestimation appears if below 0)





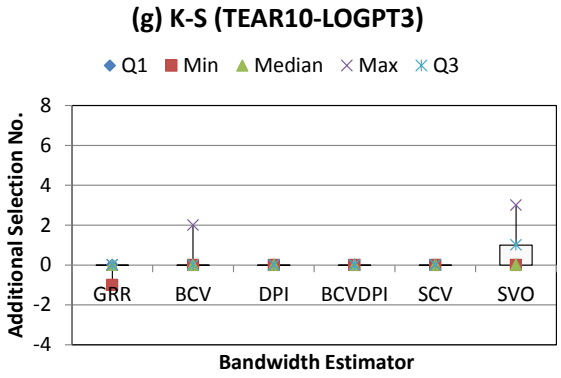


Figure B.2.2 *(Continued)*

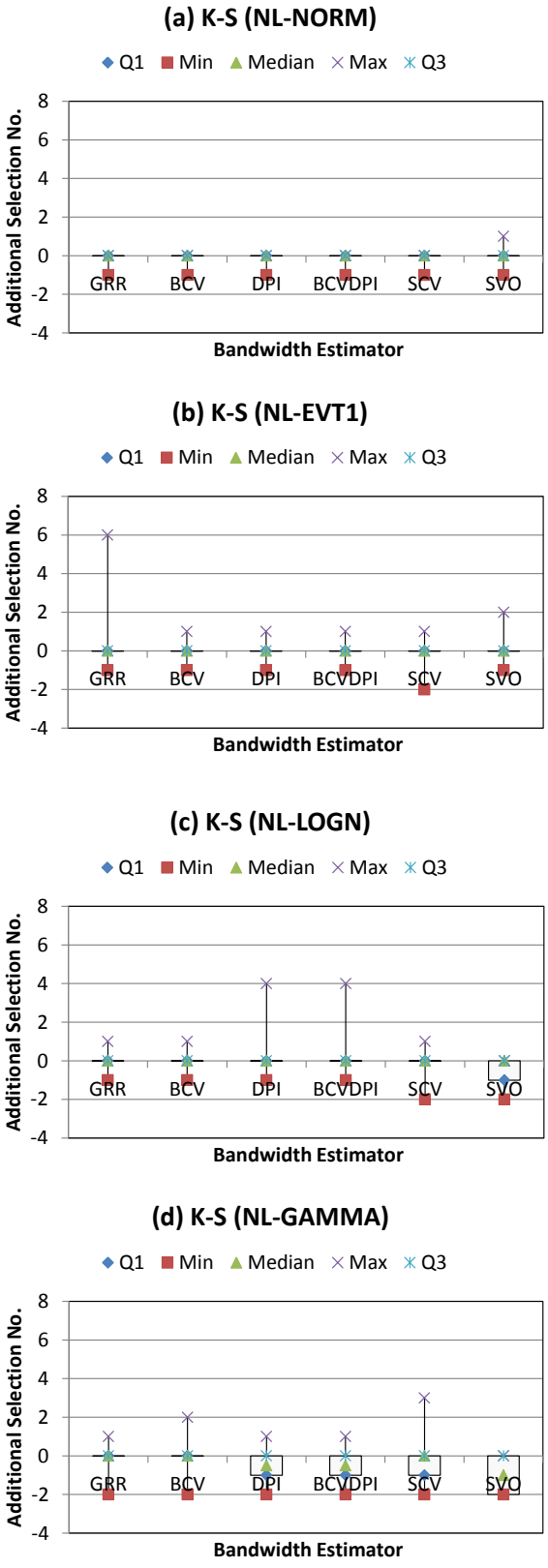
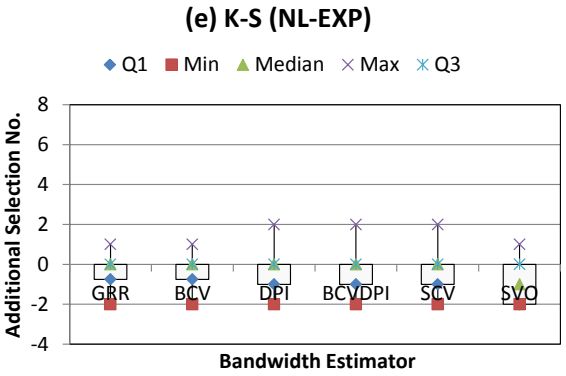
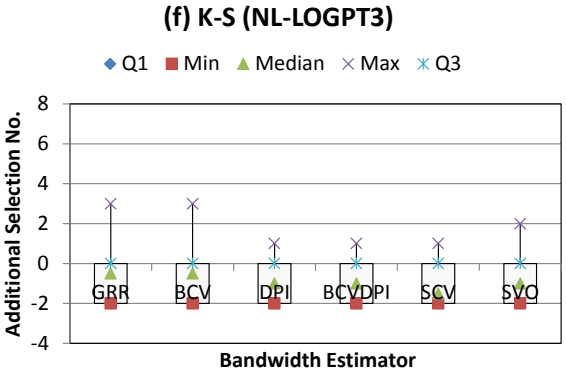


Figure.B.2.3 Number of selected additional inputs of NL model with alternative bandwidth estimators (0 indicates correct number of significant inputs; overestimation occurs if above 0; and underestimation appears if below 0)





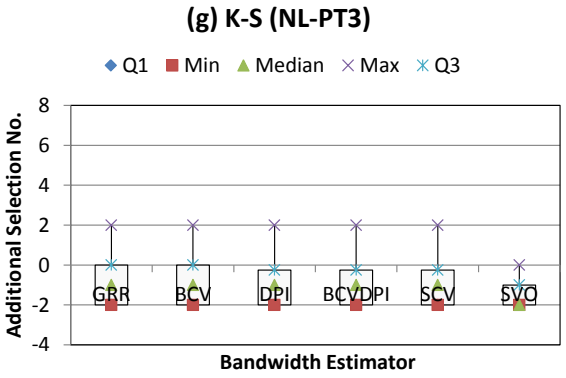


Figure B.2.3 *(Continued)*

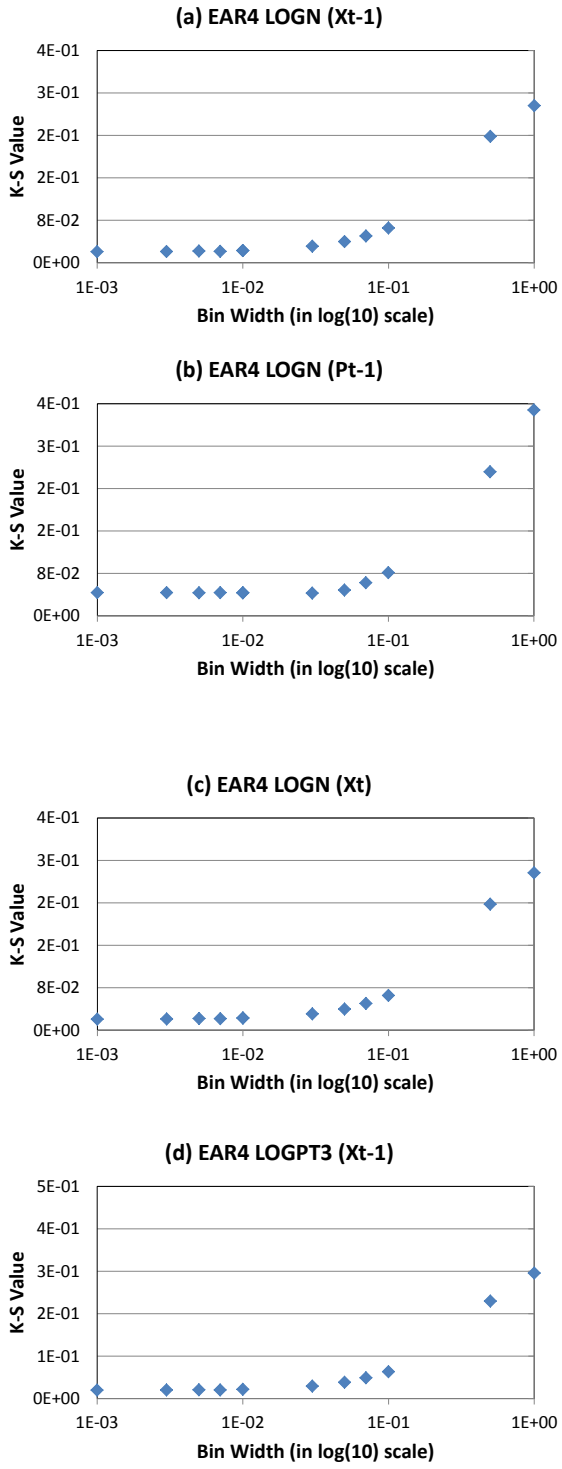


Figure.B.2.4 Sensitivity analysis of univariate histogram bin width for EAR4 model (LOGN and LOGPT3 cases; x_{t-6}, p_{t-1} and x_t)

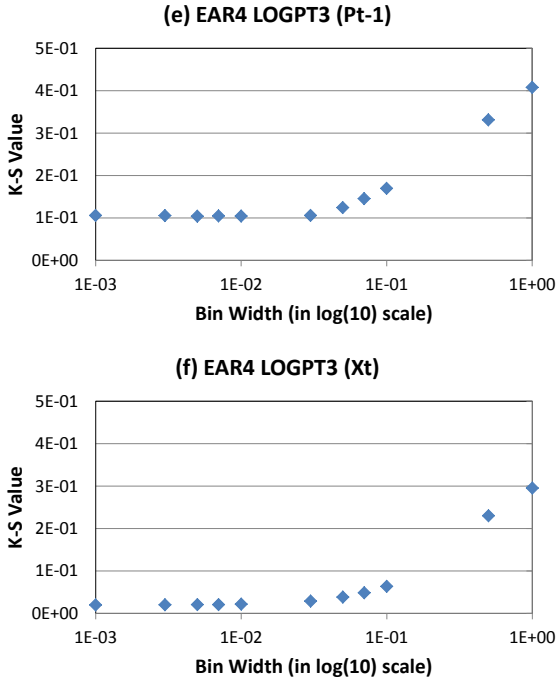


Figure B.2.4 *(Continued)*

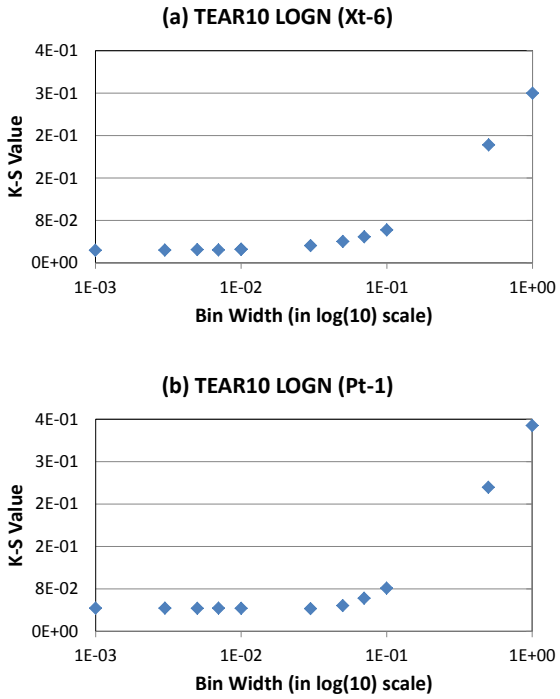
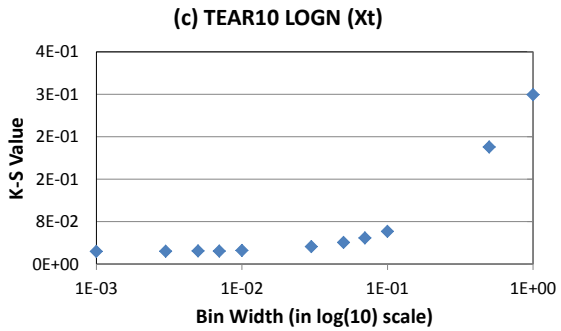
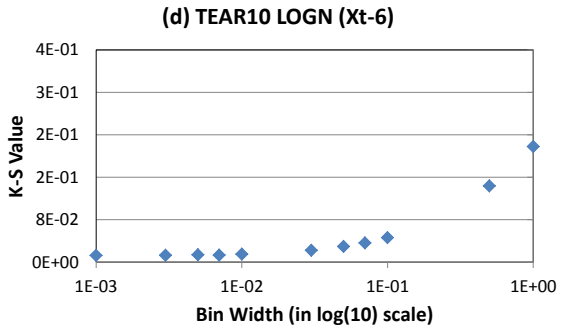
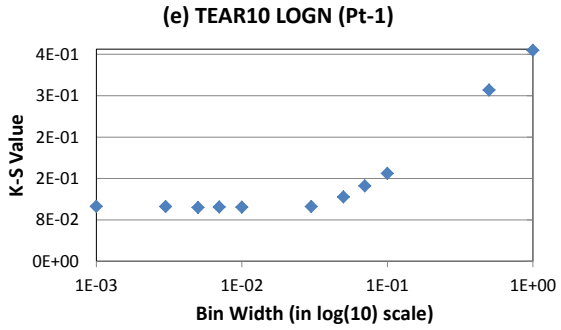


Figure.B.2.5 Sensitivity analysis of univariate histogram bin width for TEAR10 model (LOGN and LOGPT3 cases; x_{t-1}, p_{t-1} and x_t)







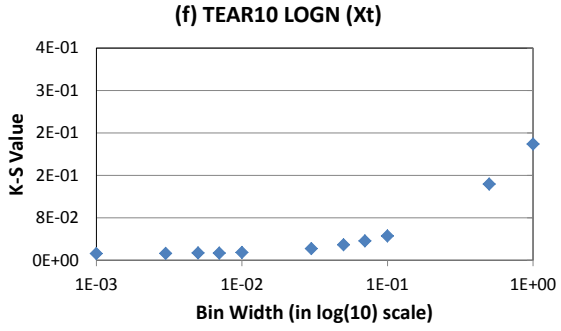


Figure B.2.5 *(Continued)*

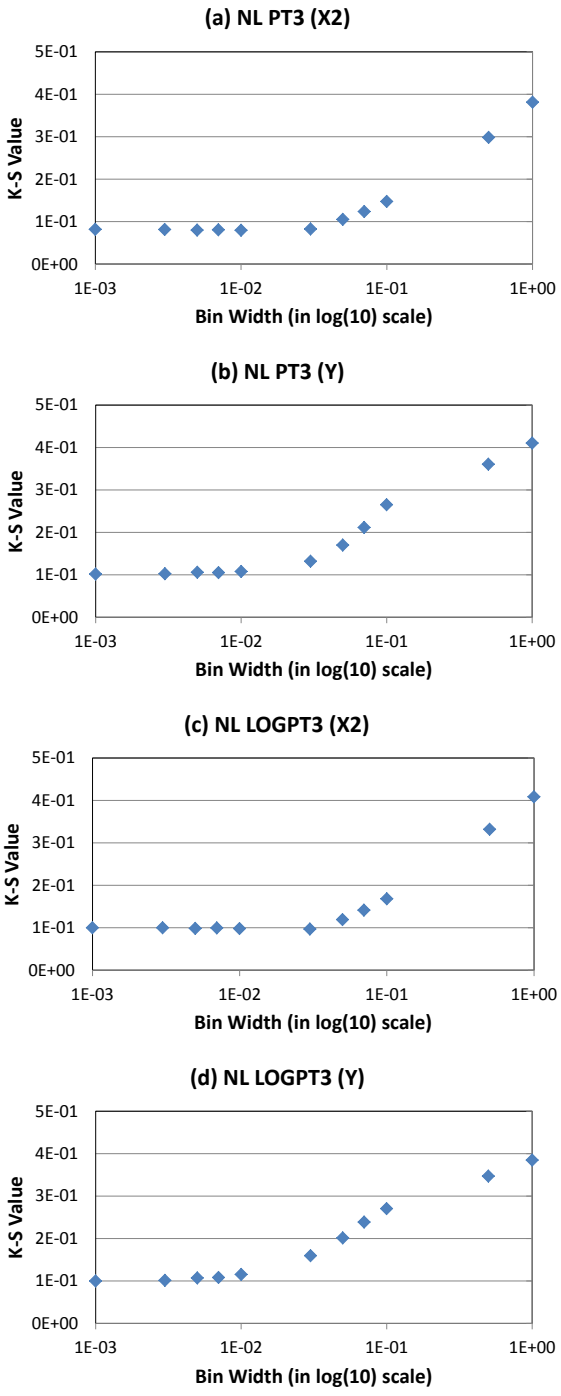


Figure.B.2.6 Sensitivity analysis of univariate histogram bin width for NL model (PT3 and LOGPT3 cases; x_2 and y)

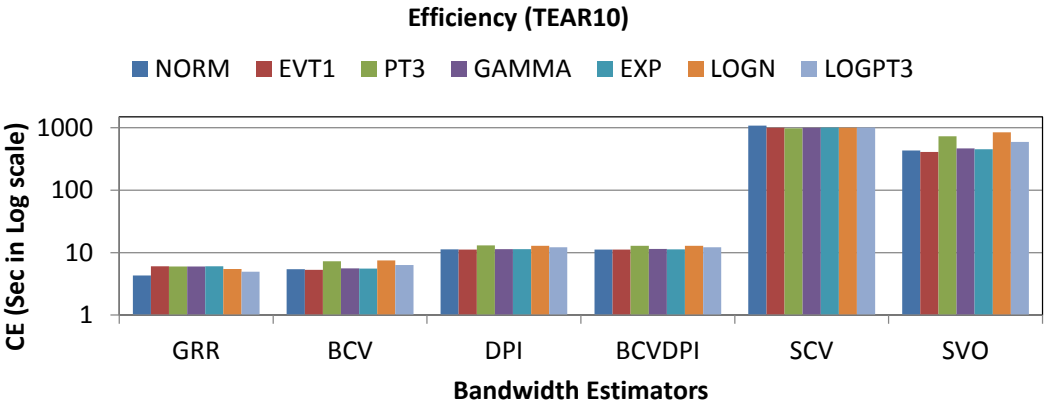


Figure. B.2.7 Computational efficiency of TEAR10 model with different bandwidth estimators

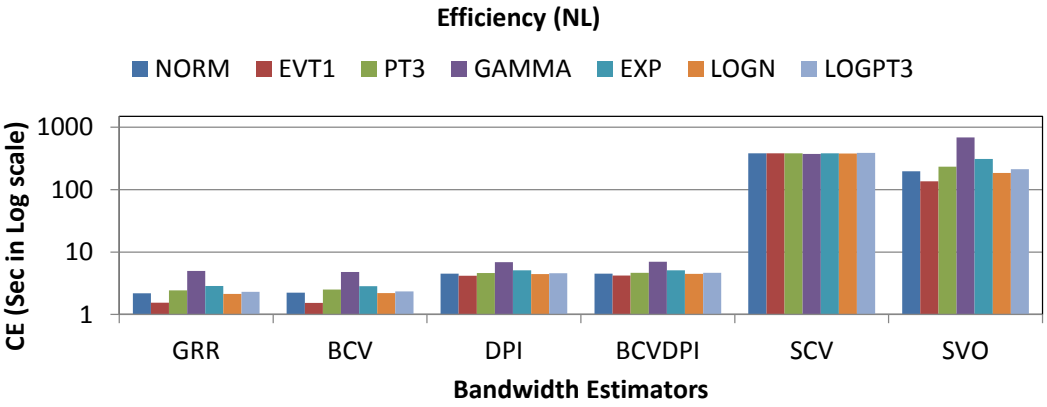


Figure. B.2.8 Computational efficiency of NL model with different bandwidth estimators

APPENDIX-C Supplementary Material from Paper 3 (Chapter 4)

C.1 Mathematical explanation and derivations

Explanation of Bivariate Reflection Correction

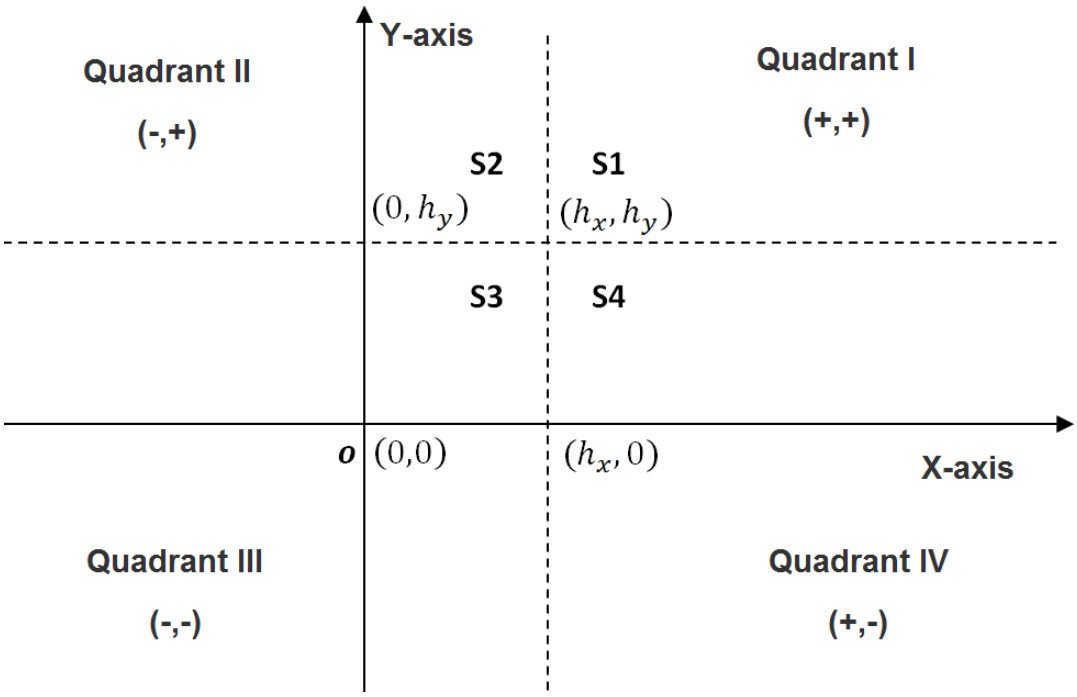


Figure. C.1.1 Quadrants of Bivariate Reflection Correction

As mentioned in Section 2, let: $\boldsymbol{X} = [X_1 \dots X_m]^T$ 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 $j = 1, \dots, n$, where n is the number of observations, $\boldsymbol{X}^j = [X_1^j \dots X_m^j]^T$ 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. C.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. (4.1) and (4.2), 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 to those in S2, however, the impact due to the boundary issue is only on the y-axis, therefore the corresponding expression is

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

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. (4.7).

Derivation of local linear polynomial regression

Let $\hat{s}_r = n^{-1}\sum_{j=1}^{n}(X_i^j - X_i)^r\, K_h\left(X_i^j - X_i\right)$, $\hat{t}_r = n^{-1}\sum_{j=1}^{n}(X_i^j - X_i)^r\, K_h\left(X_i^j - X_i\right)y^j$, and $\hat{y}(\boldsymbol{X}; p, h)_{LP} = \boldsymbol{e}_1^T\begin{bmatrix}\hat{s}_0 & \cdots & \hat{s}_p\\ \vdots & \ddots & \vdots\\ \hat{s}_p & \cdots & \hat{s}_{2p}\end{bmatrix}^{-1}\begin{bmatrix}\hat{t}_0\\ \vdots\\ \hat{t}_p\end{bmatrix}$

Then for $\hat{y}(\boldsymbol{X}; 1, h)_{LLP}$,

$\boldsymbol{e}_1^T = [1, 0]$

$\hat{s}_0 = n^{-1}\sum_{j=1}^{n} K_h\left(X_i^j - X_i\right)$

$\hat{s}_1 = n^{-1}\sum_{j=1}^{n}(X_i^j - X_i)^1\, K_h\left(X_i^j - X_i\right)$

$\hat{s}_2 = n^{-1}\sum_{j=1}^{n}(X_i^j - X_i)^2\, K_h\left(X_i^j - X_i\right)$

$\hat{t}_0 = n^{-1}\sum_{j=1}^{n} K_h\left(X_i^j - X_i\right)y^j$

$\hat{t}_1 = n^{-1}\sum_{j=1}^{n}(X_i^j - X_i)^1\, K_h\left(X_i^j - X_i\right)y^j$

$$\hat{y}(\boldsymbol{X}; 1, h)_{LLP} = [1, 0] \times \begin{bmatrix}\hat{s}_0 & \hat{s}_1\\ \hat{s}_1 & \hat{s}_2\end{bmatrix}^{-1} \times \begin{bmatrix}\hat{t}_0\\ \hat{t}_1\end{bmatrix}$$

$$= [1, 0] \times \frac{1}{\hat{s}_0\hat{s}_2 - \hat{s}_1{}^2} \times \begin{bmatrix}\hat{s}_2 & -\hat{s}_1\\ \hat{s}_1 & \hat{s}_0\end{bmatrix} \times \begin{bmatrix}\hat{t}_0\\ \hat{t}_1\end{bmatrix}$$

$$= \frac{\hat{s}_2\hat{t}_0 - \hat{s}_1\hat{t}_1}{\hat{s}_0\hat{s}_2 - \hat{s}_1{}^2}$$

$$= \frac{\hat{s}_2 n^{-1} \sum_{j=1}^{n} K_h\left(X_i^j - X_i\right) y^j - \hat{s}_1 n^{-1} \sum_{j=1}^{n} (X_i^j - X_i)^1 K_h\left(X_i^j - X_i\right) y^j}{\hat{s}_0 \hat{s}_2 - \hat{s}_1{}^2}$$

$$= n^{-1} \sum_{j=1}^{n} \frac{\left[\hat{s}_2 - \hat{s}_1 (X_i^j - X_i)\right] K_h\left(X_i^j - X_i\right) y^i}{\hat{s}_2 \hat{s}_0 - \hat{s}_1 \hat{s}_1}$$

which results in Eq. (4.19).

Derivation of local quadratic polynomial regression

Let $\hat{s}_r = n^{-1} \sum_{j=1}^{n} (X_i^j - X_i)^r K_h\left(X_i^j - X_i\right)$, $\hat{t}_r = n^{-1} \sum_{j=1}^{n} (X_i^j - X_i)^r K_h\left(X_i^j - X_i\right) y^j$, and $\hat{y}(\boldsymbol{X}; p, h)_{LP} = \boldsymbol{e}_1^T \begin{bmatrix} \hat{s}_0 & \cdots & \hat{s}_p \\ \vdots & \ddots & \vdots \\ \hat{s}_p & \cdots & \hat{s}_{2p} \end{bmatrix}^{-1} \begin{bmatrix} \hat{t}_0 \\ \vdots \\ \hat{t}_p \end{bmatrix}$

Then for $\hat{y}(\boldsymbol{X}; 2, h)_{LQP}$,

$\boldsymbol{e}_1^T = [1, 0, 0]$

$\hat{s}_0 = n^{-1} \sum_{j=1}^{n} K_h\left(X_i^j - X_i\right)$

$\hat{s}_1 = n^{-1} \sum_{j=1}^{n} (X_i^j - X_i)^1 K_h\left(X_i^j - X_i\right)$

$\hat{s}_2 = n^{-1} \sum_{j=1}^{n} (X_i^j - X_i)^2 K_h\left(X_i^j - X_i\right)$

$\hat{s}_3 = n^{-1} \sum_{j=1}^{n} (X_i^j - X_i)^3 K_h\left(X_i^j - X_i\right)$

$\hat{s}_4 = n^{-1} \sum_{j=1}^{n} (X_i^j - X_i)^4 K_h\left(X_i^j - X_i\right)$

$\hat{t}_0 = n^{-1} \sum_{j=1}^{n} K_h\left(X_i^j - X_i\right) y^j$

$\hat{t}_1 = n^{-1} \sum_{j=1}^{n} (X_i^j - X_i)^1 K_h\left(X_i^j - X_i\right) y^j$

$\hat{t}_2 = n^{-1} \sum_{j=1}^{n} (X_i^j - X_i)^2 K_h\left(X_i^j - X_i\right) y^j$

$$\hat{y}(\boldsymbol{X}; 1, h)_{LQP} = [1, 0, 0] \times \begin{bmatrix} \hat{s}_0 & \hat{s}_1 & \hat{s}_2 \\ \hat{s}_1 & \hat{s}_2 & \hat{s}_3 \\ \hat{s}_2 & \hat{s}_3 & \hat{s}_4 \end{bmatrix}^{-1} \times \begin{bmatrix} \hat{t}_0 \\ \hat{t}_1 \\ \hat{t}_2 \end{bmatrix}$$

$$= \frac{(\hat{s}_2 \hat{s}_4 - \hat{s}_3 \hat{s}_3) \hat{t}_0 - (\hat{s}_1 \hat{s}_4 - \hat{s}_2 \hat{s}_3) \hat{t}_1 + (\hat{s}_1 \hat{s}_3 - \hat{s}_2 \hat{s}_2) \hat{t}_2}{det \begin{bmatrix} \hat{s}_0 & \hat{s}_1 & \hat{s}_2 \\ \hat{s}_1 & \hat{s}_2 & \hat{s}_3 \\ \hat{s}_2 & \hat{s}_3 & \hat{s}_4 \end{bmatrix}}$$

$$= \frac{\begin{array}{c}(\hat{s}_2 \hat{s}_4 - \hat{s}_3 \hat{s}_3) n^{-1} \sum_{j=1}^{n} K_h\left(X_i^j - X_i\right) y^j - (\hat{s}_1 \hat{s}_4 - \hat{s}_2 \hat{s}_3) n^{-1} \\ \sum_{j=1}^{n} (X_i^j - X_i)^1 K_h\left(X_i^j - X_i\right) y^j + (\hat{s}_1 \hat{s}_3 - \hat{s}_2 \hat{s}_2) n^{-1} \sum_{j=1}^{n} (X_i^j - X_i)^2 K_h\left(X_i^j - X_i\right) y^j\end{array}}{(\hat{s}_2 \hat{s}_4 - \hat{s}_3 \hat{s}_3) \hat{s}_0 - (\hat{s}_1 \hat{s}_4 - \hat{s}_2 \hat{s}_3) \hat{s}_1 + (\hat{s}_1 \hat{s}_3 - \hat{s}_2 \hat{s}_2) \hat{s}_2}$$

$$= n^{-1} \sum_{j=1}^{n} \frac{\begin{bmatrix} (\hat{s}_2\hat{s}_4 - \hat{s}_3\hat{s}_3) - (\hat{s}_1\hat{s}_4 - \hat{s}_2\hat{s}_3)\left(X_i^j - X_i\right) + \\ (\hat{s}_1\hat{s}_3 - \hat{s}_2\hat{s}_2)(X^i - X)^2 \end{bmatrix} K_h\left(X_i^j - X_i\right) y^i}{[\hat{s}_0(\hat{s}_2\hat{s}_4 - \hat{s}_3\hat{s}_3) - \hat{s}_1(\hat{s}_4\hat{s}_1 - \hat{s}_3\hat{s}_2) + \hat{s}_2(\hat{s}_1\hat{s}_3 - \hat{s}_2\hat{s}_2)]}$$

which results in Eq. (4.20).

C.2 Supplementary figures and tables

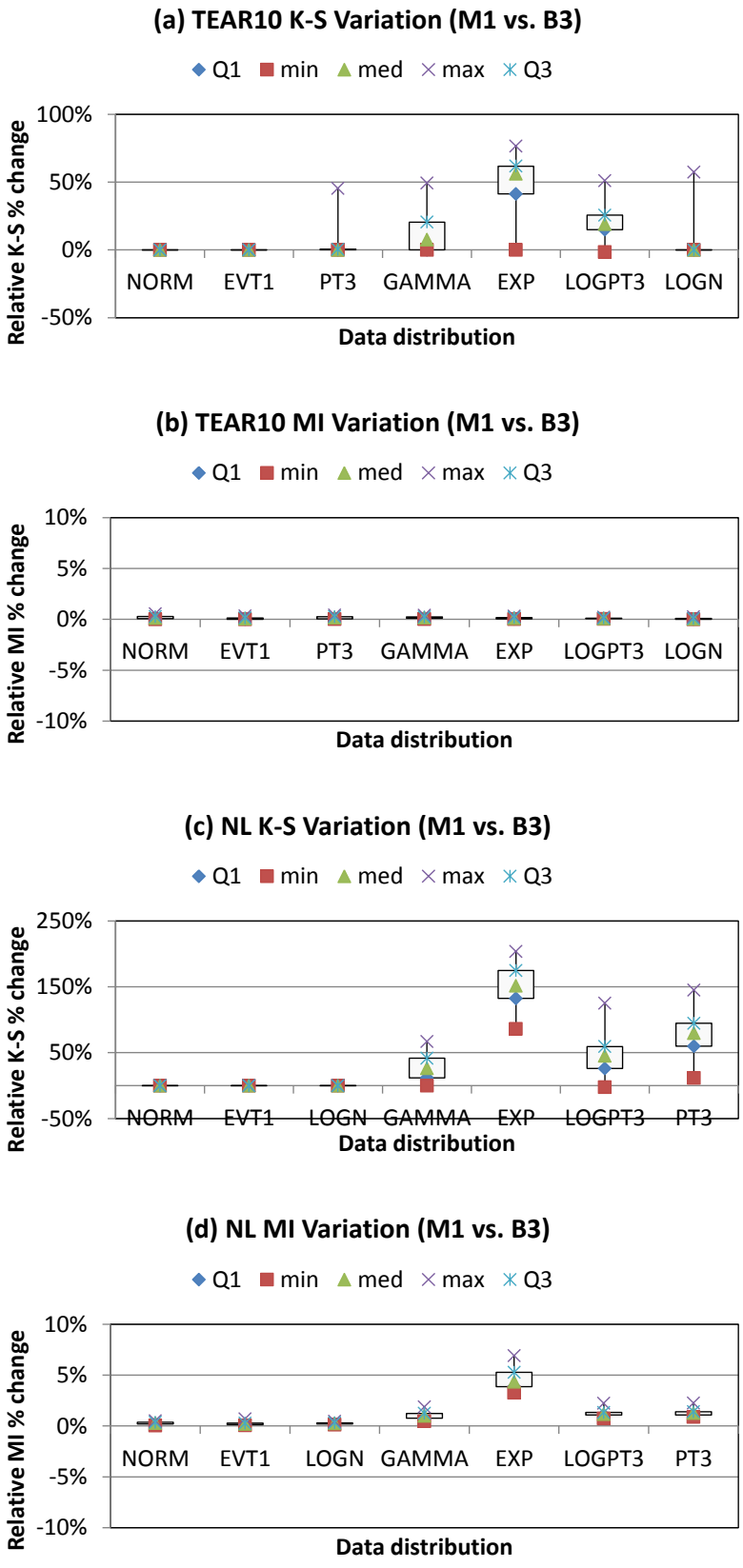


Figure. C.2.1. Relative change of K-S and MI in-between M1 and B3 (TEAR10 and NL)

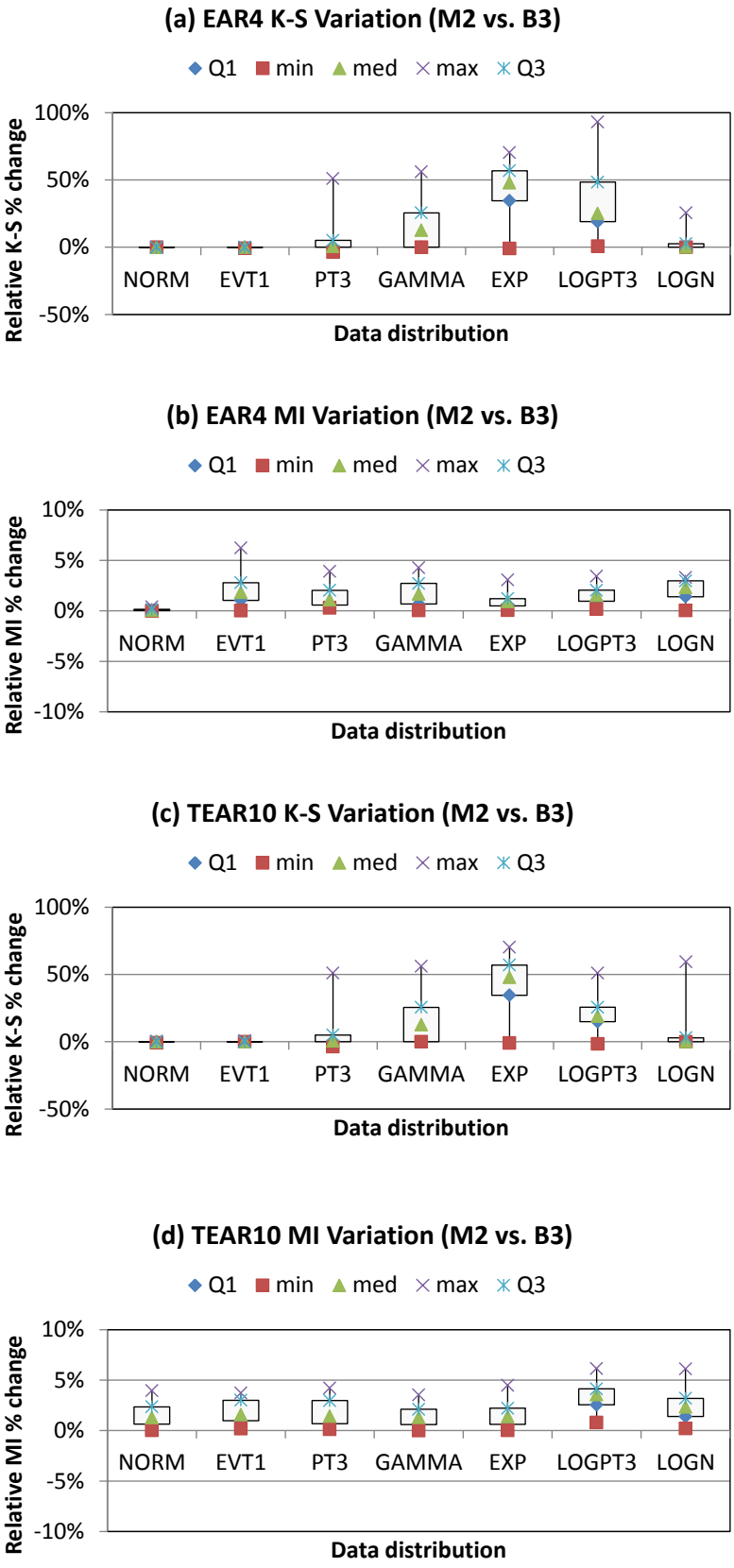
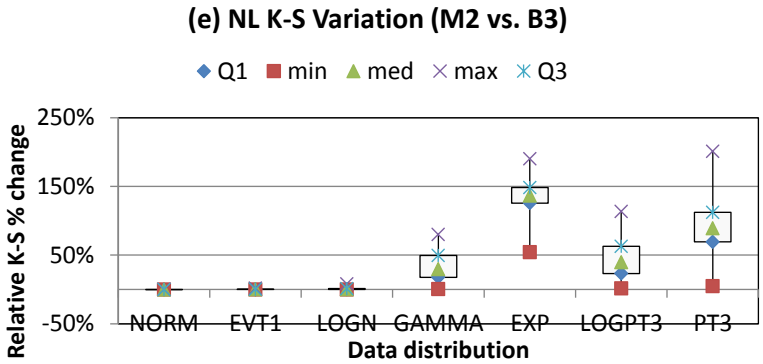


Figure. C.2.2. Relative change of K-S and MI in-between M2 and B3 for EAR4, TEAR10 and NL models



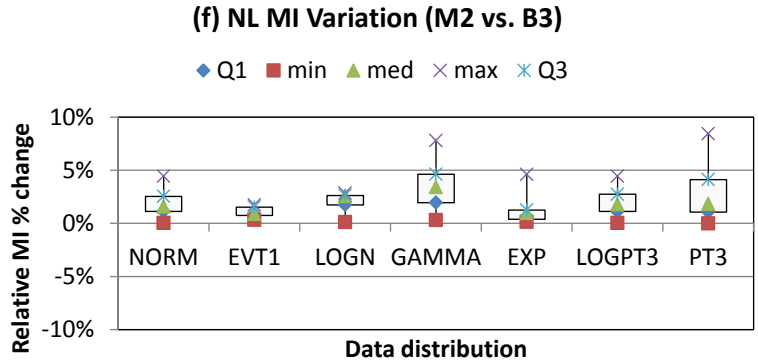


Figure. C.2.2. *(Continued)*

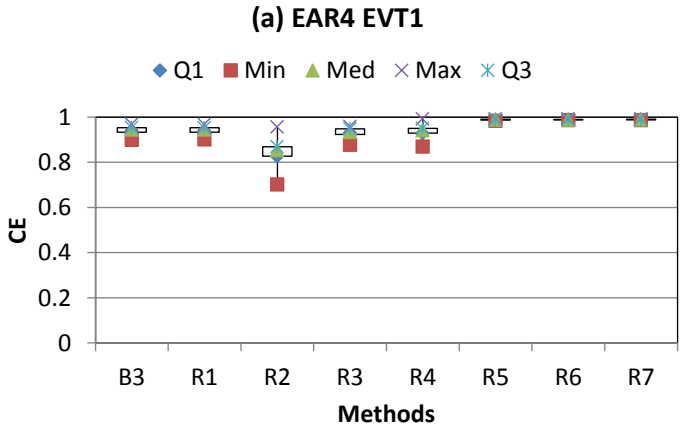
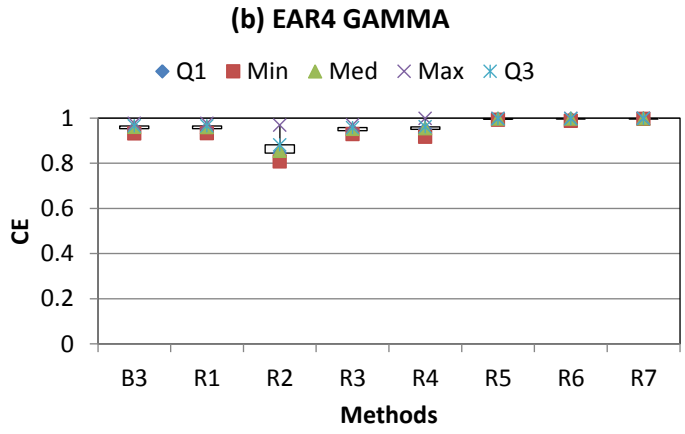
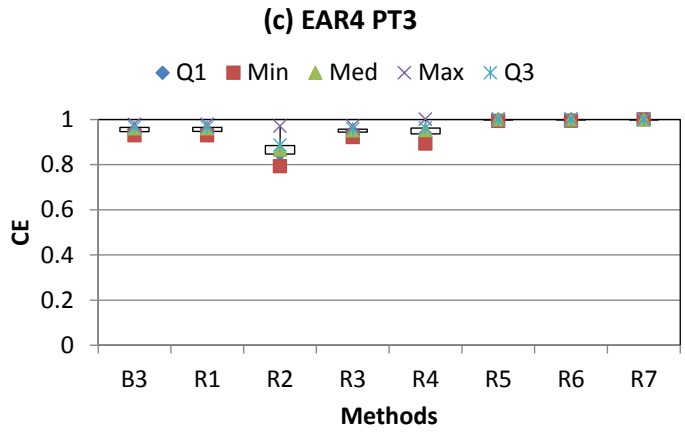


Figure. C.2.3. Accuracy of residual estimation with alternative estimators for EAR4 model (other 4 cases)





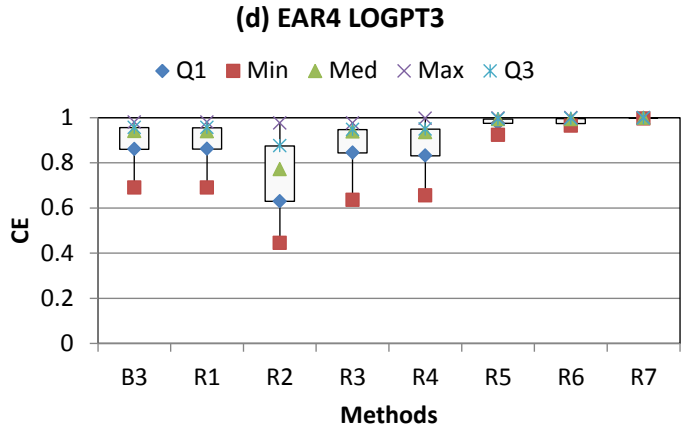


Figure. C.2.3. *(Continued)*

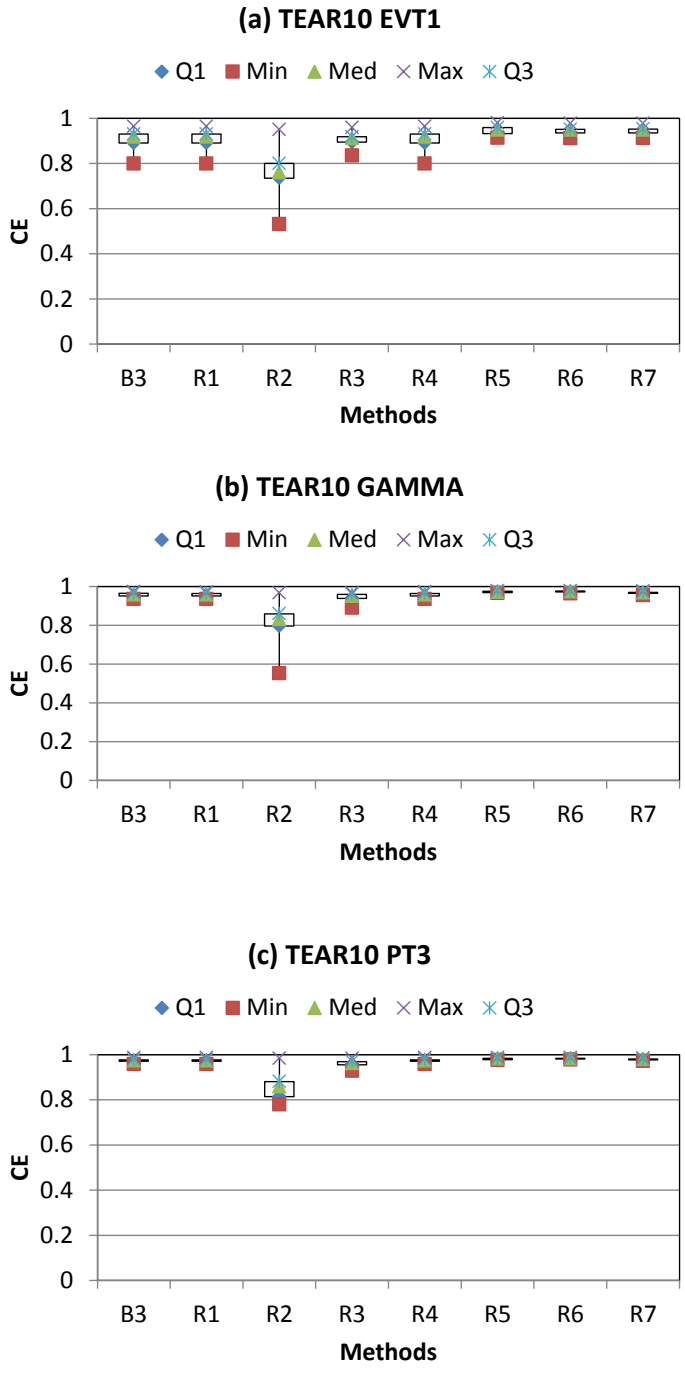


Figure. C.2.4. Accuracy of residual estimation with alternative estimators for TEAR10 model (other 4 cases)

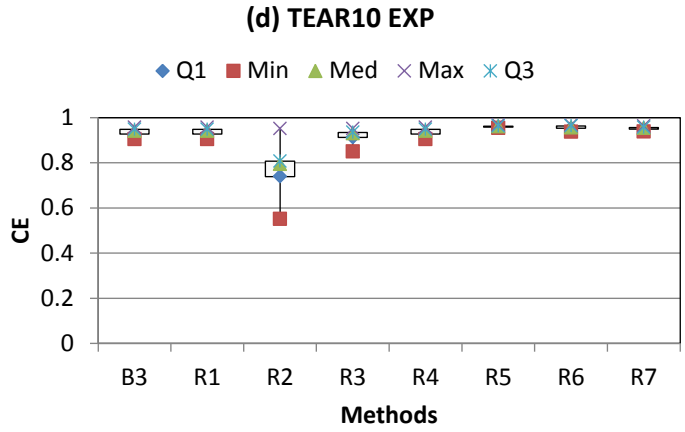


Figure. C.2.4. *(Continued)*

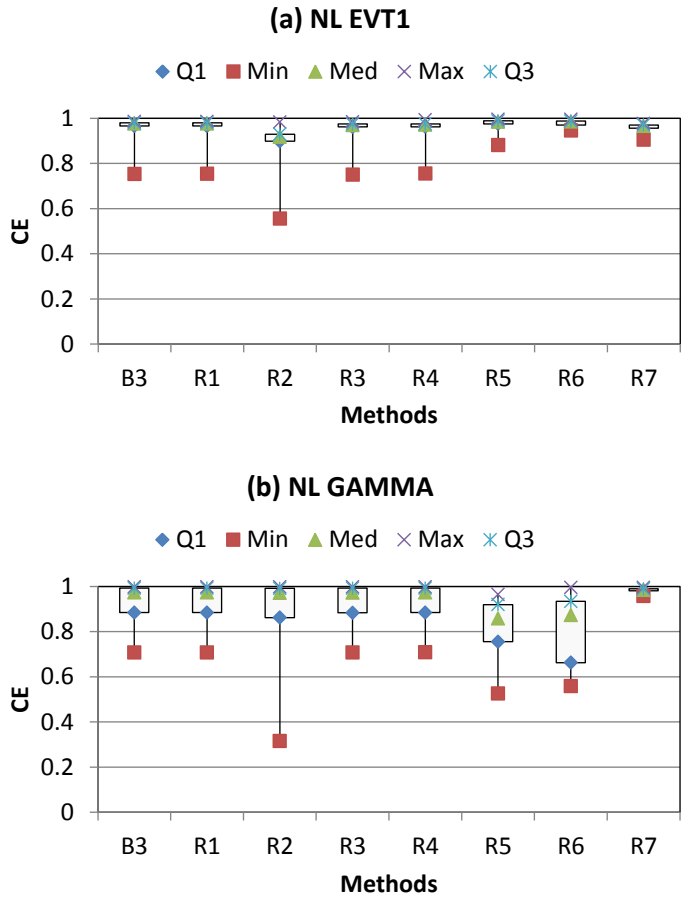
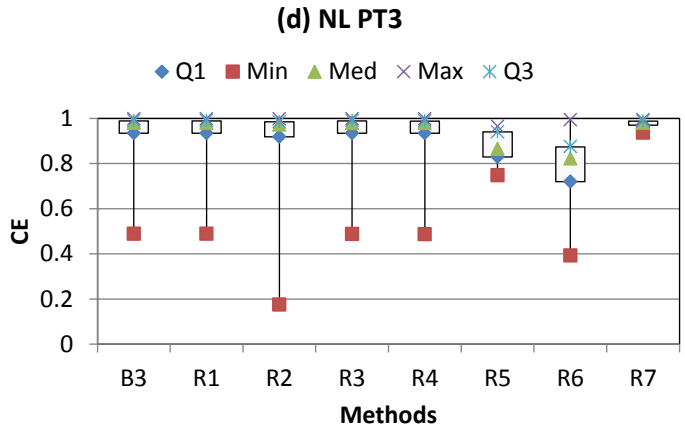


Figure. C.2.5. Accuracy of residual estimation with alternative estimators for NL model (other 4 cases)



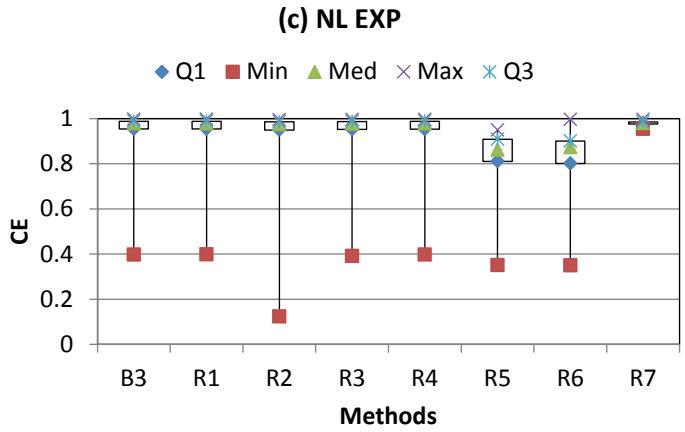


Figure. C.2.5. *(Continued)*

APPENDIX-D Copy of Publications

D.1 Copy of Paper 1 from Chapter 2 (as published)

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: 10.1016/j.envsoft. 2014.1005.1010.

Environmental Modelling & Software 59 (2014) 162–186

Contents lists available at ScienceDirect

Environmental Modelling & Software

journal homepage: www.elsevier.com/locate/envsoft

Selection of smoothing parameter estimators for general regression neural networks – Applications to hydrological and water resources modelling

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

School of Civil, Environmental and Mining Engineering, The University of Adelaide, Adelaide, South Australia 5005, Australia

ARTICLE INFO

Article history:
Received 23 December 2013
Received in revised form
5 May 2014
Accepted 6 May 2014
Available online xxx

Keywords:
General regression neural networks
Smoothing parameter estimators
Artificial neural networks
Multi-layer perceptrons
Extreme and average events
Hydrology and water resources

ABSTRACT

Multi-layer perceptron artificial neural networks are used extensively in hydrological and water resources modelling. However, a significant limitation with their application is that it is difficult to determine the optimal model structure. General regression neural networks (GRNNs) overcome this limitation, as their model structure is fixed. However, there has been limited investigation into the best way to estimate the parameters of GRNNs within water resources applications. In order to address this shortcoming, the performance of nine different estimation methods for the GRNN smoothing parameter is assessed in terms of accuracy and computational efficiency for a number of synthetic and measured data sets with distinct properties. Of these methods, five are based on bandwidth estimators used in kernel density estimation, and four are based on single and multivariable calibration strategies. In total, 5674 GRNN models are developed and preliminary guidelines for the selection of GRNN parameter estimation methods are provided and tested.

Software availability

Software name: GRNNs
Developer: Xuyuan Li, Postgraduate Student, the University of Adelaide, School of Civil, Environmental & Mining Engineering, Adelaide, SA 5005, Australia
Phone: +61 8 8313 1575
Fax: +61 8 8303 4359
Email: xli@civeng.adelaide.edu.au
Hardware requirements: 64-bit AMD64, 64-bit Intel 64 or 32-bit ×86 processor-based workstation or server with one or more single core or multi-core microprocessors; all versions of Visual Studio 2012, 2010 and 2008 are supported except Visual Studio Express; 256 MB RAM
Software requirements: PGI Visual Fortran 2003 or later version
Language: English
Size: 4.74 MB
Availability: Free to download for research purposes from the following website:http://www.ecms.adelaide.edu.au/civeng/research/water/software/generalised-regression-neural-network/

* Corresponding author. Tel.: +61 8 8313 1575; fax: +61 8 8303 4359.
E-mail addresses: xli@civeng.adelaide.edu.au, xliadelaide@gmail.com, li-xu-yuan@163.com (X. Li), aaron.zecchin@adelaide.edu.au (A.C. Zecchin), holger.maier@adelaide.edu.au (H.R. Maier).

http://dx.doi.org/10.1016/j.envsoft.2014.05.010
1364-8152/

1. Introduction

Over the last two decades, artificial neural networks (ANNs) have been used extensively in the field of hydrological and water resources modelling, and their popularity is still increasing (Maier et al., 2010; Abrahart et al., 2012; Wu et al., 2014). In the vast majority of these applications, multi-layer perceptrons (MLPs) have been used as the most common model architecture (Maier et al., 2010; Wu et al., 2014). While the use of MLPs has generally resulted in good model performance, their development is complicated by the fact that there are no rigorous methods for determining an appropriate model structure. Determination of the optimal number of hidden nodes is especially difficult, unless sophisticated Bayesian approaches are used (Kingston et al., 2008; Zhang et al., 2011), which are computationally demanding and require substantial technical expertise to implement. Therefore, the optimal model structure is generally determined by trial and error (Maier et al., 2010; Wu et al., 2014). This process usually involves a number of

steps, including (i) selection of a trial model structure, (ii) calibration of the model with the selected structure, and (iii) evaluation of the predictive performance of the calibrated model. These steps are repeated for models with different trial structures and the model structure that results in the best predictive performance is selected. Consequently, the model structure that is found to be optimal is a function of a number of factors, including:

(i) *The trial model structures selected for evaluation:* As the potential number of different model structures is generally large, the performance of a subset of all possible structures is usually evaluated. This can be achieved using different approaches, including ad-hoc, stepwise (e.g. constructive, pruning) or global approaches (Maier et al., 2010). Consequently, as different approaches generally result in the evaluation of different model structures, the structure obtained is a function of the adopted approach.

(ii) *The calibration method used:* The predictive performance of a model with a particular structure is a function of the quality of the calibration (training) process. Finding the combination of model parameters (connection weights) that gives the best predictive performance for a given network structure is complicated by the presence of a large number of local optima in the error surface (Kingston et al., 2005a). This is particularly the case if gradient-based calibration (training) methods are used (Maier and Dandy, 1999), such as the most commonly used back-propagation algorithm (Maier et al., 2010; Wu et al., 2014). In addition to the choice of calibration (training) methods, the parameters that control the searching behaviour of these methods (e.g. learning rate and momentum when the back-propagation algorithm is used) can also have a significant impact on the best predictive model performance obtained for a particular model structure (Maier and Dandy, 1998a,b). Consequently, unless the predictive performance that corresponds to the global optimum in the error surface can be identified for all models with different structures, it is not possible to identify which model structure results in the best predictive performance with certainty. As a result, the optimal model structure obtained is a function of the quality of the model calibration process.

(iii) *The calibration data used:* The available data are generally split into different subsets for calibration (training) and validation, which can be done using a number of different methods (see Maier et al., 2010). Consequently, which data points are included in the different subsets can vary, depending on which data division method is used (Bowden et al., 2002; May et al., 2010; Wu et al., 2012, 2013). This can also have an impact on which model structure is found to result in the best predictive performance. This is because different data points will result in different error surfaces during calibration, thereby potentially affecting calibration difficulty [see (ii)] and producing different global and local optima, which is likely to change which model structure results in the lowest error.

Given the factors described above, it is generally not possible to isolate the impact of model structure on the predictive performance of MLPs, making it difficult to know which model structure should be used. In addition, the trial-and-error process generally used to determine the optimal structure of MLPs is computationally expensive, as it necessitates the development of a potentially large number of models.

Although there are other alternative ANN based approaches, including Radial Basis Functions (RBFs) (Buhmann, 2003), Recurrent Neural Networks (RNNs) (Williams and Zipser, 1989) and Probabilistic Neural Networks (PNNs) (Specht, 1990), General regression neural networks (GRNNs) (Specht, 1991) provide an alternative ANN model structure that has been shown to perform well in a number of studies in water resources applications (e.g. Bowden et al., 2005b, 2006; Gibbs et al., 2006; Cigizoglu and Alp, 2006) and overcomes the shortcomings associated with MLPs discussed above, as the structure of GRNNs is fixed (Bowden et al., 2005a). This removes the ambiguity associated with determining which model structure is optimal. In addition, it increases the computational efficiency of the model development process, as there is no need to develop a number of models with different structures in order to determine which is optimal.

However, a potential issue with the application of GRNNs to hydrological and water resources problems is that there has been limited work on determining which smoothing parameter estimation methods should be adopted. As GRNNs are essentially a Nadaraya-Watson kernel regression method (Cai, 2001), parameter estimation only involves the determination of optimal values of one or more smoothing parameters, also known as kernel bandwidths. However, this is not a trivial issue, as illustrated by the vast amount of literature on kernel bandwidth estimation as applied to density estimation (e.g. Rudemo, 1982; Bowman, 1984; Scott and Terrell, 1987; Park and Marron, 1990; Hall et al., 1992; Wand and Jones, 1995). Overestimating the smoothing parameter can result in over-smoothing of the estimated density (i.e. kernel based probability density function (PDF)). In this case, the detailed local information (for instance the variation of daily rainfall in hydrological applications) will not be captured in the estimated density. In contrast, if values of the smoothing parameter are underestimated, the general trend of the estimated density (for instance the overall rainfall trend within a given time period) can be disturbed by localised features or noise.

Among the extensive literature on smoothing parameter (or kernel bandwidth) estimation in other areas of research, such as mathematics and statistics, there are a number of different approaches to obtaining optimal estimates of kernel density, which are based on assumptions about the form of the PDF and different fitness function types (i.e. the objective function on which the estimator is based). Consequently, their relative merits for determining the optimal values of the smoothing parameters for water resources GRNN models are likely to vary from case study to case study, depending on the distribution of the data and the modelling objective function used. However, the relationship between the performance of GRNNs with smoothing parameters obtained using different kernel density estimation methods and the properties of the water resources data used to develop them has not been considered previously, making it difficult to know which methods to use for particular case studies.

Therefore, the objectives of the current study are: (i) to compare the performance, in terms of both predictive accuracy and computational cost, of GRNN models for which smoothing parameters have been estimated using a range of methods, as well as that of a benchmark MLP model, for case studies with data that have varying degrees of normality, linearity and different modelling objectives (e.g. matching average or extreme events); and (ii) to develop and test empirical guidelines for the selection of the most appropriate methods for GRNN smoothing parameter estimation based on the properties of the available data (i.e. degree of normality and non-linearity) and the modelling objective.

The remainder of this paper is organised as follows. A brief introduction to GRNNs is provided in Section 2, followed by the

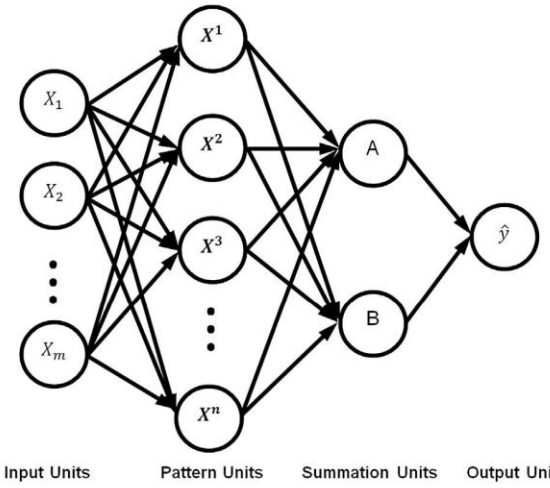


Fig. 1. General architecture of a GRNN [based upon Gibbs et al. (2006)].

Methodology in Section 3. Results and discussion are given in Section 4, and conclusion and recommendations are provided in Section 5.

2. General regression neural networks

According to Bowden et al. (2005a), GRNNs can be treated as supervised feedforward ANNs with a fixed model architecture. The general architecture of GRNNs is illustrated in Fig. 1.

Let: $\boldsymbol{X} = [X_1 \ldots X_m]^{\mathrm{T}}$ be the input, where m is the number of inputs; $(\boldsymbol{X}^j, y^j)$ be the observed pairs of input and output data (the patterns) for j=1,…,n, where n is the number of observations, $\boldsymbol{X}^j = [X_1^j \ldots X_m^j]^{\mathrm{T}}$ are the observed input data and y^j are the observed output data; and $\hat{y}$ be the GRNN estimate of the actual output y. If the joint density $f(\boldsymbol{X},y)$ is known, the conditional expectation of output y given input $\boldsymbol{X}$ is given as

$$E[y|\boldsymbol{X}] = \frac{\int_{-\infty}^{\infty} y f(\boldsymbol{X},y)\mathrm{d}y}{\int_{-\infty}^{\infty} f(\boldsymbol{X},y)\mathrm{d}y} \tag{1}$$

The joint density $f(\boldsymbol{X},y)$ in Eq. (1) is generally unknown, however, the empirical joint density of the observed input/output pairs $(\boldsymbol{X}^j, y^j)$, $j = 1, \ldots, n$ can be estimated by the Gaussian kernel-based estimator as

$$\hat{f}(\boldsymbol{X},y) = \frac{1}{2\pi^{(m+1)/2} h^{(m+1)}} \frac{1}{n} \sum_{j=1}^{n} \exp\left[-\frac{\left(\boldsymbol{X}-\boldsymbol{X}^j\right)^{\mathrm{T}}\left(\boldsymbol{X}-\boldsymbol{X}^j\right)}{2h^2}\right] \exp\left[-\frac{\left(y-y^j\right)^2}{2h^2}\right] \tag{2}$$

where h is the kernel smoothing parameter (Parzen, 1962; Cacoullos, 1966). Note that this approximation is commonly known as Parzen window density estimation. It is valid, however, only if the underlying density is continuous and the first partial derivative at any $\boldsymbol{X}$ is small. Specht (1991) combined the conditional expectation of y [Eq. (1)] with the Parzen window density estimation $\hat{f}(\boldsymbol{X},y)$ [Eq. (2)] to obtain the following estimator for y

$$\hat{y}(\boldsymbol{X},h) = \frac{\sum_{j=1}^{n} y^j \exp\left(-\frac{D_j^2(\boldsymbol{X})}{2h^2}\right)}{\sum_{j=1}^{n} \exp\left(-\frac{D_j^2(\boldsymbol{X})}{2h}\right)} \tag{3}$$

where D_j^2 is the scalar function

$$D_j^2 = \left(\boldsymbol{X}-\boldsymbol{X}^j\right)^{\mathrm{T}}\left(\boldsymbol{X}-\boldsymbol{X}^j\right) \tag{4}$$

which measures the Euclidian distance between the input $\boldsymbol{X}$ and the observed data points $\boldsymbol{X}^j$. Within this equation, the smoothing parameter h is the only unknown parameter that needs to be obtained by training (calibration).

With respect the GRNN formulation, the expression in Eq. (3) can be implemented by the four-unit (or layer) parallel network shown in Fig. 1. The GRNN consists of input, pattern, summation and output units that are fully connected. According to Specht (1991), the input units are formed by the elements of the input vector $\boldsymbol{X}$, and these then feed into each of the pattern units. The pattern units record D_j^2, the sum of squared (or absolute) difference between an input vector $\boldsymbol{X}$ and the observed data $\boldsymbol{X}^j$, and then feed into a nonlinear activation function [e.g. the exponential function as in Eq. (3)] before passing into the summation units. The summation units contain two parts, A and B, which correspond to the numerator and denominator in Eq. (3), respectively. Part A (the numerator) contains a dot product between the observed output records y^j and the weights $\exp(-D_j^2(\boldsymbol{X})/2h^2)$ from the pattern units, while part B (the denominator) only includes the weights from the pattern units. The quotient of parts A and B is the predicted output $\hat{y}$.

In Fig. 1, the model architecture of GRNNs is fixed by the fact that the number of input nodes is determined by the number of inputs m; the number of pattern nodes depends on the size of the observed input data n; and the nodes in the summation units always consist of a denominator node and a numerator node.

Within this study, a slightly generalised version of the GRNN estimator in Eq. (3) is considered, namely

$$\hat{y}(\boldsymbol{X},h) = \frac{\sum_{j=1}^{n} y^j \exp\left(-\frac{1}{2}\sum_{i=1}^{m}\frac{\left(X_i-X_i^j\right)^2}{h_i^2}\right)}{\sum_{j=1}^{n} \exp\left(-\frac{1}{2}\sum_{i=1}^{m}\frac{\left(X_i-X_i^j\right)^2}{h_i^2}\right)} \tag{5}$$

where the primary difference between Eqs. (3) and (5) is the adoption of a unique smoothing parameter h_i for each dimension of the input space $i = 1, \ldots, m$. The advantage of this form of the GRNN is that it enables an independent scaling of the kernel smoothing, as opposed to a common smoothing, along each dimension of the input space.

3. Methodology

The approach to the systematic assessment of the performance of GRNNs with different bandwidth estimators is illustrated in Fig. 2. As can be seen, there are four main steps: (i) procurement of input and output data with different degrees of normality and nonlinearity; (ii) estimation of the optimal GRNN smoothing parameter (bandwidth) for these different input or output data using a number of different smoothing parameter estimators; (iii) development

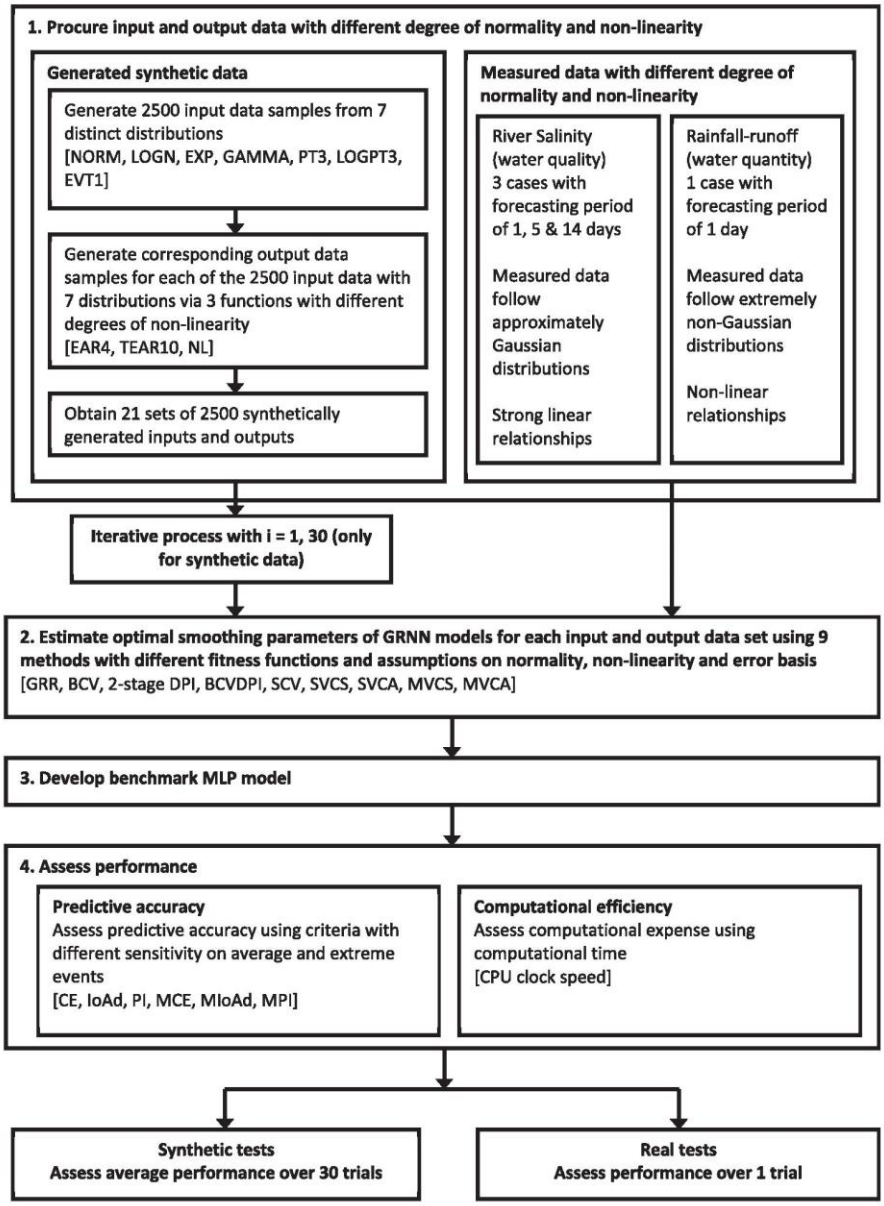


Fig. 2. Overview of proposed assessment approach.

of benchmark MLP models; and (iv) assessment of model performance. Details of each of these steps are given in the subsequent sections.

3.1. Procurement of input/output data with different degrees of normality and non-linearity

As can be seen from Fig. 2, two different approaches to procuring input and output data with different degrees of normality and non-linearity were used, including the generation of synthetic data and the use of measured data, as outlined below.

3.1.1. Synthetically generated data

Procurement of the synthetic data involved the generation of input data from distributions with differing degrees of normality, and the subsequent generation of the corresponding output data using synthetic models with different degrees of non-linearity. Data were generated from seven distinct distributions, 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) (see Fig. 2). These distributions were used because they are the most commonly adopted distributions in hydrological problems (Chow et al., 1988), and have the ability to generate data with a large range of skewness and kurtosis, which are measures of the degree of non-normality (Bennett et al., 2013). The properties of each distribution are given in Tables 1 and 2. For each distribution, an additional 25 data points were generated for each of the exogenous inputs in the time series models, as the first 25 points were rejected in order to prevent initialisation effects (May et al.,

Table 1
Details of the simulated input distributions for the time series models (EAR4, TEAR10).

Distribution	Key parameters	s	k	Normality
NORM	Mean = 3.0; sd = 1.0	0.000	−0.013	High
GAMMA	Shape = 2.0; scale = 1.0	1.370	2.638	High
LOGN	Mean = 0.5; sd = 1.0	5.326	53.694	Low
EXP	Rate = 1.0	2.132	7.219	Moderate
PT3	Shape = 2.5; scale = 3.0; location = 2.0	1.251	2.381	High
LOGPT3	Shape = 0.5; scale = 0.2; location = 2.0	4.792	43.265	Low
EVT1	Shape = 0.0; scale = 0.5; location = 10.0	1.198	2.880	High

(Key parameters in the table are used to simulate the exogenous input variable; the skewness and kurtosis shown in the table are the averaged values of all input and output data).

Table 2
Details of the simulated input distributions for the nonlinear model (NL).

Distribution	Key parameters	s	k	Normality
NORM	Mean = 3.0; sd = 1.0	1.826	5.158	High
GAMMA	Shape = 2.0; scale = 1.0	10.520	192.091	Low
LOGN	Mean = 0.5; sd = 0.4	5.389	47.767	Low
EXP	Rate = 1.0	14.029	334.408	Low
PT3	Shape = 0.5; scale = 1.0; location = 0.5	16.271	514.270	Low
LOGPT3	Shape = 0.5; scale = 0.2; location = 0.5	14.261	390.522	Low
EVT1	Shape = 0.1; scale = 0.0; location = 10.0	1.788	9.807	Moderate

(Key parameters in the table are used to simulate each of the input variables; the skewness and kurtosis shown in the table are the averaged values of all input and output data).

2008). All data sets were split into training (60%), testing (20%) and validating sets (20%) using the DUPLEX method (see May et al., 2010), in accordance with the guidelines suggested by Wu et al. (2013).

The synthetic models used to produce the output data included a linear exogenous auto-regressive time series model (EAR4), a threshold exogenous auto-regressive time series model (TEAR10), and a nonlinear input–output function (NL) (see Fig. 2), as they represent relationships with increasing degrees of non-linearity and are based on synthetic models used in previous studies (May et al., 2008; Bowden et al., 2005a; Galelli and Castelletti, 2013). The equation for the linear exogenous auto-regressive time series of order four (EAR4) is given by

$$x_t = 0.6x_{t-1} - 0.4x_{t-4} + p_{t-1} + 0.1\varepsilon_t \quad (6)$$

where x_t is the output time series; x_{t-n} is the input time series with lag n; p_{t-n} is the exogenous input with lag n; and $0.1\varepsilon_t$ is the introduced error term. The equation for the nonlinear exogenous auto-regressive time series model of order ten (TEAR10) is given by

$$x_t = \begin{cases} -0.5x_{t-6} + 0.5x_{t-10} - 0.3p_{t-1} + 0.1\varepsilon_t; & x_{t-6} \le 0 \\ 0.8x_{t-10} - 0.3p_{t-1} + 0.1\varepsilon_t; & \text{otherwise} \end{cases} \quad (7)$$

and the equation for the nonlinear input–output function (NL) is given by

$$y = (x_2)^3 + x_6 + 5\sin(x_9) + 0.1\varepsilon_t \quad (8)$$

The first two synthetic models [Eqs. (6) and (7)] were modified versions of the synthetic models used in May et al. (2008) and the third synthetic model [Eq. (8)] was modified from the one used in Bowden et al. (2005a). For the first two synthetic models, the modifications include the introduction of an independent lagged input p_{t-1} into all exogenous AR models, and the p_{t-1} were sampled from the distributions outlined in Table 1. For the third synthetic model, the significance (coefficient) of each input was slightly modified and each input was sampled based on the distributions outlined in Table 2. In addition, the error term $0.1\varepsilon_t$ was added to all models to introduce noise into the models without obscuring the influence of the actual independent variables. The noise term ε_t followed the standard normal distribution $N(0,1)$.

3.1.2. Real case studies

In order to further test the impact of the degree of normality and non-linearity of the data on the predictive performance and computational efficiency of the different GRNN parameter estimation methods investigated, as well as the performance of the empirical guidelines for the selection of the most appropriate methods for GRNN smoothing parameter estimation developed based on the results from the synthetic data, two case studies with data with different degrees of normality and non-linearity were selected. The first case study was concerned with forecasting salinity in the River Murray in South Australia one, five and 14 days in advance and the second with the prediction of runoff in the Kentucky River basin in the USA one day in advance. The data division procedure used for both real case studies was identical to the one used for the synthetic case studies (see Section 3.1.1).

The salinity case has been studied extensively in the context of ANN modelling (e.g. Maier and Dandy, 1996; Maier and Dandy,

Table 3
Inputs and outputs used to forecast salinity at Murray Bridge 1, 5, & 14 days in advance.

Case no.	Inputs				Output			
	Location	Variable	Abbreviation	Lags	Location	Variable	Abbreviation	Forecasting period
1	Murray Bridge	Salinity	MBS	1	Murray Bridge	Salinity	MBS	1
	Mannum	Salinity	MAS	1				
2	Murray Bridge	Salinity	MBS	1	Murray Bridge	Salinity	MBS	5
	Mannum	Salinity	MAS	1				
3	Mannum	Salinity	MAS	1	Murray Bridge	Salinity	MBS	14
	Morgan	Salinity	MOS	1				
	Waikerie	Salinity	WAS	1, 5				
	Loxton	Salinity	LOS	1				
	Lock 7 Lower	Flow rate	L7F	1				
	Lock 1 Upper	River level	L1UL	1				

Table 4
Inputs and output used to model rainfall–runoff from the Kentucky River basin.

Inputs				Output			
Location	Variable	Abbreviation	Lags	Location	Variable	Abbreviation	Forecasting period
Manchester Hyden Jackson Heidelberg Lexington Airport	Mean daily effective rainfall	P	0,1,2	Lock & Dam 10	Mean daily runoff	Q	1
Lock & Dam 10	Mean daily runoff	Q	1,2				

2000; Bowden et al., 2005b; Kingston et al., 2005b; Fernando et al., 2009). According to Maier and Dandy (1996), salinity in the River Murray is a function of upstream inflows of salinity, flow, river level and groundwater level. Maier and Dandy (2000) also found that different combinations of inputs contribute to the output during different forecasting periods. In line with this finding, different GRNNs were developed in this study to predict salinity at Murray Bridge one, five and 14 days in advance (Table 3). Different input variables with different lags (Table 3) were associated with each output in a given forecasting period, where the inputs were selected from previous studies (e.g. Maier and Dandy, 1996; Maier and Dandy, 2000; Kingston et al., 2005b). All data covered the period 1987–1990, and were the same as the data used by Maier and Dandy (1996, 2000).

Analysis of the input data shows that the salinity based inputs are approximately normally distributed (average $s = -1.11$ & $k = 0.319$), although distributions of some lagged inputs have multiple peaks and the distribution of the water level based input is mildly non-Gaussian (average $s = 5.96$ & $k = 2.57$). According to Bowden (2003), the input and output data contain strongly linear components. Consequently, the data for this case study are close to mildly non-normal and the relationship to be modelled is close to linear.

The rainfall-runoff problem from the Kentucky River basin has also been extensively studied in the ANN literature (e.g. Jain and Srinivasulu, 2004; Srinivasulu and Jain, 2006; Bowden et al., 2012; Wu et al., 2013). The catchment area is approximately 10240 km^2 and the average daily total rainfall measurements come from five rain gauges located at Manchester, Hyden, Jackson, Heidelberg, and Lexington Airport. The average daily streamflow at Lock and Dam 10 are used as the output. Jain and Srinivasulu (2004) suggested five significant inputs [i.e. lagged effective rainfall $P(t)$, $P(t-1)$, $P(t-2)$ and lagged runoff $Q(t-1)$, $Q(t-2)$]. Therefore, the effective rainfall, with lags from the present day to two days prior, and the flow with lags of the first two days, were adopted as inputs (Table 4). The data used in this paper were identical to the 13 years of training data (1960–1972) utilised by Jain and Srinivasulu (2004).

Analysis of the input and output data shows that the distributions of lagged effective rainfall and flow are extremely non-Gaussian (averaged $s = 5.11$ & $k = 34.8$). Although the linearity of the rainfall–runoff problem in the Kentucky River basin has not previously been analysed, the general rainfall-runoff problem is well recognised as being highly nonlinear (e.g. Hu et al., 2001; Coulibaly et al., 2001; Dawson et al., 2002; Jain and Indurthy, 2003), and therefore the data are likely to contain a strong nonlinear structure. Consequently, the data for this case study are considered to be highly non-normal and the relationship to be modelled is likely to be highly non-linear.

3.2. Estimation of GRNN smoothing parameters using different estimation methods

The parameters for all of the GRNN models for the synthetic tests and real case studies were estimated using nine methods. Of these methods, five are adopted from the literature on kernel bandwidth selection for kernel density estimation, and four are based on single and multivariable calibration optimisation strategies. The methods adopted from the kernel density estimation literature are: the Gaussian reference rule (GRR); biased cross validation (BCV); 2-stage direct plug-in (DPI); a combination of BCV and DPI (BCVDPI); smoothed cross validation (SCV). The methods based on calibration optimisation strategies are as follows: single variable calibration with squared error as the objective function (SVCS); single variable calibration with mean absolute error as the objective function (SVCA); multi-variable calibration with squared error as the objective function (MVCS); and multi-variable calibration with mean absolute error as the objective function (MVCA) (Fig. 2). These methods were selected as they are based on different fitness functions and assumptions of normality and error basis, as shown in Table 5. Details of these smoothing parameter estimators are given in the following subsections.

3.2.1. Gaussian reference rule (GRR)

The GRR based smoothing parameter estimator is the most commonly used estimator. It is based on minimising the asymptotic

Table 5
Selected smoothing parameter estimators with different fitness functions and assumptions of normality and error basis.

Applied method	Fitness function	Dependence on		Sensitive to event	No. of smoothing parameters	Optimizer
		Normality	Error basis			
GRR	AMISE	High	Mean	Average	Single	None
BCV	AMISE	High	Mean	Average	Multiple	GSS
2-stage DPI	AMISE	Low	Mean	Average	Multiple	None
BCVDPI	AMISE	Low	Mean	Average	Multiple	GSS
SCV	EMISE	Low	Mean	Average	Multiple	GSS
SVC	MAE/RMSE	None	Mean/squared	Average/Extreme	Single	GSS
MVC	MAE/RMSE	None	Mean/squared	Average/Extreme	Multiple	PSO

(GSS refers to the golden section search algorithm (Press et al., 1992); PSO stands for the particle swarm optimisation algorithm (Poli et al., 2007); MAE is the mean absolute error; RMSE denotes the root mean squared error).

mean integrated squared error (AMISE) under the integrability assumption of an unknown probability function f of the given data (Scott, 1992; Wand and Jones, 1995). Under these assumptions, the derived AMISE has the expression

$$\mathrm{AMISE}\left\{\widehat{f}(\cdot;h)\right\} = (nh)^{-1}R(K) + \frac{1}{4}h^4\mu_2(K)^2\widetilde{R(f'')} \tag{9}$$

where K is the kernel function; $R(K) = \int [K(x)]^2 \mathrm{d}x$ is the integrated square of the kernel function; $\mu_2(K) = \int x^2 K(x)\mathrm{d}x$ is the second moment of K; and $\widetilde{R(f'')}$ represents the approximation of the integrated squared second derivative of f. By assuming that the data follow a Gaussian distribution, and adopting a Gaussian kernel, the GRR based smoothing parameter estimator that minimises the AMISE is derived as

$$\widehat{h}_{\mathrm{GRR},i} = \left(\frac{4}{m+2}\right)^{1/(m+4)} \sigma_i n^{-1/(m+4)} \tag{10}$$

where σ_i is the sample standard deviation of the X_i^j (usually standardised first). As outlined in Table 5, this approach depends heavily on the Gaussian assumption.

3.2.2. Biased cross validation (BCV)

As with the GRR, the BCV (Scott and Terrell, 1987) based smoothing parameter estimation method aims to minimise the AMISE, and is based on the assumption that the data are normally distributed. However, as the BCV is a combination of cross-validation and 'plug-in' bandwidth selection described by Wand and Jones (1995), it is potentially more robust than the GRR based approach through optimisation. The AMISE is expressed as follows by substituting the estimated $\widetilde{R(f'')}$ into Eq. (9)

$$\begin{aligned}\mathrm{AMISE}_{\mathrm{BCV},i}(h) =& (nh)^{-1}R(K) + \frac{1}{4}h^4\mu_2(K)^2 n^{-2}\sum\sum_{p\neq q} \\ &\times (K' \times K')\left(X_i^p - X_i^q\right)\end{aligned} \tag{11}$$

where * indicates the convolution operation. The BCV smoothing parameter is then given as

$$\widehat{h}_{\mathrm{BCV},i} = \arg\min_h\{\mathrm{AMISE}_{\mathrm{BCV},i}(h)\} \tag{12}$$

As illustrated in Table 5, the underlying assumptions for the estimator $\widehat{h}_{\mathrm{BCV},i}$ are similar to $\widehat{h}_{\mathrm{GRR},i}$ [Eq. (10)], however $\widehat{h}_{\mathrm{BCV},i}$ is determined by minimising the $\mathrm{AMISE}_{\mathrm{BCV},i}(h)$ through an optimisation process [in the current study, the golden section search (GSS) (Press et al., 1992) was used].

3.2.3. Two-stage direct plug-in (DPI)

The motivating idea behind the DPI (Park and Marron, 1992) is to approximate the unknown term $\widetilde{R(f'')}$ with $\widehat{\varphi}_r(g)$ [which is a pilot kernel estimation of the r-th order integrated squared density derivative; g is the pilot kernel bandwidth; L is the pilot kernel; and r is the stage number into Eq. (9)] to obtain a computable form for the asymptotically optimal bandwidth. By minimising AMISE [Eq. (9)] and replacing $\widetilde{R(f'')}$ with a pilot kernel bandwidth estimation $\widehat{\varphi}_4(g)$, the DPI based smoothing parameter expression, for each input dimension i, becomes

$$\widehat{h}_{\mathrm{DPI},i} = \left[\frac{R(K)}{[\mu_2(K)]^2\widehat{\varphi}_4(g)n}\right]^{1/5} \tag{13}$$

where $\widehat{\varphi}_4(g) = n^{-1}\sum_{i=1}^n \widehat{L}^{(4)}(X^i;g)$ represents the fourth order integrated squared density derivative, which is approximated by the pilot kernel L, with the corresponding pilot bandwidth as g (Hall and Marron, 1987; Jones and Sheather, 1991). The asymptotic mean squared error (AMSE) based optimal overall pilot bandwidth g is

$$g = \left[\frac{k!L^{(r)}(0)}{-\mu_k(L)\widehat{\varphi}_{r+k}n}\right]^{1/(r+k+1)} \tag{14}$$

where k is the order of the pilot kernel L; r is the stage number of L; $\mu_k(L) = \int u^k L(u)\mathrm{d}u$ is the k-th moment of L. The stage number r determines how many kernel estimations are required to approximate $\widehat{\varphi}_4(g)$ based upon the higher order integrated squared density derivative. Although it has been found that more stages can result in a better estimation when using the DPI, the improvement comes at a significant cost in terms of computational efficiency (Wand and Jones, 1995). The commonly suggested number of stages is $r = 2$ (Park and Marron, 1992), which was adopted in this study. For a 2-stage DPI based estimator, the corresponding fitness function and assumptions on linearity and error basis are identical to those for the GRR and BCV based approaches, while the dependence on the Gaussian assumption is effectively reduced by the pilot kernel based fourth order integrated squared density derivative, as shown in Table 5.

3.2.4. Combination of biased cross validation and two-stage direct plug-in (BCVDPI)

The BCVDPI estimator is a combination of the BCV and 2-stage DPI, and is achieved by replacing the estimated term $\widetilde{R(f'')}$ in Eq. (8) with the 2-stage DPI based $\widehat{\varphi}_4(g)$ as follows

$$\mathrm{AMISE}_{\mathrm{BCVDPI},i}(h) = (nh)^{-1}R(K) + \frac{1}{4}h^4\mu_2(K)^2\widehat{\varphi}_4(g)_{\mathrm{DPI}} \tag{15}$$

Although the BCVDPI has no closed form (it requires the solution of an optimisation problem), it inherits the positive attributes of a reduced dependence on the Gaussian assumption in comparison to the DPI. The optimal smoothing parameter by minimising $\mathrm{AMISE}_{\mathrm{BCVDPI},i}(h)$ can be expressed, for each input dimension i, as

$$\widehat{h}_{\mathrm{BCVDPI},i} = \arg\min_h\{\mathrm{AMISE}_{\mathrm{BCVDPI},i}(h)\} \tag{16}$$

The fitness function and assumptions of the BCVDPI based approach are identical to those of the 2-stage DPI approach. The main difference between these two approaches is that the former uses GSS based optimisation due to the biased cross-validation procedure, while the latter does not.

3.2.5. Smoothed cross validation (SCV)

The concept behind SCV is very similar to that underpinning the DPI approach, except that SCV attempts to minimise the exact MISE, rather than the AMISE [Eq. (9)] used in the DPI method. The MISE can also be approximated as

$$\mathrm{MISE}\left\{\widehat{f}(\cdot;h)\right\} \approx (nh)^{-1}R(K) + \int (K_h \times f - f)(x)^2\mathrm{d}x \tag{17}$$

By replacing $\int (K_h \times f - f)(x)^2\mathrm{d}x$ with $\widehat{\mathrm{ISB}}(h)$, where $\widehat{\mathrm{ISB}}(h)$ is an estimation of the integrated squared bias, Eq. (16) can be re-written as

$$\mathrm{EMISE}_{\mathrm{SCV},i}(h) = (nh)^{-1}R(K) + \widehat{\mathrm{ISB}}(h) \tag{18}$$

where $\widehat{\mathrm{ISB}}(h)$ is given by

$$\widehat{\mathrm{ISB}}(h) = n^{-2}\sum_{p=1}^{p=1}\sum_{q=1}^{n}\left(K_h \times K_h \times L_g \times L_g - 2 \times K_h \times L_g \times L_g + L_g \times L_g\right) \times \left(X_i^p - X_i^q\right) \tag{19}$$

where K_h and L_g are Gaussian kernels with kernel bandwidth h and pilot kernel bandwidth g, respectively (Hall et al., 1992; Wand and Jones, 1995). The pilot kernel bandwidth g is a function of a series of pilot kernel bandwidths, each estimated based upon sequentially higher order integrated squared density derivatives (Wand and Jones, 1995). The optimal smoothing parameter is determined by finding the parameter $\widehat{h}_{\mathrm{SCV},i}$, which minimises $\mathrm{EMISE}_{\mathrm{SCV},i}(h)$ through optimisation (GSS), as shown in Eq. (20) for the i-th input

$$\widehat{h}_{\mathrm{SCV},i} = \arg\min_h\left\{\mathrm{EMISE}_{\mathrm{SCV},i}(h)\right\} \tag{20}$$

Although the assumptions with regard to normality, linearity, and error basis of the SCV based method are very similar to those of the 2-stage DPI based approach (Table 5), the fitness function of the SCV method is based upon an exact, rather than asymptotic, estimation of MISE. Therefore, the predictive accuracy of SCV is expected to be the same as or better than that of the DPI approach (Wand and Jones, 1995).

3.2.6. Single variable calibration (SVC) and multi-variable calibration (MVC)

The most commonly applied trial and error approaches to bandwidth estimation can be classified as single variable calibration (SVC) and multi-variable calibration (MVC). The SVC estimator assumes that a common smoothing parameter is applicable to all input vectors, which increases computational efficiency compared with the MVC estimator, for which smoothing parameter estimates have to be obtained for each input vector, but at the cost of potential reductions in modelling accuracy and flexibility (Gibbs et al., 2006). The fitness function used to define the SVC and MVC estimators can be either extreme event oriented (e.g. squared error) or average event oriented (e.g. mean absolute error) (Dawson et al., 2007). The combination of different optimisation algorithms and modelling objectives results in four smoothing parameter estimators, namely SVCS, SVCA, MVCS, and MVCA. The mathematical formulations of these four estimators can be written as

$$\widehat{h}_{\mathrm{SVCS}} = \arg\min_h\left\{\sum_{j=1}^{n}\left[y^j - \widehat{y}\left(\boldsymbol{X}^j, h\right)\right]^2\right\} \tag{21}$$

$$\widehat{h}_{\mathrm{SVCA}} = \mathrm{argmin}_h\left\{\sum_{j=1}^{n}\left|y^j - \widehat{y}\left(\boldsymbol{X}^j, h\right)\right|\right\} \tag{22}$$

$$\widehat{\boldsymbol{h}}_{\mathrm{MVCS}} = \arg\min_{\boldsymbol{h}}\left\{\sum_{j=1}^{n}\left[y^j - \widehat{y}\left(\boldsymbol{X}^j, \boldsymbol{h}\right)\right]^2\right\} \tag{23}$$

$$\widehat{\boldsymbol{h}}_{\mathrm{MVCA}} = \arg\min_{\boldsymbol{h}}\left\{\sum_{j=1}^{n}\left|y^j - \widehat{y}\left(\boldsymbol{X}^j, \boldsymbol{h}\right)\right|\right\} \tag{24}$$

where $\widehat{y}\left(\boldsymbol{X}^j, \boldsymbol{h}\right)$ is the GRNN prediction based upon the bandwidth vector $\boldsymbol{h}=[h_1 \ldots h_m]^T$. The optimal single smoothing parameter in Eqs. (21) and (22) is achieved by minimising the errors (either squared errors or mean absolute errors) between the observed data y^j and the predictions $\widehat{y}\left(\boldsymbol{X}^j, h\right)$. In contrast, the optimal bandwidth matrix in Eqs. (23) and (24) is obtained by minimising the errors (either squared errors or mean absolute errors) between the observed records y^i and the predictions $\widehat{y}\left(\boldsymbol{X}^j, \boldsymbol{h}\right)$. Unlike the previous methods, the fitness functions of the SVC and MVC based approaches depend only upon the calibration error between observed and predicted output data. Consequently, these approaches are independent of Gaussian assumptions (Table 5). In this research, GSS was used to obtain the bandwidths of the SVC estimators, while the evolutionary strategy particle swarm optimisation (PSO) algorithm (Poli et al., 2007), which was written in Fortran, was used for this purpose for the MVC approaches.

3.3. Development of benchmark MLP model

In order to assess the performance of the different GRNN models in absolute terms, standard MLPs were developed as benchmarks using the systematic approach outlined in Wu et al. (2014). The model inputs/outputs and training, testing and validation data were identical to those used in the development of the GRNN models. A single hidden layer was used and the optimal number of hidden nodes was determined by trial and error, considering a range of 0–5. The optimal number of hidden nodes for the different models was as follows: 2 (EAR4), 2 (TEAR10), 3 (NL), 3 (river salinity 1 day), 3 (river salinity 5 day), 4 (river salinity 14 day), and 4 (flow 1 day), respectively. The back-propagation (BP) algorithm (with learning rate of 0.1 and momentum of 0.1) was used for calibration.

3.4. Model performance assessment

As mentioned in the Introduction and shown in Fig. 2, model performance criteria included predictive accuracy and computational efficiency. The specific measures adopted to assess these two aspects of performance are outlined in the subsequent sections.

3.4.1. Predictive accuracy

As discussed in Bennett et al. (2013), careful selection of appropriate predictive performance measures is extremely important. In this study, predictive accuracy was characterised by six dimensionless criteria (listed in Fig. 2), commonly used as evaluation metrics for hydrological prediction problems (Dawson et al., 2007; Krause et al., 2005; Bennett et al., 2013). These criteria include the coefficient of efficiency (CE), the index of agreement (IoAd), the persistence index (PI), and modified forms of CE, IoAd, and PI. These measures were chosen because: they are commonly used in hydrology; they have clear cut-off points to distinguish different extents of accuracy (good, satisfactory, or poor); and they are sensitive to different types of events, which assists performance characterisation with respect to the modelling objective. Particularly, CE compares the performance of the model to a model that only contains the mean of the observations; IoAd compares the sum of squared error to the potential error; and PI compares the sum of squared error to the error based on the predictions of previous observations (Bennett et al., 2013). In order to be able to assess the impact of the modelling objective on model performance, modified versions of these metrics were also used, in which squared error terms are replaced with absolute error terms (see Krause et al., 2005).

Although predictive accuracy was assessed using all of the six performance metrics mentioned above, only the performance

based on the averaged IoAd and modified IoAd (MIoAd) is presented in the body of the paper, while the performance based on the other metrics can be found in the Appendix (Figs. A.1, A.3, & A.5). IoAd is a measure of the overall agreement between the observed and modelled records, and is expressed as

$$\mathrm{IoAd} = 1 - \frac{\sum_{j=1}^{n}\left(y^j - \hat{y}^j\right)^2}{\sum_{j=1}^{n}\left(\left|\hat{y}^j - \bar{y}\right| + \left|y^j - \bar{y}\right|\right)^2} \tag{25}$$

where y^j is the individual observation, $\hat{y}^j$ is the corresponding approximation and $\bar{y}$ is the sample mean of the observations. IoAd is sensitive to the mean and variance differences between the observed and modelled records; however, it is insensitive to systematic positive or negative errors. Good performance corresponds to IoAd values greater than or equal to 0.9, and model performance with an IoAd less than 0.8 is considered to be poor (Dawson et al., 2007).

The adopted MIoAd is very similar to Eq. (25), except that the squared error terms are replaced by the absolute value in both the numerator and denominator, so that performance becomes average event, rather than extreme event, sensitive. Details of the derivations and applications of the MIoAd can be found in Krause et al. (2005).

The reason for detailing the sensitivity of the performance criteria to the average trends and extreme events is so that an assessment of the impact of the error basis of the fitness functions used by the different smoothing parameter estimators on the performance of the GRNN models with different modelling objectives can be made.

3.4.2. Computational efficiency

Computational efficiency was measured by computational time (CT) (measured by a dual processor 2.6 GHz Intel Machine), which was based on the average CPU clock speed (in seconds), as shown in Fig. 2.

3.5. Test regime

The test regime was implemented in accordance with Fig. 2. Overall, 630 synthetic data sets with 1,575,000 data points were generated, which consisted of 30 replicates of time series generated using 3 different models, for each of which input data were generated from 7 different distributions. Each of the 630 data sets was then divided into training, testing and validation sets and used to calibrate and validate 9 GRNN models, each using 1 of 9 different smoothing parameter estimation techniques, resulting in a total of 5670 GRNN models for the synthetic data. In addition to the experiments with the synthetic data, 4 experiments were conducted with the real data, 3 for the salinity data with different forecasting periods and 1 for the rainfall runoff data. MLPANNs were also developed for each of the 30 replicates of the synthetic data sets and for the 4 experiments with real data. As part of the model development process, the residuals of the training data of all GRNNs and MLPs were checked for replicative validity [see Appendix Figs. A.2, A.4 & A.6] in accordance with the recommendations of Wu et al. (2014). The residuals were generally 'white noise', indicating that all models can be considered replicatively valid. The performance of all 5674 models was assessed using the 6 selected predictive accuracy criteria, as well as computational time. Because of the large computational requirements, all tests were coded in PGI Visual Fortran 2008 and run on a Linux 2.6.32.2 operating system. The software used for conducting the numerical experiments is available for others to use, as per the details in the Software Availability at the beginning of this paper.

4. Results and discussion

4.1. Synthetic case studies

The predictive accuracy for the validation data and computational efficiency of all GRNN models for the synthetic data are summarised in Figs. 3 and 4, respectively. The key findings in relation to the impact of the degree of normality, the degree of non-linearity and the modelling objective on GRNN performance (predictive accuracy and computational efficiency) for the different smoothing parameter estimators are presented in Section 4.1.1, with the results of the comparison with the MLP benchmark models summarised in Section 4.1.2. Preliminary empirical guidelines for the selection of the most appropriate GRNN smoothing parameter estimator based on the properties of the data and the modelling objective derived from the results of the experiments on the synthetic data sets are presented in Section 4.1.3.

4.1.1. Performance of different smoothing parameter estimation methods

Overall, the results indicate that the predictive performance of the GRNN models reduces as the degree of non-Gaussianity in the data increases, especially when the GRR, BCV, DPI, BCDPI and SCV methods were used for smoothing parameter estimation. This suggests that the DPI (or BCVDPI) and SCV methods are not consistently effective in improving the predictive performance of GRNN models for non-Gaussian data compared with using the GRR, despite their reduced reliance on the normality assumption and their increased computational cost. In fact, in many instances, use of these parameter estimation methods resulted in a decrease in predictive performance compared with that obtained using the GRR, particularly for the more extreme distributions (i.e. LOGPT3, EXP, LOGN in Fig. 3).

In contrast, use of the SVCS/SVCA and MVCS/MVCA methods was generally successful in terms of improving the predictive performance of the GRNN models for data with high degrees of non-normality compared with the models for which the GRR was used for smoothing parameter estimation. In fact, when the SVCS/SVCA and MVCS/MVCA methods are used, there is very little degradation in predictive performance with an increase in the non-normality of the data. This is most likely because these smoothing parameter estimation techniques do not rely on any Gaussian assumptions. This makes use of the SVCS/SVCA approaches a particularly attractive option for highly non-Gaussian data, on account of their much smaller computational cost compared with the MVCS/MVCA methods.

While the trends described above apply to all three synthetic data sets, they manifest themselves more strongly for the non-linear (NL) case. This suggests that the combination of non-linear and non-Gaussian data has the potential to result in a marked degradation in the predictive performance of GRNNs, unless the SVCS/SVCA or MVCS/MVCA methods are used. It should also be noted that for the NL case, there was a noticeable improvement in predictive performance when the MVCS/MVCA approach was used instead of the SVCS/SVCA method. However, this improvement was achieved at a significantly increased computational cost.

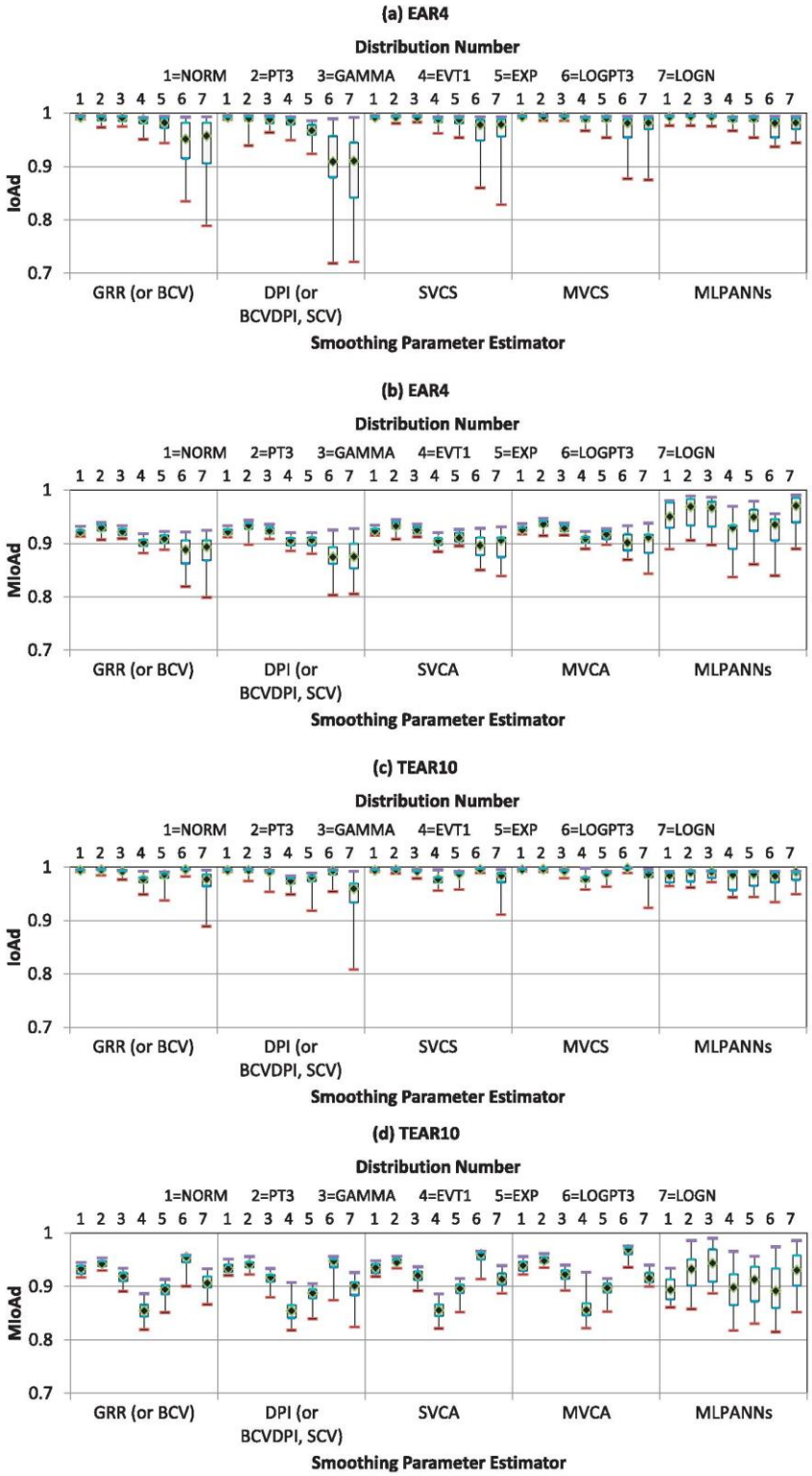


Fig. 3. Predictive accuracy for the validation data of MLPs and GRNNs for different synthetic data-generating models and distributions for which optimal parameters have been obtained using different methods.

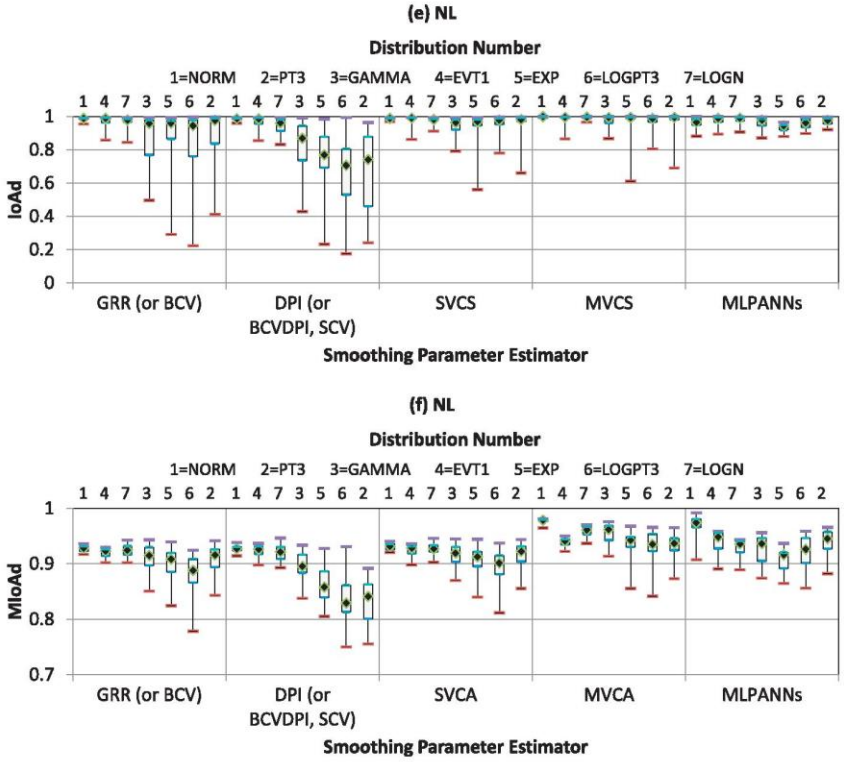


Fig. 3. (*continued*).

4.1.2. Comparison with MLP

In the vast majority of cases, the predictive performance of the MLP models was similar to that of the GRNN models for which the SVCS/SVCA and MVCS/MVCA methods were used for smoothing parameter estimation, although the MLPs performed slightly better than the best-performing GRNNs in some instances. In addition, for Gaussian or nearly Gaussian data, the predictive performance of the GRNNs for which the GRR was used for smoothing parameter estimation was very similar to that of the MLPs. Consequently, the results suggest that if a bandwidth estimation technique is used that is appropriate for the distribution of the data, the predictive performance of GRNNs is very similar to that of MLPs. In addition, this can generally be achieved at a much reduced computational cost, unless the MVCS/MVCA bandwidth estimation technique is used. Furthermore, use of GRNNs eliminates the uncertainty associated with the selection of an appropriate MLP model geometry.

4.1.3. Suggested rules and guidelines for use

Based on the findings of the 5670 computational experiments with the synthetically generated data, a set of preliminary empirical guidelines has been developed for selecting the most appropriate smoothing parameter estimation technique based on the degree of normality and degree of non-linearity of the data, as well as the modelling objective (Fig. 5). It should be noted that the smoothing parameter estimation techniques included in the suggested guidelines represent reasonable trade-offs between predictive accuracy and computational efficiency, although it is acknowledged that which trade-offs are optimal is also a function of case-study dependent circumstances and/or user preferences.

Based on Fig. 5, the preliminary empirical guidelines for selecting an appropriate method for estimating the parameter(s) of GRNNs can be grouped into a number of scenarios, as explained below:

Scenario 1: If the problem has input/output data that are mainly mildly non-Gaussian (average $s < 5$ & $k < 30$), the GRR (or BCV) smoothing parameter estimator is recommended, irrespective of linearity and model objective, as these methods are observed to provide good accuracy for these cases at a comparatively high computational efficiency.

Scenario 2: If (i) inputs and outputs are extremely non-Gaussian (average $s > 5$ & $k > 30$) and (ii) the modelling objective is to capture extreme events for a linear or non-linear problem, the use of SVCS or MVCS is suggested. However, this observed increase in predictive accuracy comes at the cost of significantly decreased computational efficiency (particularly for the MVCS).

Scenario 3: If the problem is as in Scenario 2 (extremely non-Gaussian data & linear or non-linear problem), but with a modelling objective that is average magnitude event sensitive, SVCA or MVCA should be adopted.

4.2. Real case studies

The results for the two real case studies are given in Figs. 6 and 7. Fig. 6 (a), (b), and (c) shows the predictive accuracy for the validation data of river salinity at Murray Bridge 1, 5, and 14 days in advance and the corresponding computational efficiency

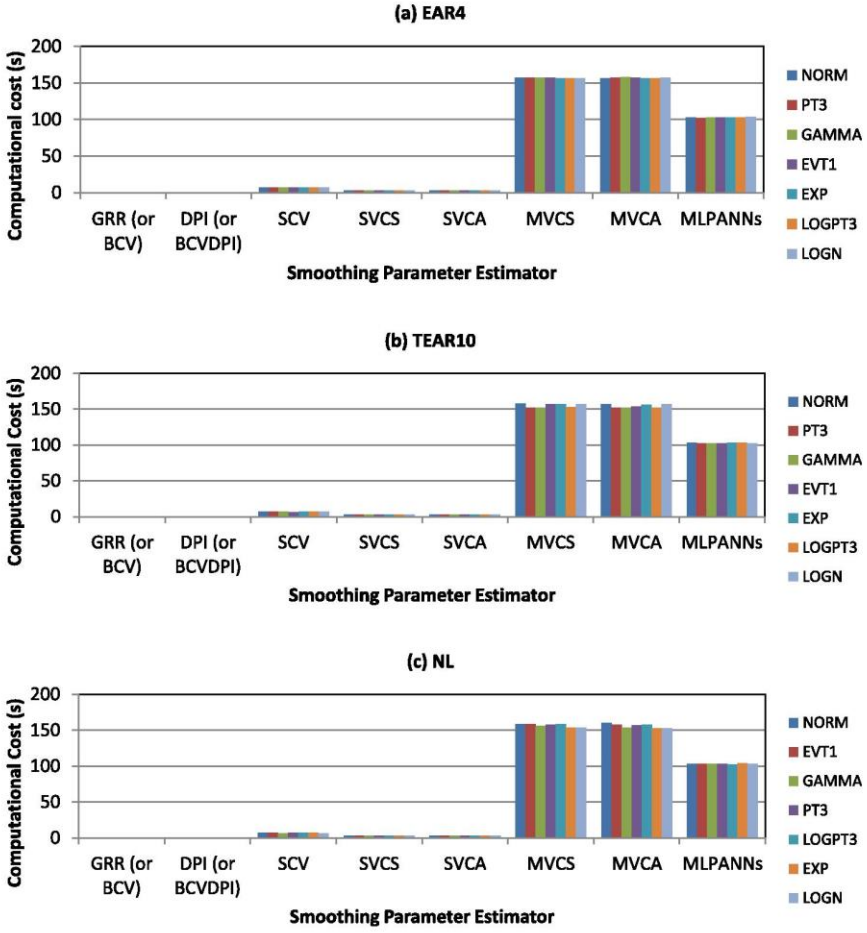


Fig. 4. Computational efficiency of MLPs and GRNNs for different synthetic data-generating models and distributions for which optimal parameters have been obtained using different methods.

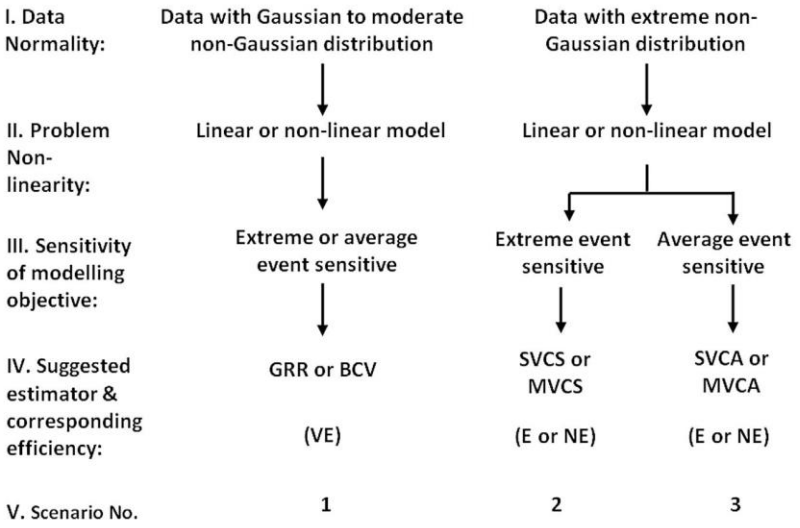


Fig. 5. Suggested smoothing parameter estimators under different problem situations (VE = comparatively very computationally efficient, E = comparatively moderately computationally efficient, and NE = comparatively not computationally efficient).

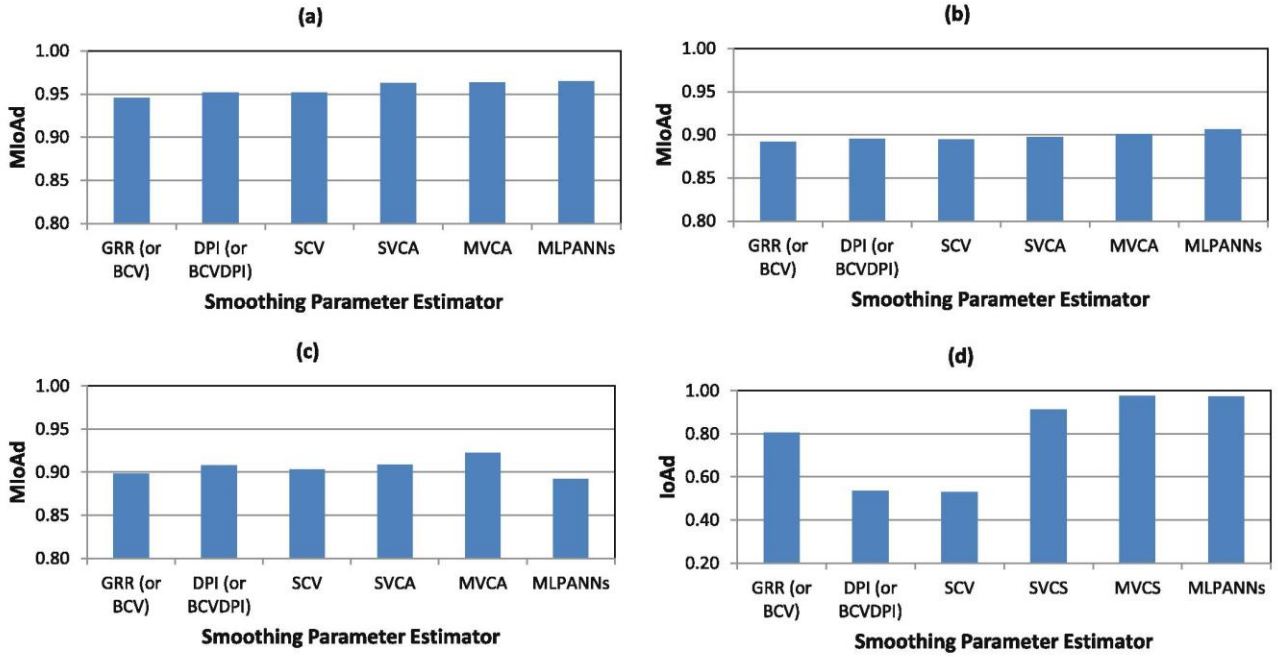


Fig. 6. Predictive accuracy of MLPs and GRNNs with different smoothing parameter estimators for the validation data for the real case studies ((a), (b), and (c): river salinity at Murray Bridge 1, 5, and 14 days in advance; (d): runoff at Lock and Dam 10 in the Kentucky River basin 1 day in advance).

is illustrated in Fig. 7 (a), (b), (c). Fig. 6 (d) displays the predictive accuracy for the validation data of runoff at Lock and Dam 10 in the Kentucky River basin 1 day in advance and the corresponding computational efficiency is given in Fig. 7 (d).

4.3. River salinity at Murray Bridge

By considering the properties of the data for the salinity case study (Table 3), and the modelling objective of capturing the

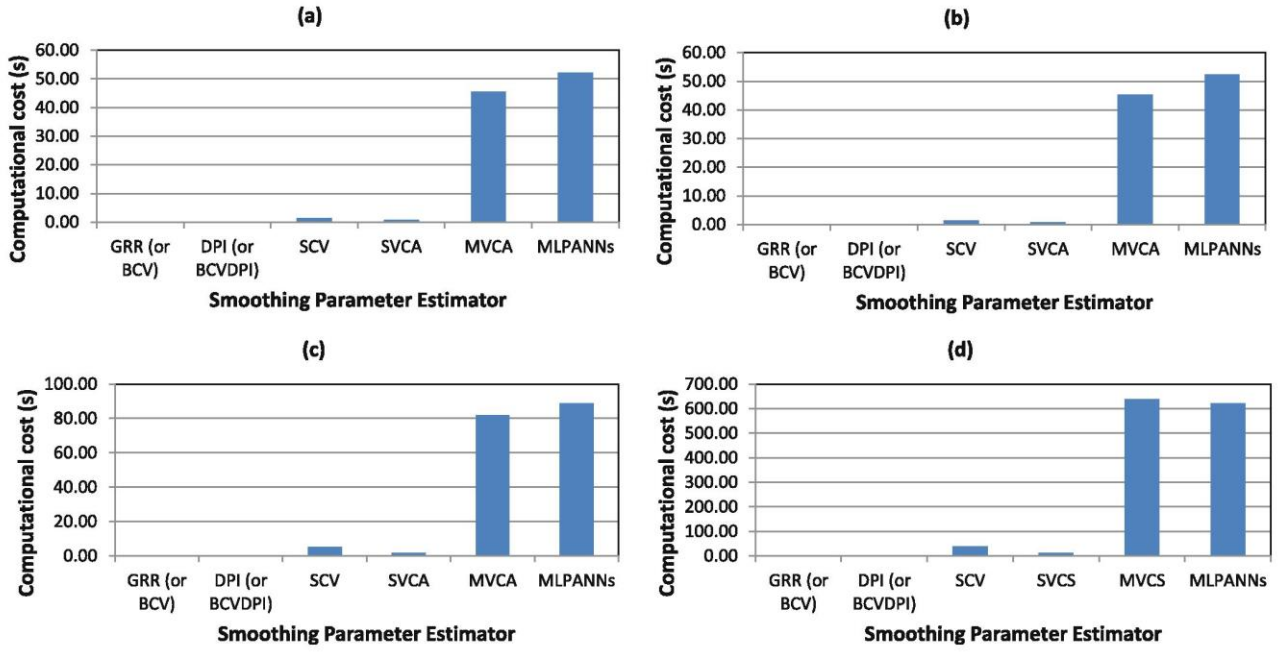


Fig. 7. Predictive efficiency of MLPs and GRNNs with different smoothing parameters for the validation data for the real case studies ((a), (b), and (c): river salinity at Murray Bridge 1, 5, and 14 days in advance; (d): runoff at Lock and Dam 10 in the Kentucky River basin 1 day in advance).

averaged salinity trends, this case study corresponds to Scenario 1 in Fig. 5. Given this, the predictive performance of the GRNNs developed using the GRR or BCV based methods was expected to be superior in terms of an appropriate trade-off between predictive accuracy and computational efficiency. This is confirmed by the results, which indicate that predictive performance was not affected significantly by using the different smoothing parameter estimation methods. Although the methods that have reduced reliance on the Gaussian assumption result in a slight increase in predictive performance, this is probably not outweighed by the additional computational costs incurred. However, as mentioned previously, the method that is considered most appropriate is case study and user dependent. For example, if high predictive accuracy was critical in this case and computational efficiency was not an issue, the MVCA based approach would be preferable. As was the case for the synthetic case studies, the predictive performance of the GRNNs is very similar to that of the MLPs, but at a significantly reduced computational cost.

4.4. Rainfall–runoff in Kentucky River basin

By considering the properties of the data for the rainfall-runoff case study (Table 4), and the modelling objective of capturing extreme events, this case study corresponds to Scenario 2 in Fig. 5. Given this, the predictive performance of the GRNNs developed using the SVCS and MVCS based methods was expected to be superior.

As shown in Fig. 6(d), the predictive performance of the GRNNs developed using the SVCS and MVCS based methods was indeed significantly better than that of the GRNNs developed using the other parameter estimation methods and was as good as that of the MLPs. In this case, the SVCS method provided the best trade-off between predictive accuracy and computational efficiency. However, if predictive accuracy was critical, the large increase in computational cost incurred [Fig. 7 (d)] for a small increase in predictive accuracy [Fig. 6 (d)] when using the MVCS method might be warranted.

5. Summary and conclusions

Artificial neural networks (ANNs) have been used extensively for hydrological and water resources modelling over the last two decades. In the vast majority of studies, multi-layer perceptrons (MLPs) have been used as the ANN model architecture. However, obtaining the optimal structure of such models is not an easy task. By using general regression neural networks (GRNNs) as the ANN model architecture, this problem can be overcome, as GRNNs have a fixed model structure. However, there has been limited investigation into the best way to estimate the parameters of GRNNs. In order to address this shortcoming, the performance of nine different GRNN parameter estimation methods was assessed in terms of accuracy and computational efficiency for data with distributions of varying degrees of normality and non-linearity on both synthetic and measured data. In addition, the impact of the objective function on model performance was assessed. In total, 5674 GRNN models were developed as part of the computational experiments conducted. As a way of benchmarking, the predictive performance and computational efficiency of the GRNN models was also compared with that of MLP models.

The main results from the synthetic case studies show that:

1. The predictive performance of GRNNs developed using the GRR, BCV, DPI, BCVDPI, and SCV based methods was generally influenced by the distribution of the input/output data because of their dependence on the Gaussian assumption (assuming the underlying density follows a normal distribution).
2. Compared to the GRNNs developed using the GRR, use of the DPI, BCVDPI, and SCV based methods did not effectively improve predictive performance, despite their decreased dependence on the Gaussian assumption and increased computational cost.
3. The predictive accuracy of GRNNs developed using the SVCA/SVCS and MVCA/MVCS based methods was relatively insensitive to the distribution of the input/output data because of their independence of the Gaussian assumption.
4. There is a distinct trade-off between predictive accuracy and computational efficiency for the methods investigated, with a reduction in computational efficiency for the methods that are least affected by the Gaussian assumption (i.e. SVCA/SVCS and MVCA/MVCS) by several orders of magnitude.
5. If an appropriate smoothing parameter estimation technique is used, the predictive performance of the GRNN models is very similar to that of the MLPANN models, although slightly worse in some instances. However, the computational cost of developing the GRNN models is generally significantly less. In addition, there is no uncertainty in relation to the selection of the most appropriate model structure.

Based on the general observations of the relationship between the performance of the different GRNN parameter estimation methods and the properties of the data and modelling objectives, preliminary empirical guidelines for selecting the GRNN parameter estimation method that represents good trade-offs between predictive accuracy and computational efficiency were developed.

The validity of the guidelines was tested and confirmed for two case studies with real data, including the forecasting of salinity in the River Murray in South Australia and a rainfall-runoff study in the Kentucky River basin in the USA.

While the results of this study provide useful insights and guidance on the selection of appropriate parameter estimation methods for GRNNs, further research into the possibility of improving the predictive performance of some of the methods that rely on the Gaussian assumption to some degree is warranted, as these methods are much more computationally efficient than the methods that are found to perform well with extremely non-Gaussian data in this study. In particular, the stage number used in the DPI, BCVDPI, and SCV methods may not be sufficient to describe extreme distributions with data accumulated at the boundary and a long tail. The boundary issue (Scott, 1992; Karunamuni and Alberts, 2005), as another critical issue with the same importance as the bandwidth, needs to be studied further for problems that contain extreme data distributions.

Acknowledgements

This research was aided by the suggestions and the original code of GRNN from Dr. Rob May and Dr. Greer Humphrey.

Appendix

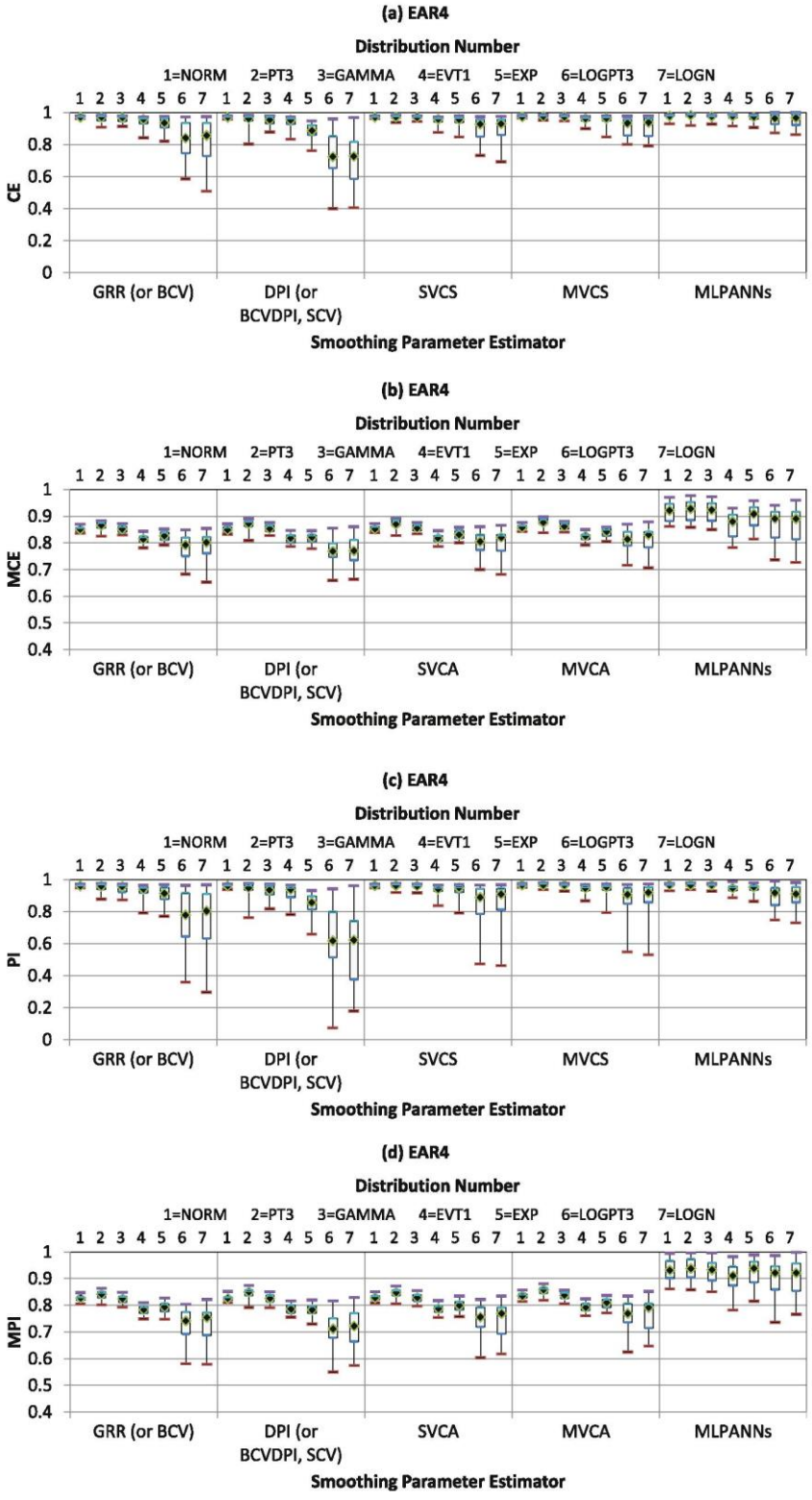
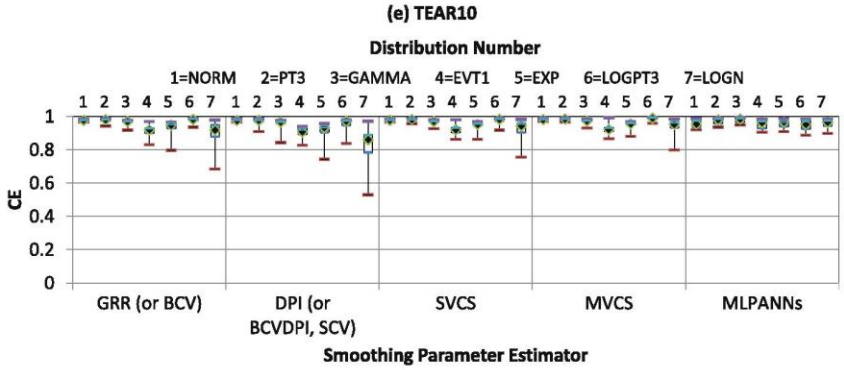
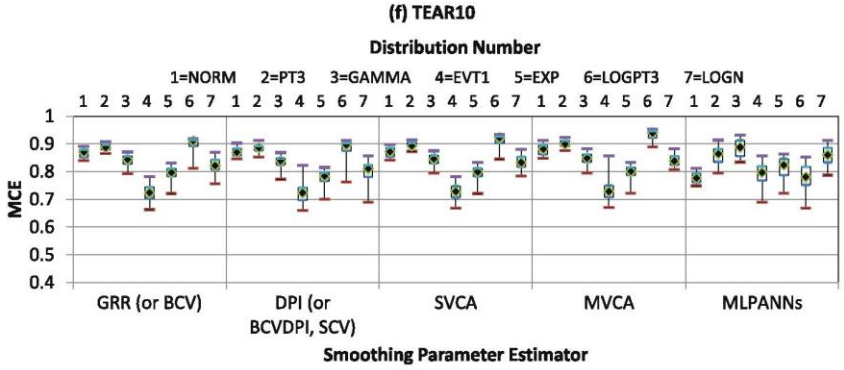
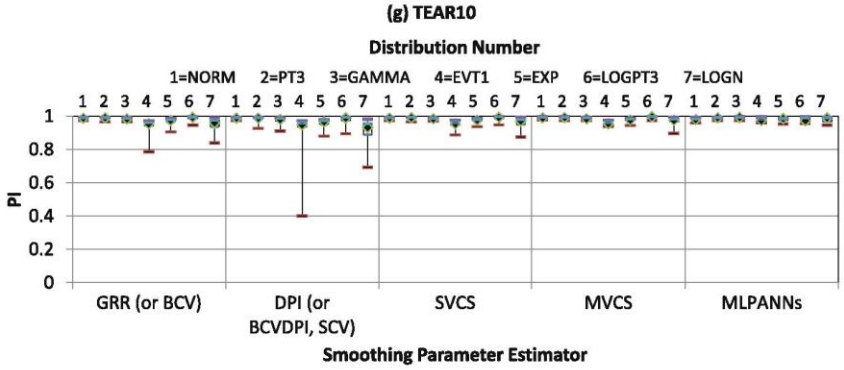


Fig. A.1. Predictive accuracy for the validation data of MLPs and GRNNs, measured by CE, MCE, PI & MPI, for different synthetic data-generating models and distributions for which optimal parameters have been obtained using different methods.







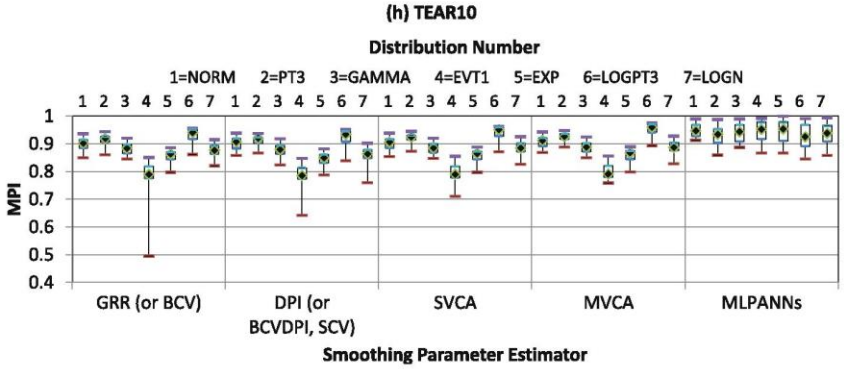
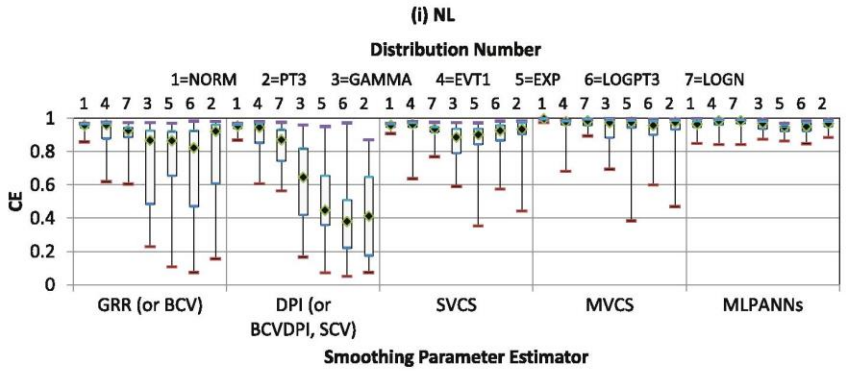
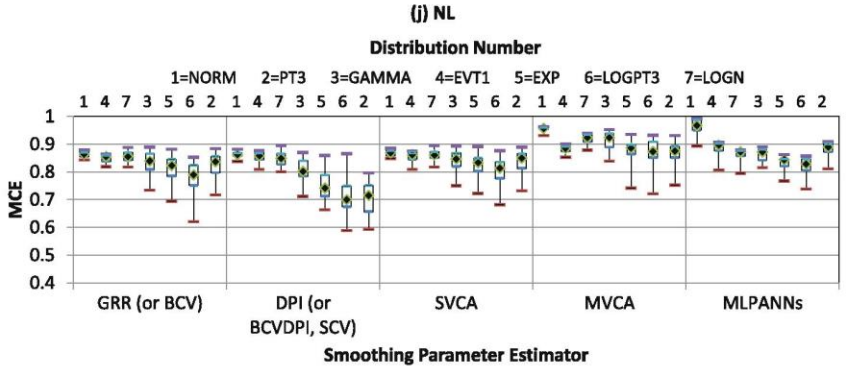
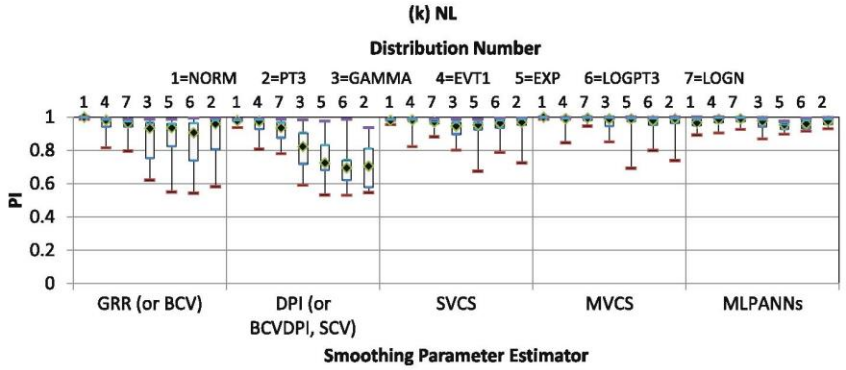


Fig. A.1. (*continued*).







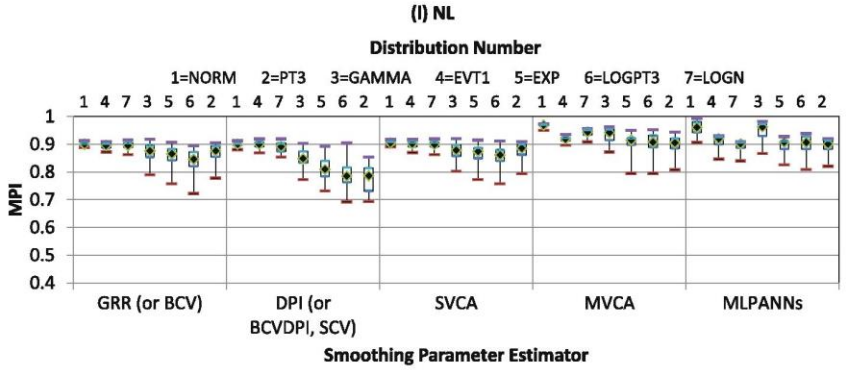


Fig. A.1. (*continued*).

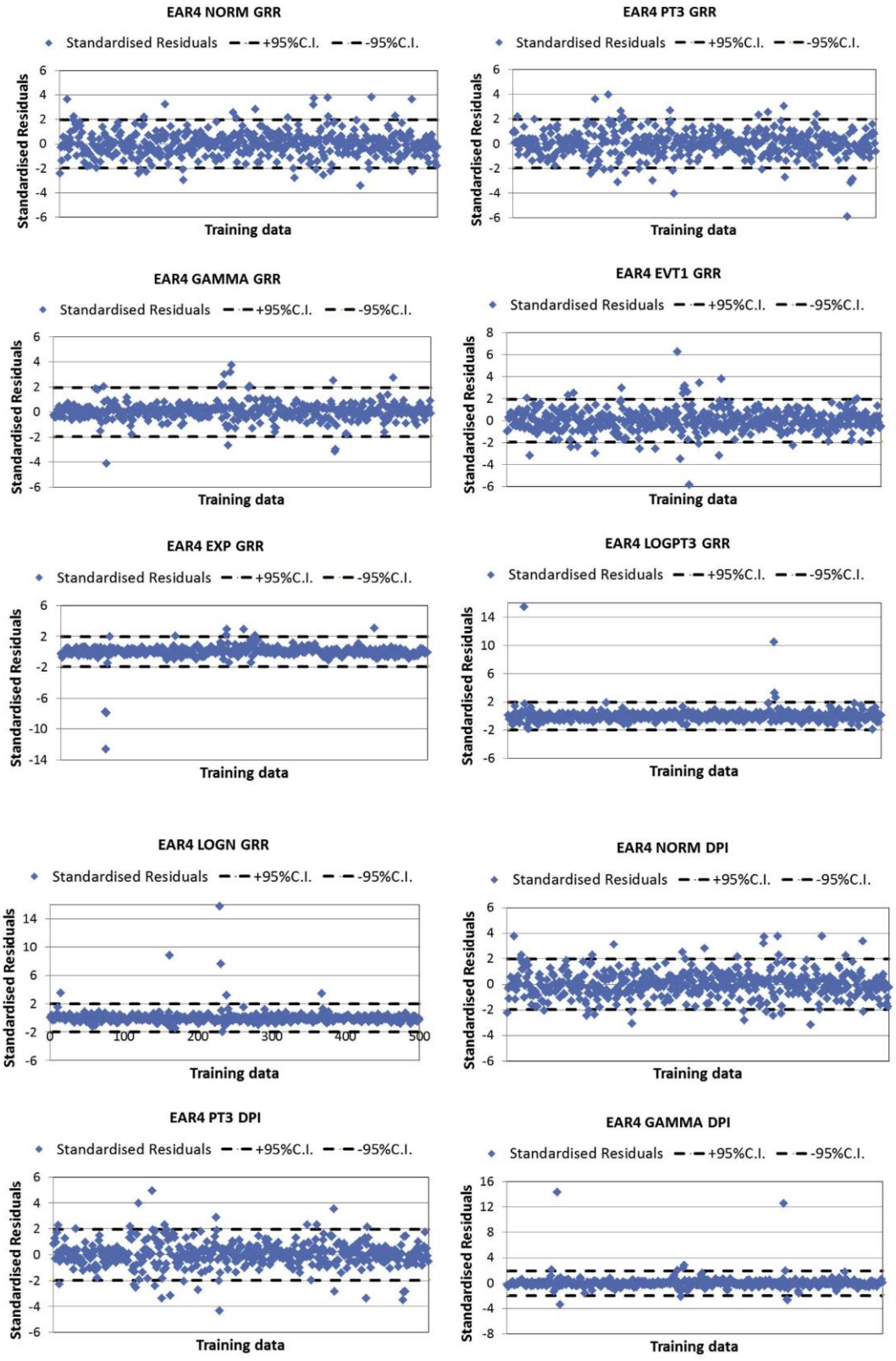


Fig. A.2. Standardised residuals for the training data of MLPs and GRNNs with different smoothing parameters for EAR4 model with different distributions (performance of the BCV was similar to that of the GRR; performance of the BCVDPI and SCV was similar to that of the DPI; similar plots were also observed for TEAR10 & NL models).

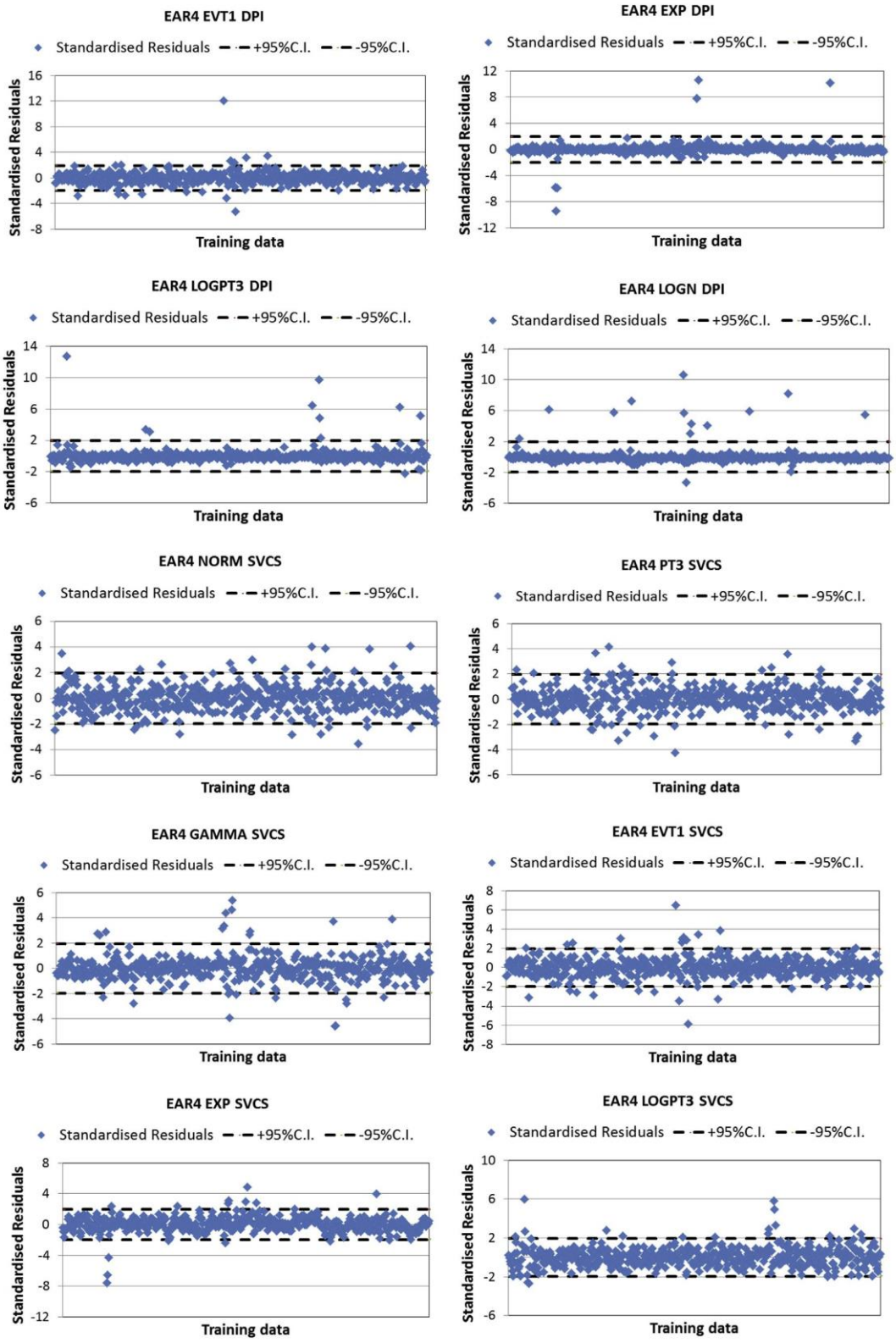


Fig. A.2. (*continued*).

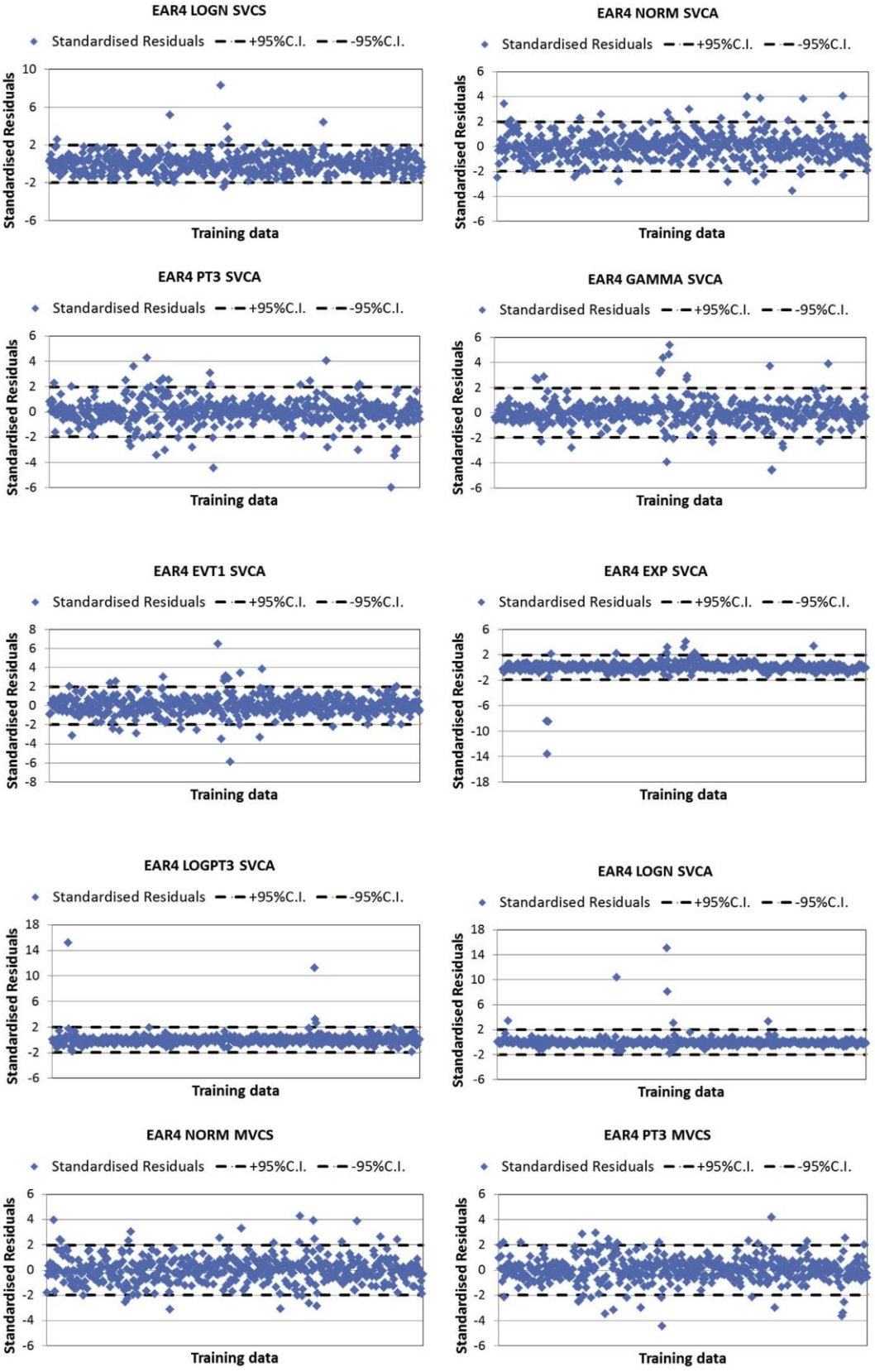


Fig. A.2. (*continued*).

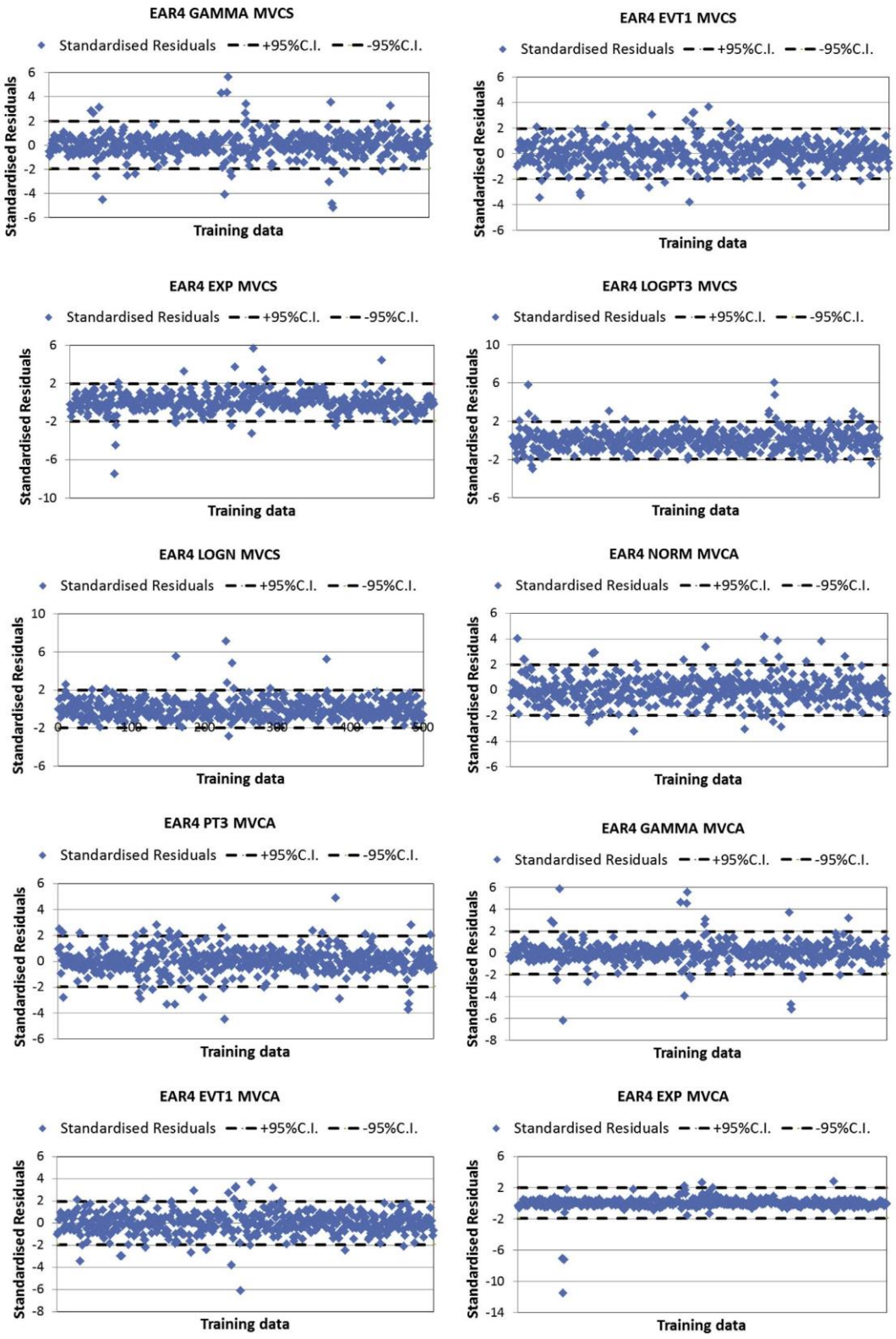


Fig. A.2. (*continued*).

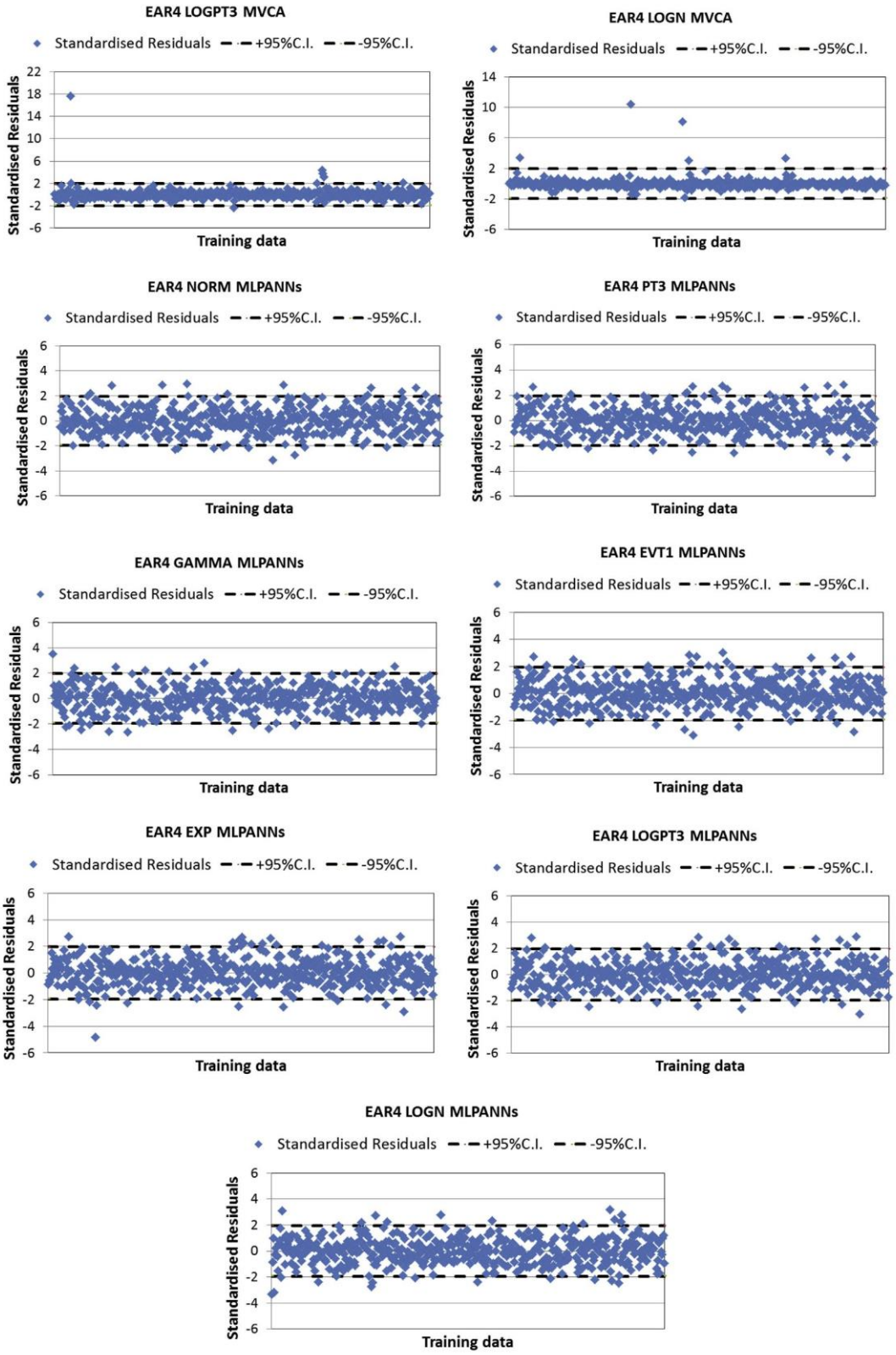


Fig. A.2. (*continued*).

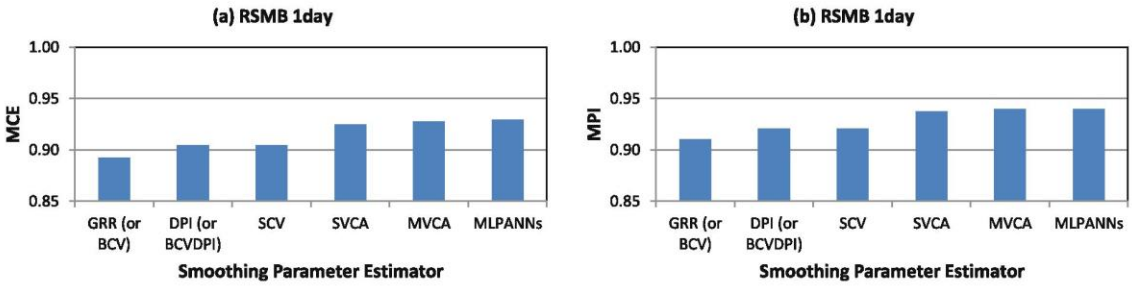


Fig. A.3. Predictive accuracy for the validation data of MLPs and GRNNs with different smoothing parameters for river salinity at Murray Bridge 1 day in advance (similar plots were also observed for 5 days & 14 days in advance).

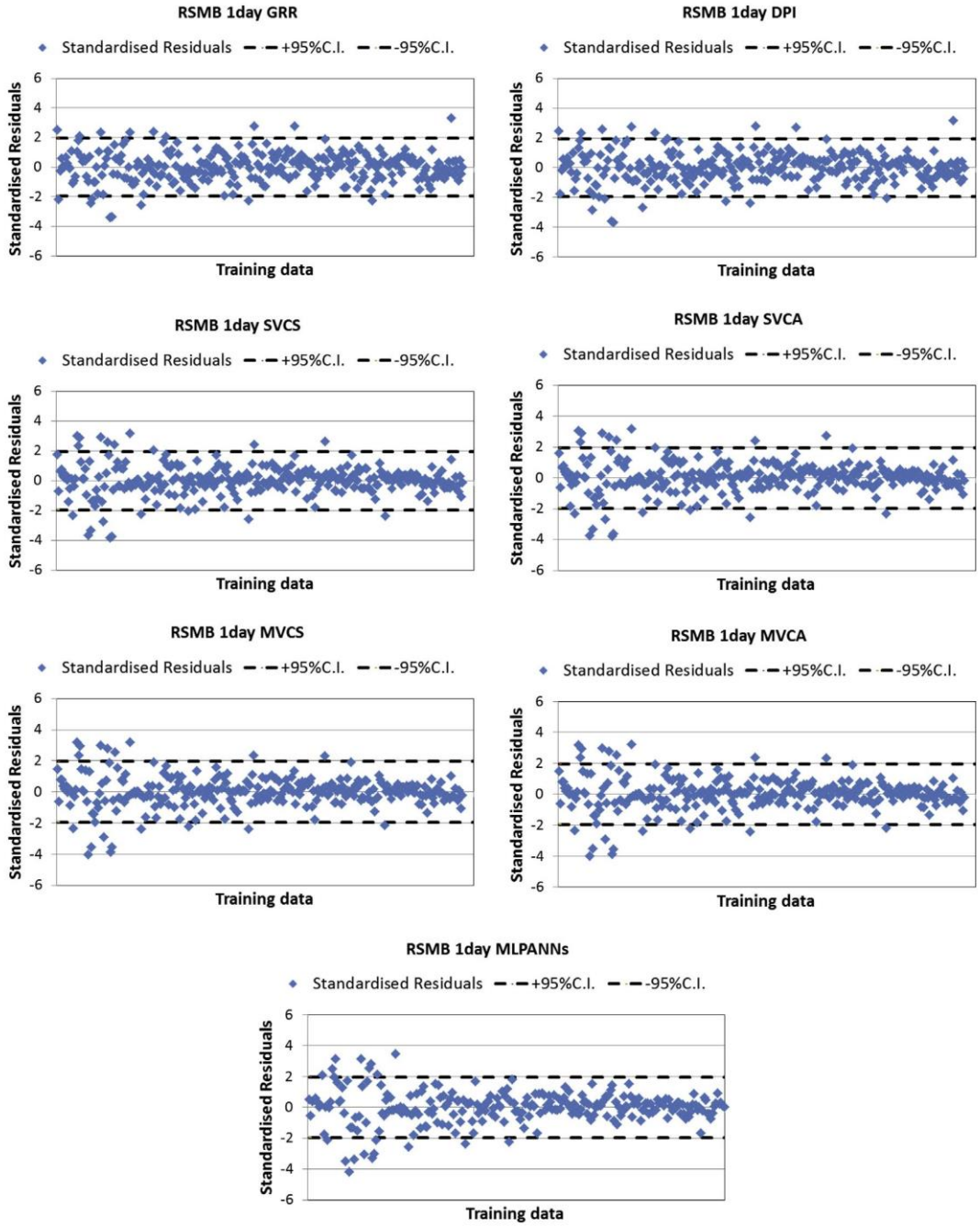


Fig. A.4. Standardised residuals for the training data of MLPs and GRNNs with different smoothing parameters for river salinity at Murray Bridge 1 day in advance (plots of the BCV were similar to those of the GRR; plots of the BCVDPI and SCV were similar to those of the DPI; similar plots were also observed for 5 days & 14 days in advance).

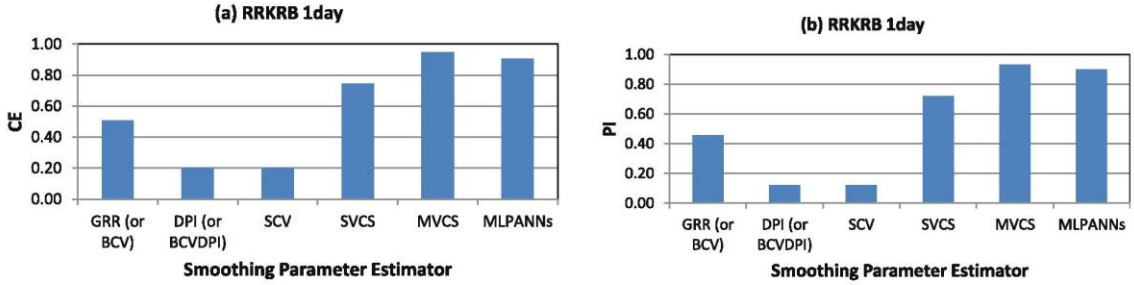


Fig. A.5. Predictive accuracy for the validation data of MLPs and GRNNs with different smoothing parameters for runoff at Lock and Dam 10 in the Kentucky River basin1 day in advance.

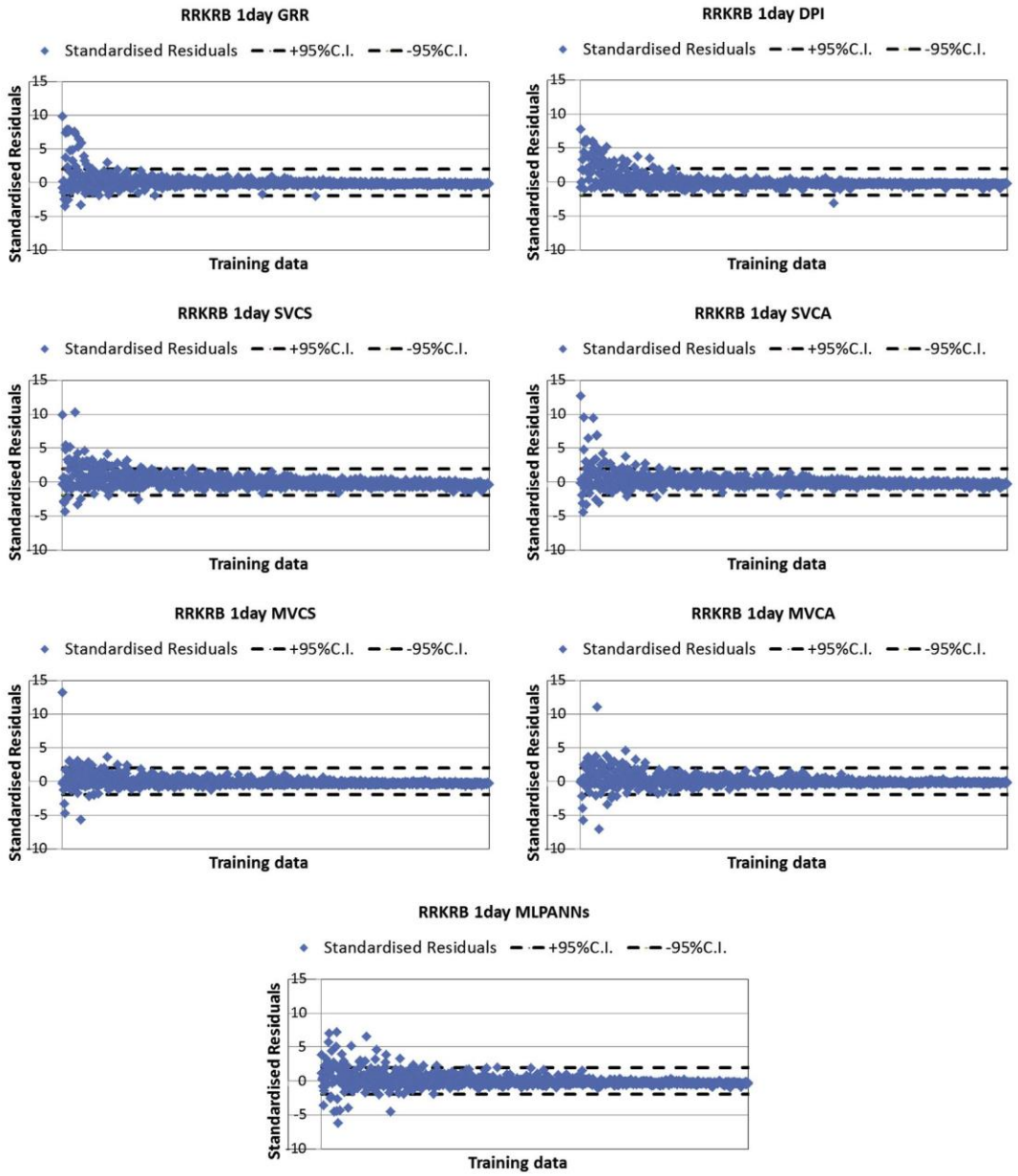


Fig. A.6. Standardised residuals for the training data of MLPs and GRNNs with different smoothing parameters for runoff at Lock and Dam 10 in the Kentucky River basin1 day in advance (plots of the BCV were similar to those of the GRR; plots of the BCVDPI and SCV were similar to those of the DPI).

References

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. Prog. Phys. Geogr. 36 (4), 480–513.

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. Environ. Model. Softw. 40, 1–20.

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 Resour. Res. 48 (10) http://dx.doi.org/10.1029/2012WR011984.

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. J. Hydrol. 301 (1–4), 75–92.

Bowden, G.J., Maier, H.R., Dandy, G.C., 2002. Optimal division of data for neural network models in water resources applications. Water Resour. Res. 38 (2), 2.1–2.11.

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. J. Hydrol. 301 (1–4), 93–107.

Bowden, G.J., Nixon, J.B., Dandy, G.C., Maier, H.R., Holmes, M., 2006. Forecasting chlorine residuals in a water distribution system using a general regression neural network. Math. Comput. Model. 44 (5), 469–484.

Bowman, A.W., 1984. An alternative method of cross-validation for the smoothing of density estimates. Biometrika 71 (2), 353–360.

Buhmann, M.D., 2003. Radial Basis Functions: Theory and Implementations, vol. 12. cambridge University Press.

Cacoullos, T., 1966. Estimation of a multivariate density. Ann. Inst. Stat. Math. 18 (1), 179–189.

Cai, Z., 2001. Weighted Nadaraya–Watson regression estimation. Stat. Prob. Lett. 51 (3), 307–318.

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. Adv. Eng. Softw. 37 (2), 63–68.

Coulibaly, P., Bobée, B., Anctil, F., 2001. Improving extreme hydrologic events forecasting using a new criterion for artificial neural network selection. Hydrol. Process. 15 (8), 1533–1536.

Dawson, C.W., Abrahart, R., See, L., 2007. HydroTest: a web-based toolbox of evaluation metrics for the standardised assessment of hydrological forecasts. Environ. Model. Softw. 22 (7), 1034–1052.

Dawson, C.W., Harpham, C., Wilby, R., Chen, Y., 2002. Evaluation of artificial neural network techniques for flow forecasting in the River Yangtze, China. Hydrol. Earth Syst. Sci. 6 (4), 619–626.

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. J. Hydrol. 367 (3–4), 165–176.

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

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. Math. Comput. Model. 44 (5), 485–498.

Hall, P., Marron, J.S., 1987. Estimation of integrated squared density derivatives. Stat. Probab. Lett. 6 (2), 109–115.

Hall, P., Marron, J., Park, B.U., 1992. Smoothed cross-validation. Probab. Theory Relat. Fields 92 (1), 1–20.

Hu, T., Lam, K., Ng, S., 2001. River flow time series prediction with a range-dependent neural network. Hydrol. Sci. J. 46 (5), 729–745.

Jain, A., Indurthy, S.K.V.P., 2003. Comparative analysis of event-based rainfall-runoff modeling techniques—deterministic, statistical, and artificial neural networks. J. Hydrol. Eng. 8 (2), 93–98.

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 Resour. Res. 40 (4), W04302.

Jones, M., Sheather, S., 1991. Using non-stochastic terms to advantage in kernel-based estimation of integrated squared density derivatives. Stat. Prob. Lett. 11 (6), 511–514.

Karunamuni, R.J., Alberts, T., 2005. On boundary correction in kernel density estimation. Stat. Methodol. 2 (3), 191–212.

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

Kingston, G.B., Maier, H.R., Lambert, M.F., 2005a. Calibration and validation of neural networks to ensure physically plausible hydrological modeling. J. Hydrol. 314 (1), 158–176.

Kingston, G.B., Maier, H.R., Lambert, M.F., 2008. Bayesian model selection applied to artificial neural networks used for water resources modeling. Water Resour. Res. 44 (4), W04419.

Krause, P., Boyle, D., Bäse, F., 2005. Comparison of different efficiency criteria for hydrological model assessment. Adv. Geosci. 5, 89–97.

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

Maier, H.R., Dandy, G.C., 1998a. The effect of internal parameters and geometry on the performance of back-propagation neural networks: an empirical study. Environ. Model. Softw. 13 (2), 193–209.

Maier, H.R., Dandy, G.C., 1998b. Understanding the behaviour and optimising the performance of back-propagation neural networks: an empirical study. Environ. Model. Softw. 13 (2), 179–191.

Maier, H.R., Dandy, G.C., 1999. Empirical comparison of various methods for training feedforward neural networks for salinity forecasting. Water Resour. Res. 35 (8), 2591–2596.

Maier, H.R., Dandy, G.C., 2000. Application of artificial neural networks to forecasting of surface water quality variables: issues, applications and challenges. In: Govindaraju, R.S., Rao, A.R. (Eds.), Artificial Neural Networks in Hydrology. Kluwer Academic Publishers, The Netherlands, pp. 595–605.

Maier, H.R., Jain, A., Dandy, G.C., Sudheer, K., 2010. Methods used for the development of neural networks for the prediction of water resource variables in river systems: current status and future directions. Environ. Model. Softw. 25 (8), 891–909.

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

May, R.J., Maier, H.R., Dandy, G.C., Fernando, T., 2008. Nonlinear variable selection for artificial neural networks using partial mutual information. Environ. Model. Softw. 23 (10), 1312–1326.

Park, B.U., Marron, J., 1992. On the use of pilot estimators in bandwidth selection. J. Nonparametric Stat. 1 (3), 231–240.

Park, B.U., Marron, J.S., 1990. Comparison of data-driven bandwidth selectors. J. Am. Stat. Assoc. 85 (409), 66–72.

Parzen, E., 1962. On estimation of a probability density function and mode. Ann. Math. Stat. 33 (3), 1065–1076.

Poli, R., Kennedy, J., Blackwell, T., 2007. Particle swarm optimization. Swarm Intell. 1 (1), 33–57.

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. Scand. J. Stat. 9 (2), 65–78.

Scott, D.W., Terrell, G.R., 1987. Biased and unbiased cross-validation in density estimation. J. Am. Stat. Assoc. 82 (400), 1131–1146.

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

Specht, D.F., 1990. Probabilistic neural networks. Neural Netw. 3 (1), 109–118.

Specht, D.F., 1991. A general regression neural network. Neural Netw. IEEE Trans. 2 (6), 568–576.

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

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

Williams, R.J., Zipser, D., 1989. A learning algorithm for continually running fully recurrent neural networks. Neural Comput. 1 (2), 270–280.

Wu, W., Dandy, G.C., Maier, H.R., 2014. 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. Environ. Model. Softw. 54, 108–127.

Wu, W., May, R.J., Dandy, G.C., Maier, H.R., 2012. A method for comparing data splitting approaches for developing hydrological ANN models. In: International Congress on Environmental Modelling and Software (6th: 2012: Leipzig, Germany) iEMSs 2012.

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 Resour. Res. 49 (11), 7598–7614.

Zhang, X., Liang, F., Yu, B., Zong, Z., 2011. Explicitly integrating parameter, input, and structure uncertainties into Bayesian neural networks for probabilistic hydrologic forecasting. J. Hydrol. 409 (3), 696–709.

D.2 Copy of Paper 2 from Chapter 3 (as published)

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 Software* 65 15-29 DOI: 10.1016/j.envsoft.2014.11.028

Environmental Modelling & Software 65 (2015) 15–29

Contents lists available at ScienceDirect

Environmental Modelling & Software

journal homepage: www.elsevier.com/locate/envsoft

Improved PMI-based input variable selection approach for artificial neural network and other data driven environmental and water resource models

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

School of Civil, Environmental and Mining Engineering, The University of Adelaide, Adelaide, South Australia 5005, Australia

ARTICLE INFO

Article history:
Received 27 June 2014
Received in revised form
27 November 2014
Accepted 28 November 2014
Available online

Keywords:
Artificial neural networks
General regression neural networks
Partial mutual information
Kernel bandwidth
Kernel density estimation
Environment
Hydrology and water resources
Input variable selection

ABSTRACT

Input variable selection (IVS) is one of the most important steps in the development of artificial neural network and other data driven environmental and water resources models. Partial mutual information (PMI) is one of the most promising approaches to IVS, but has the disadvantage of requiring kernel density estimates (KDEs) of the data to be obtained, which can become problematic when the data are non-normally distributed, as is often the case for environmental and water resources problems. In order to overcome this issue, preliminary guidelines for the selection of the most appropriate methods for obtaining the required KDEs are determined based on the results of 3780 trials using synthetic data with distributions of varying degrees of non-normality and six different KDE techniques. The validity of the guidelines is confirmed for two semi-real case studies developed based on the forecasting of river salinity and rainfall-runoff modelling problems.

1. Introduction

Artificial neural networks (ANNs) have been applied successfully and extensively to environmental (e.g. Adeloye et al., 2012; Ibarra-Berastegi et al., 2008; Luccarini et al., 2010; Maier and Dandy, 1997b; Maier et al., 2004; Millie et al., 2012; Muñoz-Mas et al., 2014; Ozkaya et al., 2007; Pradhan and Lee, 2010; Young et al., 2011) and water resources (e.g. 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., 2014) problems over the last two decades. One of the most important steps in the ANN model development process is the selection of appropriate inputs (e.g. Galelli et al., 2014; Humphrey et al., 2014; Maier et al., 2010; May et al., 2011, 2008b; Wu et al., 2014). According to Bowden et al. (2005a), if potential inputs that have a pronounced relationship with the modelled output are not included in the model, the performance of the resulting model will be compromised. Conversely, if redundant or superfluous inputs are included, computational efficiency is decreased, calibration becomes more difficult and model parameters are less well defined, potentially making model validation in terms of physical plausibility, as well as knowledge extraction, problematic (Dawson et al., 2014; Galelli et al., 2014; Haimi et al., 2013; Humphrey et al., 2014; Maier et al., 2010; May et al., 2011; Mount et al., 2013).

Given the importance and likely impact of input variable selection (IVS), it is somewhat surprising that in most studies, ad-hoc approaches are used (Maier et al., 2010; Wu et al., 2014). However, a number of quantitative approaches to IVS for ANN water resources models have already been developed and utilized, such as sensitivity analysis (Maier and Dandy, 1997a; Jain et al., 1999), the Gamma test (Agalbjörn et al., 1997; Noori et al., 2011), partial mutual information (PMI) (Bowden et al., 2005a), hybrid independent component analysis and input variable selection filter (Trappenberg et al., 2006), principal component analysis (Hu et al., 2007), use of the Box–Jenkins method (Box et al., 2013), cross-correlation analysis (Chua and Wong, 2010), distributed evaluation of local sensitivity analysis (Rakovec et al., 2014), recursive variable selection (RVS) embedded in dynamic emulation models

* Corresponding author. Tel.: +61 8 8313 1575; fax: +61 8 8303 4359.
E-mail addresses: xliadelaide@gmail.com, xli@civeng.adelaide.edu.au (X. Li), holger.maier@adelaide.edu.au (H.R. Maier), aaron.zecchin@adelaide.edu.au (A.C. Zecchin).

http://dx.doi.org/10.1016/j.envsoft.2014.11.028
1364-8152/

(Castelletti et al., 2012a,b), and tree-based iterative input variable selection (Galelli and Castelletti, 2013). Among these, PMI IVS is one of the most promising approaches, as it has a number of desirable properties, such as the ability to account for input relevance, the ability to cater to both linear and non-linear input–output relationships and the ability to determine the relative contribution (significance) of selected inputs (May, 2010). In addition, it has already been applied successfully to a number of studies (e.g. Bowden et al., 2005a,b; Fernando et al., 2009; He et al., 2011; May et al., 2008a,b; Wu et al., 2013).

However, current implementations of PMI IVS approaches are not without their limitations. Generally, kernel density estimation (KDE) is used to approximate the probability density function (PDF) needed for the calculation of MI (Bowden et al., 2005a,b; He et al., 2011; May et al., 2008a,b; Sharma, 2000a,b). One of the reasons for this is that simple methods exist for KDE that are a function of only a single parameter, the kernel bandwidth, otherwise termed the smoothing parameter (Scott, 1992; Wand and Jones, 1995). While many methods exist for estimating the bandwidth, in almost all existing PMI IVS studies dealing with environmental and water resources problems (e.g. Bowden et al., 2005a,b; May et al., 2008a,b; He et al., 2011) the Gaussian reference rule (GRR) is used for this purpose. The inherent limitation of this implementation of the PMI algorithm is that the input/output data are assumed to follow a Gaussian distribution. However, this is unlikely to be the case, as the distribution of most environmental and water resources data is generally far from normal. As a result, use of the GRR for determining the bandwidth for the KDE needed for MI estimation is likely to result in inaccurate IVS for data that are highly non-Gaussian (Galelli et al., 2014; Humphrey et al., 2014), and over-smoothed bandwidths have been found to result in more accurate MI estimates for such data (Harrold et al., 2001). Consequently, there is a need to investigate the effectiveness of alternative approaches to estimating the bandwidth in PMI IVS so that the performance of this commonly-used algorithm can be improved for data that follow non-Gaussian distributions.

In order to overcome the limitations of existing PMI IVS implementations outlined above, the objectives of the current study are: 1) to assess if, and to what degree, the performance of PMI IVS can be improved for data with different degrees of normality by using alternative bandwidth estimators with reduced reliance on the assumption that the data are normally distributed; and 2) to develop and test a set of preliminary guidelines for selecting the most appropriate bandwidth estimator for data with different degrees of normality. Consequently this paper makes a specific contribution in terms of improving the performance of the PMI algorithm for data that are encountered most commonly in practice.

The remainder of this paper is organised as follows. A detailed explanation of PMI IVS is provided in Section 2, followed by the methodology for meeting the objectives in Section 3. The results are presented and discussed in Section 4. The developed guidelines are validated on the semi-real studies in Section 5, before a summary and conclusions are given in Section 6.

2. PMI IVS

Although PMI IVS has been described in Sharma (2000a), Bowden et al. (2005a), May et al. (2008b, 2011), and He et al. (2011), the implementation of the KDE in 2-D used in this paper has not been explained clearly thus far in this field of research. Consequently, the overall procedure, mathematical details, and relevant assumptions of the PMI IVS algorithm implemented in this paper are discussed in detail below for the sake of completeness. As illustrated in Fig. 1, the *first step* is to procure candidate inputs **X** and output(s) y from the available data in accordance with an understanding of the system. Let: $\mathbf{X} = [X_1 \ldots X_m]^T$ be the input, where m is the number of inputs; $(\mathbf{X}^j, y^j)$ be the observed pairs of input and output data for $j = 1, \ldots, n$, where n is the number of observations, $\mathbf{X}^j = [X_1^j \ldots X_m^j]^T$ are the observed input data and y^j are the observed output data.

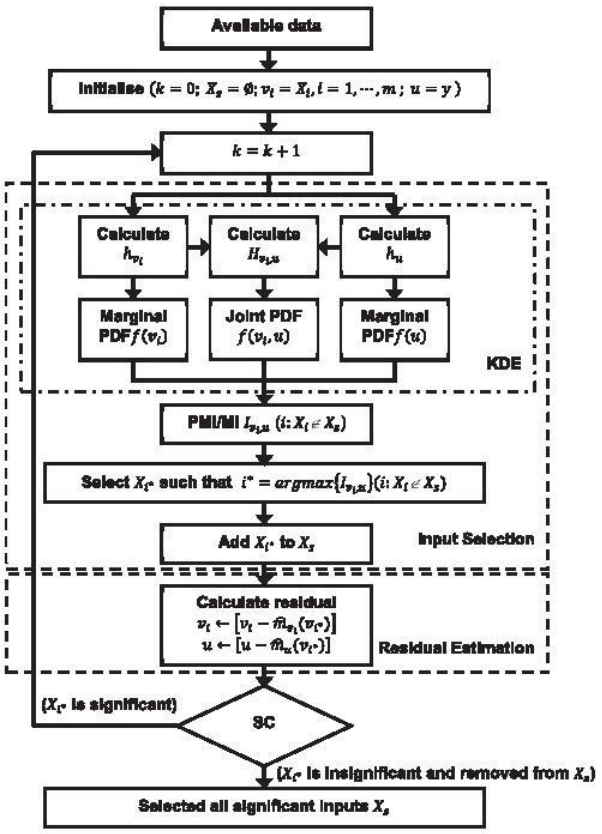


Fig. 1. Procedure of PMI IVS adopted in this study (the superscript is omitted, as all operations are performed over the input data $j = 1, \ldots, n$).

The *second step* is to estimate the marginal PDF of each individual input $f(X_i)$ and the output $f(y)$. The PDF is approximated by KDE in accordance with

$$\hat{f}(X_i) = \frac{1}{n}\sum_{j=1}^{n} K_h\left(X_i - X_i^j\right) \tag{1}$$

The kernel type K_h used in Eq. (1) is the most commonly used Gaussian kernel since the selection of kernel type has negligible impact on the accuracy of KDE (May et al., 2008b; Scott, 1992; Wand and Jones, 1995). The expression of the 1D Gaussian kernel is

$$K_h(X) = \frac{1}{\left(\sqrt{2\pi}|h|\right)}\exp\left(-\frac{X^2}{2h^2}\right) \tag{2}$$

In Eq. (2), h is the univariate kernel bandwidth, which determines the accuracy of the KDE (Duong and Hazelton, 2003; Scott, 1992; Wand and Jones, 1995). This single dimensional bandwidth, used for the marginal PDF estimation, directly contributes to the bandwidth matrix used for the joint PDF estimation (as explained later). As mentioned previously, in most studies, the GRR has been used for the estimation of the kernel bandwidth in PMI IVS due to its

high computational efficiency, ease of implementation, and reasonable stability (Bowden et al., 2005a; He et al., 2011; Huang and Chow, 2005; May et al., 2008b).

The *third step* is to calculate the joint PDF $f(X_i, y)$ between the i-th input and the output, which requires the development of a 2-D bandwidth matrix for the joint KDE. The currently used bivariate bandwidth matrix for standardised data is

$$\mathbf{H} = h_i^2 \begin{bmatrix} S_{x,i}^2 & S_{xy,i} \\ S_{xy,i} & S_y^2 \end{bmatrix} \tag{3}$$

where $S_{x,i}^2$ is the sample variance of the input X_i; $S_{xy,i}$ is the covariance between input X_i and output y, S_y^2 is the sample variance of the output y, and $h_i (h_i = h_{x,i} = h_y)$ is the estimated 1-D kernel bandwidth if the data are standardised, or for non-standardised data

$$\mathbf{H} = \begin{bmatrix} h_{x,i}^2 & \rho_{xy,i} h_{x,i} h_y \\ \rho_{xy,i} h_{x,i} h_y & h_y^2 \end{bmatrix} \tag{4}$$

(known as a hybrid class of bandwidth matrix), where $\rho_{xy,i}$ is the correlation coefficient between input X_i and output y. According to Wand and Jones (1993), the diagonal terms of the bandwidth matrix adjust the shape of the joint PDF, while the off-diagonal terms control the orientation. The empirical joint density of the i-th input X_i and the output y can be estimated by the Gaussian kernel-based estimator as

$$\hat{f}(X_i, y) = \frac{1}{n} \sum_{j=1}^{n} K_{\mathbf{H}} \left(\begin{bmatrix} X_i \\ y \end{bmatrix} - \begin{bmatrix} X_i^j \\ y^j \end{bmatrix} \right) \tag{5}$$

where the multivariate kernel is given by

$$K_{\mathbf{H}}(\mathbf{X}) = \frac{1}{\left(\sqrt{(2\pi)^m |\mathbf{H}|} \right)} \exp\left[-\frac{1}{2} \mathbf{X}^T \mathbf{H}^{-1} \mathbf{X} \right] \tag{6}$$

It should be noted that this approximation is commonly known as the Parzen window density estimation (Cacoullos, 1966; Parzen, 1962). This is valid, however, only if the underlying density is continuous and the first partial derivative at any $\mathbf{X}$ is small.

According to Shannon (1948), MI is then approximated as

$$I_{X_i,y} \approx \frac{1}{n} \sum_{j=1}^{n} \log \left[\frac{f\left(X_i^j, y^j\right)}{f\left(X_i^j\right) f\left(y^j\right)} \right] \tag{7}$$

(marginal PDFs $f(X_i^j)$ and $f(y^j)$ are as defined in Eq. (1)) in the *fourth step*. The input with the greatest MI value is the most significant input among the candidate inputs. The significant inputs are selected by means of these four steps during the first run of the algorithm and added to the significant input set X_s, that is, the set is updated to include $X_{i^*} \in X_s$ where $i^* = \text{argmax}\{I_{v_i,u}\}$.

In order to remove any redundant information, residual estimation is required in the *fifth step*. Residual estimation is at the core of the 'partial' aspect of PMI IVS and the mutual information shared between the residual inputs and output is called PMI (the term used after the 1st iteration of the PMI IVS). Typically, a general regression neural network (GRNN) (Specht, 1991) is used as the residual estimator in PMI IVS (e.g. May et al., 2008b; He et al., 2011). The residual estimator is used to update the inputs and output by removing the influence of the selected input variables. The updated input is defined as the difference between the current value of the unselected inputs v_i and the estimation of v_i based on the selected input X_{i^*} and is given by

$$v_i^j \leftarrow v_i^j - \hat{m}_{v_i}\left(X_{i^*}^j\right) \tag{8}$$

where $\hat{m}_{v_i}(X_{i^*}^j)$ is the residual estimate of v_i based on X_{i^*} which removes the shared information between the selected input $X_{i^*}^j$ and the remaining inputs v_i. Similarly, the updated output is

$$u^j \leftarrow u^j - \hat{m}_u\left(X_{i^*}^j\right) \tag{9}$$

where $\hat{m}_u(X_{i^*}^j)$ is the residual estimate of u based on X_{i^*}, which again eliminates the shared information between the selected inputs X_{i^*} and the output u.

The *sixth step* is to judge the selected input against the chosen stopping criterion. Potential stopping criteria include the bootstrapping, the tabulated critical values, the Akaike information criterion (AIC), and the Hampel test, as discussed and tested in May et al. (2008b). After updating the input and output variables based on the selected input variable, the corresponding PMI is estimated as

$$I_{v_i,u} \approx \frac{1}{n} \sum_{j=1}^{n} \log \left[\frac{f\left(v_i^j, u_i^j\right)}{f\left(v_i^j\right) f\left(u_i^j\right)} \right] \tag{10}$$

based on Eqs. (7) to (9). If the PMI value of the selected input is still significant according to the applied termination criterion, the above steps are repeated, as shown in Fig. 1, until all significant inputs X_s have been determined. In this way, the algorithm can accommodate a large number of potential input variables, as demonstrated in Fernando et al. (2009).

3. Methodology

The adopted procedure for assessing if, and to what degree, the performance of PMI IVS can be improved for data with different degrees of normality by using alternative bandwidth estimators is outlined in Fig. 2. This proposed approach contains three main steps: (i) generation of input/output data for a range of distributions (with different degrees of normality); (ii) estimation of the kernel PDF and MI for these data using a number of different kernel bandwidth estimators; (iii) assessment of the performance of the IVS process.

3.1. Generation of 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 with different degrees of normality were generated from distributions with differing degrees of normality, and the corresponding output data were obtained by substituting the generated inputs into synthetic models. Seven distinct distributions were used for input data generation, 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). The degree of normality of the input/output data was measured using skewness and kurtosis in accordance with Bennett et al. (2013). The properties of each distribution are listed in Tables 1 and 2. Although time series of different lengths (i.e. 500, 1000, and 2000) were considered in preliminary tests, their impact

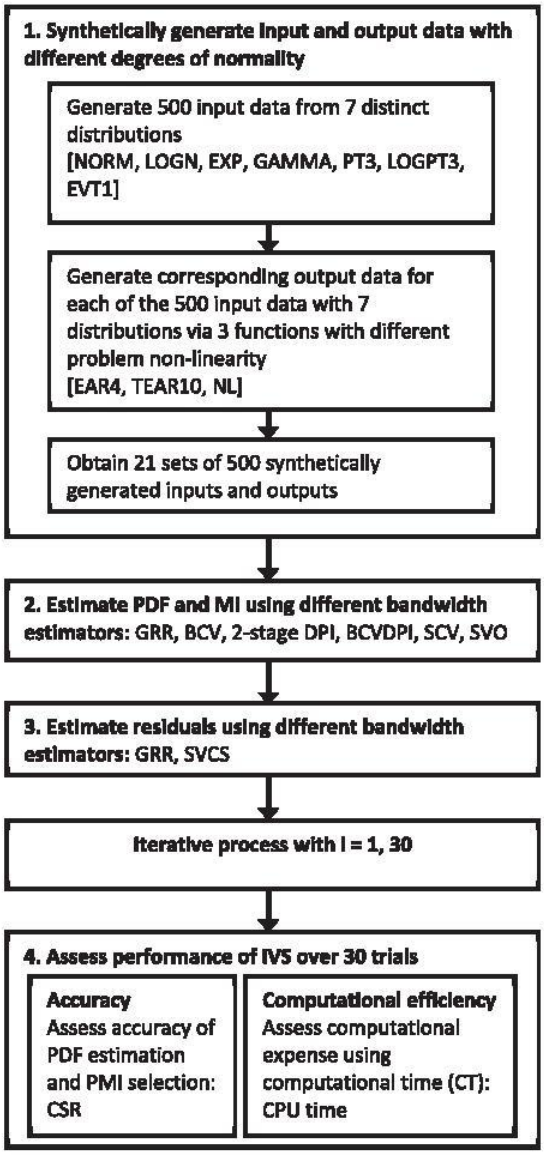


Fig. 2. Outline of the proposed experimental approach.

Table 1
Details 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
NORM	Mean = 3.0; sd = 1.0	0.000	−0.013	High
GAMMA	Shape = 2.0; scale = 1.0	1.370	2.638	High
LOGN	Mean = 0.5; sd = 1.0	5.326	53.694	Low
EXP	Rate = 1.0	2.132	7.219	Moderate
PT3	Shape = 2.5; scale = 3.0; location = 2.0	1.251	2.381	High
LOGPT3	Shape = 0.5; scale = 0.2; location = 2.0	4.792	43.265	Low
EVT1	Shape = 0.0; scale = 0.5; location = 10.0	1.198	2.880	High

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

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

Distribution	Key parameters	s	k	Normality
NORM	Mean = 3.0; sd = 1.0	1.826	5.158	High
GAMMA	Shape = 2.0; scale = 1.0	10.520	192.091	Low
LOGN	Mean = 0.5; sd = 0.4	5.389	47.767	Low
EXP	Rate = 1.0	14.029	334.408	Low
PT3	Shape = 0.5; scale = 1.0; location = 0.5	16.271	514.270	Low
LOGPT3	Shape = 0.5; scale = 0.2; location = 0.5	14.261	390.522	Low
EVT1	Shape = 0.1; scale = 0.0; location = 10.0	1.788	9.807	Moderate

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

on the results was found to be insignificant. Therefore 500 data points were generated and the first additional 25 points were rejected in order to prevent initialisation effects (May et al., 2008b).

The three synthetic models used for generating the known outputs, given a set of inputs, included a linear exogenous auto-regressive time series model (EAR4), a threshold exogenous auto-regressive time series model (TEAR10), and a non-linear input–output model (NL), as they are representative of general water engineering problem scenarios with increasing degrees of problem non-linearity and are based on those used for this purpose in previous studies (Bowden et al., 2005b; Galelli and Castelletti, 2013; Li et al., 2014; May et al., 2008b). The equation of the EAR4 model is given by

$$x_t = 0.6x_{t-1} - 0.4x_{t-4} + p_{t-1} + 0.1\varepsilon_t \tag{11}$$

where x_t stands for the output time series; x_{t-n} represents the input time series with lag n; p_{t-n} is the exogenous input with lag n; and $0.1\varepsilon_t$ is the introduced error term (as explained later). The equation for the TEAR10 model is given by

$$x_t = \begin{cases} -0.5x_{t-6} + 0.5x_{t-10} - 0.3p_{t-1} + 0.1\varepsilon_t; & x_{t-6} \le 0 \\ 0.8x_{t-10} - 0.3p_{t-1} + 0.1\varepsilon_t; & \text{otherwise} \end{cases} \tag{12}$$

and the equation for NL is given by

$$y = (x_2)^3 + x_6 + 5\sin(x_9) + 0.1\varepsilon_t \tag{13}$$

The first two synthetic models (Eqs. (11) and (12))were modified from those used in May et al. (2008b) through the introduction of an independent lagged input p_{t-1} into all exogenous AR models, and the p_{t-1} were sampled from the distributions outlined in Table 1. The third synthetic model (Eq. (13))was modified from the one used in Bowden et al. (2005a) through a slight adjustment of the significance (coefficient) of each input, and each input was sampled based on the distributions outlined in Table 2. For all three synthetic models, the error term $0.1\varepsilon_t$ was added to introduce noise without obscuring the influence of the actual independent variables. The noise term ε_t followed a standard normal distribution $N(0,1)$. In addition, for each synthetic model, 22 redundant or irrelevant input variables were included, so that the effectiveness of PMI IVS could be tested.

3.2. Estimation of PDF and MI using different bandwidth estimators

The kernel bandwidths used to estimate the PDF and MI for the synthetic and semi-real data sets were approximated by six different bandwidth estimators, including the Gaussian reference

rule (GRR), biased cross validation (BCV), 2-stage direct plug-in (DPI), a combination of BCV and DPI (BCVDPI), smoothed cross validation (SCV) and single variable optimisation (SVO) (Fig. 2). These bandwidth estimators were selected because they have distinct dependence on the Gaussian assumption. The mathematical details of each method are given in the following sections.

3.2.1. Gaussian reference rule (GRR)

As the most commonly used bandwidth estimator, the GRR is applied as the benchmark approach in this study. It approximates the bandwidth by minimising the asymptotic mean integrated squared error (AMISE) between the unknown probability function f of the given data and the KDE $\widehat{f}(\cdot;h)$ under the integrability assumption of f, in accordance with Scott (1992) and Wand and Jones (1995). The expression of AMISE is given as

$$\mathrm{AMISE}\left\{\widehat{f}\left(\cdot;h\right)\right\} = (nh)^{-1}R(K) + \frac{1}{4}h^4\mu_2(K)^2\widetilde{R(f'')} \tag{14}$$

where K is the kernel function; $R(K)$ is the integrated square of the kernel function; $\mu_2(K)$ is the second moment of K; and $\widetilde{R(f'')}$ is the integrated squared second derivative of f. According to Wand and Jones (1995), although it is ideal to determine the bandwidth by directly investigating the mean squared error (MSE) (summation of bias and variance), its expression depends on the bandwidth in a complicated way, which makes it difficult to interpret the impact of the bandwidth on the performance of the KDE. Consequently, AMISE was developed with consideration of the bias and the variance of the approximated kernel density function $\widehat{f}(\cdot;h)$ (assuming that the bandwidth approaches 0 at a rate slower than n^{-1} and K has a finite 4th moment and symmetry about origin) to overcome such issues and the optimal univariate bandwidth with respect to the AMISE can be derived as

$$\widehat{h}_{\mathrm{GRR},i} = \left(\frac{3}{4}\right)^{\frac{1}{5}}\sigma n^{\frac{-1}{5}} \tag{15}$$

by assuming that the data follow a Gaussian distribution and by adopting a Gaussian kernel. A detailed derivation of Eq. (15) is given in Wand and Jones (1995) and Scott (1992).

3.2.2. Biased cross validation (BCV)

Although the BCV based bandwidth estimator also minimises the AMISE, and depends on the Gaussian assumption through minimising the AMISE under the assumption of normally distributed data, it is a combination of a cross-validation and 'plug-in' approach, which is potentially more stable than the GRR (Scott and Terrell, 1987) as its asymptotic variance is considerably lower. The BCV is achieved via replacing the unknown $\widetilde{R(f'')}$ in Eq. (14) by a cross-validation kernel estimator $\widetilde{R(f'')} = n^{-2}\sum\sum_{p\neq q}(K''*K'')(X_i^p - X_i^q)$ and the optimal bandwidth is then determined by minimising the approximation of the AMISE with the cross-validation term. Therefore its expression is given as

$$\widehat{h}_{\mathrm{BCV},i} = \mathrm{argmin}_h\left\{(nh)^{-1}R(K) + \frac{1}{4}h^4\mu_2(K)^2n^{-2}\sum\sum_{p\neq q}\left(K''*K''\right) \times \left(X_i^p - X_i^q\right)\right\} \tag{16}$$

where K'' denotes the second derivative of kernel K and $*$ is the convolution operation and the golden section search (GSS) method (Press et al., 1992) was applied for the purpose of univariate optimisation in the current study. A detailed derivation of Eq. (16) is given in Wand and Jones (1995).

3.2.3. 2-Stage direct plug-in (DPI)

As with the GRR and BCV based approaches, the DPI estimates the optimal bandwidth by minimising the AMISE. For univariate KDE, the optimal bandwidth for Eq. (14) can be derived as $[(R(K))/(\mu_2(K)^2\widetilde{R(f'')}n)]^{\frac{1}{5}}$ in accordance with Wand and Jones (1995). The DPI is then established through replacing the unknown $\widetilde{R(f'')}$ in $[(R(K))/(\mu_2(K)^2\widetilde{R(f'')}n)]^{\frac{1}{5}}$ by a pilot kernel estimation of the r-th order integrated squared density derivative $\widehat{\varphi}_r(g)$ (where g is the pilot kernel bandwidth; L is the pilot kernel; and r is the stage number), according to Park and Marron (1992). Hence the univariate bandwidth estimator of DPI becomes

$$\widehat{h}_{\mathrm{DPI},i} = \left[\frac{R(K)}{\mu_2(K)^2\widehat{\varphi}_4\left(g\right)n}\right]^{\frac{1}{5}} \tag{17}$$

where $\widehat{\varphi}_4(g)$ is the fourth order integrated squared density derivative, which is approximated by the pilot kernel L with a pilot bandwidth g (Hall and Marron, 1987; Jones and Sheather, 1991). Although the pilot kernel L can be identical to the Gaussian kernel K, the pilot bandwidth g is estimated by minimising the asymptotic mean squared error (AMSE), resulting in

$$g = \left[\frac{K!L^{(r)}(0)}{-\mu_k(L)\widehat{\varphi}_{r+k}(g)n}\right]^{\frac{1}{r+k+1}} \tag{18}$$

where k represents the order of the pilot kernel L (normally $k = 2$); r is the stage number of L; and $\mu_k(L)$ is the k-th moment of L. Although the stage number r determines how many kernel estimations are required to approximate $\widehat{\varphi}_4(g)$ based upon the higher order integrated squared density derivative and more stages can result in a better estimation, determination of the optimal stage number is not trivial and there is a trade-off between an increase in accuracy and computational efficiency (Wand and Jones, 1995). Consequently, the stage number used for the current study was two, as suggested by Aldershof (1991) and Park and Marron (1992), which results in a desirable balance between the effectiveness and computational cost of the pilot kernel. The motivation behind the DPI is that the dependence of the Gaussian assumption is attenuated by introducing the pilot kernel estimation with $r > 0$, which makes the estimation more sensitive to the actual distribution. In fact, the GRR can be treated as a special case of the DPI with $r = 0$. A detailed derivation of Eqs. (17) and (18) can be found in Wand and Jones (1995).

3.2.4. Combination of BCV and DPI (BCVDPI)

The BCVDPI is simply a combination of the BCV and the DPI based approaches. The motivation behind this method is to maintain the advantage of low asymptotic variance in BCV, while adding the feature of reduced Gaussian dependence from the pilot kernel estimator used in DPI. Hence, the BCVDPI is implemented by replacing the cross-validation kernel estimator $n^{-2}\sum\sum_{p\neq q}(K''*K'')(X_i^p - X_i^q)$ in $\widehat{h}_{\mathrm{BCV},i}$ (Eq. (16))with the $\widehat{\varphi}_4(g)$ used in $\widehat{h}_{\mathrm{DPI},i}$ (Eq. (17)), resulting in the following expression.

$$\widehat{h}_{\mathrm{BCVDPI},i} = \mathrm{argmin}_h\left\{(nh)^{-1}R(K) + \frac{1}{4}h^4\mu_2(K)^2\widehat{\varphi}_4(g)_{\mathrm{DPI}}\right\} \tag{19}$$

As such, the BCVDPI inherits the reduced dependence on the Gaussian assumption from the 'plug-in' term $\widehat{\varphi}_4(g)$ and the optimal bandwidth is approximated by minimising the AMISE, which was obtained for the BCVDPI in this study by optimisation with the GSS.

3.2.5. Smoothed cross validation (SCV)

Although the concept behind the SCV based bandwidth estimator is similar to that underpinning the aforementioned four approaches, SCV aims to minimise the exact MISE (EMISE), rather than the AMISE used in the other four methods. The main difference between the EMISE and AMISE is that the former estimates MISE as a summation of the exact integrated squared bias and the approximation of the integrated variance of $\hat{f}(\cdot;h)$, while the later approximates MISE by integrating MSE (summation of bias and variance) with the integrability assumption and the asymptotic feature of the integrated squared bias. The EMISE derived for SCV is given as

$$\mathrm{EMISE}_{\mathrm{SCV},i}\left(h\right) = (nh)^{-1}R(K) + \widehat{\mathrm{ISB}}\left(h\right) \tag{20}$$

where the exact integrated squared bias $\widehat{\mathrm{ISB}}(h)$ is estimated by

$$\widehat{\mathrm{ISB}}(h) = n^{-2}\sum_{p=1}^{n}\sum_{q=1}^{n}\left(K_h{*}K_h{*}L_g{*}L_g - 2{*}K_h{*}L_g{*}L_g + L_g{*}L_g\right)\left(X_i^p - X_i^q\right) \tag{21}$$

where K_h and L_g are the Gaussian kernels with kernel bandwidth h and pilot kernel bandwidth g, respectively (Hall et al., 1992; Wand and Jones, 1995). g is a function of a series of pilot kernel bandwidths, each estimated based upon sequentially higher order integrated squared density derivatives, and up to the 10th order was applied in this study based on Wand and Jones (1995). The SCV based optimal univariate bandwidth is then determined as

$$\hat{h}_{\mathrm{SCV},i} = \mathrm{argmin}_h\left\{\mathrm{EMISE}_{\mathrm{SCV},i}(h)\right\} \tag{22}$$

A detailed derivation of Eq. (22) can be found in Wand and Jones (1995). Although the dependence on the Gaussian assumption of SCV is also reduced by introducing the pilot kernel estimation, which is similar to that of the DPI, the predictive accuracy of the former is expected to be the same as or better than that of the latter due to minimising EMISE, rather than AMISE.

3.2.6. Single variable optimisation (SVO)

Unlike the five estimators mentioned above, SVO, developed in this paper, determines the best bandwidth by minimising the Kolmogorov–Smirnov (K–S) statistic (Parsons and Wirsching, 1982) between the empirical and estimated cumulative density functions (CDFs). This method does not depend on the Gaussian assumption, nor the approximation of the MISE. The optimal univariate kernel bandwidth is determined as

$$\hat{h}_{\mathrm{SVO},i} = \mathrm{argmin}_h\left\{\sup_{j=1\ldots n}\left|F_{\mathrm{emp}}\left(X_i^j\right) - F_{\mathrm{est}}\left(X_i^j\right)\right|\right\} \tag{23}$$

where $F_{\mathrm{emp}}(X_i^j)$ is the empirical CDF of the input variable estimated by a histogram; $F_{\mathrm{est}}(X_i^j)$ is the estimated kernel-based CDF of the input variable; and *sup* represents the supremum function. The adopted optimiser was the GSS. The performance of the empirical histogram is a function of the histogram bin width, therefore a number of bin widths (from 0.001 to 1.0) were tested via sensitivity analysis. Although alternative ways can be used to estimate the histogram bin width for each case, the results of the sensitivity analysis (as shown in Appendix A Figs. A.4to A.6) suggest that a bin width of 0.01 was adequate for the purposes of this study.

It should be noted that the introduced kernel bandwidth estimators were implemented directly for the estimation of the univariate marginal PDF, which then extended to the bivariate joint PDF in conjunction with the bandwidth matrix, as mentioned in Section 2 (as in Eqs. (3)to (6)).

3.3. Performance assessment

As mentioned in the Introduction and described in Fig. 2, PMI performance was assessed based on selection accuracy and computational efficiency. Selection accuracy was characterised by the correct selection rate (CSR), which corresponds to the percentage of times the correct inputs are selected in the 30 independent trials with different instances of a particular data set, as was done in May et al. (2008b) and Galelli and Castelletti (2013). In addition, the degree of over- and under-estimation of the correct inputs was also assessed, in order to provide additional information on selection accuracy (see Galelli et al., 2014).

Computational efficiency was measured using the average CPU time (measured by a dual processor 2.6 GHz Intel Machine).

3.4. Test regime

The software used for conducting the numerical experiments was coded in Fortran 90/95 and run on a Linux 2.6.32.2 operating system. As outlined in Fig. 2, 630 synthetic data sets were generated, which consisted of a combination of 30 replicates, for each of the three synthetic models with input data generated from the seven distributions. For the 630 data sets, each of the 6 different kernel bandwidth estimators was used for KDE, resulting in a total of 3780 tests for the synthetic case studies.

The residual estimation required for PMI estimation (see Section 2) was carried out using a GRNN, as was the case in previous studies (e.g. Bowden et al., 2005a; May et al., 2008b; Fernando et al., 2009). The empirical guidelines proposed by Li et al. (2014) for identifying the most appropriate bandwidth estimation approach based on the distributional properties of the data were used in order to isolate the impact of different bandwidth estimators for residual estimation on IVS accuracy as much as possible. Details of the GRNN bandwidth estimators used for the different datasets resulting from the application of these empirical guidelines are given in Table 3.

The Akaike Information Criterion (AIC) (Akaike, 1974) was used as the stopping criterion (i.e. to decide when to stop adding inputs to the selected set) because it offers a trade-off between model accuracy and generalisation ability (Akaike, 1974; Bennett et al., 2013; Dawson et al., 2007; May et al., 2008b), has been found to perform well compared with alternative stopping criteria (May et al., 2008b) and has been successfully applied to a number of previous studies using PMI IVS (e.g. May et al., 2008a,b; He et al., 2011; Wu et al., 2013). The AIC stopping criterion for PMI IVS is computed as

$$\mathrm{AIC} = n \times \ln\left[\frac{1}{n}\sum_{j=1}^{n}\left(y^j - \hat{y}^j\right)^2\right] + 2k \tag{24}$$

where $\hat{y}^j$ denotes the estimated output and k is the number of effective inputs, measured by the trace of the $n \times n$ hat-matrix in KDE (May et al., 2008b). The performance of all 3780 synthetic tests was assessed against the performance criteria detailed in Section 3.3.

4. Results and discussion

Within the following, Section 4.1 focuses on assessing the selection accuracy of the PMI IVS methods with different bandwidth estimators applied to the synthetic data sets, and Section 4.2 focusses on computational efficiency. The empirical guidelines for

Table 3
GRNN bandwidth estimation techniques used for residual estimation during the PMI IVS process (based on the guidelines from Li et al., 2014).

Synthetic data set 1	EAR4						
Data distribution	NORM	EVT1	PT3	GAMMA	EXP	LOGN	LOGPT3
Bandwidth estimator	GRR	GRR	GRR	GRR	GRR	SVCS	SVCS
Synthetic data set 2	TEAR10						
Data distribution	NORM	EVT1	PT3	GAMMA	EXP	LOGN	LOGPT3
Bandwidth estimator	GRR	GRR	GRR	GRR	GRR	SVCS	SVCS
Synthetic data set 3	NL						
Data distribution	NORM	EVT1	LOGN	PT3	EXP	LOGPT3	GAMMA
Bandwidth estimator	GRR	GRR	SVCS	SVCS	SVCS	SVCS	SVCS

GRR denotes for Gaussian reference rule; SVCS stands for single variable calibration with squared error based fitness function.

the selection of the most appropriate bandwidth estimators for PMI IVS are presented in Section 4.3.

4.1. Selection accuracy

The accuracy of the PMI algorithm with alternative bandwidth estimators for the three synthetic models is summarised in Figs. 3 to 5. As can be seen from Fig. 3, for the EAR4 model, the use of alternative bandwidth estimators did not result in any significant improvement in CSR when the input/output data followed Gaussian or nearly Gaussian distributions (averages < 1.3 and $k < 3$; i.e. NORM, EVT1, and PT3). For instance, the CSRs when the GRR was used were all above 96.7% for the NORM, EVT1, and PT3 distributions, indicating very high selection accuracy. This result can be explained by the fact that the alternative bandwidth estimators did not provide a significant improvement in KDE accuracy compared with the GRR, as assessed using the Kolmogorov–Smirnov (K–S) statistic (Parsons and Wirsching, 1982), as shown in Figs. 6(a) to (c). This is not surprising, as the Gaussian assumption used in the KDE is consistent with the actual input/output data distributions, which resulted in an insignificant difference between the empirical and estimated CDFs (Figs. 6(a) to (c)). To better understand the causes for these findings, the predictive accuracy of the GRNN models used for residual estimation at each step of the PMI process was assessed using the coefficient of efficiency (CE) (Fig. 7), which measures the difference in predictive performance of the model and a model that only contains the mean of the observations (Bennett et al., 2013). As can be seen, the predictive accuracy of the GRNN models was very high, as indicated by CE values close to 1. Consequently, errors in residual estimation were unlikely to contribute to any inaccuracies in PMI IVS.

For data that were moderately non-Gaussian (average $1.3 < s < 5$ and $3 < k < 30$; i.e. GAMMA and EXP), the alternative bandwidth estimators (DPI, BCVDPI, SCV, and SVO) increased the CSR (Fig. 3). For example, for data following the EXP distribution, use of the GRR resulted in a CSR of 86.7%, whereas the CSRs for the alternative bandwidth estimators were much higher at 96.7% (SVO), 93.3% (SCV and DPI) and 90.0% (BCVDPI). As can be seen from Figs. 3, 6(e), and 6(f), the trend in improvement in CSR for the different bandwidth estimation techniques is matched by a similar trend in KDE accuracy, suggesting that the improved KDE has a direct impact on CSR. This is because the DPI, BCVDPI, SCV, and SVO based estimators have a reduced dependence on the assumption that the data follow a Gaussian distribution compared with the GRR. As was the case for the data that followed mildly non-Gaussian distributions, the accuracy of the GRNNs used for residual estimation was very high (Fig. 7), suggesting that the residual estimation step in the PMI process was unlikely to have any negative impact on CSR.

When the average distributions of the input/output data were extremely non-Gaussian (average $s > 5$ and $k > 30$; i.e. LOGN and LOGPT3), use of the alternate bandwidth estimators still resulted in a noticeable improvement in CSR (Fig. 3). However, this improvement was less pronounced for the most extreme distribution (LOGPT3), increasing CSR from 43.3% when the GRR was used to just over 60% when the DPI, BCVDPI, SCV and SVO were used. This is significantly lower than the CSR (over 90%) obtained for all other distributions. The reason for this is likely to be a combination of inaccuracy in KDE, as well as residual estimation. As can be seen in Fig. 6(g), although the use of SVO resulted in improved KDE, the K–S statistic is still outside the 95% confidence limits. In addition, there are significant errors in residual estimation, as shown in Fig. 7, even though the bandwidth estimator was based on the empirical guidelines suggested by Li et al. (2014). As seen in the LOGN and LOGPT3 boxplots in Fig. 7, despite the relatively high median, very low CE values were obtained for some of the 30 trials, which is likely to have a negative impact on CSR. These residual estimation inaccuracies are most likely caused by boundary issues (Scott, 1992; Karunamuni and Alberts, 2005), as discussed in Li et al. (2014), which occur when a symmetrical kernel is applied at a bounded and unsymmetrical boundary, resulting in an under-estimated density near the boundary.

It should also be noted that while the results suggest that improved accuracy in KDE results in improved PMI selection accuracy, consideration of the average ratio of the bandwidths of the

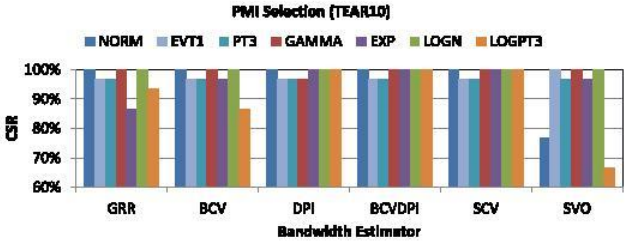


Fig. 4. Correct selection rate of TEAR10 model with alternative bandwidth estimators.

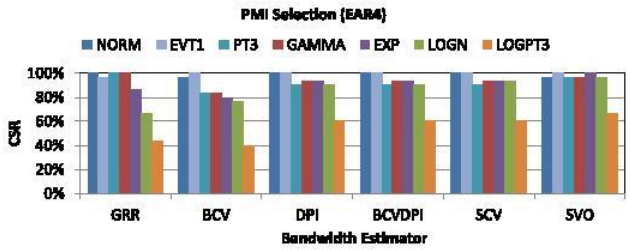


Fig. 3. Correct selection rate of EAR4 model with alternative bandwidth estimators.

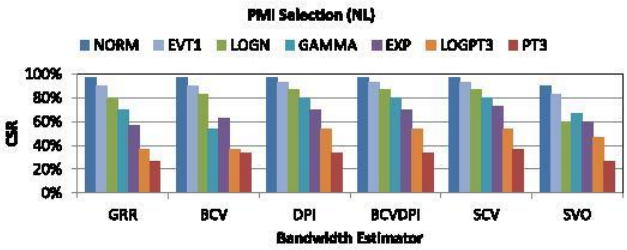


Fig. 5. Correct selection rate of NL model with alternative bandwidth estimators.

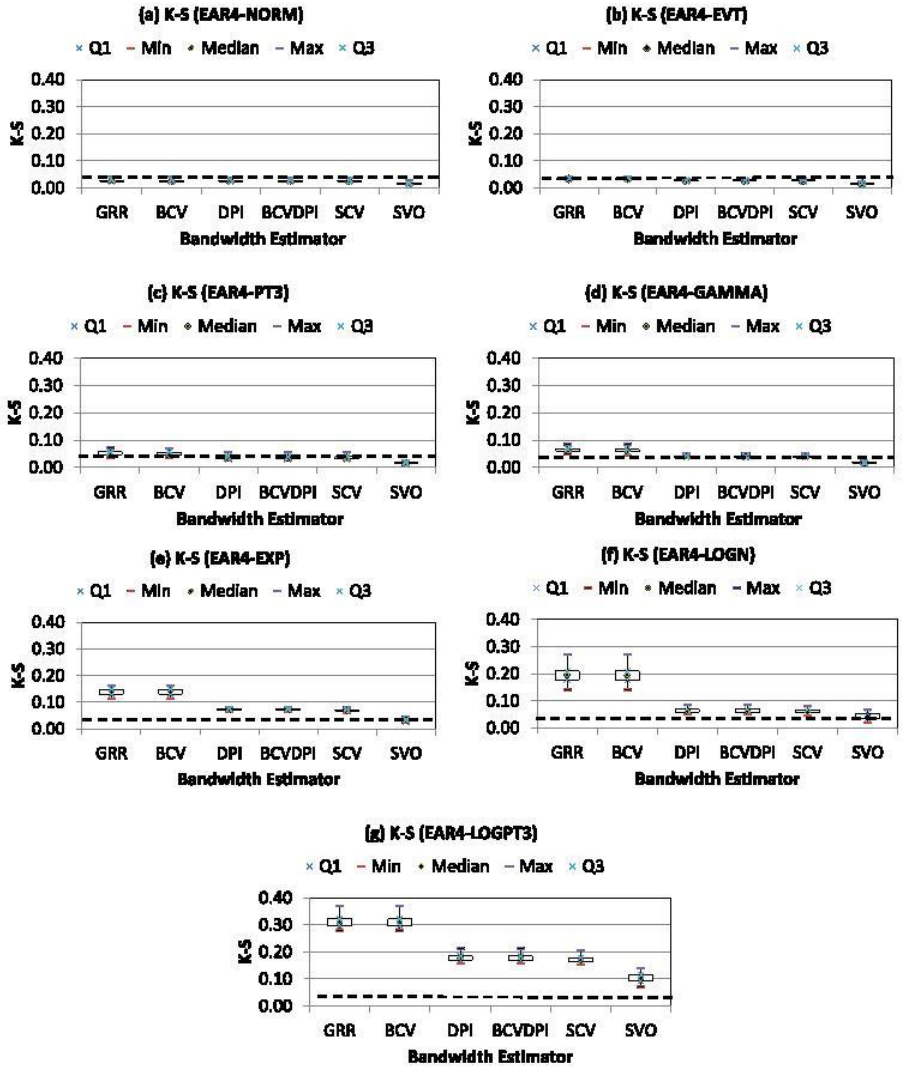


Fig. 6. KDE accuracy measured by K–S statistics for EAR4 & TEAR10 models (The dashed line indicates the 95% confidence interval for kernel density estimation based on the Kolmogorov–Smirnov (K–S) statistic (Parsons and Wirsching, 1982)).

30 replicates used in the MI calculation (see Eq. (25)) is also informative.

$$\text{Ratio of the bandwidths} = \frac{\widehat{h}_{\text{pro},i}}{\widehat{h}_{\text{GRR},i}} \tag{25}$$

where $\widehat{h}_{\text{pro},i}$ stands for the estimated bandwidth based on the proposed bandwidth estimators and $\widehat{h}_{\text{GRR},i}$ is the estimated bandwidth based on the GRR (Eq. (15)). As part of an empirical study on the effect of different bandwidth ratios on the accuracy of MI estimation, Harrold et al. (2001) found that for highly non-Gaussian data, an over-smoothed bandwidth performs best, with an optimal bandwidth ratio of 1.5. This general finding is confirmed by the

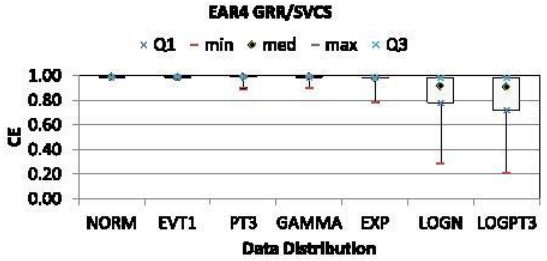


Fig. 7. Residual accuracy measured by CE for EAR4 model.

Table 4
Average ratio of different kernel bandwidths under different distribution scenarios for EAR4 model.

	NORM	EVT1	PT3	GAMMA	EXP	LOGN	LOGPT3
GRR	–	–	–	–	–	–	–
BCV	0.964	0.954	0.997	0.984	1.033	1.007	0.997
DPI	0.958	0.886	1.039	0.971	1.265	1.716	1.804
BCVDPI	0.958	0.886	1.039	0.971	1.265	1.716	1.804
SCV	0.971	0.856	1.046	0.967	1.268	1.737	1.804
SVO	0.493	0.418	0.810	0.791	1.190	1.399	1.497

The average ratio is between each of the alternative kernel bandwidth estimators and the GRR.

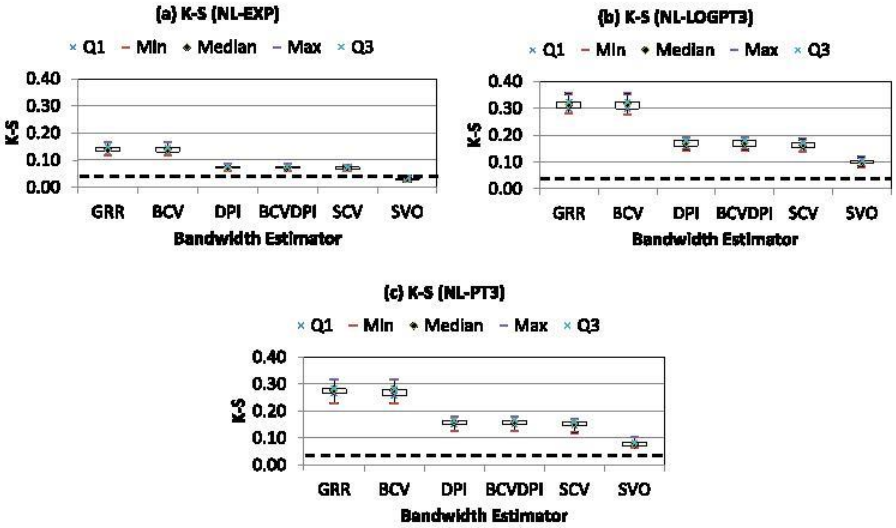


Fig. 8. KDE accuracy measured by K–S statistics for NL model (the dashed line indicates the 95% confidence interval for kernel density estimation based on the Kolmogorov–Smirnov (K–S) statistic (Parsons and Wirsching, 1982)).

results of this study (Table 4), which show that bandwidth ratios increase with the degree of non-Gaussianity for the bandwidth estimators that result in more accurate KDE. In addition, the GRR based PMI IVS is found to mainly underestimate the correct number of significant inputs (shown in Appendix A, Fig. A.1) for the non-Gaussian cases (e.g. LOGN and LOGPT3), which is consistent with the results (i.e. NL and Bank cases) in Galelli et al. (2014). This can be ascribed to the underestimated bandwidth, as the severity of underestimating the correct number of significant inputs is proportional to the bandwidth ratio outline in Table 4. However, alternative bandwidth estimators (i.e. DPI, BCVDPI, SCV, and SVO) tend to correct such underestimation with increased bandwidths, which sometimes even result in slight overestimation.

The general trends observed for the EAR4 model were confirmed by those obtained for the TEAR10 and NL models, except for the comparatively low accuracy when SVO was used for the NORM and LOGPT3 distributions for the data generated from the TEAR10 model and the overall reduction in CSR for the data generated from the NL model. Even the alternative bandwidth estimators (i.e. DPI, BCVDPI, SCV, and SVO) were found to tend to underestimate the correct number of significant inputs, as shown in Appendix A Fig. A.3. This observation is likely to be the result of the combined effect of the reduced KDE and residual estimation accuracy due to boundary issues, particularly influenced by increased problem non-linearity, as discussed below. For example, the non-Gaussianity of the NL model, as measured by skewness and kurtosis, is much more severe than that of the EAR4 and TEAR10 models (as shown in Tables 1 and 2), suggesting increased potential impact of boundary issues on KDE and residual estimation. For kernel based PDF and MI estimation, the corresponding accuracy of the KDE of the NL model is generally slightly worse than that of the EAR4 and TEAR10 models, as indicated by the K–S values in Figs. 6 and 8. For residual estimation, the overall accuracy of the NL model was found to be significantly less than that of the EAR4 model, as shown in Figs. 7 and 9. This can be explained by the fact that the univariate GRNN used for residual estimation is essentially a Nadaraya–Watson regression and therefore the corresponding bias is a function of the regression function $m(X_i)$ and the probability density function $f(X_i)$ with respect to input X_i. According to Fan (1992), Ruppert and Wand (1994), and Masry (1996), this bias increases as the boundary issue becomes severe. Consequently, the accuracy of residual and PMI estimation is likely to be compromised as the influence of boundary issues increases with increasing problem non-linearity and non-normality.

4.2. Computational efficiency

The computational efficiency of different bandwidth estimators used for the EAR4 model is given in Fig. 10. The GRR based method was found to be the most efficient overall. This can be explained by the fact that the only unknown parameter is the size of the applied data after standardisation (May et al., 2008b). The computational expense of the BCV approach was close to that of the GRR because the fitness functions used are identical, although the BCV requires an additional iterative optimisation process. The average runtimes

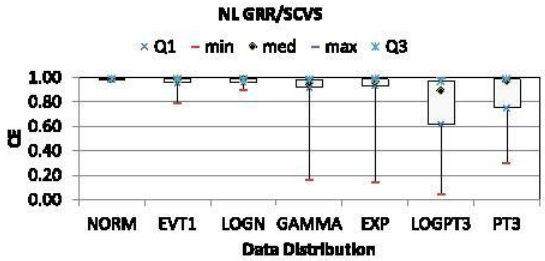


Fig. 9. Residual accuracy measured by CE for NL model.

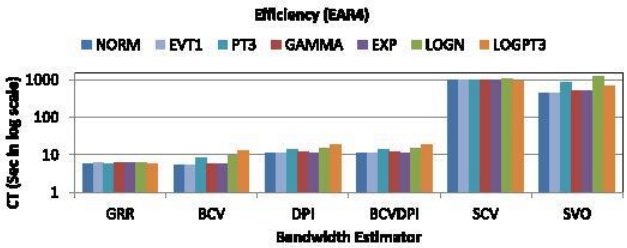


Fig. 10. Computational efficiency of EAR4 model with different bandwidth estimators.

I. Data Distribution:	Gaussian or nearly Gaussian distribution ($s<1.5$ & $k<3$)	Slight to moderate non-Gaussian distribution ($1.5<s<5$ & $3<k<30$)	Extreme non-Gaussian distribution ($s>5$ & $k>30$)
II. *Suggested bandwidth for residual & corresponding efficiency:	GRR based GRNN (VE)	GRR based GRNN (VE)	SVCS based GRNN (NE)
III. **Suggested bandwidth for KDE & MI & corresponding efficiency:	GRR or BCV (VE)	DPI or BCVDPI (E)	DPI or BCVDPI (E)
IV. Scenario No.:	1	2	3

Fig. 11. Suggested bandwidth estimators under different distribution scenarios (VE = comparatively very computationally efficient, E = comparatively moderately computationally efficient, and NE = comparatively not computationally efficient; *recommendation based on Li et al., 2014; ** recommendation based on present study).

for both DPI and BCVDPI were double that required by the GRR. This is because of the additional time required for the estimation of the pilot bandwidths during each iteration of the MI estimation (Wand and Jones, 1995). The efficiency of using SVO for bandwidth estimation is significantly less than that of the methods discussed thus far, with an average runtime of 667s, which is over 110 times greater than that associated with the GRR. The increased computational requirements of SVO are a result of the need to estimate the

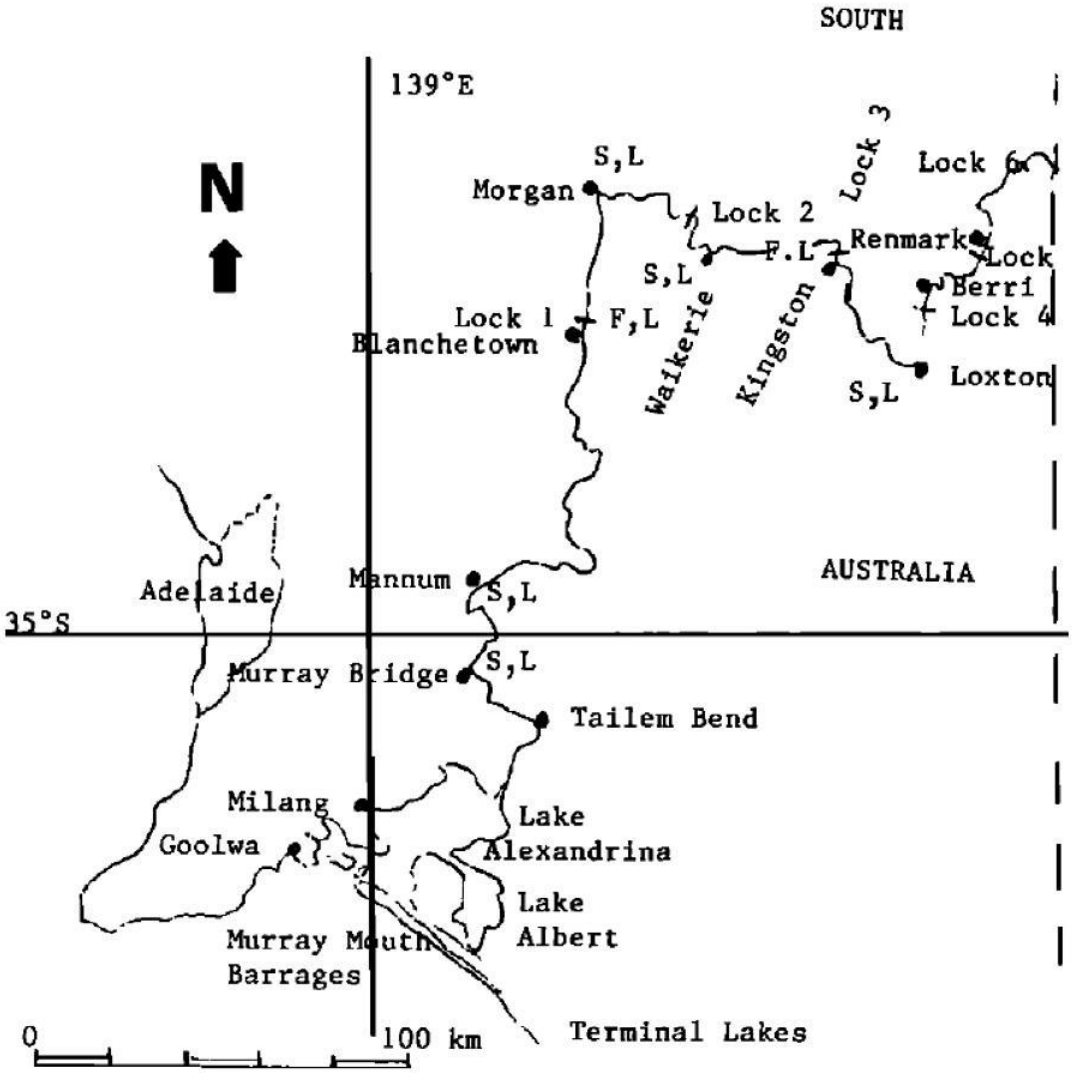


Fig. 12. The River Murray in South Australia (Maier and Dandy, 1996).

Table 5
Candidate inputs and output for the salinity case study.

Candidate inputs				Output			
Location	Variable	Abbreviation	Lags	Location	Variable	Abbreviation	Forecasting period
Mannum	Salinity	MAS	1,3,5,7,9	Murray Bridge	Salinity	MBS	14
Morgan	Salinity	MOS	1,3,5,7,9				
Waikerie	Salinity	WAS	1,2,3,4,5				
Loxton	Salinity	LOS	1,2,3,4,5				
Murray Bridge	Salinity	MBS	1,3,5,7,9				
Lock 1 Upper	River level	L1UL	−3,−1,1,3,5				

fitness function for each trial bandwidth during the optimisation process. Use of the SCV method was most inefficient, with an average runtime of over 160 times greater than that for the GRR. The inefficiency of SCV can be ascribed to the need to approximate a high order integrated squared density derivative during each iteration of the MI estimation (Wand and Jones, 1995), as well as the optimisation searching process. These findings were supported by the results for the TEAR10 and NL models (see Figs. A.7 and A.8 in Appendix A).

4.3. Suggested rules and guidelines

The preliminary empirical guidelines for selecting the most appropriate kernel bandwidth estimation technique based on the degree of normality of the data (according to the findings of the 3780 computational experiments with the synthetically generated data) are given in Fig. 11. It should be noted that the proposed guidelines represent reasonable trade-offs between selection accuracy and computational efficiency, although it is acknowledged that the best trade-off is also a function of case-study dependent features and user preferences.

As can be seen in Fig. 11, the preliminary empirical guidelines can be categorised into three scenarios, as described below:

Scenario 1: If most of the input/output data follow Gaussian or nearly Gaussian distributions (average $s < 1.3$ and $k < 3$), the GRR is suggested for residual estimation and the GRR (or BCV) is recommended for MI estimation, as these methods are able to provide good selection accuracy at a comparatively greater computational efficiency.
Scenario 2: If the input/output data are mainly moderately non-Gaussian (average $1.3 < s < 5$ and $3 < k < 30$), the GRR is suggested for residual estimation and the DPI (or BCVDPI) is recommended for MI estimation, so that selection accuracy can be improved with only a small reduction in computational efficiency, in comparison with using the GRR and BCV.
Scenario 3: If the input/output data are mainly extremely non-Gaussian (average $s > 5$ and $k > 30$), the SVCS is suggested for residual estimation and the DPI (or BCVDPI) is recommended for MI estimation. While these methods will decrease computational efficiency significantly, they are also likely to result in a marked increase in selection accuracy.

5. Testing of proposed rules and guidelines

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

5.1. River salinity at Murray Bridge

The study area of the first semi-real case is illustrated in Fig. 12. According to Maier and Dandy (1996), 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 at locations specified in Table 5). Consequently, these six inputs were used to generate the corresponding outputs (MBS + 13). Other redundant or irrelevant candidate inputs listed in Table 5 were also introduced for the purpose of testing the effectiveness of PMI IVS.

In order to generate the known outputs from the real inputs, standard multilayer perceptron (MLP) artificial neural networks (ANNs) were developed using the approach outlined in Wu et al. (2014). The historical records from 1987 to 1990 were split into training (60%), testing (20%) and validating sets (20%) using the

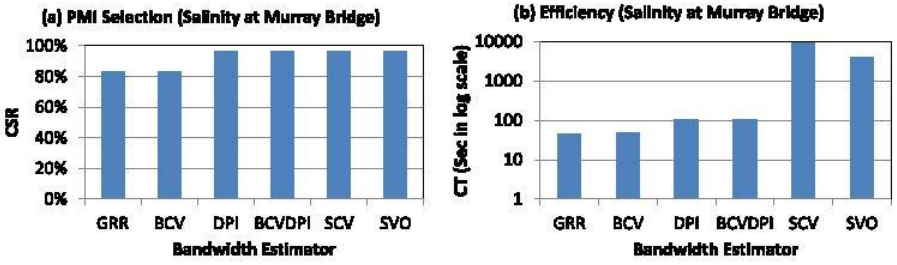


Fig. 13. Correct selection rate and efficiency of salinity forecast at Murray Bridge with proposed and alternative bandwidth estimators.

DUPLEX method (see May et al., 2010), in accordance with the guidelines suggested by Wu et al. (2013). A single hidden layer was used and the optimal number of hidden nodes was determined by trial and error, considering a range of 0 to 6. The optimal model structure was found to be 6-4-1. The back-propagation algorithm (with learning rate of 0.1 and momentum of 0.1) was used for model calibration. The test inputs were then re-simulated 30 times based on the real observations in order to obtain data sets that contained a certain degree of variation, while still maintaining the major time patterns and data distributions. This enabled IVS performance to be evaluated over 30 independent trials. The corresponding output was obtained by substituting the simulated inputs into the trained ANN model. The input/output data contain strongly linear components and follow a mildly non-Gaussian distribution, according to Bowden (2003), Wu et al. (2013) and Li et al. (2014). Consequently, this study corresponds to *Scenario 2* in Fig. 11. Given this, the selection performance of the PMI using the DPI (and BCVDPI) for KDE and the GRR for residual estimation was expected to be superior in terms of an appropriate trade-off between selection accuracy and computational efficiency.

Based on the results in Fig. 13, this was observed to be the case. The CSR resulting from the use of the proposed approach was 96.7%, compared with 83.3% when the GRR and BCV approaches were used for KDE. Although use of the SCV and SVO methods also resulted in a CSR of 96.7%, the associated computational cost was significantly greater. Consequently, the DPI/BCVDPI based method

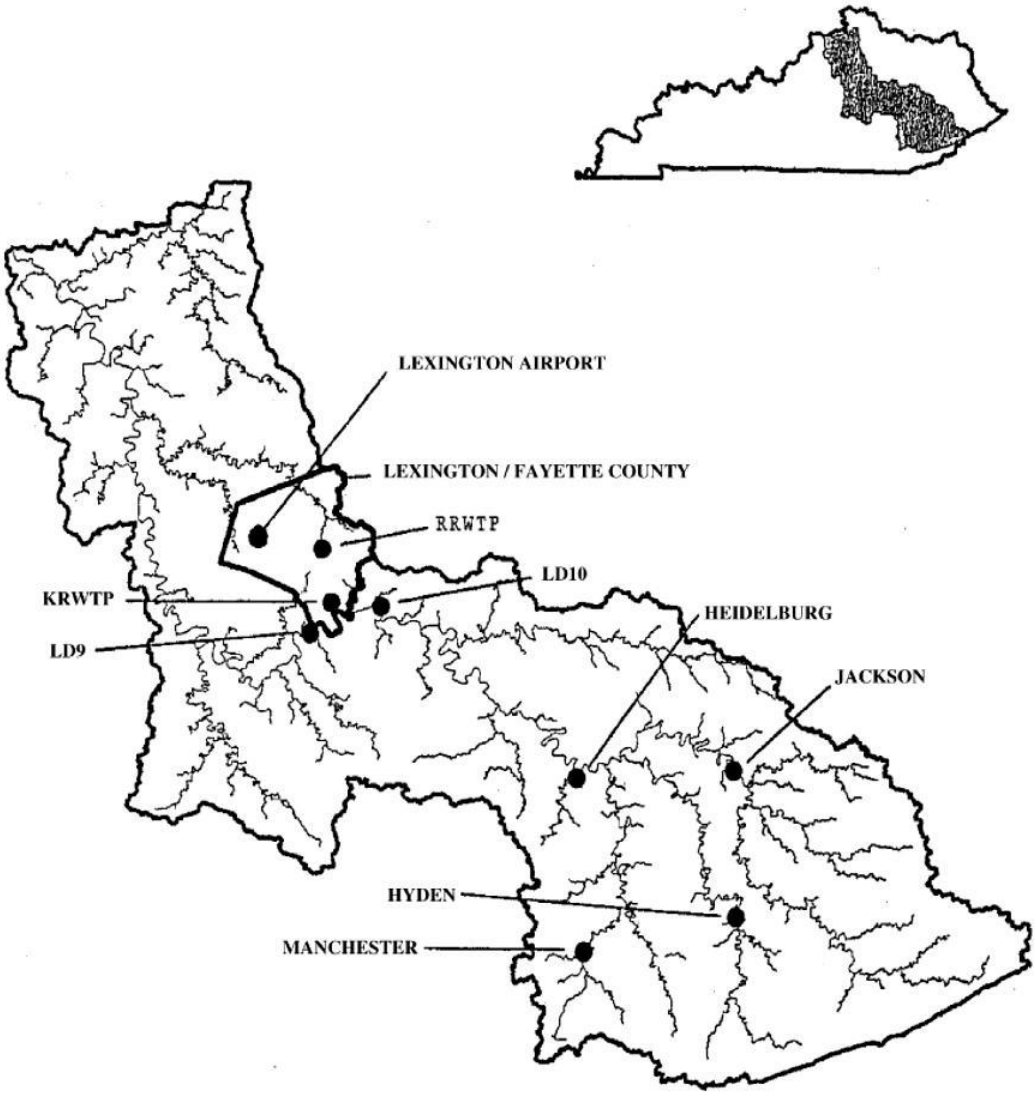


Fig. 14. The Kentucky River Basin in USA (Jain and Srinivasulu, 2004).

Table 6
Candidate inputs and output used for the rainfall-runoff case study.

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

provided a good trade-off between selection accuracy and computational efficiency for this study, as suggested by the proposed guidelines (Fig. 11).

5.2. Rainfall-runoff in Kentucky River Basin

The second semi-real data set is concerned with rainfall-runoff modelling in the Kentucky River Basin in the USA (Fig. 14). The output variable for this case study is the forecast flow at Lock and Dam 10 one day in advance (Jain and Srinivasulu, 2004). The corresponding inputs, including average daily effective rainfall and runoff with specific lag time (i.e. average daily effective rainfall: P(t), P(t-1) and average daily runoff: $Q(t-1)$, $Q(t-2)$ at locations specified in Table 6), together with other redundant or irrelevant candidate inputs, are summarized in Table 6, which are the same as those used by Bowden (2003), Wu et al. (2013) and Li et al. (2014).

The historical rainfall-runoff records from 1960 to 1972 were used for developing the MLP-ANNs using the approach described for the salinity case study. The optimal model structure was determined as 4-4-1. Thirty sets of inputs and outputs were generated using the procedure described for the salinity case study. It should be noted that the input/output data contain non-linear components and follow extremely non-Gaussian distributions, as discussed by Wu et al. (2013), Li et al. (2014) and Galelli et al. (2014). Consequently, this study corresponds to *Scenario 3* in Fig. 11. Given this, the selection performance of the PMI using the DPI (and BCVDPI) for KDE was expected to be superior in terms of an appropriate trade-off between selection accuracy and computational efficiency.

As indicated in Fig. 15(a), use of the approach suggested in the proposed guidelines derived from the synthetic data (i.e. DPI with SVCS) clearly results in the best CSR, with an accuracy of 96.7%. This is much higher than the CSR of 77.8% when the 'standard' approach (i.e. GRR with GRR) is used. While this increased selection accuracy comes at a significant increase in computational cost (i.e. 68 times more computationally expensive), as shown in Fig. 15(b), this still seems to provide the best trade-off between selection accuracy and computational efficiency, as suggested by the proposed guidelines (Fig. 11).

6. Summary and conclusions

Input variable selection (IVS), as one of the most important steps in the development of ANN and other data driven environmental and water resources models, determines the quality and quantity of information used in the modelling process. Partial mutual information (PMI) is one of the most promising approaches to IVS, as it is able to account for the relevance and redundancy of all candidate inputs and can be used for both linear and non-linear problems. However, one disadvantage of using PMI is that it requires kernel density estimates (KDEs) of the data to be obtained, which can become problematic when the data are non-normally distributed, as is often the case for environmental and water resources problems. However, this is an issue that has been ignored in previous studies on the application of PMI IVS, in which the Gaussian reference rule (GRR) has generally been used to obtain the required KDEs. This is likely to result in a reduced CSR for data that are non-Gaussian, as shown by Galelli et al. (2014) and Humphrey et al. (2014).

In order to develop an improved approach to PMI IVS for data that are non-normally distributed, the selection performances of PMI with six different kernel bandwidth estimators for KDE were assessed in terms of selection accuracy and computational efficiency for input/output data with distinct degrees of normality on three synthetic data sets. The results from the 3780 trials with the synthetic data were used to develop empirical guidelines for the choice of the most appropriate bandwidth estimation techniques for data with different degrees of non-normality. The validity of these guidelines was then tested on the two semi-real data sets.

The results of the synthetic case studies suggest that the use of GRR-based bandwidth estimators only results in good input selection accuracy if the input/output data follow Gaussian or nearly Gaussian distributions, which is in line with the results obtained by Galelli et al. (2014) and Humphrey et al. (2014). As a result of their reduced dependence on the Gaussian assumption, DPI, BCVDPI, SCV, and SVO based bandwidth estimators generally result in marked improvements in CSR for problems with data that follow non-Gaussian distributions. However, there is a distinct trade-off between selection accuracy and computational efficiency.

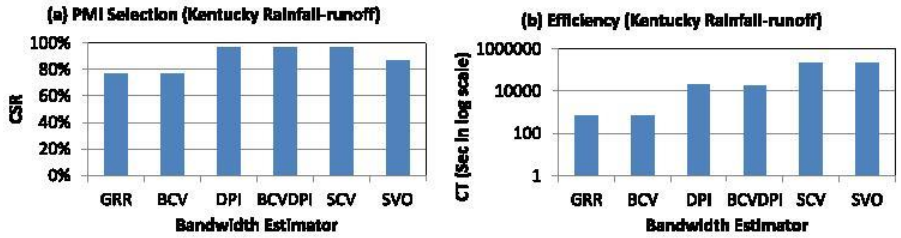


Fig. 15. Correct selection rate and efficiency of flow forecast at Kentucky River Basin with proposed and alternative bandwidth estimators.

One of the major outcomes of this paper is the development of the empirical guidelines based on the synthetic tests. As shown in Fig. 11, the suggested bandwidth estimators for KDE used in the MI calculation should be used in conjunction with the bandwidth estimators for residual estimation suggested by Li et al. (2014). The results for the two semi-real data sets, which follow mildly and extremely non-Gaussian distributions, support the validity of the proposed guidelines for the selection of appropriate bandwidth estimation methods for data with different degrees of non-normality. It should be noted that the proposed guidelines are valid for environmental and water resource applications with data that have distributional properties similar to those provided in the guidelines, and that the implementation of the guidelines is also likely to benefit other data-driven environmental and water resources models, even though they were only tested on MLPs.

Although the results of this study indicate that the use of alternative bandwidth estimators can result in significant improvements in PMI IVS for data that are non-normally distributed, these improvements were not as pronounced for extremely non-Gaussian data and the non-linear synthetic case study. This is likely due to boundary issues associated with KDE for highly non-Gaussian data (Karunamuni and Alberts, 2005; Scott, 1992). Consequently, future research should focus on potential improvements to IVS accuracy as a result of the consideration of such boundary issues. In addition, 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. Finally, there is a need to assess the performance of the proposed modifications to the implementation of the PMI algorithm on a broader set of data and against that of other IVS algorithms (see Galelli et al., 2014).

Acknowledgements

This research was aided by the suggestions from Prof. A. Sharma and the original code from Dr. R.J. May (GRR based PMI) and Dr. G.B. Humphrey (GRR based GRNN). The authors would also like to thank the three anonymous reviewers, whose input has improved the quality of this paper significantly.

Appendix A. Supplementary data

Supplementary data related to this article can be found at http://dx.doi.org/10.1016/j.envsoft.2014.11.028.

References

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. Prog. Phys. Geogr. 36 (4), 480–513.

Adeloye, A.J., Rustum, R., Kariyama, I.D., 2012. Neural computing modeling of the reference crop evapotranspiration. Environ. Model. Softw. 29 (1), 61–73.

Agalbjörn, S., Končar, N., Jones, A., 1997. A note on the gamma test. Neural Comput. Appl. 5 (3), 131–133.

Akaike, H., 1974. A new look at the statistical model identification. IEEE Trans. Automat. Control 19 (6), 716–723.

Aldershof, B.K., 1991. Estimation of Integrated Squared Density Derivatives. The University of North Carolina, Chapel Hill.

ASCE, 2000a. Artificial neural networks in hydrology II: hydrology applications. Hydrol. Eng. 5 (2), 124–137.

ASCE, 2000b. Artificial neural networks in hydrology. I: Preliminary concepts. J. Hydrol. Eng. 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. Environ. Model. Softw. 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., 2003. Data transformation for neural network models in water resources applications. J. Hydroinform. 5 (4), 245–258.

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. J. Hydrol. 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. J. Hydrol. 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 Resour. Res. 48 (10) http://dx.doi.org/10.1029/2012WR011984.

Box, G.E., Jenkins, G.M., Reinsel, G.C., 2013. Time Series Analysis: Forecasting and Control. John Wiley & Sons, Hoboken, New Jersey.

Cacoullos, T., 1966. Estimation of a multivariate density. Ann. Inst. Stat. Math. 18 (1), 179–189.

Castelletti, A., Galelli, S., Ratto, M., Soncini-Sessa, R., Young, P., 2012a. A general framework for dynamic emulation modelling in environmental problems. Environ. Model. Softw. 34, 5–18.

Castelletti, A., Galelli, S., Restelli, M., Soncini-Sessa, R., 2012b. Data-driven dynamic emulation modelling for the optimal management of environmental systems. Environ. Model. Softw. 34, 30–43.

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

Chua, L.H.C., Wong, T.S.W., 2010. Improving event-based rainfall-runoff modeling using a combined artificial neural network-kinematic wave approach. J. Hydrol. 390 (1–2), 92–107.

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. Environ. Model. Softw. 22 (7), 1034–1052.

Dawson, C.W., Mount, N., Abrahart, R.J., Louis, J., 2014. Sensitivity analysis for comparison, validation and physical-legitimacy of neural network-based hydrological models. J. Hydroinform. 16 (2), 407–424.

Dawson, C.W., Wilby, R., 2001. Hydrological modelling using artificial neural networks. Prog. Phys. Geogr. 25 (1), 80–108.

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. J. Am. Stat. Assoc. 87 (420), 998–1004.

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. J. Hydrol. 367 (3–4), 165–176.

Galelli, S., Castelletti, A., 2013. Tree-based iterative input variable selection for hydrological modeling. Water Resour. Res. 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. Environ. Model. Softw. 62, 33–51.

Haimi, H., Mulas, M., Corona, F., Vahala, R., 2013. Data-derived soft-sensors for biological wastewater treatment plants: an overview. Environ. Model. Softw. 47, 88–107.

Hall, P., Marron, J.S., 1987. Estimation of integrated squared density derivatives. Stat. Probab. Lett. 6 (2), 109–115.

Hall, P., Marron, J.S., Park, B.U., 1992. Smoothed cross-validation. Probab. Theory Relat. Fields 92 (1), 1–20.

Harrold, T., Sharma, A., Sheather, S., 2001. Selection of a kernel bandwidth for measuring dependence in hydrologic time series using the mutual information criterion. Stoch. Environ. Res. Risk Assess. 15 (4), 310–324.

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. J. Hydrol. 400 (1–2), 10–23.

Hu, T., Wu, F., Zhang, X., 2007. Rainfall-runoff modeling using principal component analysis and neural network. Nord. Hydrol. 38 (3), 235–248.

Huang, D., Chow, T.W.S., 2005. Effective feature selection scheme using mutual information. Neurocomputing 63, 325–343.

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: Ames, D.P. (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. Environ. Model. Softw. 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 Resour. Res. 40 (4), W04302.

Jain, S., Das, A., Srivastava, D., 1999. Application of ANN for reservoir inflow prediction and operation. J. Water Resour. Plan. Manag. 125 (5), 263–271.

Jones, M., Sheather, S., 1991. Using non-stochastic terms to advantage in kernel-based estimation of integrated squared density derivatives. Stat. Probab. Lett. 11 (6), 511–514.

Karunamuni, R.J., Alberts, T., 2005. On boundary correction in kernel density estimation. Stat. Methodol. 2 (3), 191–212.

Kingston, G.B., Lambert, M.F., Maier, H.R., 2005. Bayesian training of artificial neural networks used for water resources modeling. Water Resour. Res. 41 (12), W12409.
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. Environ. Model. Softw. 59, 162–186.
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. Environ. Model. Softw. 25 (5), 648–660.
Maier, H.R., Dandy, G.C., 1997a. Determining inputs for neural network models of multivariate time series. Computer-Aided Civ. Infrastruct. Eng. 12 (5), 353–368. http://dx.doi.org/10.1111/0885-9507.00069.
Maier, H.R., Dandy, G.C., 1997b. Modelling cyanobacteria (blue-green algae) in the River Murray using artificial neural networks. Math. Comput. Simul. 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. Environ. Model. Softw. 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 Resour. Res. 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. Environ. Model. Softw. 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. Environ. Model. Softw. 19 (5), 485–494.
Masry, E., 1996. Multivariate local polynomial regression for time series: uniform strong consistency and rates. J. Time Ser. Anal. 17 (6), 571–599.
May, R.J., 2010. Developing Artificial Neural Networks for Water Quality Modelling and Prediction. School of Civil, Environmental & Mining: The University of Adelaide, pp. 41–47.
May, R.J., Dandy, G.C., Maier, H.R., 2011. Review of input variable selection methods for artificial neural networks. In: InTech (Ed.), Artificial Neural Networks—Methodological Advances and Biomedical Applications, pp. 19–44. Rijeka, Croatia.
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. Environ. Model. Softw. 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 Netw. 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. Environ. Model. Softw. 23 (10), 1312–1326.
Millie, D.F., Weckman, G.R., Young II, W.A., Ivey, J.E., Carrick, H.J., Fahnenstiel, G.L., 2012. Modeling microalgal abundance with artificial neural networks: demonstration of a heuristic 'Grey-Box' to deconvolve and quantify environmental influences. Environ. Model. Softw. 38, 27–39.
Mount, N., Dawson, C., Abrahart, R., 2013. Legitimising data-driven models: exemplification of a new data-driven mechanistic modelling framework. Hydrol. Earth Syst. Sci. 17 (7), 2827–2843.
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 (Salmo trutta L.) in Iberian rivers. Environ. Model. Softw. 59, 30–43.
Noori, R., Karbassi, A.R., Moghaddamnia, A., Han, D., Zokaei-Ashtiani, M.H., Farokhinia, A., GhafariGousheh, M., 2011. Assessment of input variables determination on the SVM model performance using PCA, Gamma test, and forward selection techniques for monthly stream flow prediction. J. Hydrol. 401, 177–189.
Ozkaya, B., Demir, A., Bilgili, M.S., 2007. Neural network prediction model for the methane fraction in biogas from field-scale landfill bioreactors. Environ. Model. Softw. 22 (6), 815–822.
Park, B.U., Marron, J., 1992. On the use of pilot estimators in bandwidth selection. J. Nonparametr. Stat. 1 (3), 231–240.
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. Microelectron. Reliab. 22 (2), 163–167.
Parzen, E., 1962. On estimation of a probability density function and mode. Ann. Math. Stat. 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. Environ. Model. Softw. 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.
Rakovec, O., Hill, M., Clark, M., Weerts, A., Teuling, A., Uijlenhoet, R., 2014. Distributed Evaluation of Local Sensitivity Analysis (DELSA), with application to hydrologic models. Water Resour. Res. 50 (1), 409–426.
Ruppert, D., Wand, M.P., 1994. Multivariate locally weighted least squares regression. Ann. Stat. 22 (3), 1346–1370.
Scott, D.W., 1992. Multivariate Density Estimation and Visualization. Handbook of Computational Statistics. Springer. New York, USA.
Scott, D.W., Terrell, G.R., 1987. Biased and unbiased cross-validation in density estimation. J. Am. Stat. Assoc. 82 (400), 1131–1146.
Shannon, C.E., 1948. A mathematical theory of communication. Bell Syst. Techn. J. 33 (27), 379–423, 623–656.
Sharma, A., 2000a. Seasonal to interannual rainfall probabilistic forecasts for improved water supply management: Part 1–A strategy for system predictor identification. J. Hydrol. 239 (1–4), 232–239.
Sharma, A., 2000b. Seasonal to interannual rainfall probabilistic forecasts for improved water supply management: Part 3–A nonparametric probabilistic forecast model. J. Hydrol. 239 (1–4), 249–258.
Specht, D.F., 1991. A general regression neural network. IEEE Trans. Neural Networks 2 (6), 568–576.
Srinivasulu, S., Jain, A., 2006. A comparative analysis of training methods for artificial neural network rainfall–runoff models. Appl. Soft Comput. 6 (3), 295–306.
Trappenberg, T., Ouyang, J., Back, A., 2006. Input variable selection: mutual information and linear mixing measures. IEEE Trans. Knowledge Data Eng. 18 (1), 37–46.
Wand, M.P., Jones, M.C., 1993. Comparison of smoothing parameterizations in bivariate kernel density estimation. J. Am. Stat. Assoc. 88 (422), 520–528.
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. Environ. Model. Softw. 55, 107–119.
Wu, W., Dandy, G.C., Maier, H.R., 2014. 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. Environ. Model. Softw. 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 Resour. Res. 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. Environ. Model. Softw. 26 (10), 1199–1210.