An information theoretic approach to select alternate subsets of predictors for data-driven hydrological models

*R. Taormina^a, S. Galelli^a, G. Karakaya^b, S.D. Ahipasaoglu^a

^a*Pillar of Engineering Systems and Design, Singapore University of Technology and Design, Singapore.*

^b*Department of Business Administration, Middle East Technical University, Ankara, Turkey.*

Abstract

This work investigates the uncertainty associated to the presence of multiple subsets of predictors yielding data-driven models with the same, or similar, predictive accuracy. To handle this uncertainty effectively, we introduce a novel input variable selection algorithm, called Wrapper for Quasi Equally Informative Subset Selection (W-QEISS), specifically conceived to identify all alternate subsets of predictors in a given dataset. The search process is based on a four-objective optimisation problem that minimises the number of selected predictors, maximises the predictive accuracy of a data-driven model and optimises two information theoretic metrics of relevance and redundancy, which guarantee that the selected subsets are highly informative and with little intra-subset similarity. The algorithm is first tested on two synthetic test problems and then demonstrated on a real-world streamflow prediction problem in the Yampa River catchment (US). Results show that complex hydro-meteorological datasets are characterised by a large number of alternate subsets of predictors, which provides useful insights on the un-

*Corresponding author. Tel.: +65 6499 4786; fax: +65 6499 4786.
E-mail address: riccardo.taormina@sutd.edu.sg

derlying physical processes. Furthermore, the presence of multiple subsets of predictors—and associated models—helps find a better trade-off between different measures of predictive accuracy commonly adopted for hydrological modelling problems.

Key words: Input Variable Selection, Information Theory, Data-driven models, Extreme Learning Machines, Neural Networks

1. Introduction

Data-driven hydrological models are the product of an ‘inductive’ process, in which a functional relationship between a subset of predictors, or input variables, and an output variable is inferred from observational data. This process involves four main stages (Young, 2001): identification of the possible predictors, specification of an appropriate model structure, estimation of the parameters that best characterise this structure (calibration), and predictive validation on a dataset different from the one used in calibration. An additional step is the quantification of the uncertainty associated to the model, which generally depends on multiple sources—e.g., measurement errors in the observational data or uncertainty in the model structure, parameters and selected predictors. Uncertainty in the model structure and parameters has been extensively addressed during the past decade; available methods include Bayesian techniques (Khan and Coulibaly, 2006; Kingston et al., 2008; Zhang et al., 2009, 2011), bootstrapping (Srivastav et al., 2007; Sharma and Tiwari, 2009; Tiwari and Chatterjee, 2010), fuzzy theory (Shrestha and Solomatine, 2006; Alvisi and Franchini, 2011; Taormina and Chau, 2015a) and ensemble modelling (Dawson et al., 2002; Parasuraman and Elshorbagy,

2008), for instance. All these methods study the effect of different structures and parameterizations to issue an uncertainty estimation of the output in the form of confidence intervals. Alternatively, one can directly use the time series of residuals to create a model of the predictive uncertainty (see Pianosi and Raso (2012), and references therein).

The uncertainty in the selected predictors is often addressed by means of input variable selection (Guyon and Elisseeff, 2003), whose aim is to identify a combination of predictors that best characterises the input-output relationship being modelled. In particular, the objective of variable selection is to simplify the model identification process by avoiding the interference of redundant and irrelevant predictors, leading to parsimonious, cost-effective and easy to interpret models. Variable selection algorithms can be categorised in *wrapper* and *filter* approaches, depending on the way the selection process is carried out (Maier et al., 2010). Wrappers use global optimization techniques to select the combination of predictors that maximizes the accuracy of a given model structure. Filters conduct a local search usually informed by a statistical measure of significance (e.g., linear correlation, mutual information (MacKay, 2003)) that quantifies the relevance of each predictor with respect to the output. Filters are computationally less expensive than wrappers. On the other hand, wrappers explore the search space in a more exhaustive way and are thus likely to identify more accurate models. Variable selection algorithms are used in a variety of water resources modelling problems, such as flood regionalization (Wan Jaafar et al., 2011), statistical downscaling (Phatak et al., 2011), streamflow and water quality modelling

44 (Bardsley et al., 2015; Li et al., 2015; Creaco et al., 2016), medium-term
 45 hydro-climatic forecasts (Sharma, 2000; Noori et al., 2011) and forecast of
 46 urban water demand (Quilty et al., 2016) (see Galelli et al. (2014) for a
 47 review). Whilst all these modelling problems are characterized by the p-
 48 resence of alternate subsets of predictors, the majority of variable selection
 49 algorithms yields one subset of predictors, thereby providing a narrow view
 50 of this source of uncertainty. Such issue may sharpen as larger observational
 51 datasets become available, owing to advances in remote sensing techniques
 52 (Ceola et al., 2015), ubiquitous mobile technology (Overeem et al., 2013)
 53 and crowdsourcing (Fraternali et al., 2012; Mount et al., 2016). Only re-
 54 cently, some techniques designed to identify a few subsets of predictors with
 55 the same (or similar) information content have been proposed (Sharma and
 56 Chowdhury, 2011; Taormina and Chau, 2015b; Creaco et al., 2016).

57

58 To get around the uncertainty in the selected predictors, we introduce
 59 a novel variable selection algorithm, called Wrapper for Quasi Equally In-
 60 formative Subset Selection (W-QEISS), which is specifically conceived to
 61 identify all equally informative subsets of predictors in a given hydrological
 62 dataset—in the context of this study, we say that two subsets are quasi e-
 63 qually informative if they have (almost) the same predictive accuracy with
 64 respect to a given model class (Karakaya et al., 2016). W-QEISS builds
 65 on the formulation of a four-objective optimization problem that maximizes
 66 the predictive accuracy of a pre-selected model, minimizes the number of
 67 predictors, and optimizes two information theoretic metrics of relevance and
 68 redundancy. The optimization of the predictive accuracy ensures that the

69 interaction between the model structure and data is maximized, while the
 70 minimization of the cardinality is aimed at simplifying the final models. The
 71 adoption of relevance and redundancy metrics guarantees that the selected
 72 subsets are highly informative and with low intra-subset similarity. These
 73 metrics provide a generic measure of dependence between variables, so they
 74 can characterize any functional relationship (Sharma and Mehrotra, 2014).
 75 Extreme Learning Machines (Huang et al., 2012)—a learning algorithm for
 76 single-layer feedforward neural networks—are used as model structure, s-
 77 ince they have high predictive accuracy and limited computational burden
 78 as compared to other non-linear data-driven techniques. This aspect is par-
 79 ticularly relevant for wrappers, which repeat the calibration and validation
 80 process several times to identify the best subset of predictors. The global
 81 optimization problem is solved by means of Borg MOEA (Hadka and Reed,
 82 2013), a Multi-Objective Evolutionary Algorithm designed to handle multi-
 83 modal problems and expensive objective function evaluations. We show that
 84 the identification of quasi equally informative subsets of predictors yields t-
 85 wo main advantages. First, it determines the relative importance of each
 86 predictor, thus assisting in the interpretation of the underlying physical pro-
 87 cesses. Second, it helps find a better trade-off between multiple measures of
 88 predictive accuracy commonly used to assess the performance of hydrological
 89 models.

90

91 In the remainder of the paper, we first formulate the quasi equivalence of
 92 subsets, give some background on information theoretic criteria and describe
 93 W-QEISS algorithm. We then assess its performance on two synthetic (lin-

ear and non-linear) test problems and demonstrate it on a real-world case study of streamflow prediction in the Yampa River catchment (US), whose discharge is contributed by both rainfall and snowmelt processes. A comparison between the subsets of predictors selected by W-QEISS and four popular variable selection algorithms is included as benchmarking exercise. Concluding remarks and future research directions are outlined in Section 6.

2. Methods and Tools

2.1. Quasi equivalence of subsets

In this study, we assume that for a given dataset there can be multiple subset of predictors resulting in models having the same (or similar) predictive accuracy (Karakaya et al., 2016). We define the *quasi equivalence of subsets* as follows.

Definition 1. Let $\bar{f}(\cdot)$ be a metric of predictive accuracy taking values between 0 and 1, where 0 indicates that a model does not have any predictive skill while 1 denotes perfect predictive accuracy. Given a subset of predictors S_j , we say that another subset S_i is δ -quasi equally informative to S_j if the two subsets have almost equal predictive accuracy with respect to a given model class, i.e., subset S_i is δ -quasi equally informative to subset S_j if $\bar{f}(S_i) \geq (1 - \delta)\bar{f}(S_j)$ for $0 \leq \delta \leq 1$.

The presence of quasi equivalent subsets depends on the specific interactions between the variables in the observational dataset at hand. As discussed in Yu and Liu (2004), predictors can be classified depending on their relevance with respect to the output variable as strongly relevant, weakly relevant and

117 irrelevant. A predictor is strongly relevant if it cannot be removed from a
 118 subset without affecting its predictive accuracy, while it is irrelevant if it
 119 is not necessary with respect to the modelling task. A predictor is weakly
 120 relevant if there exists a combination of other relevant predictors that sub-
 121 sume the same information about the output variable and the rest of the
 122 predictors—this combination is named *Markov blanket*. A weakly relevant
 123 predictor can be either redundant or necessary in the modelling process. It
 124 is redundant if its Markov blanket is included in the selected subset, while it
 125 is necessary (i.e., weakly relevant but non redundant) in the opposite case.
 126 A set of predictors can thus be divided into four disjoint parts on the ba-
 127 sis of their relevance and redundancy (Figure 1): strongly relevant, weakly
 128 relevant but non redundant, weakly relevant and redundant, irrelevant pre-
 129 dictors. Since different partitions of weakly relevant predictors can result
 130 during the selection process, different quasi equally informative subsets can
 131 potentially be identified (Liu et al., 2015). The identification of these subsets
 132 requires specific metrics of relevance and redundancy and a search algorithm,
 133 especially when the dimension of the observational dataset precludes an
 134 exhaustive search in the predictors space.

135 2.2. Background on information theoretic criteria

136 Information theoretic criteria (e.g., Mutual Information, Partial Mutual
 137 Information) are widely used for quantifying the dependence between two or
 138 multiple variables. Their main advantage is that they do not assume any hy-
 139 pothesis on the underlying functional relationship (e.g., linear dependence),
 140 so they are adaptable to different modelling contexts (MacKay, 2003). These
 141 criteria are based on the concept of Shannon entropy (Shannon, 1948), which

142 is a measure of the uncertainty associated with a random variable. For a con-
 143 tinuous random variable X , the entropy $H(X)$ is defined as

$$H(X) = - \int p(x) \log p(x) dx, \quad (1)$$

144 where $p(x)$ is the probability density function of X . Similarly, the joint
 145 entropy $H(X, Y)$ of two continuous random variables X and Y is defined as

$$H(X, Y) = - \int \int p(x, y) \log p(x, y) dx dy, \quad (2)$$

146 where $p(x, y)$ is the joint probability density function of X and Y . Given any
 147 univariate sample of X and bivariate sample of (X, Y) , the discrete version
 148 of Eqs. 1 and 2 can be rewritten as

$$H(X) = - \sum_{x \in \mathbf{X}} p(x) \log p(x), \quad (3)$$

149 and

$$H(X, Y) = - \sum_{x \in \mathbf{X}} \sum_{y \in \mathbf{Y}} p(x, y) \log p(x, y), \quad (4)$$

150 where $\mathbf{X}$ is the set of all possible values of X , with probability mass function
 151 $p(x) = Pr(X = x)$, $x \in \mathbf{X}$, and $p(x, y)$ the joint probability mass function
 152 $p(x, y) = Pr(X = x, Y = y)$, $x \in \mathbf{X}$, and $y \in \mathbf{Y}$.

153

154 The Mutual Information between two variables quantifies the information
 155 obtained about one variable through the other one. For two continuous

156 random variables X and Y , the Mutual Information $MI(X, Y)$ is defined as

$$MI(X, Y) = \int \int p(x, y) \log \left[\frac{p(x, y)}{p(x)p(y)} \right] dx dy. \quad (5)$$

157 The MI between X and Y can also be expressed as

$$MI(X, Y) = H(X) + H(Y) - H(X, Y). \quad (6)$$

158 Note that the Mutual Information can be conveniently scaled between 0
159 and 1 to yield the Symmetric Uncertainty (Witten and Frank, 2005), which
160 is defined as

$$SU(X, Y) = \frac{2MI(X, Y)}{H(X) + H(Y)}. \quad (7)$$

161 A value of SU equal to 0 indicates that the variables X and Y are inde-
162 pendent, whereas a value equal to 1 implies that the knowledge about one
163 variable is enough to perfectly determine the other. Mutual Information,
164 Symmetric Uncertainty and their various extensions—e.g., Partial Mutual
165 Information (Sharma, 2000; May et al., 2008; Li et al., 2015), Partial Infor-
166 mation (Sharma and Mehrotra, 2014) and Conditional Mutual Information
167 (Quilty et al., 2016)—are commonly used to create measures of dependence
168 in hydrological modelling. In the next subsection we show how the Symmet-
169 ric Uncertainty is used to define the metrics of relevance and redundancy
170 used by W-QEISS algorithm.

171 *2.3. Wrapper for Quasi Equally Informative Subset Selection*

172 W-QEISS algorithm builds on a four-objective optimization problem,
173 whose objectives, search process and implementation are outlined next.

174 *2.3.1. Objective functions*

175 Let X be the set of candidate predictors, y the output variable, and S a
 176 subset of X . The metric of relevance $f_1(S)$ to be maximized in the search
 177 process is defined as

$$f_1(S) = \sum_{x_i \in S} SU(x_i, y), \quad (8)$$

178 where $SU(x_i, y)$ is the symmetric uncertainty between the i -th predictor x_i
 179 and the output y (see Eq. 7). The metric of redundancy $f_2(S)$ to be mini-
 180 mized is

$$f_2(S) = \sum_{x_i, x_j \in S, i < j} SU(x_i, x_j), \quad (9)$$

181 where $SU(x_i, x_j)$ is the symmetric uncertainty between the i -th and j -th
 182 predictors x_i and x_j . In other words, the relevance $f_1(S)$ of a subset S is
 183 an estimate of the explanatory power of the predictors in S , whereas $f_2(S)$
 184 quantifies their level of similarity. Hence, the minimization of $f_2(S)$ ‘drives’
 185 the search process towards the selection of predictors that are mutually max-
 186 imally dissimilar (Ding and Peng, 2005).

187 The cardinality of S , or number of selected inputs, represents the third
 188 objective (to be minimized). It is defined as

$$f_3(S) = |S|. \quad (10)$$

189 The predictive accuracy (to be maximized) is defined as

$$f_4(S) = SU(y, \hat{y}(S)), \quad (11)$$

190 where $SU(y, \hat{y}(S))$ is the symmetric uncertainty between the measured out-
 191 put y and the prediction $\hat{y}(S)$ obtained using subset S (and a pre-selected
 192 model architecture). The main advantage behind the adoption of an entropy-
 193 based metric of predictive accuracy is that it synthesises the distributional
 194 information contained in the flow duration curve of the measured and predict-
 195 ed streamflow (Pechlivanidis et al., 2014)—there is indeed a growing interest
 196 on hydrologic model evaluation based on information theoretic metrics; see
 197 Weijs et al. (2010); Pechlivanidis et al. (2014); Castillo et al. (2015), and ref-
 198 erences therein. Note that the models, once identified, can still be evaluated
 199 using more traditional metrics of predictive accuracy based, for example, on
 200 the residual error between observed and predicted time series (see Section 5).

201 *2.3.2. Search process*

202 The search process is conducted in three steps (see Algorithm 1 for a
 203 tabular version of the algorithm):

204 **Step 1.** A global, multi-objective optimization algorithm is employed to find
 205 alternate subsets of predictors, that is, a set (archive) A of Pareto-
 206 efficient solutions with respect to the objective functions $f_1(\cdot)$, $f_2(\cdot)$,
 207 $f_3(\cdot)$ and $f_4(\cdot)$.

208 **Step 2.** The best accuracy value f_4^* of the subsets in A is calculated, and
 209 the δ -quasi equally informative subsets are found on the basis of f_4^* and
 210 a pre-selected value of δ .

211 **Step 3.** Each δ -quasi equally informative subset is compared against all oth-
 212 er δ -quasi equally informative subsets in A with the purpose of eliminat-
 213 ing the supersets that do not improve the accuracy of the models—for

example, consider two efficient subsets S_i and S_j , with $S_i \supset S_j$ and $f_4(S_i) \leq f_4(S_j)$. In this circumstance, the subset S_i is eliminated, since it includes all predictors contained in S_j and it does not improve the accuracy of the model. (We refer to Steps 2 and 3 as the δ *elimination procedure*.)

The choice of an adequate optimization algorithm, model architecture and entropy estimation technique is key to equip W-QEISS algorithm with three important features, namely (1) the capability of finding a large number of equally informative subsets of predictors (ideally all, for a given value of δ), (2) modelling flexibility (i.e., capability of dealing with both linear and non-linear relationships), and (3) computational efficiency. These features are guaranteed by adoption of Borg MOEA, Extreme Learning Machines and frequency histograms.

2.3.3. Implementation

Borg MOEA

In this study we employ Borg MOEA as global, multi-objective optimization algorithm (Hadka and Reed, 2013). Similar to other evolutionary algorithms, Borg evolves a set of solutions by adopting bio-inspired recombination operators such as cross-over, selection and mutation, and converges to a set of Pareto-efficient solutions. A key aspect of Borg is that these operators are adaptively selected during the search process depending on their ability to generate the best possible solutions. In addition, Borg features an ϵ -box dominance archive and exploits time continuation. The concept of

238 ϵ -box dominance refers to the subdivision of the objectives space into hyper-
 239 boxes with side-length ϵ and the selection of Pareto-efficient solutions from
 240 each ϵ -box to ensure convergence and diversity of the search process. Time
 241 continuation entails the injection of mutated solutions from the ϵ -box domi-
 242 nance archive to diversify the population, thus maintaining search diversity
 243 and avoiding getting stuck in local minima. These characteristics make Borg
 244 a very efficient algorithm, which has been found to outperform other state-
 245 of-the-art MOEAs in terms of number of solutions returned, scalability to
 246 many-objective problems, ease-of-use and overall consistency across multiple
 247 problems (Reed et al., 2013).

248 **Extreme Learning Machines**

249 Extreme Learning Machines (ELM) were first introduced by Huang et al.
 250 (2006) as an alternative learning algorithm for single-hidden layer feedfor-
 251 ward neural networks. Unlike traditional methods that seek the best com-
 252 bination of parameter values (Maier et al., 2010), ELM randomly assign the
 253 values of input weights and hidden layer biases, thus reducing the calibra-
 254 tion problem to the determination of the output layer weights. Huang et al.
 255 (2006) proved that under appropriate conditions (e.g., infinitely differentiable
 256 activation functions in the hidden layer) the calibration of the output layer
 257 weights can be performed by solving a linear system. This simplification dras-
 258 tically increases the computational efficiency of the learning process—this is
 259 particularly relevant in the context of our study, where the search for equally
 260 informative subsets of predictors is informed by the predictive accuracy of
 261 several data-driven models. Yet, ELM have been shown to achieve better
 262 performance than traditional calibration techniques (e.g., back-propagation)

263 (Huang et al., 2012).

264

265 The output of an ELM with n inputs, n_h hidden neurons and one output
 266 can be written as

$$\mathbf{y}_j = \sum_{i=1}^{n_h} \beta_i \cdot g(\mathbf{w}_i^T \mathbf{x}_j + b_i), \text{ for } j = 1, 2, \dots, N, \quad (12)$$

267 where N is the total number of available samples, $\mathbf{x}_j$ the input vector for the
 268 j -th sample and b_i the bias parameter of the i -th hidden neuron, while $\mathbf{w}_i =$
 269 $(w_{i,1}, w_{i,2}, \dots, w_{i,n})^T$ and β_i are the weights of the connections linking the
 270 i -th hidden neuron with the inputs and output, respectively. The activation
 271 function $g(\cdot)$ can be any bounded non-constant piecewise continuous function.
 272 Eq. 12 can be rewritten as $\mathbf{Y} = \mathbf{H}\boldsymbol{\beta}$, where $\mathbf{H}$ is an $N \times n_h$ matrix, whose
 273 ji -th entry is $g(\mathbf{w}_i^T \mathbf{x}_j + b_i)$. As shown in Huang et al. (2006), the smallest
 274 norm least-squares solution is unique, and it is expressed as $\hat{\boldsymbol{\beta}} = \mathbf{H}^\dagger \mathbf{Y}$, where
 275 $\mathbf{H}^\dagger$ is the Moore-Penrose generalized inverse of matrix $\mathbf{H}$.

276 Entropy estimation

277 The evaluation of the objective functions f_1 , f_2 and f_4 requires the es-
 278 timation of multiple measures of symmetric uncertainty, which involves the
 279 estimation of marginal and joint entropies (Eqs. 3 and 4). This in turn
 280 builds on the estimation of marginal and joint probability density function-
 281 s. A number of different methods available to this purpose exists, with the
 282 most common being the construction of frequency histograms (bin count-
 283 ing), kernel density estimators and average shifted histogram (Scott, 2012).
 284 As recently discussed in Gong et al. (2014), the application of these methods

to hydrological time series raises a few practical issues (e.g., the presence of many zero values or the skewness of the observational data) that may lead to errors in density and entropy estimation. In our study we adopt frequency histograms, which rely on the idea of dividing the data range in bins and then estimating the densities on the basis of the frequencies in each bin. The rationale behind this choice is that frequency histograms provide the best trade-off between computational efficiency and estimation accuracy (Gong et al., 2014).

3. Synthetic Test Problems

To assess the competence of W-QEISS algorithm in identifying quasi equally informative subsets of predictors, we first apply it to two synthetic datasets with a priori known dependence relations. The results from these experiments can be used to gather insights on how W-QEISS performances are affected by different values of the tolerance δ (as well as the number of observations and noise in the dataset), and therefore provide some empirical guidelines for real-world applications.

3.1. Problems description

We consider linear and nonlinear test problems, whose details are outlined in Table 1. The linear problem features eight candidate predictors, all sampled from a normal (Gaussian) distribution, yielding a total of $2^8 = 256$ different subsets of predictors. The output variable Y depends on four alternate subsets of predictors (i.e., $\{X1, X2, X3\}$, $\{X1, X2, X8\}$, $\{X1, X3, X6, X7\}$ and $\{X1, X6, X7, X8\}$) and an error term ε modelled as a Gaussian random variable with zero mean and standard deviation σ . Following the definitions

309 given in Section 2.1, all candidate predictors can be classified as strongly rel-
 310 evant, weakly relevant and irrelevant (see Table 1). The nonlinear problem
 311 is setup in a similar fashion. We adopt six candidate predictors, for a total
 312 of $2^6 = 64$ different subsets of predictors, and we model the output variable
 313 Y as a function of four alternate subsets of predictors (i.e., $\{X1, X2, X3\}$,
 314 $\{X2, X3, X5\}$, $\{X3, X5, X6\}$ and $\{X1, X3, X6\}$) and an error term ε (with
 315 zero mean and standard deviation σ). The W-QEISS algorithm should ide-
 316 ally return only the optimal, alternate subsets, since subsets lacking relevant
 317 predictors are characterized by lower values of relevance $f_1(\cdot)$ and predictive
 318 accuracy $f_4(\cdot)$, while supersets of the optimal combinations are penalized in
 319 the search process by higher values of redundancy $f_2(\cdot)$ and cardinality $f_3(\cdot)$.

320 3.2. *Experimental setup*

321 For each model, we consider six values of the dataset length N (i.e.,
 322 50, 100, 200, 500, 1000 and 2000) and five values of σ (i.e., 0, 0.005, 0.01,
 323 0.025 and 0.05). The data generation process is repeated 30 times for each
 324 combination of N and σ to provide a statistical analysis of the algorithm per-
 325 formance. We thus have a total of 1,800 datasets. The W-QEISS algorithm
 326 is run on each independent instance, with the results averaged across the 30
 327 repetitions. We post-process the results using six values of δ (i.e., 0.1, 1, 2.5,
 328 5, 10 and 20 %) to control the number of quasi equivalent subsets returned
 329 by the algorithm.

330 The parameterization of Borg MOEA, Extreme Learning Machines and fre-
 331 quency histograms is carried out using available guidelines as well as a trial-
 332 and-error process. For Borg MOEA, we set the number of evaluations and
 333 the value of ϵ to 5,000 and 10^{-3} , respectively (Karakaya et al., 2016). As for

334 the Extreme Learning Machines setting, we follow the guidelines given by
 335 Huang et al. (2006). We normalize the predictors and output variables into
 336 ranges $[-1, 1]$ and $[0, 1]$, and we use a sigmoidal function for the activation
 337 function $g(\cdot)$. The number n_h of hidden neurons is set to 20, while a k -fold
 338 cross-validation process (with $k = 10$) is used to estimate the predictive ac-
 339 curacy of each model—these values are found by exhaustively exploring the
 340 combinations of 5, 10 and 20 hidden nodes and 5, 10, 15 and 20 folds. The
 341 cross-validation is repeated five times to minimize the standard deviation of
 342 the predictive accuracy estimate; the average accuracy is then taken. Finally,
 343 the implementation of the frequency histograms is based on the division of
 344 the data range into 20 bins of equal width—this value was found optimal
 345 across the values 5, 10 and 20. All experiments are carried out in MatLab on
 346 a Dual Intel Xeon CPU E5-2630 v3 @ 2.40 GHz with 32 GB RAM running
 347 Microsoft Windows 8.1. The average runtime is 4.450 ± 0.119 and $23.625 \pm$
 348 1.090 seconds for the linear and nonlinear test problems, respectively.
 349 The selection accuracy of W-QEISS is assessed by considering both the num-
 350 ber of equally informative subsets correctly specified (hereafter identified as
 351 C) and the number of sub-optimal subsets wrongly retained (W). A per-
 352 fect specification of the optimal subsets for both cases occurs when $C = 4$
 353 and $W = 0$. On the other hand, complete misspecification corresponds to
 354 $C = 0$ and $W = 252$ for the linear problem, and $C = 0$ and $W = 60$ for the
 355 nonlinear one.

356 3.3. Results

357 The results are reported in Figure 2 and 3, which illustrate the average
 358 values of equally informative subsets correctly specified (C , Figure 2) and

359 sub-optimal subsets wrongly retained (W , Figure 3) as a function of the
 360 number N of observations, standard deviation σ and tolerance δ . Results for
 361 the metric C show that the capability of W-QEISS algorithm of selecting the
 362 four alternate subsets of predictors is mostly determined by the tolerance δ .
 363 In particular, the algorithm identifies all equivalent subsets when δ is equal
 364 to 10 and 5 % for the linear and nonlinear case, respectively. This is an
 365 expected result: the search process is affected by different modelling errors
 366 (e.g., estimation of entropies, cross-validation of a data-driven model etc.),
 367 which are further exacerbated by the error term ε , so a small, non-zero value
 368 of the tolerance δ is needed to identify all equally informative subsets. The
 369 average value of C is not strongly influenced by the number of observations
 370 and error term, meaning that W-QEISS can cope well with small, noisy
 371 datasets (if δ is sufficiently large). Obviously, increasing the value of δ leads
 372 the algorithm towards the selection of more quasi equally informative subsets,
 373 with the associated risk of including some wrong subsets. This effect is clearly
 374 depicted in Figure 3, which shows the increase in W as a function of δ . In
 375 such case, the value of N and σ plays a more important role—large, less noisy
 376 datasets help reduce the value of W , especially for values of δ larger than 1 %.
 377 Overall, the results show that a perfect specification (i.e., $C = 4$ and $W = 0$)
 378 is obtained when the tolerance δ is at least 5 % and the dataset at hand is
 379 sufficiently large (i.e., $N \geq 500$) and not excessively noisy (i.e., $\sigma \leq 0.025$).
 380 This is in line with the results obtained by Galelli et al. (2014), who found
 381 that the presence of high noise is critical when evaluated in relation to the
 382 number of observations for several input variable selection algorithms.

383 4. Case study: the Yampa River

384 4.1. Study site, data availability and experimental setup

385 The Yampa River is a 400 km long stream contributing into the Colorado
386 River through its main affluent, the Green River. The Yampa River origi-
387 nates in the Rocky Mountains and it drains over 20,000 km² of northwestern
388 Colorado and southern Wyoming, passing through the cities of Steamboat
389 Springs, Hayden and Craig (Figure 4). Daily streamflow discharge data
390 (*FLOW*) are recorded at the USGS gauging station near Maybell, around
391 50 km upstream of the intersection between the Yampa River and the Little
392 Snake River. The catchment area at Maybell is of 8,762 km², around one
393 third of the total Yampa River basin. Daily hydro-meteorological data are
394 provided by eight NOAA stations (S1-S8) located both in the valleys (S1-S5)
395 and in the headwaters areas (S6-S8) (see Figure 4). Several variables are
396 recorded at each station. Daily rainfall (*PRCP*), snowfall (*SNOW*), snow
397 depth (*SNWD*), minimum (*TMIN*) and maximum (*TMAX*) temperatures
398 are gathered in the S1-S5 stations (but for *TMIN* and *TMAX* in station
399 S2). In the headwaters stations S6-S8, only the water equivalent of the snow
400 depth (*WESD*), along with *TMIN* and *TMAX*, are recorded. In addition
401 to these variables, the streamflow specific conductance (*SC*) measured at
402 Maybell is included among the potential predictors (see Table 2 for further
403 details). The specific conductance is a variable related to the concentra-
404 tion of dissolved solids (i.e., salts) in the water, and it is known to act as
405 a proxy of the discharge in snowmelt-dominated rivers (Miller et al., 2014).
406 While streamflow originating from freshly melted snow is characterized by
407 low values of *SC*, streamflow originating from groundwater flow has greater

408 concentration of dissolved salts due to longer contact times with subsurface
 409 materials. This implies that high SC is measured during the low-flow season,
 410 when deep regional baseflow is the dominant component of streamflow, while
 411 high flows induced by melting snow are characterized by lower values of SC.
 412 A previous study concerning several catchments in the Upper Colorado Riv-
 413 er Basin has shown that these dynamics are particularly noticeable in the
 414 Yampa River basin (Miller et al., 2014).

415

416 As reported in Table 2, there is a total of 34 time series (spanning from
 417 November 1998 to December 2010) that can be employed to build the pool of
 418 candidate predictors for the one day-ahead streamflow prediction of the Yam-
 419 pa River at Maybell ($FLOW_{t+1}$). In order to account for the lagged values
 420 of each variable while at the same time limiting the number of possible pre-
 421 dictors, one single predictor is built from each time series using aggregating
 422 functions. In particular, antecedent precipitation indexes (Ali et al., 2010)
 423 are built for each *PRCP* and *SNOW* time series, while moving averages are
 424 calculated for past *FLOW*, *SC*, *TMAX* and *TMIN* observations as well as
 425 for the gradients of *SNWD* and *WESD*. The gradients are preferred for the
 426 latter variables since differences between the values recorded in two adjacent
 427 days are expected to better reflect the occurrence of snowmelt. Each aggre-
 428 gated variable is computed using the observations of the three days preceding
 429 that of the forecast. For example, the predictor $TMAX_t^{S5}$ is constructed by
 430 averaging the maximum temperature measured in S5 during days t , $t - 1$ and
 431 $t - 2$. The final dataset consists of 2,240 input-output observations and 34
 432 candidate predictors, giving a total of $2^{34} \approx 1.718 \cdot 10^{10}$ combinations of pre-

dictors. (Note that the name of the station is not specified in the superscript when the variable is recorded at Maybell USGS gauging station). Figure 5 displays the value of the symmetric uncertainty (Eq. 7) estimated between each candidate predictor and the output (first row) and between each pair of predictors. As expected, the autoregressive term $FLOW_t$ appears to be strongly correlated to the output $FLOW_{t+1}$ —recall that the symmetric uncertainty varies between 0 and 1. The specific conductance SC_t and the water equivalent of the snow depth $WESD_t^{S6,S7,S8}$ also seem to carry more explanatory power than the other hydro-meteorological variables. The values of symmetric uncertainty also show the strong spatial correlation between variables of the same type (e.g., temperature).

Following the results outlined in Section 3.3, the tolerance δ of W-QEISS algorithm is set to 10 %, since this value was found to provide a reasonable trade-off between the selection of quasi equally informative subsets and the inclusion of wrong predictors. All parameters of Borg MOEA, Extreme Learning Machines and frequency histograms are not changed, with the only exception of the number n_h of hidden neurons, which is increased to 50 because of the larger number of variables involved in the modelling process (Lima et al., 2015)—a sensitivity analysis of W-QEISS performance with respect to the number of hidden neurons is reported in the supplementary material.

4.2. Benchmark algorithms and performance metrics

The results obtained by W-QEISS algorithm are compared against those achieved by four input variable selection techniques commonly employed in

458 data-driven hydrological applications, that is, Partial Correlation Input Se-
 459 lection (PCIS) (May et al., 2008), Partial Mutual Information Selection (P-
 460 MIS) (Sharma, 2000; May et al., 2008; Fernando et al., 2009), Iterative Input
 461 variable Selection (IIS) (Galelli and Castelletti, 2013b) and the modified Min-
 462 imum Redundancy Maximum Relevance (mMRMR) algorithm (Hejazi and
 463 Cai, 2009). These are filter techniques designed to identify only one specific
 464 subset of predictors—see Appendix A for additional details. Their param-
 465 eterization is set following the guidelines given in Galelli et al. (2014) and
 466 Hejazi and Cai (2009). (We also compared W-QEISS against a bi-objective
 467 wrapper designed to maximize the model predictive accuracy and and mini-
 468 mize the subset cardinality. This comparison is reported in the supplemen-
 469 tary material).

470 The comparison is carried out in terms of selected predictors (and as-
 471 sociated physical interpretability) and predictive accuracy—for consistency,
 472 an Extreme Learning Machine with 50 hidden neurons is employed to es-
 473 timate the predictive accuracy (i.e., symmetric uncertainty) associated to
 474 the subsets identified by the benchmark algorithms. Four additional metrics
 475 of predictive accuracy are used to thoroughly assess the goodness-of-fit of
 476 the models. These metrics are the Kling–Gupta Efficiency (KGE, -) (Gupta
 477 et al., 2009), the Fourth Root Mean Quadrupled Error (R4MS4E, m^3/s), the
 478 Mean Squared Logarithmic Error (MSLE, -) and the Mean Squared Deriva-
 479 tive Error (MSDE, $(\text{m}^3/\text{s})^2$). KGE efficiently accounts for the importance
 480 of various hydrograph components (e.g., correlation, bias), while R4MS4E
 481 and MSLE are used to assess model performance for high and low flows,
 482 respectively (Dawson et al., 2007; De Vos and Rientjes, 2007). MSDE is an

indicator of the fit of the shape of the hydrograph, and it specifically penalizes timing errors and noise (De Vos and Rientjes, 2007). Further details on the performance metrics are given in Appendix B.

5. Results and Discussion

5.1. Quasi Equally Informative Subsets

The four-dimensional Pareto frontier returned by W-QEISS algorithm (without δ -elimination procedure) is displayed in Figure 6. The values of relevance $f_1(\cdot)$, redundancy $f_2(\cdot)$, and cardinality $f_3(\cdot)$ are reported on the x-, y- and z-axis, while $f_4(\cdot)$ is shown using a color-scale. (Since all objective functions but for $f_3(\cdot)$ are in the range $[0,1]$, the values of $f_3(\cdot)$ are normalized in the same range). The algorithm returns a total of 1,302 quasi equally informative subsets spreading across a well defined area of the search space. The two-dimensional projections illustrate the relationships between the objective functions: large subsets are characterized by high relevance (Figure 6, bottom left), high redundancy (Figure 6, bottom right), but not necessarily by high predictive accuracy (Figure 6, top right). Note that for a given cardinality there exist several combinations of predictors yielding different values of the other objective functions.

Figure 7 shows the outcome of the δ -elimination procedure carried out for δ equal to 10 %. The procedure removes over 95 % of the subsets and returns 56 alternate subsets of predictors (10 % quasi equally informative subsets) sited in an area characterized by medium relevance ($f_1 \in [0.270, 0.679]$) and low redundancy ($f_2 \in [0.005, 0.163]$), with the number of predictors ranging

507 between 4 and 14 ($f_3 \in [0.118, 0.412]$). The values of symmetric uncertainty
 508 vary between 0.672 and 0.745. The values of the objective functions associ-
 509 ated to the subsets identified by the benchmark techniques are also displayed
 510 (see Table 3 for their numerical values). Interestingly, none of the benchmark
 511 solutions neither dominates nor overlaps with the W-QEISS ones. Indeed, it
 512 emerges that these four combinations of predictors are also located within,
 513 or at the borders, of the area occupied by W-QEISS solutions. The bench-
 514 mark solutions tend to have few inputs and low redundancy, but they are also
 515 characterized by lower relevance and predictive accuracy (with respect to the
 516 W-QEISS ones). Their values suggest that the PMIS solution is superior to
 517 the other benchmarks in terms of cardinality (three predictors selected) and
 518 predictive accuracy. The cardinality of the mMRMR solution is also equal
 519 to three, but it results in better values of relevance f_1 and redundancy f_2 .
 520 Nonetheless, its predictive accuracy is lower than that of PMIS and most
 521 of W-QEISS solutions, indicating that the mMRMR technique struggles to
 522 identify the subsets with highest explanatory power. The IIS and PCIS
 523 subsets are larger—with four and seven inputs, respectively—but they seem
 524 to carry more redundant information, resulting in lower predictive accuracy
 525 with respect to the PMIS and most of W-QEISS solutions.

526 5.2. *Analysis of selected predictors*

527 Figure 8 reports the composition of each alternate subset of predictors
 528 sorted in ascending order of cardinality and predictive accuracy. A rectangu-
 529 lar marker is placed at the intersection between the row identifying a given
 530 subset and the columns pertaining to the selected predictors. The marker
 531 color varies with the cardinality of the subset, with darker shades of gray indi-

532 cating larger subsets. Insights on predictors' relevance can be obtained from
 533 an inspection of the vertical bars traced by joining these markers across mul-
 534 tiple rows. Long and uninterrupted bars are expected for strongly relevant
 535 predictors that cannot be substituted by other input combinations without
 536 incurring into a substantial drop of predictive accuracy. Increasing gaps are
 537 found when considering predictors with progressively weaker relevance, while
 538 irrelevant inputs are recognizable by isolated markers or their total absence.
 539 This information is summarised in the topmost row of Figure 8 that reports
 540 the frequency with which each individual predictor is selected among the 56
 541 alternate subsets of predictors.

542

543 From this analysis it emerges that the autoregressive term $FLOW_t$ is the
 544 most relevant among all candidates, since it is included in all solutions—
 545 the predictor $FLOW_t$ is indeed strongly correlated to the output $FLOW_{t+1}$
 546 (see Figure 5). The specific conductance SC_t is arguably the second most
 547 significant predictor, since it is selected in about 84 % of the subsets and
 548 for each level of cardinality. As discussed in Section 4.2, SC_t is suitable
 549 for streamflow prediction in snowmelt-driven rivers because it is a proxy
 550 variable of the snowmelt process. The water equivalent of the snow depth
 551 $WESD_t^{S6,S7,S8}$ seems to carry substantial explanatory power, with a selection
 552 frequency ranging from 76.79 % to 87.50 %. It is interesting to note the con-
 553 trast with the other snow-related predictors (i.e., snowfall and snowdepth),
 554 which are at most selected in 10.71 % ($SNOW_t^{S1}$) and 17.86 % ($SNWD_t^{S3}$)
 555 of the solutions. Indeed, the water equivalent of the snow depth synthesises
 556 the information regarding the amount of snow-generated runoff flowing into

the Yampa River, so it is more informative than the snowdepth. The precipitation $PRCP_t^{S2}$ in station $S2$ is the only rainfall-related variable being frequently selected; this may be due to its central location and proximity to the gauging station. Although individual temperature predictors are selected in about one third of the subsets, most of the alternate subsets of predictors include at least one minimum or maximum temperature. This suggests that the (minimum or maximum) temperature provides some information regarding the dynamics of the snow accumulation and melting process.

The composition of the alternate subsets of predictors is somewhat reflected in the subsets identified by the benchmark algorithms, which are reported at the bottom of Figure 8. All methods select the autoregressive term $FLOW_t$, which is the first strongly relevant variable selected by W-QEISS algorithm. The specific conductance SC_t is selected by PMIS and PICS; this might explain why the data-driven models built on these subsets perform better than those associated to mMRMR and IIS. The water equivalent of the snow depth (at least one predictor between $WESD_t^{S6,S7,S8}$) is selected by all benchmark algorithms apart from PMIS, which nonetheless yields the best predictive accuracy.

5.3. Effect on multiple measures of performance

A final assessment of the predictive accuracy is carried out by calculating four additional metrics, i.e., KGE, R4MS4E, MSLE and MSDE, that account for different aspects of model performance (Section 4.2). Their values are reported in the last columns of Figure 8 and illustrated in Figure 9. Results show that the majority of the alternate subsets of predictors i-

582 identified by W-QEISS algorithm (about 75 %) yields better performance in
 583 terms of hydrograph components (KGE), high flows (R4MS4E) and shape
 584 of the hydrograph (MSDE) with respect to the benchmark solutions. Fur-
 585 thermore, the best solution found by W-QEISS (with respect to each single
 586 metric) significantly improves the benchmark solutions—from 1 to 2.5 % for
 587 KGE, 23 to 64 % for R4MS4E, 35 to 61 % for MSDE. The W-QEISS per-
 588 formance in terms of low flows is somewhat lower, with the PMIS algorithm
 589 providing indeed the best model. This fact can be attributed to the train-
 590 ing process of the Extreme Learning Machines—the underlying model used
 591 by W-QEISS algorithm to compute the predictive accuracy that informs the
 592 search process. As explained in Section 2.3.3, the calibration of the Extreme
 593 Learning Machines entails the minimization of the residual sum of squares,
 594 possibly biasing the search process towards the selection of subsets that per-
 595 form well in high-flow conditions. Nonetheless, it is still possible to choose
 596 some W-QEISS solutions that outperform all benchmarks in terms of KGE,
 597 R4MS4E and MSDE, while still providing reasonable performance on low
 598 flows—see, for example, solutions 20 and 48 in Figure 8. A comparison be-
 599 tween the measured and predicted streamflow (by these two solutions as well
 600 as the benchmarks) is given in Figure 10. The trajectories show that the
 601 two W-QEISS solutions predict well the shape of the hydrograph (in terms
 602 of correlation, bias and timing) for high-flow events (bottom panel), while
 603 the performance slightly decreases on low flows (top panel).

604 6. Conclusions

605 We sought to quantify the uncertainty associated to the selection of pre-
606 dictors yielding the same, or similar, predictive accuracy for data-driven hy-
607 drological models. To identify alternate subsets of predictors in a given
608 dataset, we developed a novel input variable selection algorithm (Wrapper
609 for Quasi Equally Informative Subset Selection) that performs the search
610 process in a four-dimensional space of relevance, redundancy, cardinality and
611 predictive accuracy. W-QEISS algorithm is designed to identify one subset
612 with the highest predictive accuracy and multiple subsets with similar infor-
613 mation content—i.e., quasi equally informative subsets. This is in contrast
614 with traditional variable selection algorithms, which are conceived to identify
615 only one subset of predictors.

616

617 Results on a streamflow prediction problem with 34 potential predictors
618 show that there exists a large number of alternate subsets of predictors, with
619 both equal and different cardinalities. Studying the structure of these sub-
620 sets reveals the relative importance of the candidate predictors, which can
621 be classified as strongly relevant, weakly relevant and irrelevant. The prac-
622 tical import of this additional information is that the modeller can gain a
623 deeper understanding of the underlying physical processes and, ultimately,
624 influence the way some processes are monitored. For example, in the context
625 of hydrological catchment modelling, where hydro-meteorological processes
626 are characterised by strong spatial dependence, knowing that one specific
627 predictor can be replaced without loss of information by one (or multiple)
628 predictor(s) may influence the design or maintenance of monitoring network-

s. Results also show that the availability of alternate subsets of predictors help find better trade-offs between multiple measures of predictive accuracy.

Further research will focus on four directions. First, the accuracy of W-QEISS algorithm may be further improved by developing novel methods for the estimation of joint and marginal entropies that are not prone to errors when dealing with the presence of many zero values or the skewness of observational data—some advances in this direction were recently shown by Gong et al. (2014) and Li et al. (2015). Second, the calibration of the underlying data-driven model (Extreme Learning Machines) could be modified to improve the model performance on low flows. This may be achieved by adopting formulations of the error term that do not bias the model evaluation process in favour of high magnitude events. Third, it would be interesting to adopt some static or dynamic model combination schemes (Devineni et al., 2008; Sharma and Chowdhury, 2011) to combine the predictions associated to the different models identified by the W-QEISS framework. Ensemble modelling may also create an opportunity to account for the uncertainty associated to the structure and parameters of the models. Another promising research direction is to investigate the scalability of W-QEISS algorithm to high dimensional problems—such as statistical downscaling, seasonal hydro-meteorological forecasts or median flood regionalization—that generally involve a large number of spatially-distributed predictors.

651 **Acknowledgement**

652 The authors are supported by the Singapore University of Technology and
653 Design under Grant SRG ESD 2013 061 and SRG ESD 2012 033. A MatLab
654 implementation of the W-QEISS algorithm is available at [https://github.](https://github.com/stefano-galelli/Matlab-Multi-objective-Feature-Selection)
655 [com/stefano-galelli/Matlab-Multi-objective-Feature-Selection](https://github.com/stefano-galelli/Matlab-Multi-objective-Feature-Selection). The
656 datasets used for this study can be accessed online at [http://waterdata.](http://waterdata.usgs.gov/nwis)
657 [usgs.gov/nwis](http://waterdata.usgs.gov/nwis) and <http://www.ncdc.noaa.gov/>.

658 **References**

- 659 Ali, S., Ghosh, N., Singh, R., 2010. Rainfall–runoff simulation using a normal-
660 ized antecedent precipitation index. *Hydrological Sciences Journal–Journal*
661 *des Sciences Hydrologiques* 55 (2), 266–274.
- 662 Alvisi, S., Franchini, M., 2011. Fuzzy neural networks for water level and dis-
663 charge forecasting with uncertainty. *Environmental Modelling & Software*
664 26 (4), 523–537.
- 665 Bardsley, W., Vetrova, V., Liu, S., 2015. Toward creating simpler hydrological
666 models: A LASSO subset selection approach. *Environmental Modelling &*
667 *Software* 72, 33–43.
- 668 Castillo, A., Castelli, F., Entekhabi, D., 2015. An entropy-based measure
669 of hydrologic complexity and its applications. *Water Resources Research*
670 51 (7), 5145–5160.
- 671 Ceola, S., Laio, F., Montanari, A., 2015. Human-impacted waters: New

672 perspectives from global high-resolution monitoring. *Water Resources Re-*
673 *search* 51 (9), 7064–7079.

674 Creaco, E., Berardi, L., Sun, S., Giustolisi, O., Savic, D., 2016. Selection of
675 relevant input variables in storm water quality modeling by multiobjective
676 evolutionary polynomial regression paradigm. *Water Resources Research-*
677 *Doi: 10.1002/2015WR017971.*

678 Dawson, C., Harpham, C., Wilby, R., Chen, Y., 2002. Evaluation of artificial
679 neural network techniques for flow forecasting in the River Yangtze, China.
680 *Hydrology and Earth System Sciences Discussions* 6 (4), 619–626.

681 Dawson, C. W., Abrahart, R. J., See, L. M., 2007. HydroTest: a web-based
682 toolbox of evaluation metrics for the standardised assessment of hydrolog-
683 ical forecasts. *Environmental Modelling & Software* 22 (7), 1034–1052.

684 De Vos, N., Rientjes, T., 2007. Multi-objective performance comparison of
685 an artificial neural network and a conceptual rainfall-runoff model. *Hydro-*
686 *logical Sciences Journal* 52 (3), 397–413.

687 Devineni, N., Sankarasubramanian, A., Ghosh, S., 2008. Multimodel ensem-
688 bles of streamflow forecasts: Role of predictor state in developing optimal
689 combinations. *Water Resources Research* 44 (9).

690 Ding, C., Peng, H., 2005. Minimum redundancy feature selection from mi-
691 croarray gene expression data. *Journal of Bioinformatics and Computa-*
692 *tional Biology* 3 (02), 185–205.

693 Fernando, T., Maier, H., Dandy, G., 2009. Selection of input variables for data driven models: An average shifted histogram partial mutual information
694 estimator approach. *Journal of Hydrology* 367 (3), 165–176.

696 Fraternali, P., Castelletti, A., Soncini-Sessa, R., Ruiz, C. V., Rizzoli, A. E.,
697 2012. Putting humans in the loop: Social computing for water resources
698 management. *Environmental Modelling & Software* 37, 68–77.

699 Galelli, S., Castelletti, A., 2013a. Assessing the predictive capability of randomized tree-based ensembles in streamflow modelling. *Hydrology and*
700 *Earth System Sciences* 17 (7), 2669–2684.

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

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

708 Gong, W., Yang, D., Gupta, H. V., Nearing, G., 2014. Estimating information entropy for hydrological data: One-dimensional case. *Water Resources*
709 *Research* 50 (6), 5003–5018.

711 Gupta, H. V., Kling, H., Yilmaz, K. K., Martinez, G. F., 2009. Decomposition of the mean squared error and nse performance criteria: Implications for
712 improving hydrological modelling. *Journal of Hydrology* 377 (1), 80–91.

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

716 Hadka, D., Reed, P., 2013. Borg: An auto-adaptive many-objective evolu-
717 tionary computing framework. *Evolutionary Computation* 21 (2), 231–259.

718 Hejazi, M. I., Cai, X., 2009. Input variable selection for water resources
719 systems using a modified minimum redundancy maximum relevance (mM-
720 RMR) algorithm. *Advances in Water Resources* 32 (4), 582–593.

721 Huang, G.-B., Zhou, H., Ding, X., Zhang, R., 2012. Extreme learning ma-
722 chine for regression and multiclass classification. *IEEE Trans. Syst. Man,*
723 *Cybern. B, Cybern. (1995-2012)* 42 (2), 513–529.

724 Huang, G.-B., Zhu, Q.-Y., Siew, C.-K., 2006. Extreme learning machine:
725 theory and applications. *Neurocomputing* 70 (1), 489–501.

726 Karakaya, G., Galelli, S., Ahipasaoglu, S. D., Taormina, R., 2016. Identifying
727 (quasi) equally informative subsets in feature selection problems for classi-
728 fication: A max-relevance min-redundancy approach. *IEEE Transactions*
729 *on Cybernetics* 46 (6), 1424–1437.

730 Khan, M. S., Coulibaly, P., 2006. Bayesian neural network for rainfall-runoff
731 modeling. *Water Resources Research* 42 (7).

732 Kingston, G. B., Maier, H. R., Lambert, M. F., 2008. Bayesian model selec-
733 tion applied to artificial neural networks used for water resources modeling.
734 *Water Resources Research* 44 (4).

735 Li, X., Maier, H. R., Zecchin, A. C., 2015. Improved PMI-based input vari-
736 able selection approach for artificial neural network and other data driv-
737 en environmental and water resource models. *Environmental Modelling &*
738 *Software* 65, 15–29.

739 Lima, A. R., Cannon, A. J., Hsieh, W. W., 2015. Nonlinear regression in
740 environmental sciences using extreme learning machines: A comparative
741 evaluation. *Environmental Modelling & Software* 73, 175–188.

742 Liu, Y., Tang, F., Zeng, Z., 2015. Feature selection based on dependency
743 margin. *IEEE Transactions on Cybernetics* 45 (6), 1209–1221.

744 MacKay, D. J., 2003. Information theory, inference and learning algorithms.
745 Cambridge university press.

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

750 May, R. J., Maier, H. R., Dandy, G. C., Fernando, T. G., 2008. Non-linear
751 variable selection for artificial neural networks using partial mutual infor-
752 mation. *Environmental Modelling & Software* 23 (10), 1312–1326.

753 Miller, M. P., Susong, D. D., Shope, C. L., Heilweil, V. M., Stolp, B. J.,
754 2014. Continuous estimation of baseflow in snowmelt-dominated streams
755 and rivers in the Upper Colorado River Basin: A chemical hydrograph
756 separation approach. *Water Resources Research* 50 (8), 6986–6999.

757 Mount, N. J., Maier, H. R., Toth, E., Elshorbagy, A., Solomatine, D., Chang,
758 F.-J., Abrahart, R., 2016. Data-driven modelling approaches for socio-
759 hydrology: Opportunities and challenges within the panta rhei science
760 plan. *Hydrological Sciences Journal* Doi: 10.1080/02626667.2016.1159683.

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

766 Overeem, A., Leijnse, H., Uijlenhoet, R., 2013. Country-wide rainfall maps
767 from cellular communication networks. *Proceedings of the National Acade-*
768 *my of Sciences* 110 (8), 2741–2745.

769 Parasuraman, K., Elshorbagy, A., 2008. Toward improving the reliability of
770 hydrologic prediction: Model structure uncertainty and its quantification
771 using ensemble-based genetic programming framework. *Water Resources*
772 *Research* 44 (12).

773 Pechlivanidis, I., Jackson, B., McMillan, H., Gupta, H., 2014. Use of an
774 entropy-based metric in multiobjective calibration to improve model per-
775 formance. *Water Resources Research* 50 (10), 8066–8083.

776 Peng, H., Long, F., Ding, C., 2005. Feature selection based on mutual infor-
777 mation criteria of max-dependency, max-relevance, and min-redundancy.
778 *IEEE Transactions on pattern analysis and machine intelligence* 27 (8),
779 1226–1238.

780 Phatak, A., Bates, B. C., Charles, S. P., 2011. Statistical downscaling of rain-
781 fall data using sparse variable selection methods. *Environmental Modelling*
782 *& Software* 26 (11), 1363–1371.

783 Pianosi, F., Raso, L., 2012. Dynamic modeling of predictive uncertainty by
784 regression on absolute errors. *Water Resources Research* 48 (3).

785 Quilty, J., Adamowski, J., Khalil, B., Rathinasamy, M., 2016. Bootstrap
786 rank-ordered conditional mutual information (broCMI)—A nonlinear in-
787 put variable selection method for water resources modeling. *Water Re-
788 sources Research* Doi: 10.1002/2015WR016959.

789 Reed, P., Hadka, D., Herman, J., Kasprzyk, J., Kollat, J., 2013. Evolutionary
790 multiobjective optimization in water resources: The past, present, and
791 future. *Advances in Water Resources* 51, 438–456.

792 Scott, D., 2012. Multivariate density estimation and visualization. In: *Hand-
793 book of Computational Statistics*. Springer, pp. 549–569.

794 Shannon, C. E., 1948. A mathematical theory of communication. *Bell Syst.
795 Tech. J.* 27, 379–423, 623–656.

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

799 Sharma, A., Chowdhury, S., 2011. Coping with model structural uncertainty
800 in medium-term hydro-climatic forecasting. *Hydrology Research* 42 (2-3),
801 113–127.

802 Sharma, A., Mehrotra, R., 2014. An information theoretic alternative to
803 model a natural system using observational information alone. *Water Re-
804 sources Research* 50 (1), 650–660.

805 Sharma, S., Tiwari, K., 2009. Bootstrap based artificial neural network
806 (BANN) analysis for hierarchical prediction of monthly runoff in Upper
807 Damodar Valley Catchment. *Journal of hydrology* 374 (3), 209–222.

808 Shrestha, D. L., Solomatine, D. P., 2006. Machine learning approaches for
809 estimation of prediction interval for the model output. *Neural Networks*
810 19 (2), 225–235.

811 Srivastav, R., Sudheer, K., Chaubey, I., 2007. A simplified approach to quan-
812 tifying predictive and parametric uncertainty in artificial neural network
813 hydrologic models. *Water Resources Research* 43 (10).

814 Taormina, R., Chau, K.-W., 2015a. Ann-based interval forecasting of stream-
815 flow discharges\ using the lube method and mofips. *Engineering Applica-*
816 *tions of Artificial Intelligence* 45, 429–440.

817 Taormina, R., Chau, K.-W., 2015b. Data-driven input variable selection for
818 rainfall–runoff modeling using binary-coded particle swarm optimization
819 and extreme learning machines. *Journal of Hydrology* 529, 1617–1632.

820 Tiwari, M. K., Chatterjee, C., 2010. Uncertainty assessment and ensem-
821 ble flood forecasting using bootstrap based artificial neural networks
822 (BANNs). *Journal of Hydrology* 382 (1), 20–33.

823 Wan Jaafar, W., Liu, J., Han, D., 2011. Input variable selection for median
824 flood regionalization. *Water Resources Research* 47 (7).

825 Weijs, S., Schoups, G., Van De Giesen, N., 2010. Why hydrological predic-
826 tions should be evaluated using information theory. *Hydrology and Earth*
827 *System Sciences* 14, 2545–2558.

- 828 Witten, I. H., Frank, E., 2005. Data Mining: Practical Machine Learning
829 Tools and Techniques, 2nd Edition. Morgan Kaufmann Publishers Inc.,
830 San Francisco, CA.
- 831 Young, P., 2001. Data-based mechanistic modelling and validation of rainfall-
832 flow processes. Model validation: perspectives in hydrological science, 117–
833 161.
- 834 Yu, L., Liu, H., 2004. Efficient feature selection via analysis of relevance and
835 redundancy. The Journal of Machine Learning Research 5, 1205–1224.
- 836 Zhang, X., Liang, F., Srinivasan, R., Van Liew, M., 2009. Estimating un-
837 certainty of streamflow simulation using bayesian neural networks. Water
838 Resources Research 45 (2).
- 839 Zhang, X., Liang, F., Yu, B., Zong, Z., 2011. Explicitly integrating param-
840 eter, input, and structure uncertainties into bayesian neural networks for
841 probabilistic hydrologic forecasting. Journal of Hydrology 409 (3), 696–709.

842 **Appendix A**

843 *A.1 Partial Mutual Information Selection (PMIS)*

844 The PMIS algorithm adopts a forward selection strategy (i.e., one vari-
845 able is selected at each iteration) based on the estimation of the Partial
846 Mutual Information (May et al., 2008). PMI is a measure of the partial
847 dependence between each candidate predictor and the output, conditional
848 on the predictors that have been selected at the previous iterations. To esti-
849 mate the PMI, the algorithm adopts a kernel density estimator and a General
850 Regression Neural Network—to account for the residual information in the
851 predictors once the effect of the already selected predictors has been taken
852 in consideration. Among the various kernel density estimators, PMIS adopts
853 the Gaussian reference bandwidth (Scott, 2012), whose accuracy depends on
854 the value of a smoothing parameter (or bandwidth) λ . The termination con-
855 dition of PMIS is based on the coefficient of determination R^2 ; the algorithm
856 stops when the selection of a further predictor leads to a decrease of R^2 in
857 the underlying General Regression Neural Network being identified.

858 *A.2 Partial Correlation Input Selection (PCIS)*

859 The PCIS algorithm adopts the same structure as the PMIS algorithm,
860 but it uses the Pearson correlation coefficient to estimate the strength of the
861 relationship between predictors and output and a multiple linear regression
862 based on least squares (in place of Partial Mutual Information and General
863 Regression Neural Network). PCIS is stopped when the selection of addition-
864 al predictors no longer results in an improvement (increase) in the Bayesian
865 Information Criterion (May et al., 2008).

866 *A.3 Iterative Input variable Selection (IIS)*

867 Similar to PMIS and PCIS, the IIS algorithm relies on a forward selection
868 strategy. The strength of the correlation between candidate predictors and
869 output is quantified with a tree-based ranking method and refined through
870 the identification of p Single-Input Single-Output (SISO) models (with the
871 best performing predictor added to the set of selected variables). The s-
872 election process continues until the accuracy of an underlying Multi-Input
873 Single-Output model, evaluated with a k -fold cross-validation, does not sig-
874 nificantly improve. The value of p and k is set to 5, while the algorithm
875 tolerance ε is equal to 0.01—this means that the IIS algorithm is stopped
876 when the selection of a further variable leads to an increase of R^2 lower than
877 0.01 (Galelli and Castelletti, 2013b). Extremely Randomized Trees are used
878 for both ranking and modelling purposes (Galelli and Castelletti, 2013a).

879 *A.4 modified Minimum Redundancy Maximum Relevance (mMRMR)*

880 This algorithm is a modified version of the well-known Minimum Redun-
881 dancy Maximum Relevance algorithm proposed by Peng et al. (2005). At
882 each iteration of its forward selection strategy, mMRMR selects the predictor
883 that maximizes the value of a mutual information-based metric accounting
884 for both relevance and redundancy of the candidate variables. (In Hejazi
885 and Cai (2009), the metric is modified through the use of a correction fac-
886 tor α to account for some biases in the estimation of the redundancy.) The
887 mMRMR then proceeds by identifying a General Regression Neural Network
888 that models the output as a function of the selected predictor(s). The search
889 process is stopped when the selection of an additional predictor leads to a
890 decrease in the Root Squared Mean Error of the neural network.

891 Appendix B

892 The Kling–Gupta Efficiency (KGE) was proposed by Gupta et al. (2009)
 893 to account for some problems associated with the use of the mean squared
 894 error and Nash–Sutcliffe efficiency when assessing the predictive accuracy
 895 of a model on various components of the hydrograph (e.g., correlation and
 896 bias). As discussed in Gupta et al. (2009), the Nash–Sutcliffe efficiency
 897 can be decomposed into three components, namely the correlation, bias and
 898 variability of the measured and predicted discharge. The KGE is built on
 899 this decomposition, and it is defined as the Euclidian distance of the three
 900 components from the ideal point, that is

$$KGE = 1 - \sqrt{(r - 1)^2 + (\alpha - 1)^2 + (\beta - 1)^2}, \quad (13)$$

901 where r is the linear correlation coefficient between predicted and observed
 902 streamflow ($\hat{y}$ and y , respectively), α a measure of relative variability defined
 903 as the standard deviation of $\hat{y}$ over the standard deviation of y (i.e., $\frac{\sigma_{\hat{y}}}{\sigma_y}$), and β
 904 a measure of variability defined as the mean of $\hat{y}$ over the mean of y (i.e., $\frac{\mu_{\hat{y}}}{\mu_y}$).
 905

906 The Fourth Root Mean Quadrupled Error (R4MS4E) is a non-negative
 907 metric that biases the model evaluation process in favour of peaks and high-
 908 magnitude events (Dawson et al., 2007). It is defined as

$$R4MS4E = \sqrt[4]{\frac{\sum_{i=1}^N (y_i - \hat{y}_i)^4}{N}}, \quad (14)$$

909 where N is the number of observations.

910

911 The Mean Squared Logarithmic Error (MSLE) adopts a logarithmic trans-
912 formation to account for the model performance on low flows. The MSLE is
913 defined as follows

$$MSLE = \frac{1}{N} \sum_{i=1}^N (\ln(y_i) - \ln(\hat{y}_i)). \quad (15)$$

914

915

916 The Mean Squared Derivative Error (MSDE) is defined as the difference
917 between the first-order derivatives of the predicted and the observed stream-
918 flow, that is

$$MSDE = \frac{1}{N} \sum_{i=1}^N \left((y_i - y_{i-1}) - (\hat{y}_i - \hat{y}_{i-1}) \right)^2. \quad (16)$$

919 This metric can be seen as an indicator of the fit of the shape of the hydro-
920 graph.

Algorithm 1 W-QEISS algorithm

Input: The dataset D (observations and output variable y) and the tolerance δ .

Output: δ -quasi equally informative subsets of predictors w.r.t. y .

Find alternative subsets with Borg MOEA

Initialize: Generate the initial population P of Borg MOEA.

repeat

for $i = 1$ **to** $|P|$ **do**

 Calculate the objective functions of the i -th subset, $S_i \in P$,

$-f_1(S_i)$: relevance of S_i , $-f_2(S_i)$: redundancy of S_i ,

$-f_3(S_i)$: cardinality of S_i , $-f_4(S_i)$: predictive accuracy of S_i .

end for

 Make necessary updates in Borg MOEA for the next iteration.

until a predetermined termination condition is satisfied

return A , the archive of the Pareto-efficient subsets w.r.t. $f_1(\cdot), f_2(\cdot), f_3(\cdot), f_4(\cdot)$.

Find δ -quasi equally informative subsets

Initialize: Get the parameter δ , set $f_4^* = \max_{S_i \in A} f_4(S_i)$ and $A_\delta = \emptyset$.

for $i = 1$ **to** $|A|$ **do**

if $f_4(S_i) \geq (1 - \delta)f_4^*$ **then**

$A_\delta = A_\delta \cup \{S_i\}$,

end if

end for

return A_δ , the set of δ -quasi equally informative subsets before elimination.

Eliminate inferior subsets

Initialize: Set $A_\delta^* = A_\delta$.

while $|A_\delta| \geq 1$ **do**

 Set $A_\delta^{temp} = A_\delta^*$ where A_δ^{temp} is the temporary set of δ -quasi equally informative subsets. $S_i \in A_\delta$,

while $|A_\delta^{temp}| \geq 1$ **do**

$S_j \neq S_i \in A_\delta^{temp}$,

if $S_i \supset S_j$ and $f_4(S_i) \leq f_4(S_j)$ **then**

$A_\delta^* = A_\delta^* - \{S_i\}$, $A_\delta^{temp} = \emptyset$,

else

$A_\delta^{temp} = A_\delta^{temp} - \{S_j\}$,

end if

end while

$A_\delta = A_\delta - \{S_i\}$,

end while

return A_δ^* , the final set of δ -quasi equally informative subsets after elimination of inferior subsets.

Figure 1: Relevance- and redundancy-based classification of predictors (adapted from Yu and Liu (2004)). The red area depicts an optimal subset, which consists of strongly relevant and weakly relevant but non redundant predictors.

Figure 2: Average number of equally informative subsets correctly specified (C) for the linear (left) and nonlinear (right) synthetic test problems.

Figure 3: Average number of sub-optimal subsets wrongly retained (W) for the linear (left) and nonlinear (right) synthetic test problems.

Figure 4: Case study map: the Yampa River.

Figure 5: Value of the symmetric uncertainty between each candidate predictor and the output (first row) and between each pair of predictors.

Figure 6: Graphical representation of the four-dimensional Pareto frontier returned by W-QEISS algorithm before δ -elimination procedure—the objective functions are relevance $f_1(\cdot)$, redundancy $f_2(\cdot)$, cardinality $f_3(\cdot)$ and predictive accuracy $f_4(\cdot)$ (upper left). Two dimensional projections of the four-dimensional Pareto frontier (top right and bottom).

Figure 7: Graphical representation of the four-dimensional Pareto frontier returned by W-QEISS algorithm after the δ -elimination procedure (with δ equal to 10 %)—the objective functions are relevance $f_1(\cdot)$, redundancy $f_2(\cdot)$, cardinality $f_3(\cdot)$ and predictive accuracy $f_4(\cdot)$ (upper left). Two dimensional projections of the four-dimensional Pareto frontier (top right and bottom). The solutions returned by PCIS, PMIS, IIS and mMRMR are also illustrated.

Figure 8: Frequency matrix of the 10 % quasi equally informative subsets of predictors, with associated metrics of predictive performance. The benchmark solutions are reported in the bottom part. The best value for each metric is in boldface type.

Figure 9: Predictive accuracy of the 10 % quasi equally informative subsets of predictors, measured in terms of Kling–Gupta Efficiency (KGE), Fourth Root Mean Quadrupled Error (R4MS4E), Mean Squared Logarithmic Error (MSLE) and Mean Squared Derivative Error (MSDE). The predictive accuracy of the benchmark solutions is included for additional comparison.

Figure 10: Comparison between the measured streamflow and the predictions issued by the benchmarks and W-QEISS solutions 20 and 48 on a low flow (top) and high flow (bottom) period. The pink lines determine the envelope for minimum and maximum discharges returned by the other 10 % quasi equally informative subsets of predictors.

Table 1: Synthetic test problems description. The error term ε is modelled as a Gaussian random variable with zero mean and standard deviation σ .

	Linear Case (8 variables)	Nonlinear Case (6 variables)
<i>Output function</i>	$Y = X1 + 0.5X2 + 0.8X3 + \varepsilon$	$Y = X1 \cdot X2 + 0.8X3 + \varepsilon$
<i>Redundancy statements</i>	$X6 + X7 = X2$ $X8 = X3$	$X5 = X1$ $X6 = X2$
<i>Variables relevance</i>	strongly relevant: $X1$ weakly relevant: $X2, X3, X6, X7, X8$ irrelevant: $X4, X5$	strongly relevant: $X3$ weakly relevant: $X1, X2, X5, X6$ irrelevant: $X4$
<i>Equivalent subsets</i>	$\{X1, X2, X3\}$ $\{X1, X2, X8\}$ $\{X1, X3, X6, X7\}$ $\{X1, X6, X7, X8\}$	$\{X1, X2, X3\}$ $\{X2, X3, X5\}$ $\{X3, X5, X6\}$ $\{X1, X3, X6\}$

Table 2: Summary of the stations and daily hydro-meteorological observations used in the study.

Station	USGS/NOAA ID.	<i>FLOW</i>	<i>PRCP</i>	<i>SC</i>	<i>SNOW</i>	<i>SNWD</i>	<i>TMIN</i>	<i>TMAX</i>	<i>WESD</i>
		m ³ /s	mm	mic.Siem./ cm	mm	mm	°C	°C	mm
Maybell	USGS 09251000	X		X					
S1	NOAA USC00055446		X		X	X	X	X	
S2	NOAA USC00053738 + USC00053742		X		X	X			
S3	NOAA USC00053867		X		X	X	X	X	
S4	NOAA USC00057936		X		X	X	X	X	
S5	NOAA USC00059265		X		X	X	X	X	
S6	NOAA USS0006J06S						X	X	X
S7	NOAA USS0006J29S						X	X	X
S8	NOAA USS0006J38S						X	X	X

Table 3: Value of the objective functions $f_1(\cdot)$, $f_2(\cdot)$, $f_3(\cdot)$ and $f_4(\cdot)$ for the benchmark solutions.

Method	Relevance $f_1(\cdot)$	Redundancy $f_2(\cdot)$	Cardinality $f_3(\cdot)$	Pred. accuracy $f_4(\cdot)$
mMRMR	0.327	0.005	0.088	0.690
PMIS	0.317	0.006	0.088	0.731
IIS	0.374	0.011	0.118	0.693
PCIS	0.443	0.016	0.206	0.711