Revista Especializada de Ingeniería
y Ciencias de la Tierra
VOL: 4 Nº 2 Enero - Julio 2025
ISSN L: 2805-1874

Significance of Wavelet Transform and Data Mining in Forecasting Sea Level Time Series at the Pacific Entrance of the Panama Canal

Importancia de la Transformada de Ondoletas y Minería de Datos para Pronosticar el Nivel del Mar de una Serie Temporal en la Entrada Pacífico del Canal de Panamá

Jose Antonio Simmonds Sheppard
Universidad de Panamá, Facultad de Ingeniería, Panamá
j.simmonds728@gmail.com
https://orcid.org/0000-0001-6180-3497

Masai Z. González
Universidad de Panamá, Departamento de Biología Marina, Panamá
masaizaret@gmail.com
https://orcid.org/0009-0006-7808-2984

Agapito Ledezma
Universidad Carlos III de Madrid, Departamento de Informática, España
ledezma@inf.uc3m.es
https://orcid.org/0000-0002-0041-6829

Juan Antonio Gómez H.
juan.gomez@up.ac.pa
https://orcid.org/0000-0002-9320-1674
Universidad de Panamá, Departamento de Biología Marina, Panamá

Recibido:8/3/2024 Aceptado: 1/5/2024
DOI https://doi.org/10.48204/reict.v4n2.6747

ABSTRACT

Understanding the tide level is crucial for safely navigating ships in harbors, managing sediment movement, and conducting environmental observations. This area has been actively researched for years, resulting in many proposed models to improve time series modeling and forecasting accuracy and efficiency. This study

uses water level data recorded every hour from 1908-2007 at the Balboa Harbor tide gauge in Panama to create a model for predicting sea level changes. The data mining tool analyzes a range of time lags in the dataset to identify the best combinations. Analysis of the time-series data suggests that using cross-validation and a more extended lag period leads to more accurate forecasting. The study also proposes denoising the time series data as a pre-processing step and utilizing mining techniques with cross-validation and attribute selection evaluators for modeling sea-level changes at coastal areas characterized by nonlinearity and chaotic climatic changes.

Keywords | Time series, Wavelets Transform, Data mining, Forecasting, Sea level

RESUMEN

Comprender el nivel de las mareas es crucial para navegar con seguridad en los puertos, gestionar el movimiento de sedimentos y realizar observaciones ambientales. Esta área ha sido investigada activamente durante años, lo que dio como resultado muchos modelos propuestos para mejorar la precisión y eficiencia de los modelos de series temporales y los pronósticos. Este estudio utiliza datos del nivel del agua registrados cada hora desde 1908 hasta 2007 en el mareógrafo del puerto de Balboa en Panamá para crear un modelo para predecir los cambios en el nivel del mar. La herramienta de minería de datos analiza una variedad de desfases temporales en el conjunto de datos para identificar las mejores combinaciones. El análisis de los datos de series temporales sugiere que el uso de la validación cruzada y un período de desfase más prolongado conduce a pronósticos más precisos. El estudio también propone eliminar el ruido de los datos de series temporales como un paso de preprocesamiento y utilizar técnicas de minería con evaluadores de selección de atributos y validación cruzada para modelar los cambios en el nivel del mar en áreas costeras caracterizadas por la no linealidad y los cambios climáticos caóticos.

Palabras claves | Serie temporal, Transformada de ondoletas, Minería de datos, Pronostico, Nivel del. mar

1. INTRODUCTION

Today, global warming is one of the environmental issues of great concern. Understanding future sea level changes is crucial for safeguarding low-lying coastal and residential areas, monitoring complex marine ecosystems, and planning and constructing coastal infrastructure. It is also essential for developing ocean energy production technologies (Rourke et al., 2010).

The prediction of sea level variations has always been the subject of intense interest to humanity, not only from a human point of view but also from an economic one. The most famous examples of flooding are in the Venice Lagoon, some of which occurred during November 1997, November 2000, and November 2001, with the surge events reaching heights between 100 and 118 cm and have been the object of intense studies with hydrodynamic models (Umgiesser et al., 2004).

The instantaneous measurements and measurements of mean sea level, dependent on time, are not seasonal, spatial, or temporal. These vary due to the synergy of specific influences in tidal changes, temperature, salinity, atmospheric pressure, and sea currents on large scales (Chen et al., 2000; Douglas et al., 2000), sometimes resulting in tidal waves and flooding. The transition from deterministic models of global ocean circulation that have a sizeable computational grid (tens of kilometers), which, among other parameters, causes sea levels at local fine scales, is a difficult task that demands enormous and extensive mathematical computation devices (Albiach et al., 2000; Monbaliu et al., 2000).

To address the prediction of changes in sea level in coastal areas, various alternative methods have been employed. Among these techniques is data mining, which does not rely on computational grids and includes approaches such as fuzzy logic.(Long & Meesad, 2014) artificial neural networks (Pashova & Popova, 2011) and genetic algorithms (Peralta et al., 2010). Artificial neural networks (ANNs) can estimate complex systems without needing prior knowledge of the internal relationships among their components. (Haykin & Network, 2004).

Some researchers already posted the question about whether the tides are changing, yet it is said that these will change only slowly over geological time. Direct comparisons of old and recent tidal observations are rare but historical; between 1761 and 1961, the oceanic semidiurnal tides, such as St Helena in the South Atlantic, were constant in amplitude to within 2 percent (Cartwright & Driver, 1971). For the French port of Brest, which is well connected to the Atlantic Ocean, there was similar stability in the tidal amplitude between 1711 and 1936 (Cartwright, 1972).

Both natural processes and engineering works can affect local tides. Siltation, changes in dredging practices for navigation, and canalization of rivers are all relevant factors. It is reported locally that there have been significant changes in tidal amplitudes. In London, more than 80 km up the River Thames from the North Sea, high water surges have increased by around 0.8m per century, whereas low water increases were only around 0.1m per century. Comparable increases have been found in northern Germany. The range increase is due to increased high water levels, while the low ones have remained roughly the same. Along the Netherlands and Belgian coasts, the tidal range has also increased locally: at Flushing, the increase in mean tidal range over the period 1900– 80 was 0.14m per century or approximately 4 percent. These increased ranges are due to changes in the coastal configurations.

In this work, it is followed a nonlinear time series forecasting analysis to understand the nonlinear dynamic behavior of the sea level rise surrounding the Pacific entrance to the Panama Canal and to provide a prediction tool that can be feasible in dealing with data that can have missing instances, irregular patterns, and that is characterized by a series whose observed values may not be extensive, has cyclic and seasonal trends or the complexity of the generating processes can be very different.

It is also interesting to explore the results' performance by running experiments that denoise the original data set versus a non-denoised version of the data, applying data mining techniques, ANN, and cross-validation (Bergmeir & Benítez, 2012). Cross-validation implies that all data sets are used in training and testing, so it is considered using cross-validation as a learner combined with attribute selection algorithms for this experiment.

2. THEORETICAL FRAMEWORK

2.1. Wavelet transformation (WT)

As described in the literature, the Fourier transform (FT) performs a harmonic decomposition and reveals the frequency content of a time signal. However, it has limitations: Besides losing time information, the signal is supposed to be stationary, meaning that the exact frequency resolution is applied for any time portion of the signal. On the other hand, the fundamental objective of the wavelet transformation, inspired by the Fourier transform and suitable for static datasets, is to achieve a complete time-scale representation of localized and transient phenomena occurring at different time scales (Labat et al., 2002). Time series data are decomposed into different components at different resolution levels using the wavelet function. Wavelet function $\psi(\tau)$ given the name "mother wavelet" (Kişi, 2009) possesses finite energy and is described mathematically by the following formula:

$$\int_{-\omega}^{+\omega} \psi(\tau) d\tau = 0 \qquad (1)$$

Where $\psi_{a,b}(\tau)$ can be expressed as:

$$\psi_{a,b}(\tau) = |a|^{-\frac{1}{2}} \psi\left(\frac{\tau - b}{b}\right) \qquad (2)$$

where a and b are real numbers; $\psi_{a,b}(\tau)$ = wavelet function; a = scale or frequency parameter; b = translation parameter. Two variables a and b determine the wavelet transformation. The variable "a" is a lengthening "$a > 1$" or reduction "$a < 1$" of the wavelet function factor $\psi(\tau)$ that fit various measures. The variable "b" can be understood as a function that has been translated or shifted throughout time $\psi(\tau)$.

For the time series $f(\tau) \in L^2(R)$ or finite energy signal (Rosso et al., 2004), the continuous wavelet transform (CWT) of time series $f(\tau)$. is defined as follows:

$$W_f(a,b) = |a|^{-\frac{1}{2}} \int_{-\omega}^{+\omega} f(\tau) * \left(\frac{\tau - b}{b}\right) d\tau \qquad (3)$$

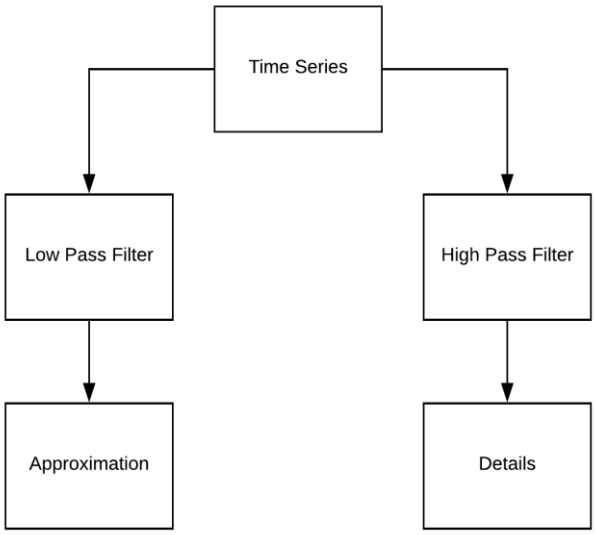
where $W_f(a,b)$ is the wavelet coefficient, ‘‘*” corresponds to the complex conjugate (Cannas et al., 2005). The mother wavelet W(τ) is the transforming function, a is the scale index parameter (the inverse of the frequency), and b is the time-shifting parameter, also known as the translation. The “CWT” calculation needs a large amount of computation time and resources. The discrete wavelet transform (DWT), however, requires less computation time and is much easier compared to the “CWT”, to execute. Dyadic scales and positions, or DWT scales, are often based on powers of two. To do this, change the wavelet representation as follows:

$$\psi_{m,n}\left(\frac{\tau - b}{a}\right) = a_0^{-\frac{m}{2}} * \left(\frac{\tau - n b_0 a_0^m}{a_0^m}\right) \qquad (4)$$

where the wavelet scale-dilation and translation are controlled by the integers m and n, respectively; a_0 is a fine specified scale step more significant than 1; and b_0 the location parameter and should be greater than zero. The standard and most straightforward choice for parameters a_0 and b_0 is 2 and $= 1$. The DWT operates two sets of functions: high-pass and low-pass filters. The original time series is passed through high-pass and low-pass filters, and detailed coefficients and approximation series are obtained (Figure 1).

Figure 1

DWT decomposition process of a time series. Own work.



2.2. Data mining and Time Series Forecasting (DM & TSF)

"Knowledge Discovery in Databases" (KDD), or data mining as it is more widely known, is extracting new and perhaps valuable information from vast volumes of data (Rushing et al., 2005).

Contrary to standard statistical methods, data mining paradigms search for appealing information without needing prior hypotheses, and patterns that can be discovered depend upon the data mining tasks used. Data mining tasks fall into two categories: descriptive tasks, which list the general characteristics of the available data, and prediction tasks, which aim to make predictions based on inference from available data. These techniques are often more powerful, flexible, and efficient for exploratory analysis than Field's statistical techniques (Bregman & Makenthun, 2006). Artificial Neural Networks, Rule Induction, and the Nearest Neighbor approach, Memory-Based Reasoning, Logistic Regression, Discriminant Analysis, and Decision Trees are the most widely used data mining techniques. Their applications would depend on the problems we are trying to solve.

The most widely employed ANN algorithm for time series forecasting is the Multilayer Perceptrons (MLPs) (Zhang, 2007). These are characterized by the feedforward architecture

comprising an output layer, an input layer, and one or more hidden layers. Every layer has connected nodes to those in the immediate next layer by acyclic links. In practical applications, it is enough to consider a single hidden layer structure (Kamruzzaman et al., 2007).

3. MATERIALS and METHODS

3.1. Study Area and Data

This study used the tide level data time series (Holgate et al., 2013) from the Balboa Harbor station tide gauge (Figure 2). This data is available to the public through the Mean Sea Level Permanent Service (http://www.psmsl.org/) database. The service is responsible for collecting, publishing, analyzing, and interpreting data on sea level gathered from the world's tide gauge network. The National Oceanography Centre (NOC), a division of the UK Natural Environment Research Council, is its home base in Liverpool. (http://www.nerc.ac.uk/).

Case data covered 1188 monthly instances from January 1908 to December 2007. The revised local reference (RLR) data are used instead of the metric. Table 1 shows a brief statistical description of the data set.

Figure 2

Balboa Harbor tide gauge location.
Source: http://motherearthtravel.com/panama/map.htm

Figure 3 shows the monthly water level time series of Balboa Harbor between 1908 and 2007, sampled hourly based on tide gauge data. High water levels, e.g., at least 7.060 m, are also shown. The high-water phenomenon has a characteristic behavior throughout the year, with June to

December being the months in which the phenomenon is more pronounced. In contrast, during the summer, it has been rarely observed.

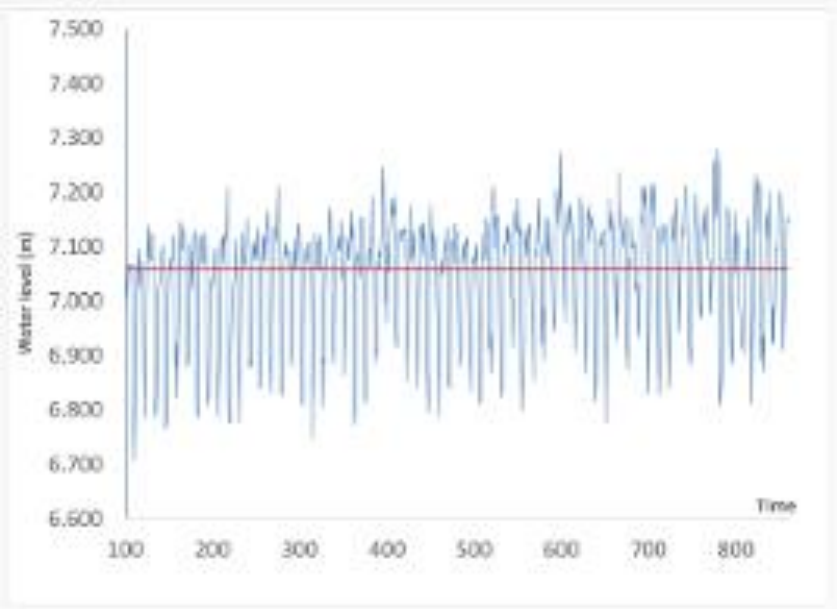
Table 1

Descriptive statistics for the Balboa Harbor water level data.

Variable	Water level
Unit	m
No. Instances	1188
Median	7.082
Mean	7.060
StdDev	0.122
Variance	0.015
Minimum	6.707
Maximum	7.439

Figure 3

Water level at the Balboa Harbor from 1908 to 2007. Own work.



4. EXPERIMENTAL SETUP

4.1. Data Structuring

The water level time series data was prepared and arranged in the following manner: Following the approach in a recent study (Simmonds et al., 2017) to forecast sea level rise at the Panama Canal Atlantic entrance, the complete Balboa dataset was considered from which prepared four sets of lags as shown in equations 6 to 9 (T = 10, T = 50, T = 60 and T = 120 respectively). Two

approaches, with three scenarios or experiments, were implemented; the first approach to forecast the water level using the original dataset without the DWT treatment and the lags as inputs to the ANN (Non-denoised data), and the second approach with the original dataset treated with the DWT (denoised data) and the lags as inputs to the ANN. The scenarios implemented for both approaches were: i) dataset without an attribute selection scheme applied, ii) using the cfs subset evaluator, and iii) using the wrapper subset evaluator (Kohavi & John, 1997) to find a predictive model that is expressed mathematically as:

$$L(K+i) = F(\text{present water level, former water level}) \quad (5)$$

where k is the time variable, i is the variable for the time step, and F is some function that defines an extensive and general class of time series.

It is known that time series forecasting models are best when the data sets are considerably large. This is also true for data mining processes, as they require large amounts of data, which is split and used for training and testing purposes, so it is derived from the properties of time series to relate the series' current values to past values as follows:

1) modeling data with lag T = 10 months

$$L(k+i) = F(L(k-i), \ldots, L(k-10), L(k+1), \ldots, L(k+n)) \quad (6)$$

2) modeling data with lag T = 50 months

$$L(k+i) = F(L(k-i), \ldots, L(k-50), L(k+1), \ldots, L(k+n)) \quad (7)$$

3) modeling data with lag T = 60 months

$$L(k+i) = F(L(k-i), \ldots, L(k-60), L(k+1), \ldots, L(k+n)) \quad (8)$$

4) modeling data with lag T = 120 months

$$L(k+i) = F(L(k-i), \ldots, L(k-120), L(k+1), \ldots, L(k+n)) \quad (9)$$

4.2. *Wavelet analysis*

In this experiment, the goal is to predict the water level one month in the future using the proposed data mining technique. On the one hand, a non-wavelet decomposition treatment of the time series data versus a wavelet decomposition analysis treatment in the latter case is evaluated. A wavelet artificial neural network approach (WANN) model has resulted, using the sub-time series components acquired by utilizing DWT on the data set. For this purpose, the wavelet transform denoised the original signal and linked it to the ANN.

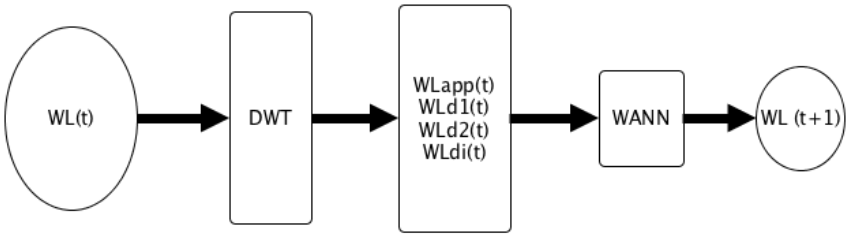
The data was prepared in the following manner, as was mentioned earlier in the previous section, we had prepared four sets of time lags from the original Non-denoised/denoised Balboa time series data; in this experiment, only time lag T = 50 was considered using the criteria that in the previous

analysis, the experimental results showed that this lag set treated with the wrapper subset algorithm for attribute selection outperformed (R = 0.9687) the scenarios under time lags T = 60 and T = 120 and only resulted in being slightly lower than the non-attribute selection scheme R(=0.9043 and 0.9055). Therefore, the denoised time series and the T = 50 lag are input to the ANN model. This wavelet-treated series with its respective time lags was used to create a hybrid model combining two methods, DWT and ANN, for a resulting WANN model that uses decomposed sub-time series, which was extracted using DWT on the tide level data. For this hybrid model, the primary time series of phenomena were decomposed into sub-time series components (details and approximation) using the sym6 wavelet family, with level = five in the MATLAB (2018a Academic version) Wavelet Toolbox. The hybrid model was then constructed, and the details and approximation of the original input time series were the input of the WANN model.

So in this experiment, the time series of the measured tide level were decomposed into several multi-frequency time series $W_{d1}(t)$, $W_{d2}(t)$, $W_{d3}(t)$, $W_{d4}(t)$, $W_{d5}(t)$, …, $W_{di}(t)$, $W_a(t)$ by DWT, in which $W_{d1}(t)$, $W_{d2}(t)$ $W_{d3}(t)$, $W_{d4}(t)$, $W_{d5}(t)$, …, $W_{di}(t)$, and $W_a(t)$ are the details and approximation of the Balboa Harbor water level time series, respectively (Figure 4). Therefore, di presents the decomposed series in level i, and a is the approximation series. Subsequently, the decomposed water level time series were employed at different levels of the time lag T = 50 as input data to the ANN technique for predicting water level in the future. These time series have various effects in the primary time series, and the role of each one is different; hence, each one's contribution to the primary time series differs from that of others. The details of the WANN model are shown in Figure 5.

Figure 5

Architecture of the resulting WANN hybrid model. Own work.



4.3. *Artificial Neural Networks*

Although the literature does not provide a specific guideline for splitting the data, it is generally agreed that most data points should be used for model building.

The regression tree M5P is the model tree learner using the M5 algorithm (Quinlan, 1992) implemented in the Weka software. The criterion employed in the M5P algorithm as a regression classifier is straightforward in carrying out the partitions. A traditional decision tree is combined

with the potential for linear regression functions at the nodes in M5P. Initially, a tree is constructed using a decision-tree induction process. Nevertheless, a splitting criterion is applied, which minimizes the intra-subset fluctuation in the water level values down each branch, as opposed to optimizing the information gain at each inner node. If the water level values of every instance that reaches a node differ significantly or there are only a few examples left, the splitting process in M5P comes to an end.

Pruning the tree back from each leaf is the second stage. A regression plane is used to transform an inner node into a leaf during pruning. Thirdly, the prediction from the leaf model is coupled with the amount that the linear model predicts for each node along the journey back to the root using a smoothing technique., to prevent abrupt discontinuities between the subtrees.

One advantage of decision tree classifications is that rules can be inferred from the trees generated that are very descriptive, helping users to understand their data. In contrast, in the case of M5P, it is not this way for linear models. Weka software can generate decision trees and rules depending on selected options. Trees and model rules were generated using 10-fold cross-validation, and the results with the best value for the Correlation coefficient (*R*) and most negligible value for the Percentage relative absolute error (%RAE) on the test dataset were selected.

It was decided to use the complete dataset to create two scenarios or experiments to find a predictive model expressed mathematically in Equation 5.

It is known that time series forecasting models are best when the data sets are considerably large; this is also true for data mining processes, as they require a large amount of data, which is split and used for training and testing purposes. As mentioned in the previous section, we employed a t-1 to t-50 lags approach as information from pass values is needed to predict future values.

The experimental models are built with the usual cross-validation approach. It is known from the literature that cross-validation is one of the most essential tools in evaluating regression and classification methods (Arlot & Celisse, 2010; Kunst & Jumah, 2004). This leads us to the justification of the challenge of forecasting performance with fewer instances. As the system's generalization is significant in time series modeling, many different techniques, of which one repeated cross-validation is the method of choice in most practical situations of data scarcity, and it shows efficiency when the amount of data for training and testing is limited.

For both scenarios mentioned above, we experimented with the following classifier schemes: a rule, a function, and a regression model tree to find which algorithm was suitable enough to deal with the data set and, therefore, perform as a feasible model predictor. Table 2 shows that the regression model tree gave the best percentage correctness.

Table 2
Statistical test results for the experiment.

Dataset	**Classifiers**		
Water level (100)	tree.M	rule	function
	0.79	0.00 *	0.76 *

As shown in the literature, all the statistics compare actual values to their estimates but do it slightly differently. They tell us "how far away" our estimated values are from the actual value θ. Sometimes square roots are used, and sometimes absolute values are used; however, the results are more affected by extreme values when utilizing square roots. For example, in RAE and RRSE, we divide those differences by the variation of θ. Since they have a scale that goes from 0 to 1, you can get a similarity scale that also ranges from 0 to 100 by multiplying this value by 100 (e.g., percentage). The values of $\sum(\theta - \theta_i)^2$ or $\sum|\theta - \theta_i|$ tell us how much θ differs from its mean value, so we could tell that it is about how much θ differs from itself, compared to variance.

For this reason, the measures are named "relative." They give us results related to the scale of θ. In this sense, the performance measure for the three classifier schemes is judged by the commonly used error measure, the root relative squared error (RRSE). The best scheme was chosen based on the results of the statistics. Model three, M5P, was chosen as the best classifier scheme in this case.

After completing the experiments, we selected the time series that was grouped into two scenarios for modeling: (i) the DWT-treated denoised data and (ii) the non-DWT-treated data series. We then ran and analyzed both cases with the selected M5P algorithm, using cross-validation with 10-fold as the testing method.

4.4. *Model Evaluation*

The measurements for functionality used for the runs were the agreement coefficient (*R*), the absolute mean error (MAE), the percentage relative absolute error (%RAE), and the index of agreement (*d*). The R, which goes from - ∞ to 1, indicates the direction and intensity of a linear relationship between two variables. If R is near to 1, the model prediction is optimal. RAE, or total absolute error, is simply a way to compare models built with larger or smaller valued data. It is scaled from 0-100. The MAE calculates the degree to which forecasts or projections match actual results. Moreover, values vary from 0 to 1, and the ideal model prediction is near 0. Higher values of the d, a bounded and nondimensional metric, indicate higher agreement between the simulated and observed values. These measures are all computed according to the following equations:

$$R = \frac{\frac{1}{N}\sum_{i=1}^{N}[(W_m)_i - \overline{W}_m]\,[(W_o)_i - \overline{W}_o]}{\sqrt{\frac{1}{N}\sum_{i=1}^{N}[(W_m)_i - \overline{W}_m]^2}\sqrt{\frac{1}{N}\sum_{i=1}^{N}[(W_o)_i - \overline{W}_o]^2}} \quad (10)$$

$$\mathrm{MAE} = \frac{1}{N}\sum_{i=1}^{N}[(W_m)_i - (W_o)_i] \quad (11)$$

$$\mathrm{RAE} = \frac{\sum_{i=1}^{N}|(W_m)_i - (W_o)_i|}{\sum_{i=1}^{N}|\overline{W}_o - (W_o)_i|} \quad (12)$$

$$d = 1.0 - \frac{\sum_{i=1}^{N}|(W_m)_i - (W_o)_i|}{\sum_{i=1}^{N}|(W_m)_i - \overline{W}_o| + |(W_o)_i - \overline{W}_o|} \quad (13)$$

where N is the total number of observations in the dataset; W_m is the predicted water level; W_o is the observed water level.

As the modeling scenarios were running, interesting patterns representing knowledge were identified. So, as patterns appeared, we ran each of the scenarios mentioned above applying parameter tuning as follows: without the aid of an attribute selection method and another test with attribute selection methods using both the cfs subset evaluator and the wrapper subset evaluator (Kohavi & John, 1997) on both approaches, to compare performance between each scenario and identify which attribute selection scheme was best at finding the adequate number of variables amongst the architecture of the selected lag and this way determine the relevance of attributes and henceforth giving the best accuracy in prediction.

5. ANALYSIS RESULTS AND DISCUSSION

5.1. Non-denoised Data

The data from the Balboa Harbor water level time series without denoising is used, with the implementation of time lags t-1 to t-50. Regression with 10-fold cross-validation is applied. Three runs corresponding to the parameter tuning scheme for each scenario were executed; the first was done by applying the M5P classifier without submitting the data set to an attribute selection algorithm. To determine the relevance of attributes, the second and third runs were implemented using the cfs subset evaluator and the wrapper subset evaluator. By evaluating each attribute's predictive power independently and the degree of redundancy between them, the cfs subset evaluator favors attribute sets that have minimal intercorrelation but strong correlations with the water level. While a classifier is also used by the wrapper subset evaluator to assess attribute sets, cross-validation is used to determine the accuracy of each set's learning scheme.

Then, to the time lags created, the data was submitted to classification with each attribute selection learner, and the results and performance of each were compared.

We report results for the runs to reveal the effectiveness of model implementation. It is unsurprising that across Table 3, model fitting and forecasting performance generally differ and show that forecasting performance increases as subsample time lags increase. However, it is announced that the model performance can be improved when attribute selection schemes are applied and the number of lags is higher. Results also indicated that the Cfs evaluator scheme performed worse than the other two approaches.

It can be noted that the variations of the MAE measures for the tree schemes applied are similar. Nevertheless, the %RAE showed a difference. It was significantly lower for the results obtained with the non-attribute selection scheme, indicating that this model implementation is adequate for the data set in question. Notwithstanding, results for run three with the wrapper attribute selection scheme can also be used as model implementation for that case.

Overall, for one month ahead, forecasting that the model executed without an attribute selection scheme applied selected 23 attributes as relevant out of 50. For cases run with the cfs subset evaluator and wrapper subset evaluator, 3 and 7 attributes were selected as relevant.

Moreover, experimental research demonstrated that a linear model with 23 characteristics was produced by the model without the application of an attribute selection technique which was represented as time lags (t-49, t-48, t-40, t-38, t-36, t-30, t-28, t-23, t-22, t-21, t-20, t-15, t-13, t-12, t-11, t-10, t-9, t-8, t-6, t-5, t-4, t-2, t-1 respectively). Initially, the model probably observed similar trends during the first two months of the series. Then, it discretized, as it skipped one month in between the sequence back in time (e.g., time lags, t-3 and 7) in search for similar patterns of the initial trend until it arrived at time lag, t-9. This addressed issue agrees that stationarity in a series is a critical factor. So, if non-stationarity cannot be removed by time series conventional methods, the model-building procedure might need a phase in the processing that establishes which components of the series to include in the modeling, as proposed by (Deco et al., 1997) or the prediction of the series might even be an impossible task (Kim et al., 2004). We also observed an increase and decrease in the consecutive months the model discretized to find similar patterns in the time series. So, for this case, the frequency of discretization observed by the model was 1, 1, 1, 5, 5, 1, 6, 1, 2, and 8 months back in time until it could arrive at the following month or series of months in which it could observe a similar pattern of the trend observed during the first two months of the series.

For model runs with the cfs subset evaluator, we observed a generated linear model that selected three attributes: three lags out of 50 back in time (t-23, t-12, and t-1, respectively). This represents twenty values less than what was selected by the first scheme (no-attribute selection). Although this number of attributes selected is lower, the previous model's performance is better in statistical terms, as it yielded a lower value of the RAE (39.732%) and a better fit of the model (R = 0.9055). As observed, this model observed similar trend patterns for the first month with the previous scheme. Notwithstanding, during the discretization process, the overall performance of this algorithm was shown to skip ten months before arriving at the following month or months, when a similar trend pattern was identified. Finally, continuing to discretize among lag values, the model took ten times more months than the previous. This is more lags back in time, where it could identify similar trend patterns for the preceding month. The experimental analysis also showed the wrapper algorithm to select seven-time lags as relevant attributes (t-48, t-38, t-23, t-14, t-12, t-11, and t-1, respectively). A quick inspection revealed that the attributes selected are almost like those selected in the previous case, which is the non-attribute selection scheme. Perhaps this scheme, the wrapper, did not consider seventeen of the lags previously selected by the non-attribute selection scheme. On the other hand, the discretization performance showed that the model needed to skip 9, 1, 8, 14, and 9 months back in time until it found the preceding month or months with similar behavior in the patterns; these concerns showed the importance of stationarity and the use of cross-validation in time series and the dependence of a model on pasts values to predict future values. In general, the non-attribute selection scheme ruled out over the other two schemes for this scenario.

Table 3

Statistical summary of results and Generated Linear Models with runs made with non-denoised dataset T= 50.

Run No.	Performance Measure *R*	MAE	Relative absolute error (%)	*d*	**Attribute Scheme No.1** Without Attribute Selection Evaluator	**Attribute Scheme No.2** Cfs Evaluator	**Attribute Scheme No.3** Wrapper Evaluator
1	**0.9055**	0.038	39.732	0.944	23	–	–
2	0.8847	0.042	43.901	0.989	–	3	–
3	0.9043	0.038	40.065	0.989	–	–	7

Generated Linear Models

1	Water Level = -0.0556 * t-49 + 0.0935 * t-48 + 0.0385 * t-40 - 0.0827 * t-38 + 0.0503 * t-36 + 0.0372 * t-30 - 0.0668 * t-28 + 0.1364 * t-23 - 0.0709 * t-22 + 0.0835 * t-21 - 0.0625 * t-20 - 0.0812 * t-15 - 0.0804 * t-13 + 0.1754 * t-12 + 0.064 * t-11 + 0.154 * t-10 - 0.0443 * t-9 - 0.0602 * t-8 - 0.0409 * t-6 + 0.1241 * t-5 - 0.0412 * t-4 - 0.037 * t-2 + 0.7071 * t-1 + 0.4206
2	Water Level = 0.0659 * t-23 + 0.2077 * t-12 + 0.3569 * t-1 + 2.5117
3	Water Level = 0.0928 * t-48 - 0.094 * t-38 + 0.1288 * t-23 - 0.1539 * t-14 + 0.1695 * t-12 + 0.1622 * t-11 + 0.658 * t-1 + 0.258

5.2. Denoised Data

Wavelet transformation is a methodology that analyses time-frequency localization with fixed window size and with changeable time-windows and frequency-windows (Meyer, 1993). By applying wavelet analysis, longer time windows can detect low-frequency information, while shorter ones are used to detect high-frequency information. Wavelet transformed signal is decomposed into components that have different scales. This offers a method for a local outlook of the signal, a multi-scale outlook, and a time-scale analysis (Misiti et al., 1996). Thus, it can perform various tasks, including detecting discontinuities in signals, detecting trends, analyzing time-frequency, denoising, and compression of signals (Hamed & Rao, 2000).

In this section, we present the statistical results for forecasting the water level 1 month in the future by combining the wavelet decomposition model (Figure 5) and ANN. The wavelet artificial neural network (WANN) model uses the sub-time series components obtained by applying the discrete wavelet transform to the Balboa water level data (Figure 6).

According to Table 4, the WANN model implemented as the second scheme in this study performed better than the scheme without the non-denoised data. Regarding R = 0.9687, MAE = 0.022, and % RAE = 23.065, it is also interesting to note that the WANN model improves R by 93

% compared to the best single ANN model. One of the main reasons for the WANN model is to eliminate noise in the input data during the pre-processing analysis.

For model runs with the non-attribute selection scheme, we observed a generated linear model that selected twenty-nine attributes, twenty-one lags out of 50 (Table 4).

Table 4

Statistical summary of results and Generated Linear Models with runs made with denoised dataset T= 50.

Run No.	Performance Measure				Attribute Scheme No.1 Without Attribute Selection Evaluator	Attribute Scheme No.2 Cfs Evaluator	Attribute Scheme No.3 Wrapper Evaluator
	R	MAE	Relative absolute error (%)	*d*			
1	0.9682	0.021	22.673	0.975	29	–	–
2	0.9277	0.033	35.486	0.000	–	4	–
3	**0.9687**	0.022	23.065	0.779	–	–	16

Generated Linear Models

1	Water Level = -0.0309 * t-49+ 0.0576 * t-47- 0.0683 * t-42 + 0.1645 * t-41 - 0.1684 * t-40 + 0.0759 * t-39 - 0.0921 * t-37 + 0.0592 * t-36 + 0.066 * t-35 - 0.0777 * t-34 + 0.0609 * t-33 - 0.0559 * t-31 + 0.1218 * t-29 - 0.1279 * t-28 + 0.1074 * t-23 - 0.0776 * t-22 + 0.0781 * t-21 - 0.0949 * t-19 + 0.1362 * t-17 - 0.182 * t-16 + 0.0517 * t-15 + 0.0869 * t-14 - 0.2328 * t-13 + 0.2209 * t-12 + 0.1052 * t-9 - 0.0642 * t-8 + 0.316 * t-3 - 0.8168 * t-2 + 1.3409 * t-1 + 0.2841
2	Water Level = -0.215 * t-43 + 0.3492 * t-23 + 0.3116 * t-12 + 0.5328 * t-1 + 0.1667
3	Water Level = -0.0528 * t-49 + 0.0755 * t-47 + 0.1097 * t-29 - 0.1412 * t-28 + 0.0974 * t-23 - 0.0987 * t-19 + 0.1559 * t-17 - 0.1924 * t-16 + 0.192 * t-14 - 0.3497 * t-13 + 0.2873 * t-12 + 0.1253 * t-9 - 0.0845 * t-8 + 0.3145 * t-3 - 0.7906 * t-2 + 1.3148 * t-1 + 0.266

6. CONCLUSIONS AND RECOMMENDATIONS

This experiment explored two approaches to forecast time series, mainly the wavelet transform implementation and using data mining with cross-validation and attribute selection algorithms.

The M5P model tree classification algorithm was used to generate model trees and linear model rules for predicting water level changes. The data from the Balboa Harbor, Panama, was obtained from the Permanent Service for Mean Sea Level database between 1908 and 2007. At the time, the series only had data until 2007 since this data was assumed to have been reduced to a common datum. (Revised Local Reference, RLR) It has been recommended to be used for time series

analysis by the PSMSL Organization rather than the metric, so we decided to use the data set for the abovementioned purpose.

The approach used to analyze the Balboa data included reviewing traditional forecasting, data mining, wavelet transform methods, as depicted by Figures 3 through 6 results, and data mining techniques for time series. We performed a comprehensive empirical study, which included implementing four sets of lagged times, t-10, t-50, t-60, and t-120, each composed of non-denoised and denoised data, respectively, as time dependencies for the data set.

Using standard 10-fold cross-validation, the forecasting reliabilities of these models were evaluated by computing some statistical measures and applying attribute selection schemes as a combination of scenario runs on the time lags implemented. As we were aware of irregularities and unique patterns in the dataset, it was decided to use attribute selection algorithms to investigate which subset of attributes produces the best cross-validated classification accuracy for the three scenarios implemented, as to allow each attribute individually from the entire dataset and run a cross-validation for each reduced version of the dataset. Once we had determined the best number of attributes in the dataset, the process was repeated with this reduced dataset to find the best number of attributes, and so on.

The trials' outcomes reveal that employing cross-validation and attribute selection schemes improves prediction proficiency. The wrapper subset evaluator outperformed the other two schemes for all executed scenarios. The best prediction was shown for time lag, t-120, which suggests that the length of the time series was essential for the model and if enhanced with the application of an attribute selection evaluator, performance can be improved significantly, as was the case with the application of the wrapper subset evaluator. Nonetheless, we need to address that as model complexity increases, so does the model prone to overfitting. Therefore, there is a chance for error to increase, as was discussed in the experimental results and analysis section. We also noticed that a certain number of lags is necessary as the data contains dependencies characterized by the nature of the time series (e.g., stationarity, seasonality, and irregularities), and time-evolving environmental effects may occur. Although not bad, the cfs subset evaluator performance did not excel over the non-attribute choice strategy for predicting the water level in statistical terms.

The shortcomings of the commonly used methods in time series forecasting are well documented, and various other methods have been proposed in the literature. For this reason, a more extensive data set, which will comprise data collected over many decades, will be needed to pursue better results.

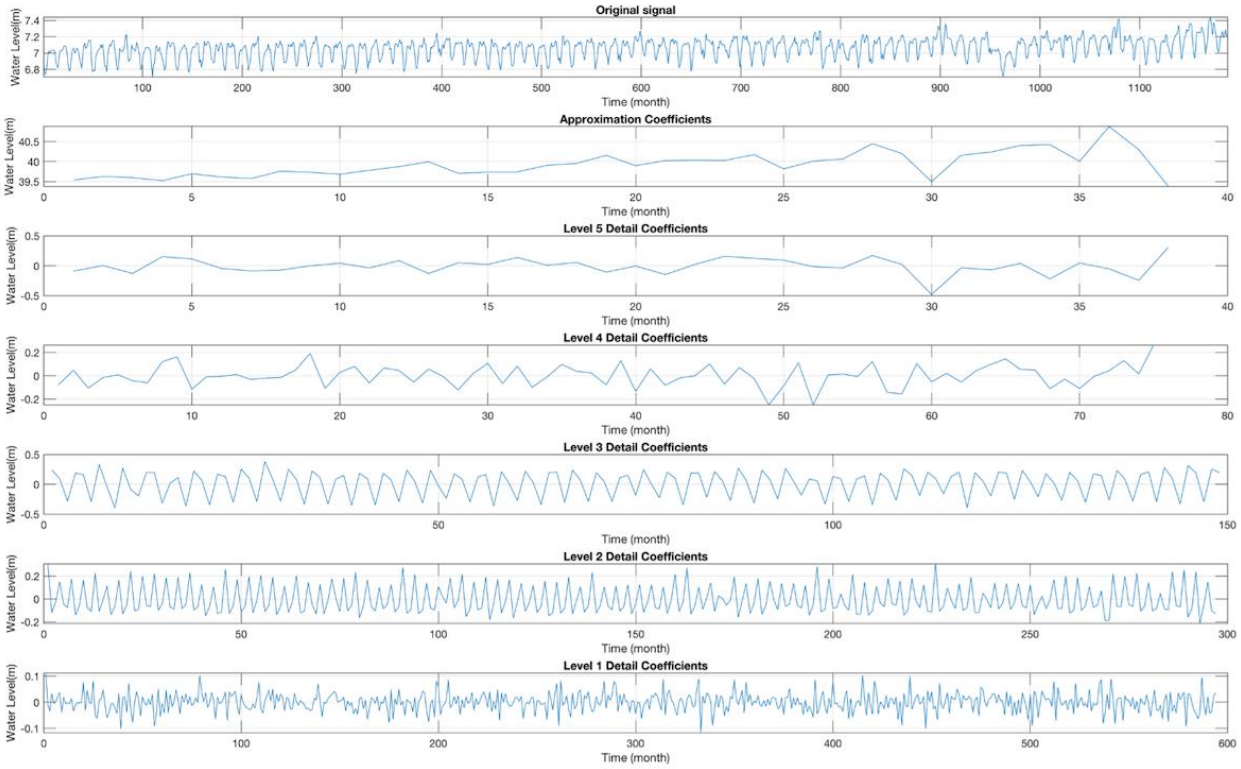
We can conclude that this work is vital for climate change studies because the variations in weather conditions in terms of water level rise in an area of study that is within the boundaries of the Panama Canal operations are not well documented and can be studied using these data mining techniques.

7. ACKNOWLEDGMENTS

The authors thank the Spanish MICINN under projects: TRA2015 63708-R, and TRA2016-78886-C3-1-R for partially supporting this work. Similarly, the authors acknowledge and thank the Vicerectoría de Investigación y Postgrado of the University of Panama, University of Panama, Departamento de Biología Marina y Limnología, University of Panama, Facultad de Ingeniería y de Ciencias de la Tierra, Departamento de Ingeniería Civil, Ingeniería Hidrológica y de Recursos Hídricos, and the Secretaría Nacional de Ciencia, Tecnología e Innovación (SENACYT) for supporting this research project.

Figure 6

Wavelet decomposition of the Balboa Harbor water level Time Series. Own work.



REFERENCES BIBLIOGRAFICAS

Albiach, J. C. C., Fanjul, E. A., Lahoz, M. G., Gomez, B. P., & Sanchez-Arevalo, I. R. g. (2000). Ocean forecasting in narrow shelf seas: application to the Spanish coasts. *Coastal Engineering*, *41*(1-3), 269-293.

Arlot, S., & Celisse, A. (2010). A survey of cross-validation procedures for model selection.

Bergmeir, C., & Benítez, J. M. (2012). On the use of cross-validation for time series predictor evaluation. *Information Sciences*, *191*, 192-213.

Bregman, J. I., & Makenthun, K. M. (2006). Environmental Impact Statements.

Cannas, B., Fanni, A., Sias, G., Tronci, S., & Zedda, M. K. (2005). River flow forecasting using neural networks and wavelet analysis. Geophys. Res. Abstr,
Cartwright, D. (1972). Secular changes in the oceanic tides at Brest, 1711–1936. *Geophysical Journal International*, *30*(4), 433-449.
Cartwright, D. E., & Driver, J. (1971). Tides and waves in the vicinity of Saint Helena. *Philosophical Transactions of the Royal Society of London. Series A, Mathematical and Physical Sciences*, *270*(1210), 603-646.
Chen, J., Shum, C., Wilson, C., Chambers, D., & Tapley, B. (2000). Seasonal sea level change from TOPEX/Poseidon observation and thermal contribution. *Journal of Geodesy*, *73*, 638-647.
Deco, G., Neuneier, R., & Schümann, B. (1997). Non-parametric data selection for neural learning in non-stationary time series. *Neural networks*, *10*(3), 401-407.
Douglas, B., Kearney, M. T., & Leatherman, S. P. (2000). Sea level rise: History and consequences. *International Geophysics Series, Academic Press London*, *75*, 97-119.
Hamed, K. H., & Rao, A. R. (2000). Trend analysis by using wavelets. In *Building Partnerships* (pp. 1-10).
Haykin, S., & Network, N. (2004). A comprehensive foundation. *Neural networks*, *2*(2004), 41.
Holgate, S. J., Matthews, A., Woodworth, P. L., Rickards, L. J., Tamisiea, M. E., Bradshaw, E., Foden, P. R., Gordon, K. M., Jevrejeva, S., & Pugh, J. (2013). New data systems and products at the permanent service for mean sea level. *Journal of Coastal Research*, *29*(3), 493-504.
Kamruzzaman, J., Sarker, R., & Begg, R. (2007). Artificial Neural Networks. *Intelligent Information Technologies: Concepts, Methodologies, Tools, and Applications: Concepts, Methodologies, Tools, and Applications*, 222.
Kim, T. Y., Oh, K. J., Kim, C., & Do, J. D. (2004). Artificial neural networks for non-stationary time series. *Neurocomputing*, *61*, 439-447.
Kişi, Ö. (2009). Neural networks and wavelet conjunction model for intermittent streamflow forecasting. *Journal of Hydrologic Engineering*, *14*(8), 773-782.
Kohavi, R., & John, G. H. (1997). Wrappers for feature subset selection. *Artificial intelligence*, *97*(1-2), 273-324.
Kunst, R. M., & Jumah, A. (2004). *Toward a theory of evaluating predictive accuracy*.
Labat, D., Mangin, A., & Ababou, R. (2002). Rainfall–runoff relations for karstic springs: multifractal analyses. *Journal of Hydrology*, *256*(3-4), 176-195.
Long, N. C., & Meesad, P. (2014). An optimal design for type–2 fuzzy logic system using hybrid of chaos firefly algorithm and genetic algorithm and its application to sea level prediction. *Journal of Intelligent & Fuzzy Systems*, *27*(3), 1335-1346.
Meyer, Y. (1993). Wavelets: algorithms & applications. *Philadelphia: SIAM (Society for Industrial and Applied Mathematics*.
Misiti, M., Misiti, Y., Oppenheim, G., & Poggi, J.-M. (1996). Wavelet toolbox. *The MathWorks Inc., Natick, MA*, *15*, 21.
Monbaliu, J., Padilla-Hernandez, R., Hargreaves, J. C., Albiach, J. C. C., Luo, W., Sclavo, M., & Guenther, H. (2000). The spectral wave model, WAM, adapted for applications with high spatial resolution. *Coastal Engineering*, *41*(1-3), 41-62.
Pashova, L., & Popova, S. (2011). Daily sea level forecast at tide gauge Burgas, Bulgaria using artificial neural networks. *Journal of sea research*, *66*(2), 154-161.
Peralta, J., Gutierrez, G., & Sanchis, A. (2010). Time series forecasting by evolving artificial neural networks using genetic algorithms and estimation of distribution algorithms. The 2010 international joint conference on neural networks (IJCNN),
Quinlan, J. R. (1992). Induction of decision trees. *Machine learning*, *1*, 181-186.
Rosso, O., Figliola, A., Blanco, S., & Jacovkis, P. (2004). Signal separation with almost periodic components: a wavelets based method. *Revista mexicana de física*, *50*(2), 179-186.
Rourke, F. O., Boyle, F., & Reynolds, A. (2010). Tidal energy update 2009. *Applied Energy*, *87*(2), 398-409. https://doi.org/10.1016/j.apenergy.2009.08.014

Rushing, J., Ramachandran, R., Nair, U., Graves, S., Welch, R., & Lin, H. (2005). ADaM: a data mining toolkit for scientists and engineers. *Computers & geosciences*, *31*(5), 607-618.
Simmonds, J. A., Gomez, J. A., & Ledezma, A. (2017). Forecasting sea level changes applying data mining techniques to the Cristobal Bay time series, Panama. *Journal of Water and Climate Change*, *8*(1), 89-101. https://doi.org/10.2166/wcc.2016.041
Umgiesser, G., Canu, D. M., Cucco, A., & Solidoro, C. (2004). A finite element model for the Venice Lagoon. Development, set up, calibration and validation. *Journal of Marine Systems*, *51*(1-4), 123-145.
Zhang, G. P. (2007). A neural network ensemble method with jittered training data for time series forecasting. *Information Sciences*, *177*(23), 5329-5346.