Available online at www.sciencedirect.com

SciVerse ScienceDirect

Fuzzy Sets and Systems 215 (2013) 90–111

FUZZY sets and systems

www.elsevier.com/locate/fss

A linguistic decision tree approach to predicting storm surge

S. Royston[a,b,*], J. Lawry[b], K. Horsburgh[a]

[a] National Oceanography Centre, Joseph Proudman Building, 6 Brownlow Street, Liverpool L3 5DA, UK
[b] Department of Engineering Mathematics, University of Bristol, Merchant Venturers Building, Woodland Road, Bristol BS8 1UB, UK

Received 8 September 2011; received in revised form 21 August 2012; accepted 2 October 2012
Available online 12 October 2012

Abstract

A linguistic decision tree algorithm (LID3) is applied to the problem of predicting storm surge. Of particular interest is the prediction of large positive storm surge for flood warning purposes. The application site is the North Sea which has a well-understood physical system for the generation and progression of storm surge, which lends itself to testing of the LID3 algorithm on a real-world prediction problem. Using available water level and meteorological data, the decision tree provides predictions of surge on the Thames Estuary up to 8 h in advance, accurate to the order of 0.1 m, which is comparable to alternative data driven methods. However, the success of the data driven approaches applied here are all limited by the sparsity of training data for extreme events (which by their nature are rare). A major benefit of the decision tree approach is the ability to make inference from the resulting IF–THEN rules of the tree structure. In this application of the LID3 algorithm, clear and plausible model rules can be deduced from the tree structure that are consistent with our understanding of the physical drivers of storm surge at this location. The label semantic framework is interpreted probabilistically, allowing the user to employ standard statistical approaches to identify statistically significant rules. It is demonstrated that the rules can successfully discriminate between surges that may pose a threat and those that should not, based on tide gauge measurements available up to 8 h prior to the surge signal reaching the Thames Estuary. This is promising for the potential application of such computationally efficient and easy to implement rule learning algorithms for the further investigation of complex environmental systems.

Keywords: Approximate reasoning; Linguistic modelling; LID3; Learning; Oceanography; Storm surge

1. Introduction

Storm surges are the sea level response to wind stress and atmospheric pressure gradient [1] and they are a critical component of total sea level during coastal flood events. Storm surges pose a considerable threat to coastal and low-lying regions, where the prolonged high water of the surge may inundate large areas rapidly, causing extensive damage and potential loss of life. Flooding caused by the landfall of tropical cyclones and storm surge from mid-latitude depressions has caused over half a million fatalities in recent history [2–6]. Major cities worldwide profit from coastal or estuarial

* Corresponding author at: National Oceanography Centre, Joseph Proudman Building, 6 Brownlow Street, Liverpool, L3 5DA, UK. Tel.: +44 151 795 4864; fax: +44 151 795 4801.
E-mail address: samyst@noc.ac.uk (S. Royston).
Abbreviations: RBM, rule based model; TWL, total water level

0165-0114/$ - see front matter © 2012 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.fss.2012.10.001

locations where the risk from coastal flooding needs to be mitigated; for example, 1.25 million people live and work in the floodplain of the Thames Estuary, UK, with an estimated property value of approximately £200 billion [7]. Although the physical forces that drive storm surge are well understood, there remain inaccuracies in forecasts based on hydrodynamic models due to inherent uncertainties in atmospheric forecasts and the parameterisation of sub-grid scale processes such as the complex interactions between tide, surge and waves, particularly in shallow, enclosed seas. Therefore, the problem of real-time prediction of storm surge lends itself to evaluating the potential of rule based models (RBMs), because the dominant driving mechanisms should be identifiable in the resultant rules and the accuracy of these RBMs can be compared against alternative data driven models.

Deterministic, hydrodynamic models are preferred for operational flood warning purposes, due to their proven accuracy given good quality meteorological forecasts. There are, however, regions of the world where the cost of developing and operating such models is prohibitive and there remain opportunities for data driven techniques to provide operational service. Although some operational systems exist that use data mining techniques, there appears to be a preference towards crisp methods. For example, at Italy's ICPSM where statistical least squares and artificial neural network (ANN) models are implemented as members of a suite of operational storm surge models for the Adriatic [8,9] and at the Texas A&M University, where a suite of ANN models run autonomously to provide total water level predictions for port and harbour sites along the Texan Gulf of Mexico [10–13].

Fuzzy approaches have been successfully applied to other dynamical environmental problems, such as rainfall-runoff and river discharge predictions in hydrological science [14–19] and radar echo classification and turbulence forecasting in the atmospheric sciences [20]. In physical ocean sciences, the application of fuzzy data mining approaches has largely been limited to forecasting wind waves in the open ocean (for example, [21–23]) and determining empirical descriptions of phenomena. For example, an adaptive-network based fuzzy inference system (ANFIS) to forecast the beach run-up from regular and irregular waves was found to match experimental observations more accurately than the classic empirical formula [24].

A fuzzy Naïve-Bayes (NB) approach was adopted by Randon et al. [25] for the same problem of forecasting storm surges in the North Sea. The fuzzy NB approach was relatively successful in predicting residual water levels at Sheerness on the Thames Estuary, UK (labelled 'S' in Fig. 1), given water level data from a tide gauge at Whitby

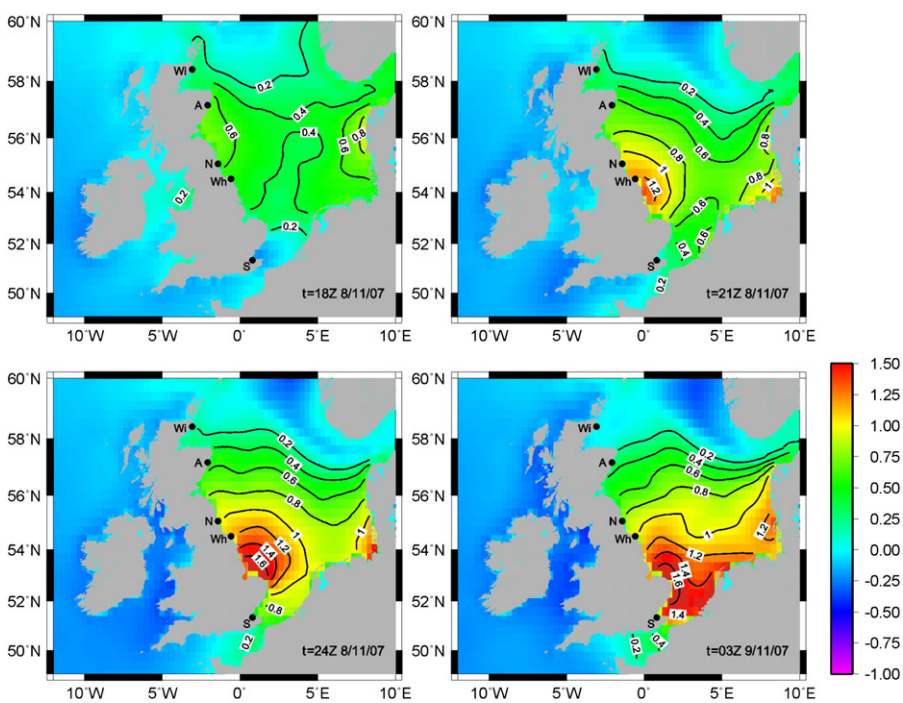

Fig. 1. Example of the cyclonic progression of storm surge around the North Sea basin, contours given in metres of storm surge, predicted by the operational UKCMF model of a storm event in November 2007 every 3 h. UK strategic national tide gauge locations are given for Wi: Wick, A: Aberdeen, N: North Shields, Wh: Whitby and S: Sheerness. Lerwick is located at 60°9′N 1°8′W on the Shetland Islands.

(labelled 'Wh' in Fig. 1), on the UK's east coast between 7.5 and 8 h in advance. However, it is difficult to extract meaningful relationships from the NB approach and so no interpretation of the model was given. Similarly, a 'model fusion' approach has been applied to the same problem, whereby an ensemble of linguistic decision tree models were combined to give a mean forecast of residual water levels [26]. Whilst, again, the model shows reasonable skill over all the data, the ensemble approach means that it is difficult to combine and interpret the rules inferred by each of the decision tree models.

Here, we combine the generalisation capability of a fuzzy approach with the transparency of a rule-based model (RBM) to investigate the applicability and success of forecasting storm surges in a well understood region, the North Sea, extending [27]. Decision tree algorithms have been shown to be as successful as alternative data mining methods for linear or semi-linear problems [28,29]. The algorithm used here, LID3, has been benchmarked on UCI repository data sets [30] and successfully applied to classifying weather radar images [31] and real-time river stage forecasting for flood purposes [32,33]. We utilise the probabilistic interpretation of the label semantic framework to employ statistical methods to significant test IF–THEN rules extracted from the decision tree structure. In this way, the added benefit of an RBM's transparency is demonstrated.

Taking into account the well-understood mechanisms of tide and surge in the North Sea and the relative success of the NB model on limited input data, this real-time storm surge prediction problem is considered ideal for testing the accuracy and interpretation of RBMs for dynamic, environmental problems.

Hereafter, Section 2 provides a brief introduction to the physical mechanisms of storm surge generation and progression. Section 3 gives an overview of the label semantic framework and LID3 decision tree algorithm. The data set used and its application in the label semantics framework are described in Section 4. Section 5 presents the results of the work and discusses both accuracy and extraction of significant IF–THEN rules from the decision tree. Conclusions from the work are given in Section 6.

2. Storm surges

The linguistic decision tree is applied to the problem of predicting storm surge at the Sheerness tide gauge, on the Thames Estuary, UK, given water level measurements from gauges around the UK's east coast and atmospheric forecasts over the southern North Sea basin. The UK strategic national tide gauge at Sheerness is of particular importance because predicted extreme sea levels here are used to determine whether or not to close the Thames Barrier which protects London from flooding. Tides and storm surges progress cyclonically around the North Sea basin as progressive coastally trapped gravity waves [1]. Storm surge can develop internally within the basin, due to consistent wind forcing, or can be externally driven by atmospheric depressions centred in the Atlantic producing gravity scale disturbances which propagate from the north into the North Sea basin [1]. Thus, although large storm surges are rare and stochastic in nature, once the forcing atmospheric circulation exists over the ocean, the development and progression of storm surge obeys physical rules such that they are, within reason, forecastable. Water level measurements from the UK's east coast gauges identify the propagation of external storm surges within the North Sea [34,35] and the development of storm surge within the North Sea basin is implied by atmospheric measurements or forecasts, where they exist. An example of the progression of a simulated storm surge wave around the North Sea basin, together with the tide gauge locations used in this study, is presented in Fig. 1. In order to give a reasonable lead time for the forecasts, of up to 8 h, the Whitby tide gauge is the most southerly gauge which is considered useful for real-time flood warning purposes [25].

It is well known within the oceanographic community that there is non-linear interaction between the tide and storm surge wherever tide has significant amplitude, such as the UK's east coast. The residual water level record, particularly at the more southerly gauges used in this work, displays clustering in the residual peaks which avoid tidal high and low water (refer to [36] for more information). The skew surge is an alternative measure of storm surge defined by the difference between the astronomical high water level and the peak observed water level for each tidal cycle. The components of total water level are illustrated in Fig. 2. As such, skew surge is a more appropriate measure for flood forecasting purposes, since it provides a single peak water level for each tidal cycle reducing the overall noise in the signal.

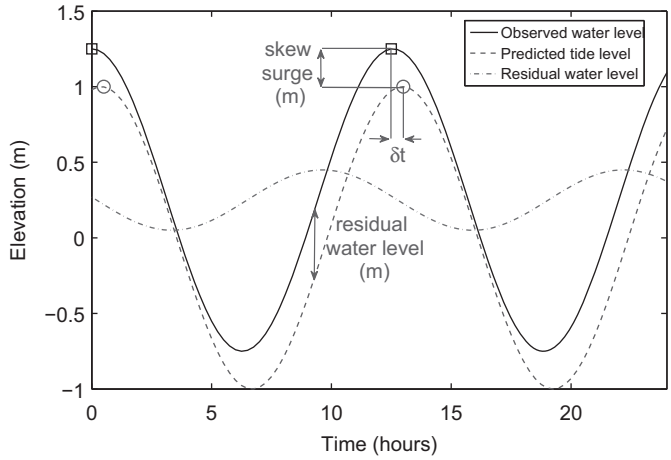

Fig. 2. Schematic of the components of total water level in a tide gauge record.

3. Method

A decision tree approach has the advantage of transparency over alternative data mining techniques and is therefore employed here, to evaluate the potential of an RBM in interpreting relationships in the data. The success of the fuzzy approach of Randon et al. [25] is noted, and thus a linguistic decision tree is used in an attempt to take into account the inherent noise and inaccuracies that exist in tide gauge measurements and the physical mechanisms leading to the observed storm surge at a specific tide gauge.

Linguistic decision trees (LDT) [30] are a tree-structured classification model based on label semantics. In this work, we employ the LID3 algorithm, where the information heuristics used for building the tree are modified from Quinlan's ID3 [37] in accordance with the label semantics framework [38,39]. The nodes of the LDT are linguistic descriptions of variables and leaves are sets of appropriate labels. In such decision trees, the probability estimates for branches across the whole tree are used for classification, instead of the majority class of the single branch into which the examples fall. Linguistic expressions such as *small*, *medium* and *large* are used to learn from data and build a linguistic decision tree guided by information based heuristics. For each branch, instead of labelling it with a certain class (such as positive or negative in binary classification) the probability of members of this branch belonging to a particular class is evaluated from a given training database. Unlabelled data is then classified by using probability estimation of classes across the whole decision tree.

3.1. Brief overview of label semantics

Label semantics [38,39] is a methodology for using linguistic expressions or fuzzy labels to describe (typically numerical) values. Label semantics proposes two fundamental and inter-related measures of appropriateness of labels as descriptions of an object or value. We begin by identifying a finite set of basic labels $LA = \{L_1, \ldots, L_n\}$ for describing elements from the underlying universe Ω. When faced with describing instance x, an agent may consider each label in LA and attempt to identify the subset of labels that are appropriate to use. Let this complete set of appropriate labels for x be denoted by $\mathcal{D}_x$. Uncertainty concerning $\mathcal{D}_x$ is then represented by the mass function m_x defined on sets of labels.

Definition 1 (*Mass functions on labels*). $\forall x \in \Omega$ a mass function on labels is a function $m_x : 2^{LA} \to [0, 1]$ such that $\sum_{S \subseteq LA} m_x(S) = 1$.

Note that there is no requirement for the mass associated with the empty set to be zero. Instead, $m_x(\emptyset)$ quantifies the agent's belief that none of the labels are appropriate to describe x.

Appropriateness measures for labels are then related to mass functions according to the rule that $\mu_{L_i}(x)$, denoting the appropriateness of L_i to describe x, corresponds to the sum of m_x over those subsets of labels containing L_i.

Definition 2 (*Appropriateness of labels*). $\forall x \in \Omega, \forall L_i \in LA, \mu_{Li}(x) = \sum_{F \subseteq LA: L_i \in F} m_x(F)$.

Here, F denotes a focal set, a subset of labels with non-zero mass. In many cases we assume that for any $x \in \Omega$ the subsets of labels for which m_x is non-zero forms a nested sequence. This is referred to as the *consonance assumption* and is particularly justifiable in cases where the appropriateness of labels is judged based on a single shared criterion. See [38] or [39] for a more detailed justification of this assumption. Making the consonance assumption means that m_x can be determined directly from the values for $\mu_{L_i}(x)$ if these are known for the basin labels $L_i \in LA$. Specifically, given appropriateness measures $\mu_{L_1}(x), \ldots, \mu_{L_n}(x)$ ordered such that $\mu_{L_i}(x) \geq \mu_{L_{i+1}}(x)$ for $i = 1, \ldots, n-1$ then assuming consonance the mass function m_x is given by

$$\begin{aligned} m_x(\{L_1, \ldots, L_n\}) &= \mu_{L_n}(x) \\ m_x(\{L_1, \ldots, L_i\}) &= \mu_{L_i}(x) - \mu_{L_{i+1}}(x) : i = 1, \ldots, n-1 \\ m_x(\emptyset) &= 1 - \mu_{L_1}(x). \end{aligned} \tag{1}$$

Appropriateness measures play a role similar to that of membership functions used in fuzzy set theory [40].

An advantage of the label semantic framework is that the basic labels in LA are building blocks for more complex compound expressions which can then also be used as descriptors. A countably infinite set of expressions, LE, can be generated through recursive applications of logical connectives to the basic labels in LA. The measure of appropriateness of an expression $\theta \in LE$ as a description of instance x is denoted $\mu_\theta(x)$ and quantifies an agent's subjective probability that θ can be appropriately used to describe x. In general, for any label expression $\theta \in LE$ an agent should be able to identify a maximal set of label sets, $\lambda(\theta)$, that are consistent with θ so that the meaning of θ can be interpreted as a constraint $\mathcal{D}_x \in \lambda(\theta)$.

Definition 3 (λ-*Sets*). $\lambda : LE \to 2^{LA}$ is defined recursively as follows: $\forall \theta, \varphi \in LE$.

(i) $\forall L_i \in LA, \lambda(L_i) = \{T \subseteq LA : L_i \in T\}$;
(ii) $\lambda(\theta \wedge \varphi) = \lambda(\theta) \cap \lambda(\varphi)$;
(iii) $\lambda(\theta \vee \varphi) = \lambda(\theta) \cup \lambda(\varphi)$;
(iv) $\lambda(\neg\theta) = \lambda(\theta)^c$;
(v) $\lambda(\theta \to \varphi) = \lambda(\neg\theta) \cup \lambda(\varphi)$.

It follows that in general, the appropriateness measure for an expression, θ, is given by

$$\mu_\theta(x) = \sum_{T \in \lambda(\theta)} m_x(T). \tag{2}$$

3.2. The LID3 algorithm

Consider a classification problem where examples are to be classified on the basis of attributes $\mathbf{x} = \langle x_1, \ldots, x_k \rangle$ as one of the t classes $C_1, \ldots, C_t$. Here we assume that $x_i \in \Omega_i = [a_i, b_i]$ where $a_i < b_i \in \mathbb{R}$ and that $LA_i = \{L_{i,1}, \ldots, L_{i,n_i}\}$ is a predefined set of labels for describing the elements of Ω_i. Furthermore, we assume that $L_{i,j}$ is defined by the appropriateness measure $\mu_{L_{i,j}} : \Omega_i \to [0, 1]$. In this context a linguistic decision tree is a probabilistic tree structured classifier with nodes corresponding to the description sets $\mathcal{D}_{x_i}$ for attributes $x_1, \ldots, x_k$. The branches of the tree that are then generated by the possible values of $\mathcal{D}_{x_i}$ corresponding to those subsets of LA_i which have non-zero mass function value for some element of Ω_i. More formally, the possible values of node $\mathcal{D}_{x_i}$ are the elements of the set $\mathcal{F}_i$ defined by

$$\mathcal{F}_i = \{F \subseteq LA_i : \exists x \in \Omega_i, m_x(F) > 0\}. \tag{3}$$

Consequently a branch B of a linguistic decision tree is a conjunction of the form

$$(\mathcal{D}_{x_{i_1}} = F_{i_1}) \wedge (\mathcal{D}_{x_{i_2}} = F_{i_2}) \wedge \cdots \wedge (\mathcal{D}_{x_{i_d}} = F_{i_d}). \tag{4}$$

where $1 \leq d \leq k$, $i_j \neq i_r$ for $j \neq r$ and $F_{i_j} \in \mathcal{F}_{i_j}$ for $j = 1, \ldots, d$. Associated with each branch B there is a conditional probability distribution on the classes $P(C_1|B), \ldots, P(C_t|B)$. Then given an instantiation of the attribute vector $\mathbf{x}$,

Jeffrey's rule of conditioning [41] is applied across the branches B of the decision tree, to obtain a probability distribution on classes conditional on $\mathbf{x}$ as follows:

$$P(C_l|\mathbf{x}) = \sum_B P(C_l|B)P(B|\mathbf{x}), \tag{5}$$

where

$$P(B|\mathbf{x}) = \prod_{j=1}^{d} m_{x_{i_j}}(F_{i_j}). \tag{6}$$

The LID3 algorithm [30] is an extension of the well-known ID3 algorithm, introduced by Quinlan [37]. LID3 infers a linguistic decision tree from a training database DB of examples each corresponding to a vector of attribute values together with their associated class:

$$DB = \{\langle \mathbf{x}^{(r)}, C^{(r)} \rangle : r = 1, \ldots, k\}. \tag{7}$$

Using this database LID3 applies the standard ID3 entropy search heuristic to identify the most informative attributes but the relevant branch and class probabilities are determined by

$$P(B) = \frac{1}{k} \sum_{r=1}^{k} P(B|\mathbf{x}^{(r)}) \tag{8}$$

and

$$P(C_l|B) = \frac{\sum_{r:C^{(r)}=C_l} P(B|\mathbf{x}^{(r)})}{\sum_{r=1}^{k} P(B|\mathbf{x}^{(r)})}. \tag{9}$$

3.3. Probabilistic decision tree: application to regression problems

In this study, the LID3 algorithm is applied so that, rather than there being classes C_l in the target data, we have t focal sets denoted $F_{l=1,\ldots,t}$, obtained from a set of labels on the output space (as described by [32,42]). The labels are defined such that the mass assignments are triangular and have full coverage on the domain of discourse. In this case, Jeffrey's rule applies to the target focal sets, F_l:

$$P(F_l|\mathbf{x}) = \sum_B P(F_l|B)P(B|\mathbf{x}), \tag{10}$$

where the probability of a label set at a branch, $P(F_l|B)$, depends on both the label set distribution of the input vector, $\mathbf{x}$, and that on the target, y, according to the standard frequentist model as follows:

$$P(F_l|B) = \frac{\sum_{i \in DB} \prod_r m_{x_r(i)}(F_j) m_{y(i)}(F_l)}{\sum_{i \in DB} \prod_r m_{x_r(i)}(F_j)} \tag{11}$$

or by Laplace's law of succession:

$$P(F_l|B) = \frac{\sum_{i \in DB} \prod_r m_{x_r(i)}(F_j) m_{y(i)}(F_l) + 1}{\sum_{i \in DB} \prod_r m_{x_r(i)}(F_j) + t} \tag{12}$$

or where no data exists in the training database, we can assume a non-informative prior, $P(F_l|B) = 1/t$.

Therefore, given an unseen attribute vector $\mathbf{x}^{(r)}$, the LID3 algorithm returns a probability distribution on each target focal set, F_l, at each appropriate leaf node. To then 'defuzzify' the predicted probability distribution on F_l to a real-valued prediction of the output variable, given a specific input, the estimate or expected value is given by

$$\hat{y} = \sum_{F_l} a_l P(F_l|\mathbf{x}), \tag{13}$$

where a_l is determined from some distribution on each target focal set, typically by taking the modal value.

3.4. Extraction of rules

The method of rule extraction is described here for the probabilistic decision tree algorithm. We employ a standard statistical approach to significance test the linguistic rules, which to our knowledge is the first attempt to significance test the IF–THEN rules in a probabilistic framework.

The forecast for a given new instance of the data vector is derived from the sum of the probabilities of each branch in the decision tree being fired by the antecedent data, x, multiplied by the probability of each target set given the branch (Eq. (10)). So for each new instance in the data vector, one can extract IF–THEN rules from the branches fired, with a probability assigned to each branch. We interrogate the decision tree structure in this way in Section 5.2 to identify the failings for a given storm surge event. However, where the size of the decision tree is large, the rules extracted in this way for each new instance can be complex, with a non-zero probability in many different branches.

An alterative approach is to ask the question "What antecedent conditions lead to a certain expression on the target?" So rather than identifying IF–THEN rules for each new instance, rules are identified by searching for the maximum probability on the focal sets given a branch:

$$\underset{B}{\operatorname{argmax}}\, P(F_l|B) \tag{14}$$

or the largest probabilities above a subjective threshold, say $P(F_l|B) > p_l$. For example, if we were interested in obtaining rules relating to the expression $\theta =$'*not small*' describing the target variable, from the label set $LA = \{small, medium, large\}$, the corresponding maximal set of label sets is given by $\lambda(\theta) = \{\{medium\}, \{large\}, \{medium, large\}, \emptyset\}$. Then the appropriateness of the expression θ to describe the target constrained by $\mathcal{D}_x \in \lambda(\theta)$, given each branch, is simply given by $\mu_\theta(y) = \sum_{F_l \in \lambda(\theta)} P(F_l|B)$.

In the example application described hereafter, it was found that those branch nodes with highest probabilities in one focal set were generally calculated from low numbers of training data, due to the tree algorithm aiming towards heterogeneity in the leaf nodes. However, for real-world applications the low number of training data points leads to uncertainty in the generalisation capability of the IF–THEN rule corresponding to that branch. To satisfy oneself that the relationships determined between the antecedent data and target variable are statistically significant, we perform significance tests on the size of training data in each branch node with a high probability in the target focal set(s) of interest. This approach is applied because we are interested in those instances in the extremes of the output distribution. If the more standard approach of pruning were to be applied, these more extreme and hence more rare instances are likely to be merged with other less extreme instances and the information would be diluted. It has been shown that the quality of probability estimates in full expanded trees is better than in pruned trees [43]; this applies to the linguistic decision tree described here where the branches correspond to linguistic descriptions of objects with probability estimates at the leaf nodes. In our proposed approach, we instead extract those rules and hence predictions that are good for elicitation through a statistical analysis of the leaf node.

Firstly, we assume a binomial probability on the linguistic labels of interest by forming the expression of interest 'y is θ' and the complementary negative expression 'y is not θ'. In the example above, where $\theta =$ '*not small*', the probability distribution at a branch node can be split into

$$\begin{aligned} p &= \sum_{F_l \in \lambda(\theta)} P(F_l|B) \\ &= P(F_{\{medium\}}|B) + P(F_{\{large\}}|B) + P(F_{\{medium,large\}}|B) + P(F_\emptyset|B) \end{aligned} \tag{15}$$

$$\begin{aligned} q &= 1 - p = 1 - \sum_{F_l \in \lambda(\theta)} P(F_l|B) \\ &= P(F_{\{small\}}|B) + P(F_{\{small,medium\}}|B) \end{aligned} \tag{16}$$

We reiterate that the branches of the linguistic decision tree (the conjunction of the form $\mathcal{D}_{x_{j_1}} = F_{j_1} \wedge \mathcal{D}_{x_{j_2}} = F_{j_2} \ldots$) are determined from the training database and that any subsequent data applied to the model structure are further samples from the population. We use the normal approximation to a binomial distribution on the assumption that given an infinite sample size with which to build the decision tree, we can obtain a large number of examples in each leaf node. That is, we have a Bernoulli experiment in that either θ or $\neg\theta$ occurs in the examples in the leaf node. There is an underlying probability of these expressions given B. Given a sample of size N of elements in B then the number of these

elements satisfying θ (lets say n) will follow a binomial distribution. As N tends to infinity this can be approximated by a normal distribution.

Therefore, given the central limit theorem, it can be expected that the probability of each focal set at each branch ($P(F_l|B)$ for each l and B) over multiple samples from the population follows a normal distribution. Therefore, the probability distributions of the summed target focal sets, p and q, also follow normal distributions. The width of the confidence interval on these probabilities, p and q, is given by

$$\text{width} = 2z_\alpha \,\text{se}(p) = 2z_\alpha\sqrt{\frac{pq}{n}} \tag{17}$$

where z_α denotes the standard normal distribution z-score for a given confidence α (the magnitude of the abscissa of the normal curve that gives an area under the curve equal to the desired confidence interval), se denotes the standard error and n denotes the sample size. Rearranging, given the sample size in a branch node, n_B, and for a desired maximum error in the probability estimate, $E = \text{width}/2$, the confidence in that probability can be derived from the z-score:

$$z_\alpha = E\sqrt{\frac{n_B}{pq}} \tag{18}$$

In the label semantics framework, this approach can be generalised for any expression on the target focal sets, 'y is θ' $\rightarrow \lambda(\theta) = F_\theta$, provided the focal sets can be split into two complementary descriptive label sets.

4. Application to the skew surge database

4.1. Data

Yearly water level data are provided by the British Oceanographic Data Centre (BODC) [44] for UK tide gauges located at Lerwick, Wick, Aberdeen, North Shields, Whitby and Sheerness (the locations of which are given in Fig. 1). Data is downloadable in ASCII fixed-format files giving the date and time of observations, total observed water level and the residual water level (being the observed total water level minus the astronomical tidal prediction). Data are flagged by BODC to indicate where values are null or missing, improbable, or interpolated. The skew surge for each tidal cycle is calculated from the maximum total water level and predicted astronomical high water, with a null value given where any data is missing from that tidal cycle.

Meteorological data have been obtained from the National Oceanography Centre's (NOC) archive of the UK's operational Coastal Monitoring and Forecasting (UKCMF) hydrodynamic model runs for the years 1999–2008 [45]. The atmospheric pressure, which accounts for an inverse barometer effect on sea level, and wind speed at 10 m elevation across the southern North Sea, whose drag at the air–sea interface forces the set-up of storm surge, has been extracted from these meteorological reanalyses as attributes for the decision tree model.

Correlation of the available atmospheric data with the skew surge record for Sheerness indicates that the peak water levels have a good correlation with indicators of the passage of low pressure systems over the period of each tidal cycle. In addition, skew surge correlates well with the north–south component of wind speed at a time after the peak water level has passed Whitby and prior to its arrival at Sheerness, which corresponds with the development of storm surge within the shallow southern North Sea. The sign-corrected square of the wind speed components possess an improved correlation, since the wind stress acting on an ocean surface is parameterised as proportional to a square-law in the wind speed [46].

Accordingly, the target or output data are taken to be the skew surge at Sheerness for each tidal cycle. The input data sets are summarised in Table 1.

4.2. Derivation of label sets

Appropriate labels were defined using expert judgement to guide the labelling of the available data given some underlying statistical distribution. In practice, the definition of appropriate labels and their distribution can be subjective and should take into account the uncertainties in the data (for example, measurement error) as well as inconsistencies in different expert opinion of how appropriate each label is to a given value. The assignment of labels and corresponding appropriateness measures is the only user-defined feature of the method. The choice of labels is in effect

Table 1
Model input data sets.

Data Set	Input attributes	Training data	Test data
Model 1:Water level data	Skew surge elevation and timing at Lerwick, Wick, Aberdeen, North Shields and Whitby. Astronomical high tide prediction at Sheerness.	02 Jan. 1980–15 Mar. 2003 (80%)	15 Mar 2003–31 Dec. 2008 (20%)
Model 2:Water level and atmospheric data	Water level data as model 1. Atmospheric pressure, p, at Sheerness at the time of high water. Sign corrected squared east–west and north–south wind speed components, u|u| and v|v|, over the southern North Sea and forecast 3 h prior to high tide at Sheerness.	02 Feb. 1999–24 Jul. 2004 (50%)	24 Jul. 2004–31 Dec. 2008 (50%)

a parameterisation on the method. The subsequent derivation of mass assignments on focal sets for each instance is then done as described in Section 3.1. We apply trapezoidal labels with 50% overlap and full coverage, leading to triangular focal sets, so that the derived mass assignments, m_y, on each data vector sum to 1. This approach allows an interpretation of mass assignments in a standard probabilistic framework, but as a consequence limits the way in which label appropriateness measures are assigned as the approach requires full coverage.

Here, all of the attributes and output are continuously valued. For each, the labels can be assigned by any method, as the assignment is user-defined and subjective. The distribution of the data may be utilised in this stage, for example, by choosing percentile bins to represent different labels with appropriateness of 1, with the 50% overlap in trapezoidal labels determining the values of appropriateness as it reduces to 0. To be less subjective, one can assume a non-informative prior of a uniform distribution. Taking the maximum and minimum values in the training data set, the continuous values can be divided uniformly between the maximum and minimum over the desired number of labels. Again, the resulting ranges define those values with appropriateness of 1, with the 50% overlap in trapezoidal labels determining the values of appropriateness as it reduces to 0. Alternatively, the user may use an expert voting scheme to identify appropriateness measures or may identify thresholds for levels of interest. For example, in this problem, the return period of skew surge (number of years) is a key statistic for flood forecasters as it indicates both the severity of a surge and the risk it poses to flood defences. For each site, the labels could relate to return periods rather than descriptive labels with the appropriateness measures derived accordingly.

Fig. 3 presents a histogram of all available Sheerness skew surge data, as described in Table 1. It shows that the data, although close to normal distribution, is highly leptokurtic. The mean displays a bias of $+0.058$ m. It has been found that the data is best approximated by a Pareto-type distribution (not shown). Table 2 presents percentiles of the example training data given in Fig. 3 for different statistical distributions, for describing the data by five labels. Discretisation of the data onto label sets based on a percentile or normal distribution biases the resulting focal sets towards the (more frequent) central values, reducing the resolution at the extremes. It is noted that the understood error in skew surge from measurement errors and errors in the harmonic analysis of the astronomical tide is around 0.1 m [47]. The majority of data lies within this understood error or displays very small skew surge and is not of interest for flood forecasting. Applying such a distribution to the labelling would lead to a tree structure containing a very small number of branches where the modal probability occurs for the target focal sets at the ends of the domain, reducing the accuracy of the tree applied to data vectors with *large positive* or *large negative* skew surge at Sheerness.

Therefore, it was found that improved accuracy for the extreme events that are of interest to flood forecasters could be gained by applying a uniform distribution to label the data, taking the maximum and minimum values from the training data set. For any data set where the extreme events are of interest, it is imperative not to bias towards the more prevalent central values and hence, the authors consider an uninformative prior to be the most appropriate choice when working within the label semantics framework. A similar argument can be made in the fuzzy framework when extremal values are of interest.

For each attribute and the output, a non-informative prior has been assumed, by taking the maximum and minimum values in the training data set and splitting the data uniformly into n bins where $\mu_{i=1,\ldots,n} = 1$ for each label

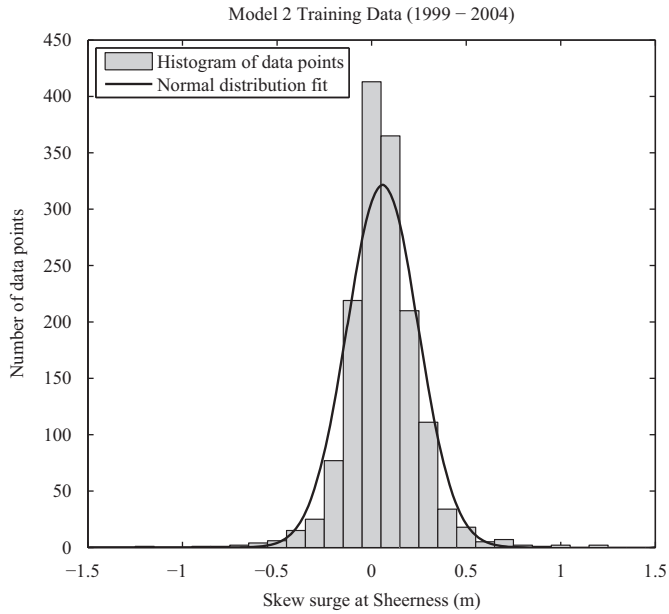

Fig. 3. Histogram of example training data (from model 2) showing a fitted normal distribution.

Table 2
Statistical distributions.

Distribution	Percentiles					
	0%	20%	40%	60%	80%	100%
Percentile	−1.197 m	−0.064 m	0.017 m	0.087 m	0.185 m	1.213 m
Normal	−1.197 m	−0.040 m	0.021 m	0.070 m	0.155 m	1.213 m
Uniform	−1.197 m	−0.715 m	0.233 m	0.249 m	0.731 m	1.213 m

L_i progressively from negative to positive. For this example, Fig. 4a presents the appropriateness measure of labels for the output, skew surge at Sheerness, given a uniform distribution with five labels. For this output, the training data set has a minimum of −1.197 m and a maximum of +1.213 m, so a uniform interval of 0.482 m has been applied such that the thresholds between labels are −0.715 m, −0.233 m, +0.249 m and +0.731 m, such that:

$$L_{1:large\ negative} : \mu_{large-ve} = \begin{cases} 1 & \text{if } y \le y_{\text{lower bound}} = -0.715\,\text{m} \\ \dfrac{(y - y_{\text{lower bound}})}{(y_{\text{upper bound}} - y_{\text{lower bound}})} & \text{if } y_{\text{lower bound}} < y < y_{\text{upperbound}} \\ & = -0.474\,\text{m} < y < -0.715\,\text{m} \\ 0 & \text{if } y \ge y_{\text{upper bound}} = -0.474\,\text{m} \end{cases}$$

and so on.

The mass assignment of focal sets is then described by the appropriateness measure from the consonance assumption, as per Eq. (1). As an example from the data shown in Fig. 4:

$$\begin{aligned} y &= 0.6, \\ m_y &:= \{s. + ve\}; \quad \mu_{s.+ve} - \mu_{l.+ve} = 1 - 0.456 = 0.544 \\ &:= \{\text{s.} + \text{ve}, \text{l.} + \text{ve}\}; \quad \mu_{l.+ve} = 0.456. \end{aligned}$$

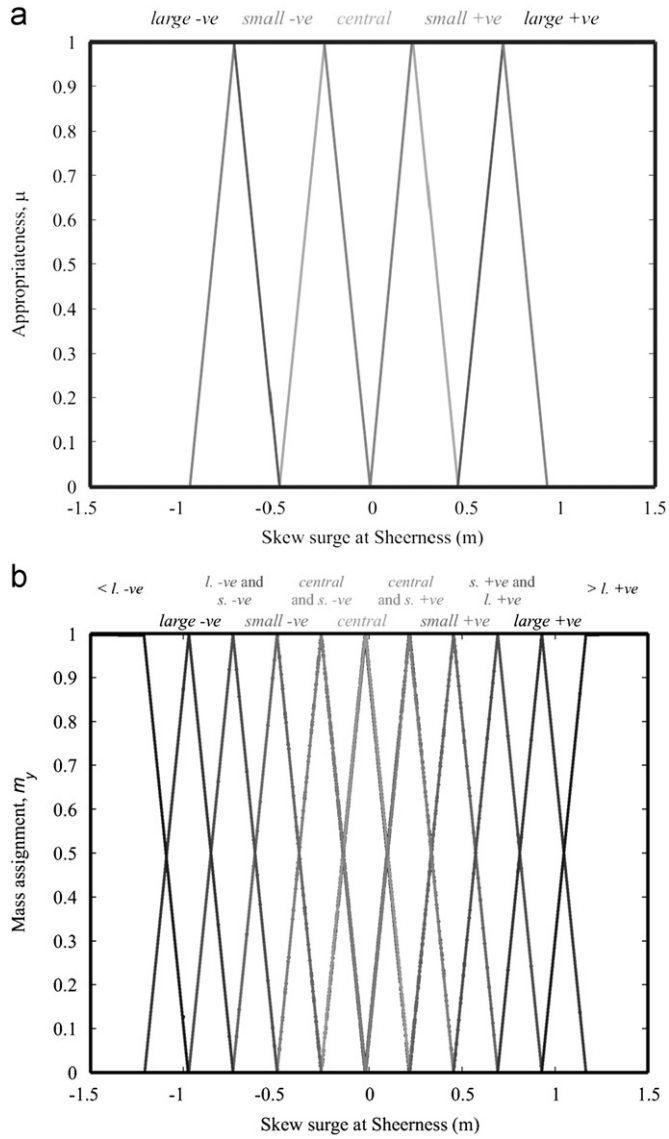

Fig. 4. Example of labels and focal label sets for the skew surge at Sheerness; (a) appropriateness, μ, of fuzzy labels L_i and (b) mass function, m_y, of the focal label sets F_i. Dots represent the training data for model 2.

It was found that accuracy was improved by providing an additional fuzzy label with an appropriateness value of 1 at the domain limits of the training data set, which can be thought of as adding linguistic labels *less than large negative* and *greater than large positive* [48]. In effect, labelling the data in this way describes flood events up to return periods of the order of the length of the training database as *greater than large positive*. By assigning trapezoidal labels with 50% overlap, the subsequent modal mass assignment in each focal set is assigned uniformly from negative to positive, with $m_j = 1$ at $y \leq -0.956$ m, $y = -0.715$ m, -0.474 m, -0.233 m, $+0.008$ m, $+0.249$ m, $+0.490$ m, $+0.731$ m and $y \geq +0.972$ m, as presented in Fig. 4b.

4.3. Model sensitivities

Sensitivity of the LID3 algorithm is constrained by only one parameter, the number of focal sets defined for each variable. It was found that increasing the resolution of the focal label sets improves accuracy, to a point, after which

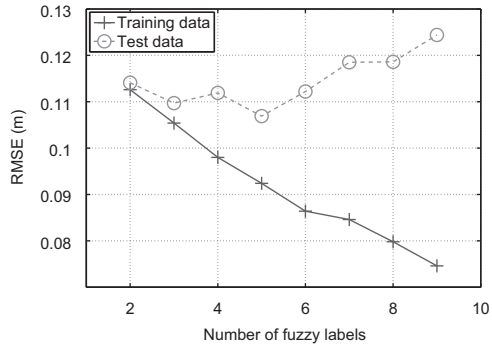

Fig. 5. Variability of accuracy measures with increasing number of fuzzy labels.

the model appears to overfit to the training data set. For this problem, the optimum number of fuzzy labels was found to be 5 for the test data set (resulting in 11 focal label sets), as shown in Fig. 5. The accuracy of the LID3 algorithm applied between three and seven number of labels is discussed in Section 5 for information. A 10-fold cross-validation was performed on the model 2 data set using the LID3 algorithm to determine variability from changes in the training and test data splits. In addition, the algorithm is boosted by the AdaBoost.RT [49] algorithm, with an error margin of $\phi = 0.1$ m relating to the understood error in skew surge from tide gauge measurements [47], to test the possibility of improving accuracy, although it is noted that such an approach reduces the interpretability of the decision tree structure.

In the problem considered here, there exist incidences of leaf nodes without data in the LID3 tree. These occur because, as the entropy-based heuristic aims to improve homogeneity in the probability distributions on the target focal sets, branches are extended by all focal label sets even though some are not represented in the training data. This occurs particularly often in the problem considered here because the training data are highly correlated, as is the case in many applications from earth sciences to biomedical sciences. In some cases, it has been noted that the sparsity of data in small disjuncts can reduce predictive accuracy and numerous attempts have been made to rectify this problem [43,50–52]. However, in the problem presented here, where a number of the inputs have high correlation and covariance, the test data vector tends to follow the same physically realistic patterns as the training data vector, within the fuzzy linguistic framework. Therefore, the way in which the probability is assigned at leaves, either by standard frequentist (Eq. (11)) or Bayesian estimators (Eq. (12)), does not have undue influence on the resulting predictions. It was found, for small disjuncts relating to more extreme and sparse events, such as a *large positive* skew surge at Sheerness, Bayesian approaches tend to give a more uniform spread in the probability estimates, which subsequently reduces the accuracy at these nodes. The LID3 algorithm with significance testing on the branches (as described in Section 3.4) provides a more holistic approach good for problems where the data of interest is prevalent in the data and for those where the data of interest lies at extremes, leading to small disjuncts.

5. Results and discussion

The decision tree structure was built using the LID3 algorithm on the databases defined in Table 1. Given the new instances in the test data set, the tree provides a probabilistic prediction in the form of membership functions on the target fuzzy label sets, Eq. (10), which are then transformed back into real-valued forecasts by Eq. (13).

5.1. Tree structure

On application of the training data to LID3, the branches of the LID3 tree can be examined to determine linguistic IF–THEN rules to check for consistency with our understanding of the physical mechanisms of storm surges. Herewith, the tree structure inferred from the data set applied to five linguistic labels is discussed, as this tree provides the best accuracy (as discussed in Section 5.2). For example, for flood warning purposes, the occurrence of very large skew surge at Sheerness is of primary importance. We can define the expression $\theta = \{$*large positive only*$\}$ which corresponds

with the label sets $\lambda(\{large\ positive\ only\}) = \lambda(\{large\ positive \wedge \neg small\ positive \wedge \neg central \wedge \neg small\ negative \wedge \neg large\ negative\}) = \{\{large\ positive\}, \{> large\ positive\}\}$. Then a probability distribution at each branch node can be defined corresponding with this expression 'y is θ' and the complementary negative expression 'y is not θ', equivalent to the sum of the probabilities on the upper two focal sets, $P(F_{l=10,11}|B)$, against the sum of the remaining focal sets, $P(\neg F_{l=10,11}|B) = P(F_{l=1,\ldots,9}|B)$.

Using this expression, the tree structure is interrogated to find the highest probability of the expression θ in the tree branches:

$$\underset{B}{\operatorname{argmax}} \sum_{l=10,11} P(F_l|B) \tag{19}$$

giving the preceding conditions in the attribute vector that lead to such an event. For example, the highest probability of a *large positive only* skew surge occurring at Sheerness in model 1 is found from the rule:

IF Whitby skew surge IS large positive ($0.719\,\text{m} < x_7 < 1.138\,\text{m}$)
AND Lerwick skew surge IS between central and small positive ($0.157\,\text{m} < x_3 < 0.420\,\text{m}$)
AND timing of the Whitby skew surge IS significantly delayed ($-48\,\text{min} < x_8 < -1\,\text{h}$)
THEN there is a strong probability ($P > 0.99$) that Sheerness skew surge will be large positive and greater.

However, the number of samples in this leaf is very small, which reduces confidence in the probability estimates drawn from the training data set. We therefore utilise the new approach outlined in Section 3.4 to obtain the following statistically significant relationships from the input data for model 1 that potentially lead to a hazardous skew surge at Sheerness, with $E = 0.1$:

Rule 1:
IF Whitby skew surge IS *large positive and greater* ($x_7 > 0.928\,\text{m}$)
THEN there is a strong probability ($P > 0.90 \pm 0.1$) that Sheerness skew surge will be *large positive and greater* (at the 90th percentile confidence level).

Rule 2:
IF Whitby skew surge IS between *small positive* and *large positive* ($0.509\,\text{m} < x_7 < 0.928\,\text{m}$)
THEN there is a strong probability ($P > 0.75 \pm 0.1$) that Sheerness skew surge will be *large positive and greater* (at the 95th percentile confidence level).

Rule 3:
IF Whitby skew surge IS between *small positive* and *large positive* ($0.509\,\text{m} < x_7 < 0.928\,\text{m}$)
AND Lerwick skew surge IS between *small negative* and *central* ($-0.106\,\text{m} < x_3 < 0.157\,\text{m}$)
THEN there is a strong probability ($P > 0.80 \pm 0.1$) that Sheerness skew surge will be *large positive and greater* (at the 85th percentile confidence level).

Rule 4:
IF Whitby skew surge IS *large positive* ($0.719\,\text{m} < x_7 < 1.138\,\text{m}$)
AND Lerwick skew surge IS *central* ($0.026\,\text{m} < x_7 < 0.288\,\text{m}$)
AND skew surge timing at Wick IS *central* ($-12\,\text{min} < x_{10} < 12\,\text{min}$)
THEN there is a strong probability ($P > 0.99 \pm 0.1$) that Sheerness skew surge will be *large positive and greater* (at the 99th percentile confidence level).

Rule 5:
IF Whitby skew surge IS *small positive* ($0.300\,\text{m} < x_7 < 0.719\,\text{m}\}$
AND skew surge timing at North Shields IS *central* ($-15\,\text{min} < x_{10} < 15\,\text{min}$)
AND North Shields skew surge IS *large positive* ($0.614\,\text{m} < x_7 < 0.976\,\text{m}$)
THEN there is a moderate probability ($P \approx 0.70 \pm 0.1$) that Sheerness skew surge will be *large positive and greater* (at the 90th percentile confidence level).

Each of these rules is consistent with the understood physical mechanisms of large positive storm surges occurring at Sheerness, with a storm surge signal apparent as an external signal in the northern gauges, reinforced by local meteorological forcing and thus growing in amplitude as it travels southwards. Fig. 6 presents a diagram of some of the tree branches, input labels and histogram of probability estimates on the fuzzy label sets of skew surge at Sheerness. The figure highlights the splitting of data resulting in rule 5. It can be seen that the subset of data at the second branch is successfully split by the entropy heuristic to differentiate between probability estimates of *small positive* (F_8) and *large positive* (F_{10}) skew surge at Sheerness given increases from *central* to *large positive* skew surge at North

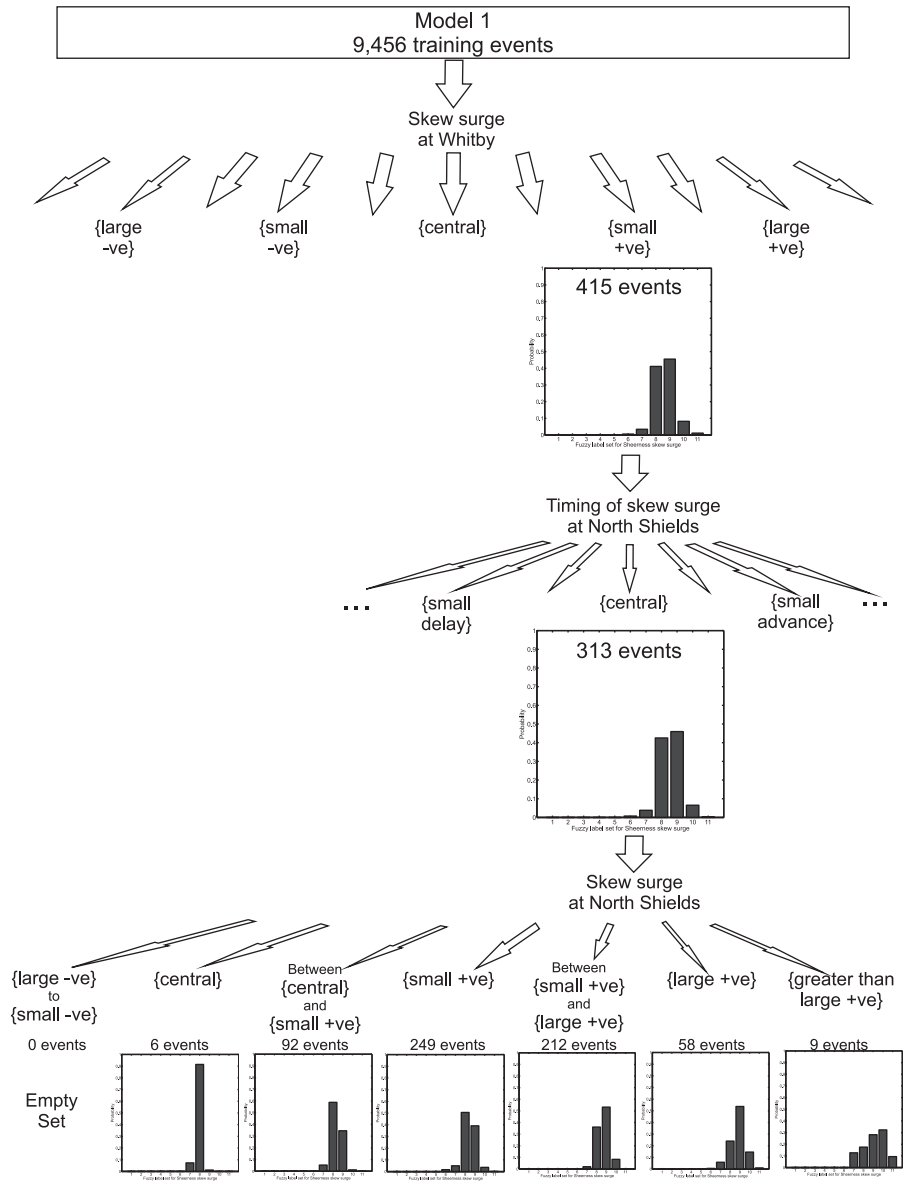

Fig. 6. Schematic of the tree branches and histograms of the probability estimates on the focal label sets for Sheerness skew surge determined by the LID3 algorithm, for rule 5, highlighting the split between leaves with *small positive* (set 8) and *large positive* (set 10) skew surge at Sheerness from a given set of inputs.

Shields. Linguistically, given a *large positive* skew surge occurring at Whitby with the skew surge at North Shields occurring *centrally* relative to tidal high water, the skew surge recorded at North Shields can be used to identify those storm surges which are likely to pose a risk at Sheerness (and hence to London) and those that are likely to be less severe.

The tree structure for model 2 is much smaller than that for model 1. The majority of information, as determined by the entropy heuristic, is obtained from the skew surge elevation at Whitby and the north–south wind speed component. These inputs can be thought of as proxies for an external (progressive) storm surge and the development of internal storm surge, respectively. Therefore, the LID3 algorithm successfully identifies the key physical drivers of large positive skew surge at Sheerness.

Table 3
Predictive accuracy.

Method	All Data			Upper 5th Percentile			CPU timing (s)
	MAE (m)	RMSE(m)	r^2	MAE (m)	RMSE (m)	r^2	
Model 1: *Water level attributes*							
LLS regression	0.091	0.123	0.54	0.137	0.180	0.22	0.2
CART	0.127	0.169	0.32	0.202	0.265	0.06	3
Fuzzy NB	0.118	0.163	0.37	0.145	0.194	0.02	1101
ANFIS	0.127	0.146	0.52	0.143	0.190	0.17	84
ANN	0.087	0.121	0.54	0.137	0.175	0.23	73
SVR	0.089	0.123	0.54	0.137	0.183	0.17	7
LID3 (4 no. labels)	0.093	0.126	0.47	0.189	0.230	0.21	1902
LID3 (5 no. labels)	0.091	0.125	0.48	0.139	0.189	0.22	4501
LID3 (6 no. labels)	0.091	0.125	0.48	0.173	0.214	0.17	10,316
LID3 (AdaBoost.RT)	0.120	0.144	0.48	0.188	0.229	0.16	26,145
Model 2: *Water level and meteorological attributes*							
LLS regression	0.077	0.104	0.67	0.113	0.168	0.23	0.01
CART	0.105	0.139	0.49	0.163	0.207	0.24	1
Fuzzy NB	0.101	0.144	0.24	0.150	0.243	0.05	29
ANFIS	0.096	0.120	0.65	0.120	0.165	0.28	12
ANN	0.073	0.102	0.69	0.115	0.161	0.32	55
SVR	0.073	0.100	0.69	0.113	0.169	0.18	0.45
LID3 (4 no. labels)	0.088	0.115	0.59	0.152	0.197	0.29	1718
LID3 (5 no. labels)	0.080	0.107	0.64	0.137	0.180	0.34	4274
LID3 (6 no. labels)	0.083	0.112	0.60	0.150	0.197	0.24	7755
LID3 (AdaBoost.RT)	0.081	0.110	0.59	0.171	0.220	0.14	26,996

5.2. Accuracy

The accuracy of the LID3 algorithm is compared against several alternative data-driven techniques: a simple linear least squares (LLS) regression; the CART algorithm; a fuzzy Naïve-Bayes (NB) approach; an Adaptive-Neuro Fuzzy Inference System (ANFIS); a feed-forward, back-propagation artificial neural network (ANN); and support vector regression (SVR). Details of these models can be found in Appendix A. The sensitivity of the algorithm to the number of labels chosen and to the effect of boosting to maximise accuracy is investigated.

Table 3 gives the mean absolute error (MAE), root mean square error (RMSE) and coefficient of determination, r^2, for the LID3 algorithm with between three and seven labels assigned to the data set and AdaBoost.RT applied against the alternative techniques, over the entire test data set and for the upper 5th percentile (corresponding to skew surges at Sheerness greater than 0.33 m), for both model 1 (using only water level attributes) and model 2 (using water level and meteorological attributes). Fig. 7 shows scatter plots of the predicted against observed skew surge from the best LID3 algorithm (with five labels assigned), for models 1 and 2.

It can be seen that the LID3 algorithm with five labels applied gives the most accurate solutions on the test data set, but that the sensitivity of the model results to the number of labels is small. As indicated in Fig. 5, the optimum number of labels, which is the only user-defined parameter of the method, is gained from a compromise between high resolution in the label intervals and good generalisation, since the larger number of labels tend to overfit to the training data set. An increase in the number of labels assigned to the data set leads to an increase in the computational effort because the number of branches at each depth increases by a power law (to the power of the depth of the tree) with the number of resulting focal sets. Whilst pruning the decision tree would reduce the errors from overfitting to the training data set, it would reduce accuracy for the more extreme events which are of interest for flood warning purposes. Probability estimation trees, which the LID3 algorithm can be thought of as a linguistic type of, have been shown to give better probability estimates as full trees rather than pruned ones [43]. It is also clear that whilst boosting improves the accuracy of the method over all of the test data set, the accuracy declines for the more rare, extreme events that are of interest here for flood forecasting purposes. This is because the algorithm subsamples the data using weightings

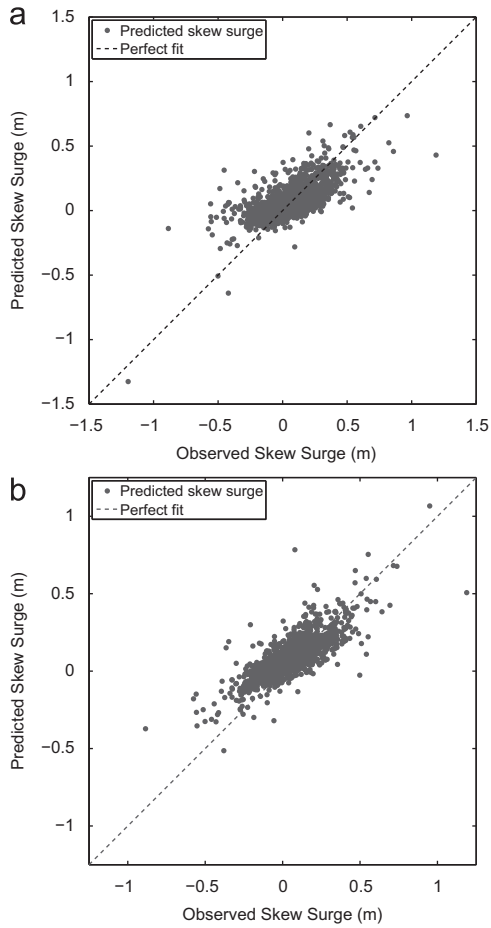

Fig. 7. Scatter plot of observed against predicted skew surge at Sheerness, predicted by the LID3 algorithm for (a) model 1 (water level) test data and (b) model 2 (including meteorological) test data.

for each learner in the ensemble to derive a final prediction. This subsampling reduces the availability of data in the extreme cases, leading to leaf nodes with little or no information and subsequently reducing accuracy here. Future work could investigate amending the boosting algorithm to weight each member of the ensemble in relation to the accuracy for these more extreme events.

The LID3 algorithm provides predictions of skew surge at Sheerness up to 8 h in advance with a comparable accuracy to the alternative methods over the training and test data set, for both model input structures. This is promising given the success of LLS regression and the more advanced methods on continuous valued data sets [53].

The inclusion of meteorological data as inputs to the data driven models improves accuracy over the whole data set for all approaches. Fig. 8 presents scatter plots of the predictions of large positive skew surge by the LID3 algorithm compared to LLS regression. For the LID3 algorithm, the standard error of the expected value of skew surge at Sheerness, determined from the probability distributions at the tree leaves, is calculated as described in Appendix B and given in Fig. 8 for information. It can be seen that, although the accuracy of the LLS regression is better than that of the LID3 algorithm over the upper data points, as shown in Table 3, for the few very large events highlighted, the LID3 algorithm does marginally better. Both methods provide reasonable forecasts of the November 2007 event, with a maximum error of 0.13 m. The LID3 algorithm has a high uncertainty associated with this estimation due to the small number of similar events occurring in the training database.

A 10-fold cross-validation was performed for the LID3 algorithm, to test the sensitivity of the algorithm to the training and test data splits. For the most successful model, the model 2 data set including meteorological inputs, described by five labels, the mean RMSE over 10-folds was 0.109 m, which compares well with the 0.107 m RMSE

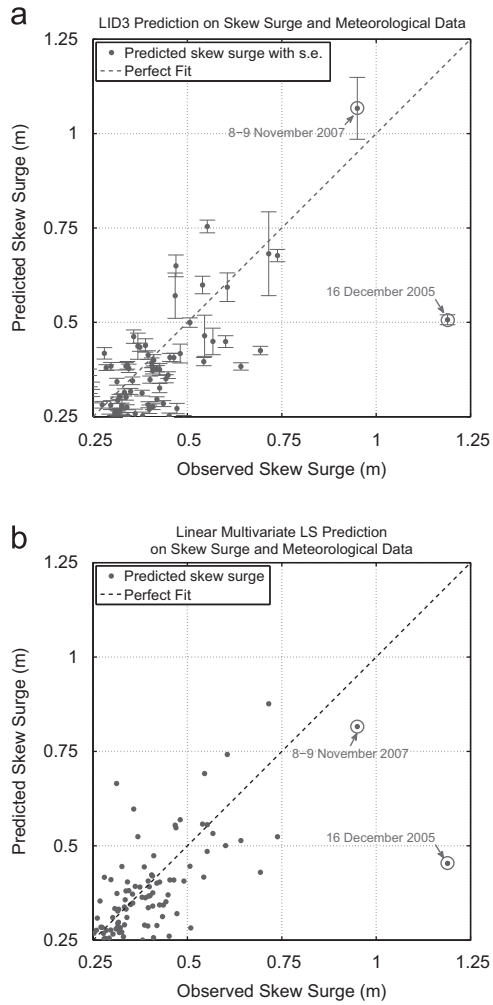

Fig. 8. Scatter plots of observed against predicted skew surge at Sheerness, for the upper 5th percentile of data, comparing (a) the LID3 algorithm against (b) LLS regression. Error bars indicate the standard error in the probability estimates at tree leaves, providing a measure of confidence in the predictions.

value determined for the main model with a 50/50% split in the data, and the standard deviation on that mean RMSE was 0.018 m, indicating that the model is robust to changes in the training set.

The largest skew surge at Sheerness in the test data set occured on 16 December 2005 and its magnitude was not predicted by the data driven methods, as shown in Fig. 8. A major benefit of the LID3 method is its transparency, which allows us to investigate this failure. Fig. 9 presents the probability estimates on the focal label sets of skew surge at Sheerness at each leaf node resulting from the attribute vector for this event, $P(F_l|B)$, and the resulting probability distribution on the these focal sets from the attribute data vector, $P\left(F_l|x^{(r)}\right)$. Significant storm surge events with this character were poorly represented in the training data set, as can be seen by the small peak in probability for the *greater than large positive* label (F_{11}) in the first histogram which is overwhelmed by the greater probabilities in the focal label sets for lesser events ($F_{l=1,\ldots,10}$). Therefore, the LID3 algorithm failed to predict this skew surge event, due to the lack of attribute data to specifically account for this character of events, a mechanism that all data driven methods are fallible to. In order to improve predictions of events such as this, input attributes would need to encompass changes in several variables both spatially and temporally (within a tidal cycle) over the North Sea.

It is noted that this ability to interrogate the tree structure aids our understanding of both the algorithm's successes and weaknesses, which is of considerable benefit to flood warning managers, and of the various mechanisms in the physical system that can lead to storm surges which pose a risk.

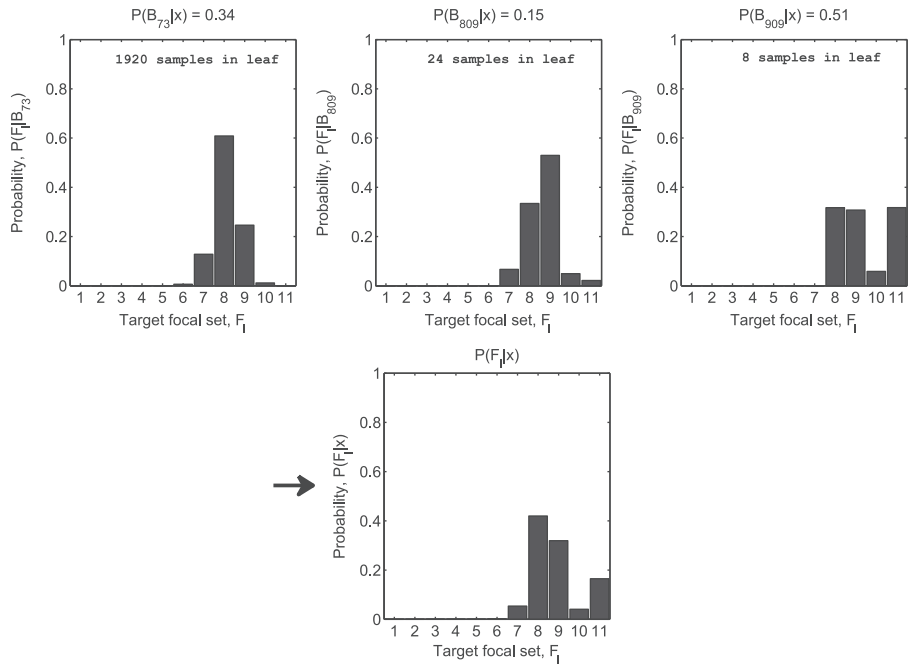

Fig. 9. Histograms of the probability estimates on the focal label sets of skew surge at Sheerness. The top row gives the probability of target focal label sets for each applicable leaf node of the tree given the December 2005 event. The lower figure presents the subsequent probability estimates on the target focal label sets for that event, from Eq. (10).

6. Conclusions

The LID3 algorithm gives comparable accuracy to alternative data driven methods over the whole data set and during more extreme storm surge events, with the accuracy of the methods approaching the measurement error in the data, of around 0.1 m [47].

The major benefit of the LID3 approach, compared with other data driven methods, is in the transparency of the tree structure. The fuzzy IF–THEN rules have been interrogated for events of interest for flood warning purposes; those skew surges at Sheerness to be appropriately labelled as *large positive only*. A new approach to test the statistical significance of the IF–THEN rules is proposed. The rules have been found to concur with our understanding of the mechanisms in the physical system which lead to large skew surge at Sheerness. Moreover, where the data driven methods fail to predict an extreme event, the inferred rules and probability estimates at leaves provide insight into the physical mechanism for that event and help identify where such events are poorly characterised and require further investigation.

It is noted that, no attempt has been made to compare the predictions from the LID3 algorithm with the operational dynamical model, as some technical issues with the data archive need to be resolved. An analysis will be the subject of a future paper.

All of the data driven techniques are fallible to the sparsity of data for extreme events, which are of interest here and in many other applications such as climate modelling, industrial control systems and biomedical investigation of rare conditions. The linguistic labelling, distribution of focal sets and LID3 algorithm were optimised to take into account the sparsity of the extreme magnitudes of skew surge in the data set, including different implementations of the probability assignments at nodes. It was found that an uninformative prior distribution was most appropriate to accurately forecast extreme output data. In this case, the relationship between antecedent data generally follows the same physically realistic patterns and as such, empty nodes from the training data set were generally not ‘fired’ by the test data set. Therefore, the application of methods such as the *m*-estimate to improve accuracy at small disjuncts did not result in significantly improved accuracy in the predictions. This is similarly the case in many applications with real-world data. It was determined that events where the data driven methods failed were generally poorly represented

in the training data set, through interpretation of the model ruleset. Although an improvement in the number and frequency of input attributes for the models, for example by using surrogate data from hydrodynamic model runs or stochastically perturbed data, may improve predictive accuracy, the sparsity of the data for such events will always limit the applicability of such methods, because an overdetermined system will result. By using the label semantics framework, the statistical significance of model rules can be derived, which quantifies confidence in both the interpreted rules and uncertainty in the predictive estimates from the model. This is of particular importance and benefit to a user interested in forecasting extreme events.

The algorithm successfully offers insight by interpreting the probabilistic rules of the tree, which identifies driving mechanisms in the data that are consistent with our understanding of the underlying physics of this system. LID3 may therefore be useful in understanding the physical systems at sites where operational hydrodynamic models do not exist or where no operational meteorological model is available or where the physical system is complex and not fully understood (such as sites with complex interactions between tides, storm surge and waves).

Acknowledgments

The British Oceanographic Data Centre provided tide gauge data as part of the function of the National Tidal & Sea Level Facility, hosted by the National Oceanography Centre and funded by the Environment Agency and the Natural Environment Research Council. Thanks is also expressed to Dr. Jane Williams of the National Oceanography Centre for preparing data from the archive of the operational storm surge numerical model.

Appendix A. Comparative data driven models

The accuracy of the LID3 algorithm in predicting the skew surge at Sheerness was compared against several popular data-driven methods, from simple and efficient linear least squares (LLS) regression to the 'state-of-the-art' support vector regression (SVR).

A multivariate LLS model was derived from the training data set for each model, with and without meteorological data.

A crisp regression tree was built using the Matlab *classregtree* function,[1] which uses the Gini impurity measure as per the CART algorithm [54]. The tree structure was built using the training data set for each model, with and without meteorological data, with pruning switched 'on' and used to forecast for the test data set.

Following the methodology described by [25], a fuzzy Naïve-Bayes approach was employed, where the probability of the class of an output, C_k, can be determined from the conditional density on the input space given C_k, denoted $f(x_1, \ldots, x_n|C_k)$ according to Bayes theorem under the naïve assumption that antecedent data is conditionally independent given the output class. To obtain the best predictive accuracy on the data set, a very high resolution of 15 fuzzy labels was required.

An adaptive-neuro-fuzzy inference system (ANFIS) was built using the Matlab Fuzzy Logic Toolbox.[2] The toolbox learns an optimum structure, in this case a Takagi–Sugeno type FIS, using a feed-forward back-propagation algorithm on the training data set. This was applied to both models, with and without meteorological data input, for differing numbers and types of membership functions and different clustering methods (the toolbox allows grid partitioning, subjective clustering and fuzzy c-means clustering in ANFIS mode). The best results were obtained from fuzzy c-means clustering into seven trapezoidal fuzzy label sets for both models.

A feed-forward back-propagation (FFBP) artificial neural network (ANN) was developed using the Matlab Neural Network Toolbox (see footnote 1). Different model architectures and transfer functions were tested, with the best accuracy obtained from a 3-layer model with eight and two hidden neurons with tan-sigmoid transfer functions in the hidden layers and one output node with a pure linear transfer function.

The LibSVM [55] freeware package was used to apply support vector regression (SVR) to this problem. Linear, polynomial, tan-sigmoid and radial basis function kernels were tested. The cost function, ε-insensitivity and spread function, γ, were set to default values of 1, 0.001 and the reciprocal of the number of attributes, respectively. The best accuracy was obtained using the radial basis function kernel.

[1] ©The Mathworks 2009.
[2] ©The Mathworks 2009.

Appendix B. Standard error calculation

The real-valued prediction, $\hat{y}$, is determined from the weighted sum of the probability of a given target focal label set given the input data vector:

$$\hat{y} = \sum_{F_l} a_l P\left(F_l|\mathbf{x}\right)$$

as given in Eq. (13). The branches of the decision tree are generated from an entropy heuristic to maximise the information gained on the target, skew surge at Sheerness. The tree structure is fixed by the training data set, which is a sample of the population of skew surge and meteorological attribute and target data. Any subsequent data applied to the model structure can be thought of as further samples from the population.

Thus, the real-valued prediction, $\hat{y}$, is the expected value of the output given a probability distribution (which is estimated from the sample training data set):

$$E[y] \approx \sum_{F_l} a_l P(F_l|\mathbf{x})$$

When a different sample data set from the same population is applied to the same tree structure, a different expected value or sample mean is anticipated. Given the central limit theorem, the expected value for $\hat{y}$ given a fixed tree structure which would be obtained from taking numerous samples of the population of skew surge at Sheerness follows a normal distribution and the variance from the sample mean can be approximated by

$$\mathrm{Var}[y] \approx \sum_{F_l} P(F_l|\mathbf{x})(a_l - E[y])^2$$

Therefore, assuming that the population size is considerably larger than the sample size (of the test data set) the standard error on the sample mean, presented in Fig. 8, can be approximated by

$$\begin{aligned} SE(\hat{y}) &\approx \frac{1}{\sqrt{\tilde{n_B}}}\mathrm{Var}[y]^{1/2} \\ &\approx \frac{1}{\sqrt{\tilde{n_B}}}\left(\sum_{F_l} P(F_l|\mathbf{x})(a_l - E[y])^2\right)^{1/2} \end{aligned}$$

where $\tilde{n_B}$ is the weighted mean of the sample sizes of all branches with non-zero mass assignments for each data vector $\mathbf{x_i}$.

References

[1] D.T. Pugh, Tides, Surge and Mean Sea-Level, John Wiley and Sons, 1987.

[2] E.N. Rappaport, Loss of life in the United States associated with recent Atlantic tropical cyclones, Bull. Am. Meteorol. Soc. 81 (9) (2000) 2065–2074.

[3] S.N. Jonkman, V.K. Vrijling, Loss of life due to floods, J. Flood Risk Manage. 1 (1) (2008) 43–56.

[4] H.M. Fritz, C.D. Blount, S. Thwin, M.K. Thu, N. Chan, Cyclone Nargis storm surge in Myanmar, Nat. Geosci. 2 (2009) 448–449.

[5] Q. Zhang, L. Wu, Q. Liu, Tropical cyclone damages in China 1983–2006, Bull. Am. Meteorol. Soc. 90(4) (2009) 489–495.

[6] Université Catholique de Louvrain, EM-DAT: The OFDA/CRED International Disaster Database, available online at ⟨http://www.emdat.be⟩, 2011.

[7] S. Lavery, B. Donovan, Flood risk management in the Thames Estuary looking ahead 100 years, Philos. Trans. R. Soc. Math. Phys. Eng. Sci. 363(1831) (2005) 1455–1474.

[8] P. Canestrelli, L. Zampato, Sea-level forecasting at the Centro Previsioni e Segnalazioni Maree (CPSM) of the Venice Municipality, in: C.A. Fletcher, T. Spencer (Eds.), Flooding and Environmental Challenges for Venice and its Lagoon: State of Knowledge, Cambridge, 2005, pp. 85–98 (Chapter 11).

[9] M. Bajo, G. Umgiesser, Storm surge forecast through a combination of dynamic and neural network models, Ocean Modelling 33 (2010) 1–9.

[10] D.T. Cox, P. Tissot, P. Michaud, Water level observations and short-term predictions including meteorological events for entrance of Galveston Bay, Texas, J. Waterw. Port Coast. Ocean Eng. 128 (1) (2002) 21–29.

[11] C. Steidley, A. Sadovski, P. Tissot, R. Bachnak, Z. Bowles, Using an artificial neural network to improve predictions of water levels where tide charts fail, in: M. Ali, F. Esposito (Eds.), Lecture Notes in Computer Science: Proceedings of the 18th International Conference on Innovations in Applied Artificial Intelligence, Springer Verlag, 2005.
[12] P.E. Tissot, J. Davis, N. Durham, W.G. Collins, Implementation of a neural network based surge prediction system for the Texas coast, in: Sixth Conference on Artificial Intelligence and its Applications to the Environmental Sciences, American Meteorological Society, 2008.
[13] Texas A&M University Division of Nearshore Research, Primary water level forecasts, available online at ⟨http://lighthouse.tamucc.edu/Forecasts/WaterLevelForecasts⟩, 2011.
[14] L. See, S. Openshaw, Applying soft computing approaches to river level forecasting, Hydrol. Sci. 44 (5) (1999) 763–778.
[15] H. Vernieuwe, O. Georgieva, B. de Baets, V.R.N. Pauwels, N.E.C. Verhoest, F.P. de Troch, Comparison of data-driven Takagi–Sugeno models of rainfall-discharge dynamics, J. Hydrol. 30 (1–4) (2005) 173–186.
[16] A. Bárdossy, Fuzzy rule-based flood forecasting, in: R.J. Abrahart, L. See, D.P. Solomatine (Eds.), Practical Hydroinformatics: Computational Intelligence and Technological Developments in Water Applications, Springer, Berlin, Heidelberg, 2008, pp. 177–187.
[17] A.P. Jacquin, A.Y. Shamseldin, Development of rainfall-runoff models using Mamdani-type fuzzy inference systems, in: R.J. Abrahart, L. See, D.P. Solomatine (Eds.), Practical Hydroinformatics: Computational Intelligence and Technological Developments in Water Applications, Springer, Berlin, Heidelberg, 2008, pp. 189–200.
[18] I.D. Cluckie, A. Moghaddamnia, D. Han, Using an adaptive neuro-fuzzy inference system in the development of a real-time expert system for flood forecasting, in: R.J. Abrahart, L. See, D.P. Solomatine (Eds.), Practical Hydroinformatics: Computational Intelligence and Technological Developments in Water Applications, Springer, Berlin Heidelberg, 2008, pp. 201–214.
[19] S. Alvisi, M. Franchini, Fuzzy neural networks for water level and discharge forecasting with uncertainty, Environ. Modelling Software 26 (2011) 523–537.
[20] J.K. Williams, C. Kessinger, J. Abernethy, S. Ellis, Fuzzy logic applications, in: S.E. Haupt, A. Pasini, C. Marzban (Eds.), Artificial Intelligence Methods in the Environmental Sciences, Springer, 2009, pp. 347–377.
[21] G. Sylaios, F. Bouchette, V.A. Tsihrintzis, C. Denamiel, A fuzzy inference system for wind-wave modeling, Ocean Eng. 36 (2009) 1358–1365.
[22] M. Özger, Significant wave height forecasting using wavelet fuzzy logic approach, Ocean Eng. 37 (2010) 1443–1451.
[23] M. Özger, Prediction of ocean wave energy from meteorological variables by fuzzy logic modeling, Expert Syst. Appl. 38 (2011) 6269–6274.
[24] R. Bahktyar, A. Yeganeh Bakhtiary, A. Ghaheri, Application of neuro-fuzzy approach in prediction of runup in swash zone, Appl. Ocean Res. 30 (1) (2008) 17–27.
[25] N.J. Randon, J. Lawry, K.J. Horsburgh, I.D. Cluckie, Fuzzy Bayesian modelling of sea-level along the East Coast of Britain, IEEE Trans. Fuzzy Syst. 16 (3) (2008) 725–738.
[26] J. Lawry, H. He, Linguistic decision trees for fusing tidal surge forecasting models, in: C. Borgelt, et al. (Eds.), Advances in Intelligent and Soft Computing: Combining Soft Computing and Statistical Methods in Data Analysis, vol. 77, Springer, 2010, pp. 403–410.
[27] S.J. Royston, K. Horsburgh, J. Lawry, Prediction of skew surge by a fuzzy decision tree, in: 90th Annual Meeting of the AMS, American Meteorological Society, 2010, available online at ⟨http://ams.confex.com/ams/90annual/techprogram/paper_159569.htm⟩.
[28] A. Elshorbagy, G. Corzo, S. Srinivasulu, D.P. Solomatine, Experimental investigation of the predictive capabilities of data driven modeling techniques in hydrology—Part 2: application, Hydrol. Earth Syst. Sci. 14 (2010) 1943–1961.
[29] J. Mahjoobi, A. Etemad-Shahidi, An alternative approach for the prediction of significant wave heights based on classification and regression trees, Appl. Ocean Res. 30 (3) (2008) 172–177.
[30] Z. Qin, J. Lawry, Decision tree learning with fuzzy labels, Inf. Sci. 172 (2005) 91–129.
[31] D.R. McCulloch, J. Lawry, M.A. Rico-Ramirez, I.D. Cluckie, Classification of weather radar images using linguistic decision trees with conditional labelling, in: IEEE International Conference on Fuzzy Systems (FUZZ-IEEE), IEEE, 2007, pp. 1–6, ISBN:1-4244-1210-2.
[32] D.R. McCulloch, J. Lawry, I.D. Cluckie, Real-time flood forecasting using updateable linguistic decision trees, in: IEEE International Conference on Fuzzy Systems (FUZZ-IEEE), IEEE, 2008, pp. 1935–1942, ISBN:978-1-4244-1818-3.
[33] J. Lawry, D.R. McCulloch, N.J. Randon, I.D. Cluckie, Artificial intelligence techniques for real-time flood forecasting, in: G. Pender, H. Faulkner (Eds.), Flood Risk Science and Management, Wiley-Blackwell, 2010(Chapter 8).
[34] J. Darbyshire, M. Darbyshire, Storm surges in the North Sea during the winter 1953–4, Proc. R. Soc. London Ser. A 235 (1201) (1956) 260–274.
[35] D.E. Cartwright, A unified analysis of tides and surges around north and east Britain, Philos. Trans. R. Soc. A 263 (1968) 1–55.
[36] K.J. Horsburgh, C. Wilson, Tide–surge interaction and its role in the distribution of surge residuals in the North Sea, J. Geophys. Res.: Oceans 112 (2007) 13.
[37] J.R. Quinlan, Induction of decision trees, Mach. Learn. 1 (1986) 81–106.
[38] J. Lawry, A framework for linguistic modelling, Artif. Intell. 155 (1–2) (2004) 1–39.
[39] J. Lawry, Studies in Computational Intelligence 12: Modelling and Reasoning with Vague Concepts, Springer, 2006.
[40] L.A. Zadeh, Fuzzy sets, Inf. Control 8 (3) (1965) 338–353.
[41] R.C. Jeffrey, The Logic of Decision, McGraw-Hill, 1965.
[42] Z. Qin, J. Lawry, Prediction trees using linguistic modelling, in: Proceedings of the International Fuzzy Systems Association World Congress, IFSA, 2005.
[43] F. Provost, P. Domingos, Tree induction for probability-based ranking, Mach. Learn. 52 (3) (2003) 199–215.
[44] BODC, British Oceanographic Data Centre UK Tide Gauge Network data, available online at ⟨www.bodc.ac.uk⟩, 2011.
[45] EA, UK Coastal Monitoring and Forecasting operational storm surge model forecasts, ⟨http://www.environment-agency.gov.uk/research/policy/116129.aspx⟩, 2011.
[46] H. Charnock, Wind stress on a water surface, Q. J. R. Meteorol. Soc. 84 (350) (1955) 639–640.
[47] J. Flowerdew, K. Horsburgh, C. Wilson, K. Mylne, Development and evaluation of an ensemble forecasting system for coastal storm surges, Q. J. R. Meteorol. Soc. 136 (2010) 1444–1456.

[48] N. J. Randon, Fuzzy and Random Set Based Induction Algorithms, Ph.D. Thesis, University of Bristol, Faculty of Engineering, 2004.
[49] D.P. Solomatine, D.L. Shrestha, AdaBoost.RT: A boosting algorithm for regression problems, in: Proceedings of the International Joint Conference on Neural Networks, IEEE, Piscataway, NJ, 2004, pp. 1163–1168.
[50] J.R. Quinlan, Improved estimates for the accuracy of small disjuncts, Mach. Learn. 6 (1) (1991) 93–98.
[51] S. Džeroski, B. Cestnik, I. Petrovski, Using the *m*-estimate in rule induction, J. Comput. Inf. Technol. 1 (1993) 37–46.
[52] S. Kunte, K.H. Upadhya, Estimating multinomial probabilities, Am. Stat. 50 (3) (1996) 214–216.
[53] Y.S. Kim, Comparison of the decision tree, artificial neural network, and linear regression methods based on the number and types of independent variables and sample size, Expert Syst. Appl. 34 (2) (2008) 1227–1234.
[54] L. Breiman, J.H. Friedman, R.A. Olshen, C.J. Stone, Classification and Regression Trees, Wadsworth Press, 1984.
[55] C.-C. Chang, C.-J. Ling, LibSVM: A library of support vector machines, available online at ⟨http://www.csie.ntu.edu.tw/∼cjlin/libsvm⟩, 2005.