SEA LEVEL PREDICTION IN THE NORTH-WESTERN BLACK SEA USING AUTOREGRESSIVE INTEGRATED MOVING AVERAGE AND MACHINE LEARNING MODELS

Dr. Maria Emanuela Mihailov[1]
Dr. Alecsandru Vladimir Chirosca[2]
Dr. Gianina Chirosca[3]
[1] National Institute for Marine Research and Development "Grigore Antipa" Constanta, Romania
[1] Maritime Hydrographic Directorate, Constanta, Romania
[2, 3] Faculty of Physics, University of Bucharest, Magurele, Ilfov, Romania

ABSTRACT

Data prediction models are essential for estimating extreme environmental changes and predicting anomalies by learning when the actual data is outside previously accepted values. This paper focuses on predicting two years of sea level in the North-Western Black Sea region. Data from the UNESCO/ IOC tsunami observation and Permanent Service for Mean Sea Level archives were analysed using Auto Regression - and Seasonal-Regression Integrated Moving Average models. This work proposes one such model obtained by using modern Machine Learning algorithms, and the results are compared with standard models such as ARIMA obtained for the same data. Using Machine Learning can produce software models ready to run with hardware using much lower specifications than those used for model training which is not the case for standard statistical models. The merged dataset in the analysed period (2006-2016) from the tide gauges along the Romanian Black Sea Coast is consistent and satisfactorily used to develop and validate a Seasonal Regression Integrated Moving Average and Machine Learning model for sea-level forecasts. The data show that the sea level evolution in cyclical changes of the other parameters that influence it. Furthermore, slight demarcation of the two models was observed between the comparison of observed and predicted values.

Keywords: *sea level, ARIMA, SARIMA, Machine Learning, Black Sea*

INTRODUCTION

The socio-economic developments and infrastructures around the world have been concentrated in the coastal area. As a nodal point for naval transport (commercial, recreational, research, or military purposes) or the richness of marine resources, this area offers many future perspectives at the interface between land and sea. The most predominant concern worldwide is the sea level rise that can negatively affect/endangering the coastal area's infrastructure and inhabitants. That implies a change in the mean sea level due to the climate change perspective, but locally, it can induce wind and storm / extreme events frequencies that contribute to coastal zone vulnerability. [1] estimate a rise in all scenarios within a range of 0.6 - 0.7m in 2100 regarding the "global mean sea level" rise. Various statistical sea level forecasting strategies using ANN (Artificial neural network), ARMA (Mixed Autoregressive-Moving Average process), ARIMA (Autoregressive Integrated

Moving Average process), ARFIMA (Autoregressive Fractionally Integrated Moving Average process) are used in the scientific community to find the best-fit model for the time series data for the sea level measurements [2], [3], [4].

In the Black Sea, few studies evaluated, estimated or modelled the sea level based on in-situ measurements or satellite altimetry for the entire Black Sea basin [5, 6] or for coastal areas: the Turkish and Georgian coast [7], the Russian coast [8], and for Romanian [9]. For the entire Black Sea basin, the estimated rate of the mean sea level rate calculated from satellite data is 7.6 ± 0.3 mm/year [6]. On the other hand, the average annual amplitude of sea-level variations is about 0.35 mm from the altimetry or 0.56 mm from the tide gauges [5], [7]. Furthermore, from available in-situ data, for the entire Black Sea coast, the sea level shows different ascendent rates/patterns (sea level rise): about 1.37 mm/year at Constanta (Romania), at Varna (Bulgaria) 1.22 mm/year, Bourgas (Bulgaria) a rate of 1.91 mm/year, Batumi (Georgia) by 1.96 mm/year, 6.68mm/year at Poti (Georgia), Sevastopol (Ukraine) 1.26 mm/year and at Tuapse (Russia) with 2.46 mm/year [9].

In the present study, the long-term time series on the regional scale (Western Black Sea coast) is used to detect a historical data pattern and then extrapolated for future forecasts. The analysis and models are performed using a Python environment. This paper is focused on the time decomposition of sea level in the Western Black Sea and the formulation of short-term forecasts using ARIMA modelling. Using data prediction models is suitable for identifying the seasonal and non-seasonal parameters that determine current and future hydrological conditions based on historical data analysis. One of the most popular time-series analyses is ARIMA. Still, the main challenge remains to set up models that accurately describe temporal hydrodynamic conditions and enable accurate forecasts. Furthermore, further observation of time series usually reveals significant, non-random relationships.

MATERIALS AND METHODS

Study area and data source. The Black Sea, a semi-enclosed basin located in Eastern Europe, connects with the Marmara Sea through the Bosporus Strait and northeast with the Sea of Azov through the Kerch Strait.

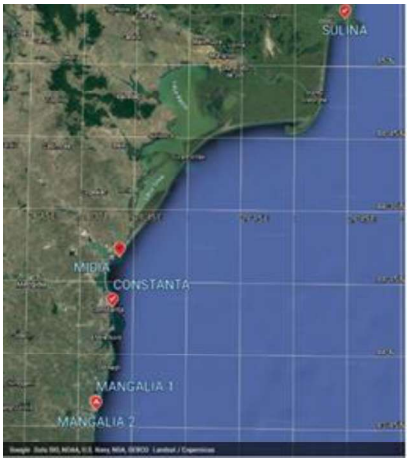

Fig. 3. *Coastal sea level tide gauges on the Romanian Black Sea shelf [10]*

In this paper, for developing sea level short-term ARIMA forecasts on the Western Black Sea coast, Constanta stations mean sea level, and "csta" and "csta2" are used for the 2006 – 2018 periods. Also, for other coastal stations as Sulina – station no. 93 (2016 - 2018), available data, are used for correlation with other sea-level coastal stations and similar for Mangalia -station no. 91(2016 – 2018) from Joint Research Centre Sea level database. All monitoring stations along the Romanian Black Sea coast are illustrated in Fig. **3**. On the Romanian Black sea coast, three organisations record and analyse the sea level for different purposes. National Institute for Marine Research and Development "Grigore Antipa" (NIMRD), "csta" station, use a systematic recording of sea level realised at Constanta (44.17N, 28.67E) through a mechanical level recorder [9] For station "csta2", since 2015, the National Institute for Earth Physics (NIEP) installed a radar tide gauge sensor OTT model RLS used to record sea level at Constanta (44.15N, 28.67E) as a component of a network for marine seismicity purposes (Fig. **3**). Furthermore, we specify that annually high precision GPS measurements at the landmarks of all tides and their zeros will be unified with the ultimate goal of updating the zeros "sea-level" of the Black Sea. These measurements are carried out by the Institute of Geodesy, Photogrammetry, Cartography and Cadastral Bucharest Romania (for NIMRD and NIEP tide gauges).

Model setup. We build prediction models using the recorded data, starting with a conservative approach using the standard seasonal ARIMA model with exogenous regressor - SARIMAX. It is appropriate for our data pattern [11], [12]. We further increase the prediction accuracy by using machine learning methods first through Long-Term Short-Term memory layers [13], followed by an encoder-decoder pattern [A4 14] that provides better accuracy than the previous two. Seasonal ARIMA (Seasonal Auto-Regressive Integrated Moving Average with eXogeneus Regressors Model) prediction model uses a set of parameters (p, d, q) presented in eq. 1 where L is the lag function. Our model was fitted against the sea level data using Akaike Information Criteria (AIC) [15] minimisation. AIC, eq. 2, evaluates

the model deviation against the actual data for the distribution probability *f* with the measured data g.

$$(p,d,q)\ x\ (P,D,Q)_S, \phi p(L) < \phi >_P (L^s)\Delta^d \Delta_s^D D_s y_t = A(t) + < \theta_q > (L) < \theta_Q > (L^s)\zeta t \quad (1)$$

$$I(g:f) = E\ log\ g(x) - Elog\ f(x) \quad (2)$$

Recurrent network design was implemented using the Keras framework with TensorFlow backend. Our model consists of 3 layers. The first two LSTM types with 36 fully connected neurons use rectified linear unit activation function and one layer with a single neuron for output (**Error! Reference source not found.**). The o ptimiser is nadam with an error estimation of type mean squared error.

Table 4. *Machine Learning model for sea-level prediction using LSTM layers*

Layer (type)	Output Shape	Parameters #
1stm_23 (LSTM)	(None, 60, 36)	5472
1stm_24 (LSTM)	(None, 36)	10512
Dense_15 (Dense)	(None, 1)	37
Total params: 16,021 Trainable params: 16,021 Non-trainable params: 0		

No overfitting was identified for our model using a window of 60 records for 200 epochs. However, introducing dropout layers increases the model uncertainty.

RESULTS AND DISCUSSIONS

Data sets. Recorded data is always susceptible to missing data points due to device malfunction or other external factors. Our dataset has only a few missing data points (11 missing data points from the total dataset of 385 records). To achieve data evaluation and build prediction models (data has to be continuous) and for the missing points, a linear approximation was performed, taking into account two value points before and after the missing data point. For such cases, a polynomial regression model was performed for the 4 data points, and thus we obtained the analytical function used to interpolate missing data. *Fig. **4*** shows how an interpolation model for one missing data point was built.

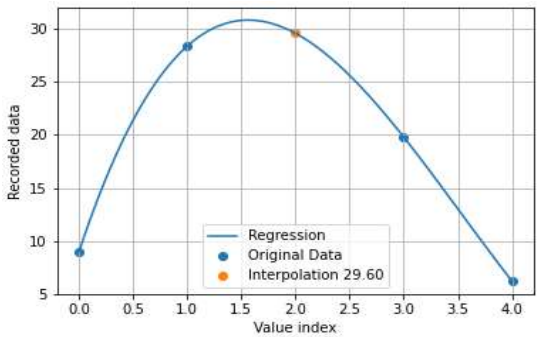

Fig. 4. *Data interpolation using a 3rd order polynomial regression. For this case, the obtained polynomial is* $1.2x^3-12.7x^2+30.9x+9$.

From the 85 years of the monthly dataset, 83 years are used for model calibration (1933 - 2016), while the last two years (2017 - 2018) validate the statistical model. All databases described (PSML, IOC) were integrated into one, and the duplicates were excluded. Box and whisker plots were realised: the median, minimum and maximum values, and the lower (25%) and upper (75%) quartiles. The quartiles delimit each box, and the whiskers represent the extreme values. The dataset contains 982 instances of approximately 40 years from 1933 to 2018 with six attributes, as shown in *Table 5*.

Table 5. *Merged sea-level data description and dataset attributes*

Variable Name	Description
YEAR	Year
MONTH	Month
STATION	Station name
LEVEL	Numerical Sea-Level
SLP	Sea level pressure
FLOW	Danube Discharge

The sea level data: the Romanian Black Sea. We consider that: if time series show gradual shifts to relatively higher or lower values over a more extended period, then it can be stated that a pattern exists in the datasets. However, limited understanding of the dynamic processes that influence the sea level rise determines low confidence in quantitative projections. Relevant processes on the western Black Sea coast are river discharge (Danube river), river and sea ice melting (during intense winters), and climate change.

The linear function y=Bx+A was used (*Fig. **5***) to determine the quantitative relationship between the river discharge and the sea- level**Error! Reference source n ot found.**. Danube river discharge and the sea level have an excellent correlation coefficient in the range of 0.64. Therefore, the slope (coefficient B) could show the river discharge influence on the mean sea level. Meanwhile, coefficient B has negative values. Considering the Constanta Port data as basic in analysing the other tide gauges - Sulina and Mangalia - the correlation was made between the Romanian coastal sea-level stations. The coefficient of correlation is very good ($r_a = 0.78$; $r_b = 0.79$, and $r_c = 0.85$), and the slope of the regression is close to unity. The following correlations are determined following the linear function (*y=Bx+A*), as *Fig. **5***:

a) $Z_{Mangalia} = 0.94\ Z_{Constanta} + 124.8$ (r = 0.79)

b) $Z_{Sulina} = 1.01\ Z_{Constanta} + 20.3$ (r = 0.78)

c) $Z_{Sulina} = 0.91\ Z_{Mangalia} - 89.0$ (r = 0.85)

were, the differences between the free terms correspond to the different zeros of the tides.

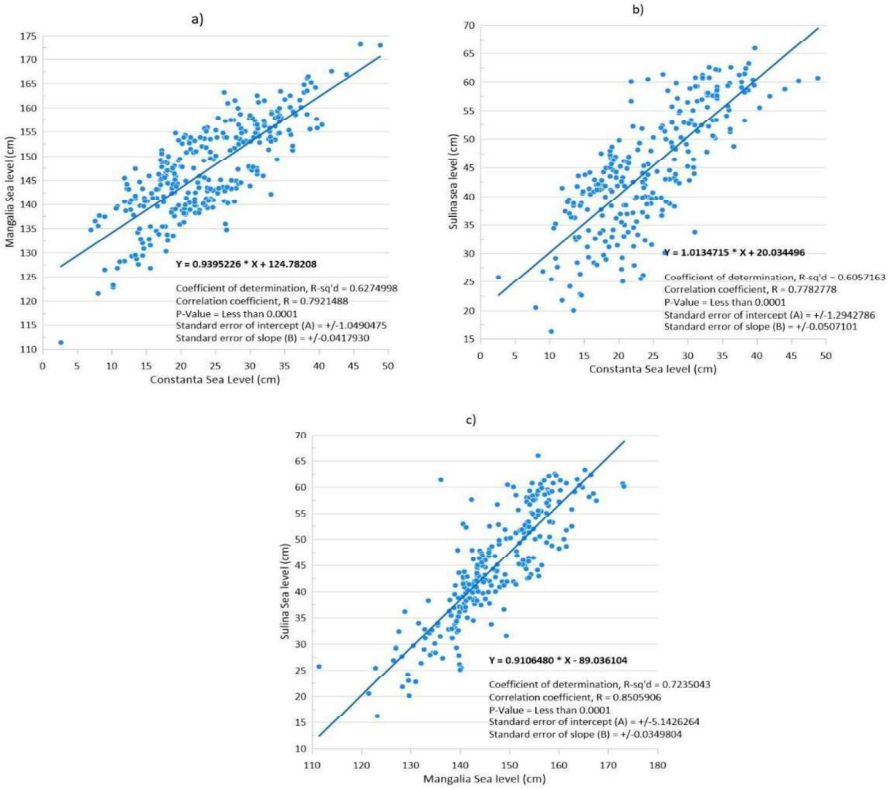

Fig. 5. *The correlation between sea level (cm) recorded at the tide gauges, 2016 - 2018 period: a) Constanta and Mangalia; b) Constanta - Sulina, and c) Mangalia and Sulina.*

Pre-processing, Model setup and analysis. We evaluated the correlation and partial correlation for our data (*Fig.* ***6***). The autocorrelation clearly shows the need for seasonal models.

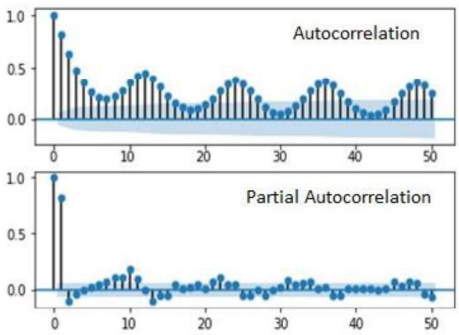

Fig. 6. *Autocorrelation and partial autocorrelation for sea level data at Constanta (1933 - 2018)*

For SARIMAX model development, we used a 12 months window. We evaluated the AIC for a large set of model parameters obtaining the minimal AIC value for the parameters collection (2,1,2) x (2,1,2,12), getting an AIC value of 6583.19. The predicted values for our model are presented in *Fig. 7*.

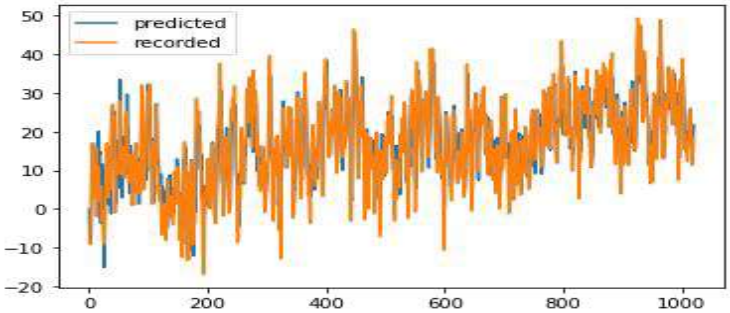

Fig. 7. *SARIMAX predicted versus measured sea-level data at Constanta (2006 - 2018)*

A time-series pattern can be identified by evaluating the monthly movements in historical data, as shown in *Fig. **8***. However, SARIMAX predictions have significant errors (RMEs) for the inflexion data points, providing a mean RME of 18, showing that more parameters (water flow, the temperature a.s.o) need to be evaluated for better accuracy. To maintain univariate data (1D input data), we employ a Machine Learning (ML) method for the same dataset using a Recurrent Neural network using the sea level monthly means (2006 – 2018). The ML model trained for 10 hours using the provided dataset (monthly averages for 2006-2015) and delivered the results presented in *Fig. **8***.

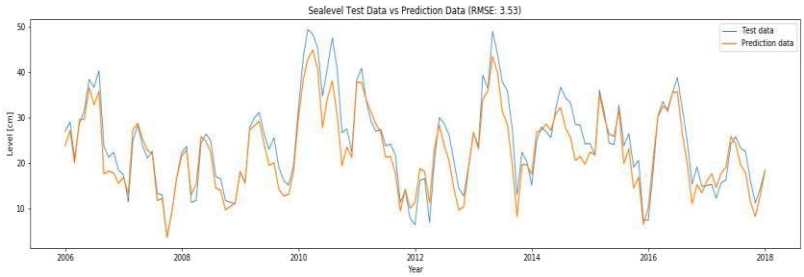

Fig. 8. *Machine Learning using LSTM layers predicted versus measured sea level data at Constanta site (2006 - 2018)*

The ML model provided a better RMS than the SARIMA model (3.53) for the provided dataset. The accuracy plot presented in **Error! Reference source not found.** shows that univariate data suggests intrinsic variations that lead to higher errors. Our model was improved using an encoder-decoder pattern for seq2seq time series prediction to assess such parameters properly.

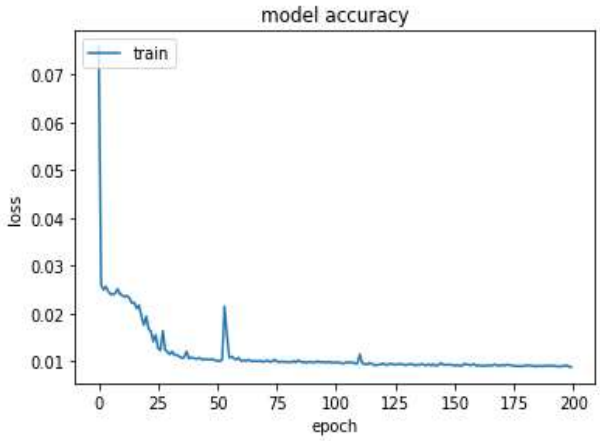

Fig. 9. *Machine Learning using LSTM accuracy plot for each epoch (each time the model sees the data).*

From the last two figures (*Fig. **8***, **Error! Reference source not found.**), the accuracy cannot be improved by increasing the epoch counts, and new models have to be used to increase the model accuracy.

CONCLUSION

Sea-level simulation shows the potential capability of machine learning algorithms to forecast based on historical time-series data. In general, satisfactory predictions with the values of correlation coefficients equal to 0.7–0.9, root mean square errors of about 10% of the tidal range and scatter indexes equal to 0.1–0.2 were produced. The validated neural methodology can be successfully applied to other coastal regions provided site-specific training and appropriately carried out validations. The neural technique was successfully implemented to predict hourly sea levels in the first set of simulations, with lead times from 1 to 24 hours, and afterwards to correct the initial simulations results. An ancillary correcting neural network improved the quality of the initial hourly simulations when assimilating the outcomes of more complex nets. Simultaneously, its application led to overfitting when the input data were coming from the network with the same number of inputs–processing–output units. ANNs were implemented to forecast sea levels in the second set of simulations, averaged over 12-h, 24-h, 5-day and 10-day periods, three-time steps ahead. This produced high-quality predictions over the first two time steps rather than over the third time interval. The latter's lack of accuracy was attributed to the reduced number of input-output training pairs and weak interrelations.

ACKNOWLEDGEMENTS

We thank Permanent Service for Mean Sea Level, UNESCO / IOC tsunami observation / alerting network and European Commission Joint Research Centre - ISPRA - Space, Security and Migration Directorate (JRC) for maintaining the

servers and the long-term data availability. MEM study has been carried out with financial support from the Monitoring Study Contract – Romanian Ministry of Water and Forests no.55/2018. Furthermore, MEM continued the work, performed under the Sectorial Research-Development Plan of the Romanian Ministry of National Defence 2020 and 2021.

REFERENCES

[1] Church JA, et al., Sea level change. Climate Change 2013: The Physical Science Basis, Contribution of Working Group I to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change, Cambridge Univ Press, Cambridge, UK, pp 1137–1215, 2013 (https://www.eea.europa.eu/data-and-maps/indicators/sea-level-rise-5/nicholls-et-al-2010-sea, accessed 1.25.21).

[2] Ercan A., Kavvas M.L., Abbasov R.K., Long-Range Dependence and Sea Level Forecasting, SpringerBriefs in Statistics. Springer International Publishing, pp 51, 2013.

[3] Fernandez Q.F.R., Montero N.B., Po III R. B., Addawe R. C., Diza H. M. R., Forecasting Manila South Harbor mean sea level using Seasonal Arima models, Journal of Technology Management and Business, vol. 5/issue 1, 2018.

[4] Srivastava P.K., Islam T., Singh S.K., Petropoulos G.P., Gupta M., Dai Q., Forecasting Arabian Sea level rise using exponential smoothing state space models and ARIMA from TOPEX and Jason satellite radar altimeter data: Forecasting Sea level rise using altimeter data, Meteorological Applications, vol. 23, pp 633–639, 2016.

[5] Avsar N.B., Jin S., Kutoglu H., Gurbuz G., Sea level change along the Black Sea coast from satellite altimetry, tide gauge and GPS observations, Geodesy and Geodynamics, vol. 7, pp 50–55, 2016.

[6] Kubryakov A.A., Stanichnyi S.V., The Black Sea level trends from tide gauges and satellite altimetry, Russian Meteorology and Hydrology, vol. 38, pp. 329–333, 2013.

[7] Avsar N.B., Jin S., Kutoglu S.H., Recent sea level changes in the Black Sea from satellite gravity and altimeter measurements. The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences, vol. XLII-3/W4, pp 83–85, 2018.

[8] Goriacikin I.N., Ivanov V.A., Black Sea Level: Past, Present, Future, EKOCI-Gidrofizica, Sevastopol, pp 210, 2006. (in Russian)

[9] Mihailov M.-E., Buga L., Spînu A.-D., Dumitrache L., Constantinoiu L.-F., Tomescu-Chivu M.-I., Interconnection between Winds and Sea Level in the Western Black Sea Based on 10 Years Data Analysis from the Climate Change Perspective, Cercetări Marine - Recherches Marines, vol .48 / issue 1, pp 171–178, 2018.

[10] https://earth.google.com/web/

[11] Vagropoulos S.I., Chouliaras G., Kardakos E.G., Simoglou C.K., Bakirtzis A., Comparison of SARIMAX, SARIMA, modified SARIMA and ANN-based models for short-term PV generation forecasting, 2016 IEEE International Energy Conference (ENERGYCON), Leuven, 2016, pp 1-6.

[12] Durbin J, Koopman S.J., Time Series Analysis by State Space Methods: Second Edition. Oxford University Press, 2012.

[13] Hochreiter S., Schmidhuber J., Long Short-Term Memory, Neural Computation, vol. 9/ issue 8, pp 1735-1780, 1997.

[14] Cho K., van Merrienboer B., Gulcehre C., Bahdanau D., Bougares F., Schwenk H., Bengio Y., Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation, Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), Doha, Qatar, 2014, pp 1724–1734.

[15] Akaike H., Akaike's Information Criterion, International Encyclopedia of Statistical Science, Lovric M. (eds) Springer Berlin Heidelberg, pp 25, 2011.