Celal Bayar University Journal of Science

Comparison of Random Forest, SVR and KNN Based Models in Sea Level Prediction for Erdemli Coast of Mersin

Yavuz Karsavran[1]*

[1]İstanbul Şişli Vocational School, Architectural Restoration Program, İstanbul, Türkiye
*karsavran@itu.edu.tr
*Orcid No: 0000-0001-5944-0658

Received: 1 November 2023
Accepted: 28 March 2024
DOI: 10.18466/cbayarfbe.1384547

Abstract

Seawater level prediction is very important in terms of future planning of human living conditions, flood prevention and coastal construction. Nevertheless, it is hard to correctly predict the daily future of sea water level because of the atmospheric conditions and effects. Therefore, Random Forest (RF), Support Vector Regression (SVR) and K-Nearest Neighbor (KNN) methods were used for the prediction of seawater level on Erdemli coast of Mersin in this study. In this paper, root mean square error (RMSE) and coefficient of determination (R^2) were applied as model evaluation criteria. In addition, 15-minute sea water level data of Erdemli Station for approximately 18 months were obtained and used as is. The results depict that Random Forest model can predict the seawater level for 1st and 2nd days with R^2 of 0.80, 0.63, respectively, KNN model can predict for 1st and 2nd days with R^2 of 0.80, 0.64, respectively, and SVR model can predict for 1st and 2nd days with R^2 of 0.77, 0.60, respectively.

Keywords: Random Forest, SVR, KNN, sea level prediction, Mersin Erdemli coast, Mediterranean.

1. Introduction

Due to climate change and human endeavors, seawater level around the World has increased significantly in recent years [1-3]. Accurate prediction of sea level circulation is an important phenomenon for coastal areas with increasing population [3-5]. Because, sea level rise destructively influences ecological habitat and social economy of coastal zones [6,7].

There are commonly two approaches to predicting water levels; physically based modeling and machine learning. Unlike physically based models that need various hydrological and geomorphological data, machine learning methods only need historical water level data to predict future vision of water level. This makes machine learning methods more cost-effective and time-efficient than physically based models. For this reason, machine learning methods are widely used in predicting seawater level [8].

Machine learning methods are generally employed in ocean engineering, particularly in predicting sea level change [3,9,10]. Imani et al. [11] predicted Caspian Sea level using satellite altimetry data and they employed SVR and gene expression programming. Kişi et al. [12] studied on daily water level prediction in Lake Urmia by applying hybrid of SVR and firefly algorithm. KNN was used to develop a model to predict the water level of the river during typhoons [13]. Khaledian et al. [14] estimated the Caspian Sea level using a 34-year water level dataset using SVR and ANN. Altunkaynak and Kartal [15] applied SVR and KNN methods to predict the sea level of the Bosphorus for up to 7 days lead time. Karsavran et al. [16] used SVR to predict seawater level oscillations of the Bosphorus. Sea level circulations of western Peninsular Malaysia were predicted using SVR [17]. Alshouny et al. [18] used both SVR and KNN methods for sea level prediction. Guyennon et al. [19] applied RF to predict the water level of Lake Bracciano. Karsavran [20] applied SVR, ANN and MLR models to forecast the Black Sea coast of Sinop.

As seen above, there are many studies in this field, but there is a lack of research on the future vision of sea water level oscillations on the Mersin Erdemli coast. In addition, comparing the prediction performances of RF, SVR and KNN methods is a new phenomenon on the Mersin coast. I evaluate the future prediction performance based on the performances of these machine learning methods.

Celal Bayar University Journal of Science
Volume 20, Issue 2, 2024, p 14-18
Doi: 10.18466/cbayarfbe.1384547

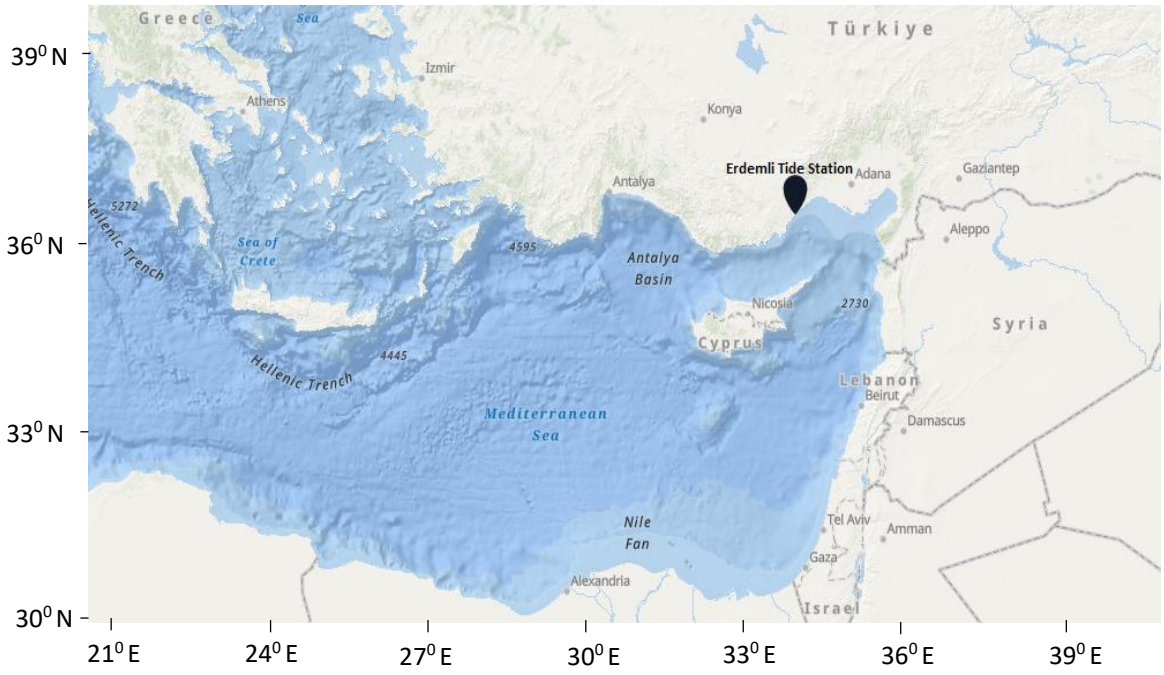

Figure 1. The location of the Erdemli Tide Gauge Station.

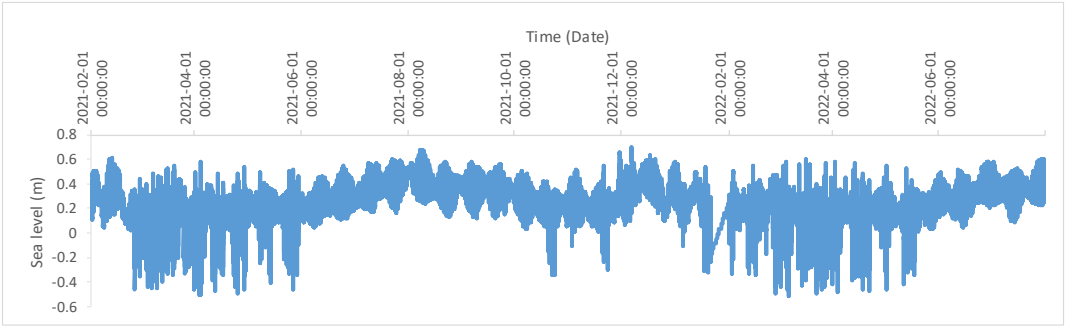

Figure 2. Time series of seawater level in Erdemli station.

2. Materials and Methods

2.1. Data and Study Area

In this study, measurements of the Erdemli tide gauge station (Figure 1), located in the northeast of the Mediterranean, were used. Turkish Sea Level Monitoring System (TUDES, https://tudes.harita.gov.tr/) provided sea level data at 15-minute time intervals. Seawater level has been measured at Erdemli Station since May 2003, and 18-month measurements from February 2021 to August 2022 were used in this study (Figure 2). Linear interpolation method was employed to estimate the missing data. In this study, 30% of the total data was used for testing and the remaining 70% was used for training all models [16]. Data separation was done randomly and the same test and training data were used for each model run.

2.2. Methods

2.2.1. Random Forest

Random Forest (RF), one of the ensemble machine learning methods, is a packing (bootstrap collection) model. RF creates multiple regression trees constructed independently using a bootstrap sample of the dataset [19,21,22]. Classification and regression trees (CART) algorithm is used to create decision trees. During creating these decision trees, a wide variety of randomly chosen variables are used by the random subspace method. Accordingly, the best-branched variable in each leaf node is decided by the random sub-space technique. The prediction results are introduced by RF according to the individual results of the decision trees and overall results are adjusted according to the average predictions of the decision trees [23].

2.2.2. Support Vector Regression

SVR is a statistical learning-based neural network used in various engineering regression problems [24]. It has a hyperplane-driven machine learning algorithm to partition data from one dimension into higher dimensional space [18]. SVR solves the regression problems with Equation 2.1:

$$f(x)=\sum_{i=1}^{n} w.\phi_i(X)+b \quad \textbf{(2.1)}$$

where w=weight, $\emptyset_i$(X)= Kernel function and b=bias. The optimal objective function is shown in Equation 2.2:

$$\min R=\frac{1}{2}w^2+C\sum_{i=1}^{n}(\xi_i+\xi_i^*) \quad \textbf{(2.2)}$$

The constraint conditions are depicted in Equation 2.3:

$$Subject\_to\begin{cases} f(x_i)-y_i\le\varepsilon+\xi_i \\ y_i-f(x_i)\le\varepsilon+\xi_i^* \\ \xi_i\ge 0,\xi_i^*\ge 0, i=1,2,...,n \end{cases} \quad \textbf{(2.3)}$$

where C= cost factor, ε= allowable error, ξ_i and ξ_i^* are relaxation numbers. Both will be greater than zero if there are some prediction errors, otherwise both will be zero [16,25].

2.2.3. K-Nearest Neighbor

K-Nearest Neighbor (KNN) method is one of the most commonly applied method in machine learning studies. The KNN method is a modeling methodology for regression and classification based on the value of the K parameter, which estimates the distance between the sample features. The distance can be estimated applying Euclidean, Minkowski and Manhattan distance equidistant formulas [26]. In addition to being a plain and easy method to put into action, the KNN is also very effective in predicting yield. It requires no assumptions about data distribution. Due to example-based learning algorithm, incremental learning is easily achieved, requiring no training before making predictions. Thus, KNN has commonly been used for various supervised learning tasks [3].

2.2.4. Model Evaluation Criteria

Model performances were acquired from two different numerical error statistics. These are the coefficient of determination (R^2) and the root mean square error (RMSE) depicted in Equation 2.4 and Equation 2.5, respectively.

$$R^2=\left[\frac{\frac{1}{n}\sum_{i=1}^{n}(WL_0(i)-WL_0^{'})(WL_f(i)-WL_f^{'})}{\sqrt{\frac{1}{n}\sum_{i=1}^{n}(WL_0(i)-WL_0^{'})^2}\cdot\sqrt{\frac{1}{n}\sum_{i=1}^{n}(WL_f(i)-WL_f^{'})^2}}\right]^2 \quad \textbf{(2.4)}$$

$$RMSE=\sqrt{\frac{1}{n}\sum_{i=1}^{n}(WL_f(i)-WL_0(i))^2} \quad \textbf{(2.5)}$$

where $WL_o(i)$ and $WL_f(i)$ are observed and forecasted seawater level, respectively. WL_0' and WL_f' shows their averages, and n is the number of data [27-28].

3. Results and Discussion

In this study, the performances of RF, KNN and SVR for sea level forecast of Mersin Erdemli coast were compared for the next 4 days. First of all, the model input set combination is decided by applying RF. Comparison of input sets for the next day (t+1) prediction results of the RF model at Erdemli Station is shown in Table 1.

The input values WL(t) and WL(t-1) increase the value of R^2 to 0.80, while the next three values WL(t), WL(t-1) and WL(t-2) reduce the performance of the RF. As a result, WL(t) and WL(t-1), which produced the highest performance, were used as inputs in all models.

After decision on the input set, seawater level for Erdemli was estimated using the RF model for lead times of 1, 2, 3 and 4 days. Finally, R^2 is 0.80 and 0.63 for lead times of 1 and 2 days, respectively (Table 2).

In the same way, the KNN was applied to estimate seawater level with related lead times [3]. Similar to the RF, the results are 0.80 and 0.64 of R^2 for lead times of 1 and 2 days, respectively (Table 2).

Additionally, the SVR model was used to predict Erdemli's sea water level with specified lead times. In this model, Radial Basis Function (RBF) was applied as the Kernel function and the C parameter of the Kernel is 1000 [16,29]. The best results of SVR model are 0.77 and 0.60 of R^2 for lead times of 1 and 2 days, respectively (Table 2).

The results of the RF, KNN and SVR models show that the RF and KNN models have similar prediction performances, while the SVR model has slightly worse prediction performance than the RF and KNN models. The RF and KNN models have the same performance in seawater level prediction with R^2=0.80 and RMSE=0.07 for 1-day lead time. Moreover, RF and KNN have similar prediction performances with R^2=0.63 and 0.64, respectively, for 2-day lead time. Similarly, the SVR

model provides prediction performance with R^2=0.77 and 0.60 for 1 and 2 days of lead time, respectively. However, all model prediction performances decrease significantly at 3 and 4 days lead time (Table 2). Accordingly, RF and KNN models are more ideal than SVR in prediction seawater level in Mersin Erdemli for next 1 and 2 days. However, all models have no ability to accurately predict the seawater level for next 3 and 4 days.

Comparison of the RF, KNN and SVR prediction performances of the water level on the Mersin coast is helpful to choose the accurate method in the machine learning methods for future studies in the Mediterranean. Additionally, this research can be used to create a warning system against sudden increases in water levels in Mersin Erdemli. Long-term projections can also be produced according to the increase in water levels in Mersin coast. The results and the approach presented and used in this paper can be applied for the analysis of such phenomena.

4. Conclusion

In this article, RF and KNN models achieved higher determination coefficient R^2 in predicting the sea water level at Mersin Erdemli coast for 1 and 2 days lead time, but SVR model gave slightly worse results than RF and KNN. However, the results of all models show that they have no capability to predict seawater level in Mersin Erdemli coast for the 3 and 4 day lead time.

I believe that the results presented here can open new insights in modeling seawater level. Especially, the machine learning methods used in this paper can be applied to the other regions of the Mediterranean coast, including but not limited to Antalya, Hatay and Adana.

Table 1. Model performance of RF for sea level t+1 based on input sets.

Input Set	Output Set	RMSE (m)	R^2
WL(t)	WL(t+1)	0.079	0.75
WL(t)WL(t-1)	WL(t+1)	0.071	0.80
WL(t)WL(t-1)WL(t-2)	WL(t+1)	0.072	0.79
WL(t)WL(t-1)WL(t-2)WL(t-3)	WL(t+1)	0.072	0.79

Table 2. Model performances with respect to lead time prediction WL(t+L)

Inputs (t = day)	Prediction (t = day)	RF		KNN		SVR	
		RMSE (m)	R^2	RMSE (m)	R^2	RMSE (m)	R^2
WL(t)WL(t-1)	WL(t+1)	0.07	0.80	0.07	0.80	0.07	0.77
WL(t)WL(t-1)	WL(t+2)	0.09	0.63	0.09	0.64	0.10	0.60
WL(t)WL(t-1)	WL(t+3)	0.11	0.48	0.11	0.49	0.11	0.45
WL(t)WL(t-1)	WL(t+4)	0.13	0.34	0.13	0.34	0.13	0.31

Acknowledgement

Thanks to Professor Tarkan Erdik for his support in this study. Also, thanks to TUDES for providing the data.

Author's Contributions

Yavuz Karsavran: Drafted and wrote the manuscript, performed the modeling and result analysis.

Ethics

There are no ethical issues after the publication of this manuscript.

References

[1]. Woodworth, PL, Hunter, JR, Marcos, M, Hughes, CW. 2021. Towards reliable global allowances for sea level rise. *Global and Planetary Change;* 203: 103522.

[2]. Yesudian, AN, Dawson, RJ. 2021. Global analysis of sea level rise risk to airports. *Climate Risk Management;* 31: 100266.

[3]. Jin, H, Zhong, R, Liu, M, Ye, C, Chen, X. 2023. Using EEMD mode decomposition in combination with machine learning models to improve the accuracy of monthly sea level predictions in the coastal area of China. *Dynamics of Atmospheres and Oceans*; 102: 101370.

[4]. Primo de Siqueira, BV, Paiva, A de M. 2021. Using neural network to improve sea level prediction along the southeastern Brazilian coast. *Ocean Model;* 168: 101898.

[5]. Zhao, J, Cai, R, Sun, W. 2021. Regional sea level changes prediction integrated with singular spectrum analysis and long-short-term memory network. *Advances in Space Research;* 68: 4534–4543.

[6]. Bernstein, A, Gustafson, MT, Lewis, R. 2019. Disaster on the horizon: The price effect of sea level rise. *Journal of Financial Economics;* 134: 253–272.

[7]. Meilianda, E, Pradhan, B, Comfort, LK, Alfian, D, Juanda, R, Syahreza, S, Munadi, K. 2019. Assessment of post-tsunami disaster land use/land cover change and potential impact of future sea-level

rise to low-lying coastal areas: A case study of Banda Aceh coast of Indonesia. *International Journal of Disaster Risk Reduction;* 41: 101292.

[8]. Zakaria, MNA, Ahmed, AN, Malek, MA, Birima, AH, Khan, M MH, Sherif, M, Elshafie, A. 2023. Exploring machine learning algorithms for accurate water level forecasting in Muda river, Malaysia. *Heliyon;* 9(7).

[9]. Ishida, K, Tsujimoto, G, Ercan, A, Tu, T, Kiyama, M, Amagasaki, M. 2020. Hourly-scale coastal sea level modeling in a changing climate using long short-term memory neural network. *Science of the Total Environment;* 720: 137613.

[10]. Accarino, G, Chiarelli, M, Fiore, S, Federico, I, Causio, S, Coppini, G, Aloisio, G. 2021. A multi-model architecture based on Long Short-Term Memory neural networks for multi-step sea level forecasting. *Future Generation Computer Systems;* 124: 1–9.

[11]. Imani M, You RJ, Kuo CY. 2014. Forecasting Caspian Sea level changes using satellite altimetry data (June 1992–December 2013) based on evolutionary support vector regression algorithms and gene expression programming. *Glob Planet Change*; 121:53–63.

[12]. Kisi O, Shiri J, Karimi S, Shamshirband S, Motamedi S, Petkovi´c D, Hashim R. 2015. A survey of water level fluctuation prediction in Urmia Lake using Support Vector Machine with firefly algorithm. *Appl Math Comput*; 270:731–743.

[13]. Paul, GC, Senthilkumar, S, Pria, R. 2018. An efficient approach to forecast water levelsowing to the interaction of tide and surge associated with a storm along the coast of Bangladesh. *Ocean Engineering;* 148; 516–529.

[14]. Khaledian, MR, Isazadeh, M, Biazar, SM, & Pham, QB. 2020. Simulating Caspian Sea surface water level by artificial neural network and support vector machine models. *Acta Geophysica*; 68: 553-563.

[15]. Altunkaynak A, Kartal E. 2021. Transfer sea level learning in the Bosphorus Strait by wavelet based machine learning methods. *Ocean Engineering*; 233: 109116

[16]. Karsavran, Y, Erdik, T. 2021. Artificial Intelligence Based Prediction of Seawater Level: A Case Study for Bosphorus Strait. *International Journal of Mathematical, Engineering and Management Sciences;* 6(5): 1242.

[17]. Balogun, AL, Adebisi, N. 2021. Sea level prediction using ARIMA, SVR and LSTM neural network: assessing the impact of ensemble Ocean-Atmospheric processes on models' accuracy. *Geomatics, Natural Hazards and Risk;* 12(1): 653-674.

[18]. Alshouny, A, Elnabwy, MT, Kaloop, MR, Baik, A, Miky, Y. 2022. An integrated framework for improving sea level variation prediction based on the integration Wavelet-Artificial Intelligence approaches. *Environmental Modelling & Software;* 152: 105399.

[19]. Guyennon, N, Salerno, F, Rossi, D, Rainaldi, M, Calizza, E, Romano, E. 2021. Climate change and water abstraction impacts on the long-term variability of water levels in Lake Bracciano (Central Italy): A Random Forest approach. *Journal of Hydrology: Regional Studies;* 37: 100880.

[20]. Karsavran, Y. 2024. Comparison of ANN and SVR based models in sea level prediction for the Black Sea coast of Sinop. *Turkish Journal of Maritime and Marine Sciences;* 1-8.

[21]. Liaw, A, Wiener, M. 2002. Classification and regression by random Forest. *R news;* 2 (3): 18–22.

[22]. Loh, WY. 2011. Classification and regression trees. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery;* 1 (1): 14–23.

[23]. Başakın, EE, Ekmekcioğlu, Ö, Özger, M. 2023. Developing a novel approach for missing data imputation of solar radiation: A hybrid differential evolution algorithm based eXtreme gradient boosting model. *Energy Conversion and Management;* 280: 116780.

[24]. Patil, SG, Mandal, S, Hegde, AV. 2012. Genetic algorithm based support vector machine regression in predicting wave transmission of horizontally interlaced multilayer moored floating pipe breakwater. *Advances in Engineering Software;* 45: 203–212.

[25]. Lin, GQ, Li, LL, Tseng, ML, Liu, HM, Yuan, DD, Tan, RR. 2020. An improved moth-flame optimization algorithm for support vector machine prediction of photovoltaic power generation. *Journal of Cleaner Production;* 253: 119966.

[26]. Li, G, Liu, F, Yang, H. 2022. Research on feature extraction method of ship radiated noise with K-nearest neighbor mutual information variational mode decomposition, neural network estimation time entropy and self-organizing map neural network. *Measurement;* 199: 111446.

[27]. Mehr AD, Kahya E, Olyaie E. 2013. Streamflow prediction using linear genetic programming in comparison with a neuro-wavelet technique. *J Hydrology;* 505: 240–249.

[28]. Karsavran, Y, Erdik, T, & Ozger, M. 2023. An improved technique for streamflow forecasting between Turkish straits. *Acta Geophysica;* 1-12.

[29]. Wang, WC, Chau, KW, Cheng, CT, Qiu, L. 2009. A comparison of performance of several artificial intelligence methods for forecasting monthly discharge time series. *Journal of Hydrology;* 374(3-4): 294-306.