Analysis of the Gulf of Mexico Coastal and Estuarine Water Levels as Random Signals

Alexey L Sadovski, G. Beate Zimmer, Blair Sterba-Boatwright,
Philippe Tissot, Ray Bachnak.
Department of Computing and Mathematical Sciences
Texas A&M University-Corpus Christi
6300, Ocean Dr., Corpus Christi, Texas 78412
USA

Abstract: This paper explores estimates for the best possible water level predictions for test stations in the shallow waters of the Gulf of Mexico. The predictions are made by Artificial Neural Networks (ANNs) and Statistical Modeling. Namely, here we use the theory of stochastic processes and spectral analysis to evaluate the best possible quality of forecasts by using the stochastic properties of the inputs as well as the given goodness criteria of predictions. As a result of such investigation we can outline limitations of the ANN predictions of water levels in the bays and estuaries of the Texas coast.

Key-Words:-Stochastic, Predictions, Random Signal, Neural Networks, Modeling, Statistics.

1 Introduction

Tides are the result of astronomical forces and tide tables reflect only these forces, not meteorological influences such as weather. In deep coastal waters, the tide tables are sufficiently accurate, but in shallow coastal waters the meteorological components of the water levels can not be ignored. The importance of accurate forecasts is stressed by NOAA emphasis on its PORTS project: "The **P**hysical **O**ceanographic **R**eal-**T**ime **S**ystem (PORTS®) is a program of the National Ocean Service that supports safe and cost-efficient navigation by providing ship masters and pilots with accurate real-time information to avoid groundings and collisions.

Since about 1990, the Texas Coastal Observation Network (TCOON), housed in the Conrad Blucher institute within the College of Science and Technology at Texas A&M University–Corpus Christi (TAMUCC) has collected and archived data on water levels, wind strength and wind direction for a network of gauging stations along the Gulf of Mexico coastline in Texas. These data are available in real time on web at http://lighthouse.tamucc.edu/~pquery. For the last five years, researchers at more or less closely affiliated with Texas A&M University – Corpus Christi have worked on using these data to generate 12-hour or 24-hour water level predictions for these stations.

A. Sadovski [1] has used regression and statistical models to obtain forecasts.
P. Tissot with P. Michaud, D. Cox and S. Duff have worked with Artificial Neural Network (ANN) models [2]. The quality of a water level forecast is measured by several criteria devised by the National Ocean Service: they define the tolerance level of a forecast as 15 cm and define any prediction that differs by more than 30cm from the actual water level as an outlier. Their assessment criteria are: the root mean square error (RMSE), the Central Frequency (i.e. the percentage of time the forecasts are within 0.15 m of the measured water level), the positive outlier frequency (POF) and the Negative Outlier Frequency (NOF) as well as the maximum duration of a positive

outlier (MDPO) and the maximum duration of a negative outlier (MDNO).
The successes of the different models are detailed in [3]. One feature was common to all models: some stations were easier to predict than others. Easy stations included Bob Hall Pier and Port Aransas on the Bay side of the barrier island outside Corpus Christi Bay, whereas for Pleasure Pier outside the Barrier Island shielding Houston Bay the predictions were less successful and forecasts for Morgans Point in the Houston Ship Channel proved to be the biggest enigma for all models.
The following central quality for forecasts with different models are taken from [3], which used models trained on one year of data prior to 2000 and tested on the three consecutive years 2000-2002, before averaging the values over the tree test years.

Table 1. Models performances for the study's stations tested on 2000-2002.

Station/Model	RMSE [m]	CF [%]	POF [%]	NOF [%]	MDPO [hrs]	MDNO [hrs]
Bob Hall Pier						
Tide table	0.114	84.2	0.35	1.65	17	54
Persistence (24 hr)	0.086	92.0	0.45	0.17	8	8
Linear Regression (24hr)	0.224	93.2	0.33	0.17	17	16
ANN (24 hr)	0.075	94.6	0.30	0.10	8	6
Port Aransas						
Tide table	0.112	83.7	0.31	1.53	19	43
Persistence (24 hr)	0.075	94.2	0.25	0.03	9	0
Linear Regression (24hr)	0.172	94.8	1.05	0.02	24	2
ANN (24 hr)	0.070	95.7	0.16	0.02	7	0
Pleasure Pier						
Tide table	0.149	72.8	1.41	3.39	28	72
Persistence (24 hr)	0.146	79.6	3.07	2.29	25	29
Linear Regression (24hr)	0.149	83.7	2.43	1.31	26	34
ANN (24 hr)	0.123	84.6	2.34	0.80	22	20
Morgans Point						
Tide table	0.174	67.3	3.88	4.56	47	74
Persistence (24 hr)	0.178	71.2	5.29	4.18	31	34
Linear Regression (24hr)	0.110	55.6	2.86	1.33	18	22
ANN (24 hr)	0.142	80.4	4.13	0.68	26	12

To improve the Morgans Point forecasts using just past and present wind and water data, we tried a range of things. P. Tissot experimented at length to determine the ideal inputs and architecture of the ANNs. He had written the ANN program used by all TAMUCC researchers involved with the ANN prediction of water levels and he tried to optimize the input parameters - namely the number of hours of past water levels used, hours of past wind data used and the number of neurons in the hidden layer - for the ANNs in an order that followed his physical intuition into the situation, while trying to use ANNs of minimal complexity, ideally with just one neuron in the hidden layer.
B. Zimmer [4] generalized this approach to Monte Carlo methods and generated 250 ANNs with randomly chosen parameters and evaluated their performance over the six test years 1999-2004. The highest Central Frequency reached was 78.16%, the lowest CF 77.56%. That permits the conclusion that the design of the network was not the hindering factor.

B. Sterba-Boatwright used Kalman Filtering on the 1998 input data. This smoothed out the training data a little bit, but in testing the ANN on unfiltered data no significant differences in the performance were obtained, indeed the Central Frequency for the test years dropped by 0.15%.
B. Zimmer [5] experimented with different training methods and training algorithms and assessments of the training, again with no significant success.
There are still avenues of ANN design and training to be explored, but all researchers involved reached the same conclusion: the forecasts for Morgans Point are not limited by our ANN design methods but by the input data.

2 The Problem:

The question at hand is why different regression and ANN models and approaches cannot improve quality of predictions for some stations such as Morgans Point. To explain it we decided to use statistical factor analysis. We applied factor analysis to the water levels over the period of 48 hours with the interval of 2 hours. The conclusion is that no more than 5 factors explain over 90% of variance for water levels for all TCOON stations. Then we compared the results of the factor analysis for shallow water stations with the results of factor analysis for deep water stations. For instance at the shallow water stations Bob Hall Pier and Morgans Point the first main component explains over 68% of the variation of the primary water levels.
Analyzing results of factor analysis for different stations we have discovered the following:

- In coastal shallow waters and estuaries the major or the first component is not a periodical component, and we call this component "weather". Other main components are periodical and we call them "astronomical".
- In off-shore deep waters, the first two or three components are astronomical components, while weather is a less dominant component.
- Our conclusion is that the prime factor affecting water levels in estuaries and shallow waters is weather.
- It has been observed also, that linear regression models for different locations have different coefficients for the same variables. We think that this difference may be explained by the geography of the place where the data is collected.

3 Primary Water Levels as a Stochastic Process.

The qualitative analysis of the covariance function of the primary water levels at the Morgans Point and Bob Hall Pier as well as the nature of the processes (excluding tropical storm surges) permit us to assume that both primary water levels are stationary stochastic processes. A stochastic process X(t) is called a stationary process if the covariance function K(t,t') of the process depends only on the difference τ between t and t' and does not depend on t and t', where

$$K(t,t') = \frac{E[(X(t)-m_t)(X(t')-m_{t'})]}{\sigma_t \sigma_{t'}}$$

and E[….] is the expected value, m_t and σ_t are the mean and the standard deviation of the stochastic process X(t) at the time t.

Verifying the stationarity for any set of measured data is nearly impossible, but we will make an attempt to it later in the paper. However, we can prove that the process under consideration is an ergodic process. By calculating running averages we can show that the time average is equal to the

ensemble average, which is supported by the graphs on Figures 4 and 5 for Morgans Point and Bob Hall Pier respectively.

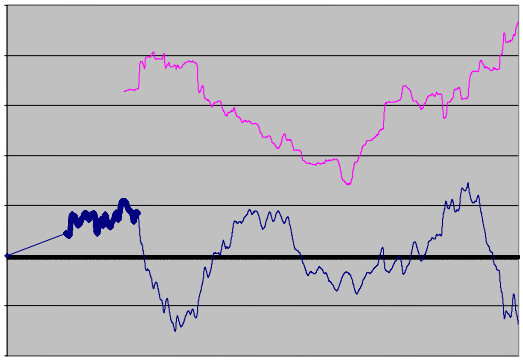

Figure 4. Mean (blue) and standard deviation (red) of moving averages of 1000 observations for Morgans Point.

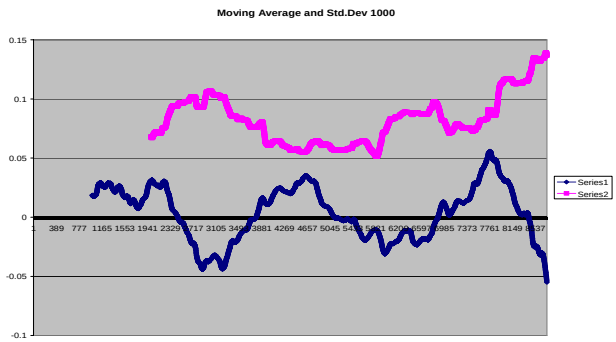

Figure 5. Mean (blue) and standard deviation (red) of moving averages of 1000 observations for Bob Hall Pier.

It is clear by looking at the data on these graphs that we cannot reject the hypothesis that the mean is zero. Indeed, the confidence interval around the constant mean m=0 will be $(-e_{\frac{\alpha}{2}}; e_{\frac{\alpha}{2}})$, where $e_{\frac{\alpha}{2}} = z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}$. For example, for level of confidence α=0.05 we have $z_{\frac{\alpha}{2}}$=1.96 and $\frac{\sigma}{\sqrt{n}}$>0.05, so the confidence interval is (-0.1; 0.1) which includes all displayed values of running averages.

Now, we know that transition functions of different ANN either linear or almost linear over interval [-1;1]. This implies that we can evaluate output standard deviation by using the following formula: If $X = aY + bZ$ then $\sigma_x = \sqrt{a^2 \sigma_y{}^2 + b^2 \sigma_z{}^2}$.

Since all the input to ANN are ergodic and assumed to be stationary and since the operations within the ANN are continuous then the output of the ANN is an ergodic process by proposition 4.3 in [6].

We understand that in order to achieve good prediction the output of the ANN should have (at most) the same standard deviation as the difference between primary water levels and harmonic water levels (which are also inputs of ANN). Detailed observation showed that for Morgans Point the covariance function K(τ) becomes insignificant for τ≥2, while for Bob Hall Pier there is significant covariance for τ up to the 16th hour. The above statement means that we deal with white noise at Morgans Point and not so white noise at Bob Hall Pier. The following table 4 provides standard deviations for the primary water levels and the difference between primary water levels and harmonic water levels for the "bad" year 1998 and the "good" year 2000.

Table 4. Standard deviations.

	1998 pwl	2000 pwl	1998 p-h	2000 p-h
Morgans Point	0.262	0.226	0.211	0.165
Bob Hall Pier	0.235	0.201	0.147	0.093

Now we are to try to demonstrate that the input data is a stationary process. With this purpose it is enough to show that for sufficiently long different periods of time autocorrelation functions are almost the same [7]. To be more precise, we cannot reject null-hypothesis that these to auto-regressions are the same. It is well known that addition or subtraction of a deterministic process to a stationary process is results in a new stationary process. Now we know that the process under consideration (primary water levels minus harmonics) is an ergodic one. Now we can verify our previous assumption that this process is a stationary process. Here we have to check hypothesis that for different and long enough realization of the process we have the same covariation (autocorrelation function). The two graphs on Figures 6 and 7 give us an excellent example that our claim is valid. To be more precise, for any given lag number both values of respective autocorrelations belong to the same confidence interval which implies that we fail to reject claim that it is the same covariance function.

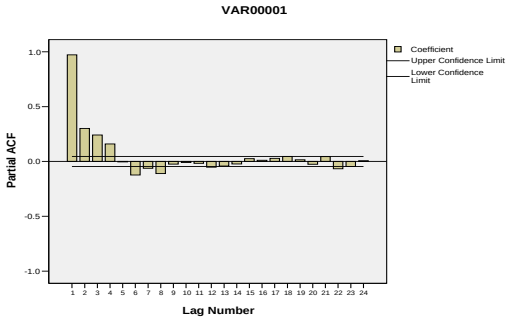

Fig 6. Autocorrelation function, the first 6 month of 1998, Bob Hall Pier

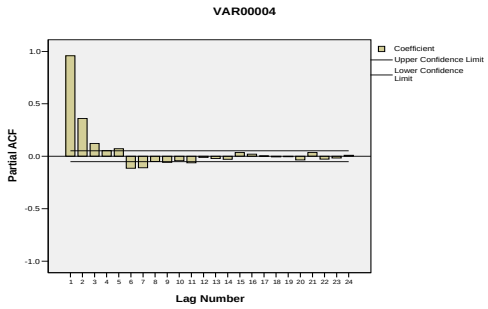

Fig 7. Autocorrelation function, the last 6 month of 1998, Bob Hall Pier

Now we can estimate the probability of the forecast hitting [-0.15;+0.15] interval around p-h where p is a primary water level and h is a harmonic water levels. Taking into consideration that approximately σ(p-h)=.21 for Morgans Point we can estimate probability by using celebrated table of Normal Distribution. As we can see from the Fig.8 There is strong correlation r=0.98 between two neighbor hours for p-h water levels. Predicting for 12 or 24 hours by using strongly correlated next to each other water levels will give us coefficients of correlation r(12)=0.785 and r(24)=0.616 respectively. It is well known that coefficient of determination r^2 gives proportion of deviation explained by correlation. The other part of the variations cannot be explained and has random nature. So we have $s = \sigma(1-r^2)$ as the standard deviation for the uncertainty of the prediction, and we can find that s(12)= .0818 as well as s(24)= 0.128

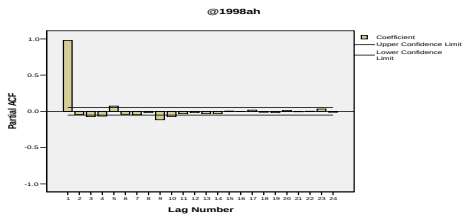

Fig.8 Autocorrelation Function for Morgans Point, 1998.

Now we can apply celebrated Normal Distribution to find probability of hitting desired interval [-0.15;+0.15] for prediction by using well known formula

$$p(t) = 2\Phi(\frac{0.15}{s(t)})$$

Here we get the following values p(12)=0.9328 for twelve hours prediction and p(24)=0.758 for twenty four hours forecast. It means that at the best we should expect success in prediction of water levels of 93% for half day forecast and 76% of successful predictions for a 24 hours forecast.

The other way to make an estimation is the following. Suppose we work with 1998 Morgans Point training data, and just the current water level (no past water levels, no wind) the ANN outperforms the persistent model with a CF of 71.2% instead of 66.5%. The weights and biases are hidden neuron w = - 0.5440 b= 2.4949 for the output neuron w = - 15.0537 and b = 13.9364. The transfer function for the hidden neuron is f(x) = 1 /(1 + exp(-x)). With one input this takes the input x in the interval [-1 , 1] and turns it into 1 / (1 + exp(0.5440 x - 2.4949)). If we graph this function on the interval [-1,1] then we could see that it is concave down, but quite close to its tangent line at x=0, whose equation is y = -0.038301 x + 1 / (1 + exp(- 2.4949)), and the maximum difference between the two functions on [-1,1] is 0.01. Now the output neuron gets to work, it multiplies the output of the hidden neuron by - 15.0537 and adds bias of 13.9364.
Because of -0.038301*- 15.0537= 0.5765822535 the standard deviation of the output data is only 57.66% of the standard deviation of the input data. For instance, for 1998 the input data have a standard deviation of 0.2135, hence the output data will have a standard deviation of 0.2135*0.5765822535=0.1231516112, and the 15 cm or 0.15 m divided by 0.1231516112 give a z-score of 1.21801861 with respect area 0.1112 to its right under the standard normal curve. So the central frequency in this case will be 1-2*.1112=0.7776 or 77.7% .At the same time the skill set calculated a CF of 71.2%, not 77.7%.

4 Conclusions:

1. We have demonstrated that the input data represents an ergodic process.
2. We have shown that primary water levels as well as difference between them and harmonic water levels are stationary stochastic processes.
3. We evaluated by different means probability of the successful forecast, which is -for 24 hours from now – is approximately equal to 75%.
4. Now we some tools to evaluate Central Frequency of the prediction by analyzing stochastic properties of the input data.

5 Acknowledgements:

The work presented in this paper is funded in part by the following federal and state agencies:
- National Aeronautic and Space Agency (NASA Grant #NCC5-517)
- Texas Research Development Grant
- National Oceanic and Atmospheric Administration (NOAA)
- Coastal Management Program (CMP).

The views expressed herein are those of the authors and do not necessarily reflect the views of NASA, TGLO, NOAA, CMP or any of their sub-agencies.

References:

[1]. Sadovski, A. L., P. Tissot, P. Michaud, C. Steidley, Statistical and Neural Network Modeling and Predictions of Tides in the Shallow Waters of the Gulf of Mexico. In "WSEAS Transactions on Systems", Issue 2, vol. 2, WSEAS Press, pp.301-307.

[2] P.E. Tissot, D.T. Cox, and P.R. Michaud, "Optimization and Performance of a Neural Network Model Forecasting Water Levels for the Corpus Christi, Texas, Estuary", 3rd Conference on the Applications of Artificial Intelligence to Environmental Science, Long Beach, California, February 2003.

[3]. Tissot P., Cox, D., Sadovski, A., Michaud, P., and Duff, S., Performance of Water Level Forecasting for the Texas Ports and Waterways, proceedings of the PORTS 2004 Conference, Houston, TX, May 23–26, 2004.

[4]. G. Beate Zimmer, Philippe E. Tissot , Jeremy S. Flores, Zack Bowles, Alexey L. Sadovski, and Carl Steidley, Water Level Forecasting along the Texas Coast: Interdisciplinary Research with Undergraduates. Proceedings of the 2005 American Society for Engineering Education Annual Conference & Exposition, 2005. Published as a CD.

[5]. G. Beate Zimmer, Jeremy S. Flores, Philippe E. Tissot, and Alexey L. Sadovski, Research with an Undergraduate Student: Using Entropy to Assess the Training of a Neural Network. Proceedings of the American Society for Engineering Education- Gulf-Southwest Section 2005 Annual Conference. Published as a CD.

[6]. Krengel, U. "Ergodic Theorems", Walter de Gruyter, 1985.

[7]. J. L. Doob, Stochastic Processes, John Wiley&Sons, p.654, 1953