Application of a linguistic decision tree model to sea level problems

Sam Royston[*,†], **Jonathan Lawry**[†], **Kevin Horsburgh**[*]

[*]National Oceanography Centre, Liverpool, UK; [†]Engineering Mathematics, University of Bristol, UK

samyst@noc.ac.uk

Abstract

Knowledge and understanding of sea level variability on varying spatial and temporal scales remains a key field of research in the earth sciences. Given the complexity of the earth system and its feedbacks, many attempts have been made to produce simplified models of this variability and there remains a potential for probabilistic and transparent data-driven techniques to further our understanding in this field. Here, a fuzzy rule-based approach, a linguistic decision tree, has been applied to very different problems in sea level science: the short-term forecasting of storm surge in the North Sea, and the replication of annual mean sea level variability on a local scale. The model's merits are proven in the storm surge problem, displaying comparable accuracy to alternative methods, with two benefits. Firstly, the model gives probabilistic estimates of the storm surge. In addition, statistically significant IF-THEN rules produced by the algorithm can be interpreted linguistically and are found to be consistent with our understanding of the physical system. The same probabilistic and transparent approach is then applied to the mean sea level problem. The algorithm identifies the data fields providing most information about the system and the rules can be interpreted to identify key drivers of sea level variability.

1 Decision tree model

The approach here is to apply a probabilistic decision tree algorithm[1], denoted LID3, to predict sea level variability. The decision tree algorithm is fuzzy and lies in a Bayesian framework, where the attribute and target data are 'fuzzified' into membership functions on descriptors, as shown in Figure 1, denoted $m_x(F_j)$ and $m_y(F_i)$ for the inputs and target respectively.

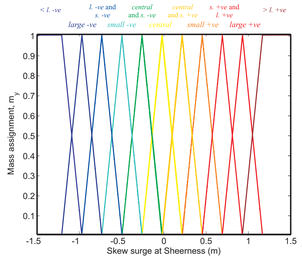

Figure 1: Examples of descriptors (label sets) for one of the data sets applied here.

A database of $i = 1, \ldots, N$ data vectors with k number of input attributes, $x_{r=1,\ldots,k}$ and the target variable, y, are used to develop the tree structure. The algorithm begins with the whole training database and identifies the input variable that minimises entropy, maximising information gain, with respect to the fuzzy descriptors of the target variable. The database is then split into branches based on the descriptors of the most informative input variable so that at each node, there exists a conditional probability distribution on the descriptors of the target, denoted $P(F_{i=1}|B), \ldots, P(F_{i=m}|B)$. We assume a non-informative prior $P(F_i|B) = 1/m$ and where information is available in the training database, use Bayes theorem to derive a conditional posterior probability according to the standard frequentist model:

$$P(F_i|B) = \frac{P(B|F_i)P(F_i)}{P(B)} \quad (1)$$

$$= \frac{\sum_{i\in DB}\prod_r m_{x_i^{(r)}}(F_j)m_{y_i}(F_i)}{\sum_{i\in DB}\prod_r m_{x_i^{(r)}}(F_j)} \quad (2)$$

where DB denotes the training database.

Given a new instantiation of the attribute vector, $\mathbf{x}^r$, Jeffery's rule of conditioning is applied across all branches of the decision tree to obtain a probability distribution on the target descriptors:

$$P(F_i|\mathbf{x}) = \sum_B P(F_i|B)P(B|\mathbf{x}) \quad (3)$$

where

$$P(B|\mathbf{x}) = \prod_{j=1}^{d} m_{x_{ij}}(F_{i_j}). \quad (4)$$

2 Storm surge

2.1 Introduction

Storm surge remains a significant hazard to coastal communities around the world. The timely and accurate forecast of storm surge has the potential to save lives and protect property and assets, via flood warning alert, flood defence and evacuation procedures. In some regions, operational systems have been developed based on hydrodynamic forecast models which solve the shallow-water equations. Although these models exhibit good accuracy for lead times up to 48 hours, computational resources limit how accurate the forecasts can be (for example by the parametrisation of sub-grid scale processes and features) and the development of operational probabilistic forecasts by ensemble. Furthermore, there are regions where fully-fledged hydrodynamic models are too expensive to develop and run operationally. Thus, the aim of this work is to investigate the applicability of an alternative data-driven approach to short-term forecasting of storm surge in a region where a fair comparison can be made against an operational forecast model.

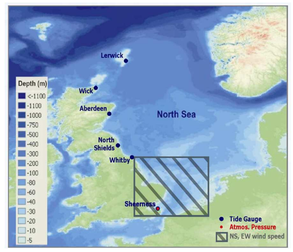

Figure 2: The North Sea site and data used.

The application site is the North Sea. The physical mechanism of storm surge is relatively well understood here, where surge and tide progress cyclonically around the basin as a coastally-trapped gravity wave. Storm surge may be of external origin, entering the basin from the north, and / or may develop within the basin due to atmospheric (pressure and wind stress) forcing. We chose the tide gauge site of Sheerness as our forecast location as this gauge is used in flood warning procedures for the operation of the Thames Barrier. Forecasts are made approximately 7.5 to 8 hours ahead using the following input data from tide gauges 'upstream' of the progressive wave and meteorological data, as shown in Figure 2.

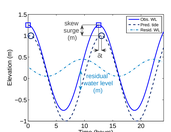

Figure 3: Schematic of the components of total water level.

Skew surge is an alternative representation of storm surge to residual water level and is used in this study. As shown in Figure 3, skew surge is the difference between the observed maximum water level and the predicted tidal high water for each tidal cycle and is considered a more useful metric for flood forecasting.

2.2 Hindcasts

The decision tree algorithm is found to exhibit accuracy comparable to the operational model and alternative data-driven techniques over all of the test data. For the larger skew surge events identified from the upper 5^{th} percentile of the test dataset, the LID3 algorithm is marginally worse than alternative methods but benefits from probabilistic forecasts. Figure 4 presents the mean forecast value and standard error bars against observed skew surge, focusing on the large positive skew surges in test dataset. This measure of uncertainty is highly beneficial to flood forecasters.

Table 1: Predictive Accuracy

Method	All Data			Upper 5^{th} Percentile		
	AAE (m)	RMSE (m)	r^2	AAE (m)	RMSE (m)	r^2
Operational model	0.076	0.097	0.52	0.130	0.157	0.21
LID3 decision tree	0.078	0.107	0.42	0.168	0.203	0.06
LLS regression	0.073	0.102	0.45	0.147	0.192	0.09
ANN (2-layer)	0.070	0.096	0.49	0.139	0.181	0.14
SVR (RBF kernel)	0.071	0.100	0.71	0.147	0.189	0.22

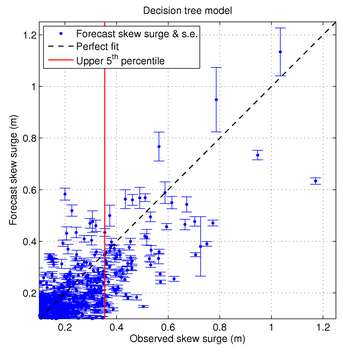

Figure 4: Scatter plot of forecast against observed skew surge, including standard error bars, from the decision tree algorithm.

2.3 IF-THEN rules

The decision tree algorithm determines a tree structure from the training database. The fuzzy IF-THEN rules of the decision tree can be interrogated and checked for consistency with our understanding of the physical system. In this example, the decision tree identifies the key mechanisms for storm surge generation given differing conditions along the north-east coast of the UK and within the basin. The tree identifies external and internal sources of storm surge generation.

Figure 5 is a schematic of one branch of the decision tree structure, showing that the tree identifies the most informative attribute to be skew surge at Whitby (the closest gauge to Sheerness), with additional information where skew surge at Whitby can be described as *central to small positive* obtained from the north-south wind speed data, the key driver in developing internal surge.

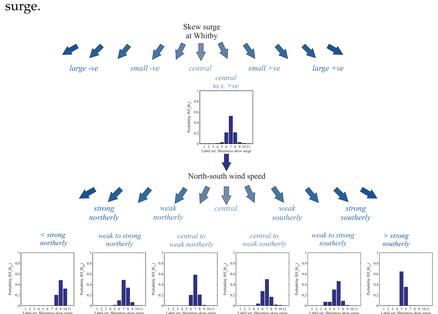

Figure 5: Decision tree structure for the central branches of the decision tree, showing consistency with our understanding of the key mechanisms of storm surge generation in the North Sea.

3 Mean sea level variability

3.1 Introduction

The decision tree algorithm is applied to the problem of determining local scale variability in mean sea level. Mean sea level from monthly periods and greater are of interest to science in the study of the effects of climate variability and in determining the potential for significant sea level rise due to atmospheric warming. The annual mean sea level (AMSL) varies considerably around the long-term mean, with the determination of trends in long records of key scientific interest. Here, we determine the success of the decision tree model in filling data gaps in the tide gauge record at Brest, using analogue tide gauge record (Newlyn) and atmospheric variables. Figure 6 presents the raw locally referenced AMSL records for Brest and Newlyn, highlighting the correlation between the two records. The rate of change in the AMSL, also presented, removes the data from the local reference frame.

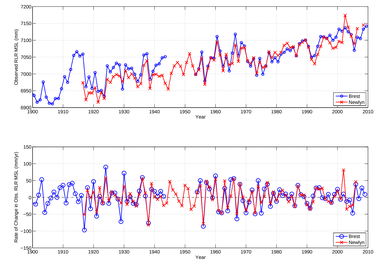

Figure 6: Raw locally referenced annual mean sea level (a) and rate of change (b) for the Brest and Newlyn tide gauge data.

Although the Newlyn record is closely correlated with the Brest record, there are periods where the rate of change in AMSL is considerably different such as 2001-2002 and in addition, there appears to be a step change in the Brest sea level record pre-war compared with post-war.

3.2 Hindcasts

We concentrate on the post-war record with the aim of hindcasting an artificial data gap, so we can quantify the algorithm's success. The decision tree model is used to forecast the rate of change in AMSL at Brest using the rate of change in AMSL at Newlyn and the rate of change in various atmospheric indices, which inform about large-scale variability such as basin-scale gyre spin-up. The annual mean sea level is then reconstructed by forward and backward integration from the start and end of the decade respectively, and the mean value is taken as the hindcast AMSL, as shown in Figure 7.

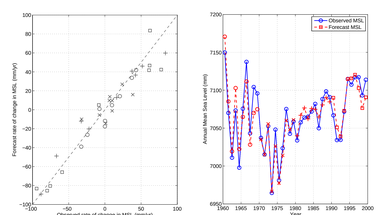

Figure 7: Forecast Brest annual mean sea level, reconstructed from the rate of change forecast by the decision tree model, as a) a scatter plot, and b) a timeseries, for four decades of test data.

Table 2: Predictive Accuracy

Method	Rate of change		MSL	
	RMSE (mm/yr)	r^2	RMSE (mm)	r^2
LID3 decision tree	12.6	0.923	14.5	0.871

3.3 IF-THEN rules

The fuzzy IF-THEN rules of the decision tree are interrogated to identify key physical drivers of the local scale variability. Figure 8 presents the statistically significant (to the 90^{th} percentile) branches of the decision tree, displaying splitting of the training data into branches with negative, near zero and positive year-to-year changes in mean sea level. It can be seen that most information comes from the Newlyn tide gauge data, which is unsurprising. Thereafter, the variability in mean sea level pressure at Brest is most informative, followed by the rate of change of atmospheric indices describing north-south and east-west pressure gradients over the north-east Atlantic basin.

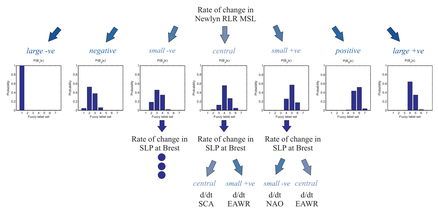

Figure 8: Schematic of the decision tree structure for the rate of change in mean sea level at Brest.

4 Conclusions

It is encouraging that the data-driven model has comparative accuracy to alternative methods in hindcasting skew surge in short-term forecasts. Similarly, reasonable accuracy has been demonstrated in data gap filling for long-term mean sea level variability. The natural error estimates given by the decision tree model are highly beneficial to flood forecasters and highlight uncertainties in mean sea level reconstruction.

The decision tree structure identifies key driving mechanisms in both studies. In the forecasting of storm surge, the presence of surge at the closest gauge to Sheerness is dominant in informing the magnitude and sign of surge at Sheerness. However, the development of internal surge (from the presence of northerly winds) and the presence of external surge (identified in the timing of surge at the northerly gauges) are also identified. This gives us confidence in the decision tree structure.

In the mean sea level problem, the decision tree model suggests that local scale information from the Newlyn tide gauge data can be supplemented with information from both local-scale and large-scale atmospheric pressure variability; in particular those indices that relate to the north-south and east-west pressure gradients experienced in the north-east Atlantic basin. These indices may be an important data source for long-term sea level studies.

References

[1] Z. Qin and J. Lawry (2005) *Information Sciences* **172**, pp.91–129

Acknowledgements

Tide gauge data was provided by BODC (www.bodc.ac.uk) and meteorological data was taken from NOC's archive of the UK's operational storm surge forecast model. Annual mean sea level data is provided by PSMSL (www.psmsl.ac.uk) and atmospheric indices are available from UEA (www.cru.uea.ac.uk) and NOAA (www.cpc.ncep.noaa.gov). S. Royston's PhD was funded by the Flood Risk Management Research Consortium.