Environment and Ecology Research 10(2): 218-224, 2022 http://www.hrpub.org
DOI: 10.13189/eer.2022.100211

Performance of Water Level Forecasting Based on Chaos Approach Using Data Splitting

Adib Mashuri[1], Nur Hamiza Adenan[2,*], Nor Suriya Abd Karim[2], Mohd Shahriman Adenan[3], Nurulhuda Che Abd Rani[4]

[1]Department of General Studies, Batu Lanchang Vocational College, 11600 Jelutong, Pulau Pinang, Malaysia
[2]Department of Mathematics, Faculty Science and Mathematics, Universiti Pendidikan Sultan Idris, 35900 Tanjong Malim, Perak, Malaysia
[3]School of Mechanical Engineering, College of Engineering, Universiti Teknologi MARA, 40450 Shah Alam, Selangor, Malaysia
[4]Department of Mathematics, Science and Computer, Politeknik Sultan Abdul Halim Mu'adzam Shah, 06900 Jitra, Kedah, Malaysia

Received November 4, 2021; Revised February 12, 2022; Accepted March 7, 2022

Cite This Paper in the following Citation Styles

(a): *[1] Adib Mashuri, Nur Hamiza Adenan, Nor Suriya Abd Karim, Mohd Shahriman Adenan, Nurulhuda Che Abd Rani , "Performance of Water Level Forecasting Based on Chaos Approach Using Data Splitting," Environment and Ecology Research, Vol. 10, No. 2, pp. 218 - 224, 2022. DOI: 10.13189/eer.2022.100211.*

(b): *Adib Mashuri, Nur Hamiza Adenan, Nor Suriya Abd Karim, Mohd Shahriman Adenan, Nurulhuda Che Abd Rani (2022). Performance of Water Level Forecasting Based on Chaos Approach Using Data Splitting. Environment and Ecology Research, 10(2), 218 - 224. DOI: 10.13189/eer.2022.100211.*

Abstract Forecasting accuracy should be prioritised in flood plain areas. This research focuses on data split of water level time series datasets in producing excellent forecasts as measured by coefficient correlation (*CC*). The datasets involved 6000 hours chosen from a recent research location at Sungai Dungun water level, which the data was split into different ratio datasets (50:50, 60:40, 70:30, 80:20, 90:10). A recent study has proved that the chaotic dynamic existed in time series data when running the data using the Cao method. The dataset used was divided into training and testing data to evaluate the performance based on the local linear approximation method. Those sets of data required a combination of parameters for prediction. In this study, the data split of water level time series data gave impacts to the combination of parameters for prediction. The result obtained was in the range of strong forecast using chaos approach with over 95% accuracy in every dataset. In addition, the dataset with a 50:50 ratio showed the highest *CC* obtained, and its values decreased in ascending order of 60:40, 70:30, 80:20, and 90:10. It showed that the splitting data of training and testing had an impact on prediction results. The higher number of training data ran, the lower number of *CC* was obtained. However, the chaos method still gives excellent prediction results, even when forecasting using different ratios of data set.

Keywords Data Splitting, Different Ratio, Water Level, Chaos Approach, Prediction

1. Introduction

The water level in the floodplain area is unpredictable and uncertain due to ongoing phenomena [1, 2]. Technically, heavy rain may cause the river to overflow since the water exceeds the river bank, leading to flood occurrence [3]. Furthermore, floods can lead to human and natural damage that causes a significant impact on the land structure, livestock, and residential area [4]. Hence, accurate flood forecasting is needed for flood control as a flood early warning signal.

There are many methods to get accurate flood forecasting. For instance, Machine Learning Method [5], Artificial Neural Network [6], autoregressive integrated moving average or ARIMA method [7] and chaos approach [8] are the methods that have been implemented currently to get accurate flood forecasting. These methods give excellent forecasting results. However, all methods except the chaos method need more than two variables to obtain accurate results. At the same time, there is only one variable needed to develop excellent flood forecasting by

using the chaos approach [9].

The chaos approach is a fundamental discovery in scientific research which has been used in many areas such as financial [10], hydrology [2], and meteorology [11]. In the hydrology area, the chaos approach has been widely implemented in forecasting sea level [12], water level [13], river flow [14] and sediment transport [15].

Table 1. Data splitting in recent studies

Researcher	Research field	Data Size	Finding
Nguyen et al. [17]	Pedology	> 500	70:30 data split presented the best performance in predicting the shear strength of soil
Rácz et al. [19]	Machine Learning	> 100	70:30 and 80:20 data split gave a better performance for larger datasets
Joseph and Vakayil [18]	Programming Algorithm	> 500	It has proposed a new "SPlit" method to split datasets optimally.

Data splitting refers to dividing one-time series data into two parts [16]. The purpose of data splitting is for cross-validation in prediction. Hence, the first part of the time series data is used to develop the prediction model, and the other part evaluates the prediction performance [16]. Data splitting can also be defined as a sample of research data divided into two datasets: training set for model training and testing set for model validation [17]. Table 1 shows all recent studies using the global prediction method that applied data splitting. The research found different results on predicting time series data using different prediction methods. Most of the research used data size of more than 100 series. Some research concluded that the best datasets used in predicting time series data were above 70% of training data and 30% of testing data. Joseph and Vakayil [18] also proposed a new method of optimal data split that gave the best prediction after finding that every time series data produced different optimal data. However, it can be concluded that using 70:30 (training: testing) data gave excellent data forecasting results. The idea of data split can be used in any time-series data and any research field. Data split usage influenced this research to investigate whether data splitting in the chaos approach gives the same result as recent research. Hence, investigating the idea of the chaos approach is the main aim of the research.

Sungai Dungun has been selected as the location of recent research on water level prediction using a different number of time series data based on the chaos approach. After the research was successfully done, the same research location was used to investigate the different impacts of the data split for water level forecasting.

Therefore, this study continues the previous study by Mashuri et al. [13], where it focused on different amounts of data. Meanwhile, this study was quite different, focusing on the same amount but the different ratio of data used. Therefore, the same amount of time series data was considered in this study by applying data splitting to evaluate the performance of prediction based on the chaos approach.

2. Data

Water level time series data is technically measured using the unit meter (m). This study used hourly time series data because it is suitable for forecasting floodplain areas [20]. Hence, this study was conducted at a station in Sungai Dungun, which is located in a floodplain area, whereby the flood often hits this station. Data used were gathered from July 2009 to March 2010, which consisted of 6000 data since this area was affected by flood at this range of time [21]. The data used is around 2009 to 2010 because this research is a continuation of reference Mashuri et al. [13]. So the data used is still the same as the data used in Mashuri et al. [13].

By referring to Fig. 1, a total of 6000 hourly time series data were used in this study. All the data were split into a specific ratio of training and testing data set in Table 2. The data were grouped into 5; SD50, SD60, SD70, SD80 and SD90. All data sets had a different ratio of training and testing data. For SD50, the ratio for training and testing was 50:50, and SD60 was 60:40. Meanwhile, SD70 consisted of 70:30, SD80 was 80:20, and SD90 was 90:10. In conclusion, the limitation of this study is only focusing on data splitting used before forecasting can be done to investigate the significant data ratio that can be used to have a better prediction.

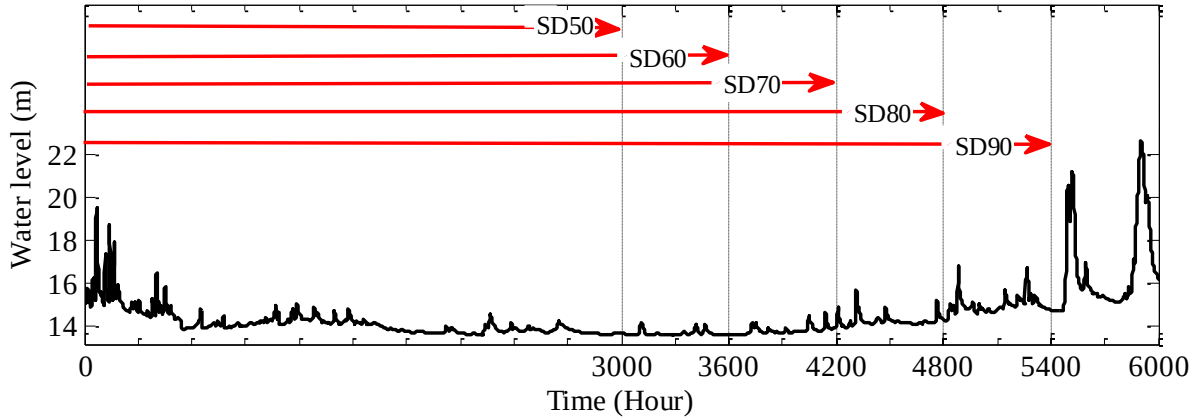


Figure 1. Hourly water level time series data at station Kampung Sungai Station in Sungai Dungun

Table 2. Data allocation for training and testing data set

No.	Data Set	Percentage (%)		Amount (hours)	
		Training	Testing	Training	Testing
1	SD50	50	50	3000	3000
2	SD60	60	40	3600	2400
3	SD70	70	30	4200	1800
4	SD80	80	20	4800	1200
5	SD90	90	10	5400	0600

3. Methodology

The X time series is recorded hourly as follows:

$$X = \{x_1, x_2, ..., x_{N-1,} x_N\} \quad (1)$$

The X time series refers to the training set of data while N is the total of data involved. An example for SD6000, the total data involved are $N = 6000$ and the x_1 value refers to the first hour. According to Takens [22], the phase space reconstruction can be developed. The phase space involves a single variable which refers to the training set that needs to be reconstructed into multi-dimensional phase space Y_i^m as follows:

$$Y_m = (x_i, x_{i+\tau}, x_{i+2\tau}, ..., x_{i+(m-1)\tau}) \quad (2)$$

where τ indicates time delay, m is the embedding dimension and $i = 1, 2,\ldots, N-(m-1)\tau$. In order to obtain the value of τ the Average Mutual Information (AMI) is used as follows:

$$I(T) = \frac{1}{N}\sum_{i=1}^{N} p(x_i, x_{i+T}) \log_2 \left[\frac{p(x_i, x_{i+T})}{p(x_i)\, p(x_{i+T})}\right] \quad (3)$$

where $p(x_i)$ and $p(x_{i+T})$ are the marginal probability of x_i and x_{i+T}. Meanwhile, $p(x_i, x_{i+T})$ is the joint probability of $p(x_i)$ and $p(x_{i+T})$ The first minimum value of $I(T)$ is considered as the value of τ.

As to gain the value of m, the Cao method was used in this study. The Cao method does not only provide the value of m, but this method can help to determine the chaotic dynamics on the training set of data [23]. Furthermore, this method does not depend on the amount of data and this suits the aim of research that involves different amounts of time series data to see the impact of parameters values and prediction accuracy. As such, Cao method is more relevant to be applied compared to the several methods used in order to determine the chaotic dynamics of a specific time series data as utilised in Lyapunov exponent method [24], phase space plot [25] and correlation dimension [26]. The Cao method involves two parameters which are $E1(m)$ and $E2(m)$. The parameter $E1(m)$ can be calculated by:

$$E1(m) = \frac{E1(m+1)}{E(m)}, \text{ and} \quad (4)$$

$$E(m) = \frac{1}{N-m\tau}\sum_{i=1}^{N-m\tau} \frac{\left\|Y_i^{m+1} - Y_n^{m+1}\right\|}{\left\|Y_i^m - Y_n^m\right\|} \quad (5)$$

where $\|\bullet\|$ refers to the Euclidean distance and Y_i^m refers to the neighbouring value for Y_i^m. If $E1(m)$ is saturated when the value m is larger than m_0, then $m_0 + 1$ is the optimum dimension value [27]. Besides identifying the value of m, Cao method is also used to determine the dynamic of this system towards chaotic dynamics or random time series data. If the value $E1(m)$ continues with

the increase of *m*, hence the time series is random. Cao [23] also introduces $E2(m)$ as follows:

$$E2(m) = \frac{E^*(m+1)}{E^*(m)}, \text{ and} \quad (6)$$

$$E^*(m) = \frac{1}{N-m\tau}\sum_{i=1}^{N-m\tau}\left|x_{i+m\tau} - x_{i+m\tau}^{NN}\right|. \quad (7)$$

If the chaotic dynamic exists in the training set of data, the value $E2(m)$ will not be fixed to 1 for any *m* or at least one *m*. Meanwhile, prediction based on chaos approach can be calculated with local linear approximation method:

$$Y_{i+1}^m = \alpha Y_i^m + \beta \quad (8)$$

where Y_{i+1}^m is the one step ahead phase space and Y_i^m is the last phase space. The constants α and β depend on the nearest neighbour, k. In this research, the number of $k = 2m,$, where m refers to the embedding dimension.

4. Result and Discussion

Based on the latest research of Sungai Dungun time series data by Mashuri et al. [13], the chaotic dynamic existed when running the data using the Cao method. Table 3 shows the outcomes of time delay, τ using the AMI method and embedding dimension, m using the Cao method. According to the respected data set, the resulting combination of τ and m (τ, m) is displayed. The results found that using different ratios of time series data had an impact on the value of τ. It can be seen that the increase of training data resulted in a higher value. For the value of *m*, this study found that *m* = 6 for all data sets except SD50, which was 5.

From Table 3, the slight difference can be seen in the determination of the value of τ and the value of *m* where the values τ obtained from *I*(*T*) and *E*2(*m*). In addition, the difference in ratio of datasets used also affects the parameter's determination. The time series datasets proved by the presence of chaotic behaviour as in Mashuri et al. [13] by using parameter values τ obtained.

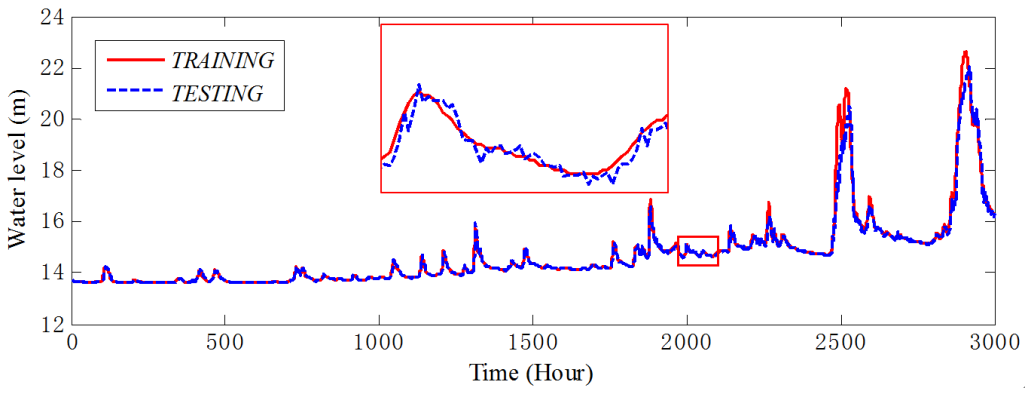
The forecasting of water levels at Sungai Dungun was conducted using the local linear approximation method. The performance accuracy for the prediction was represented by the value of the correlation coefficient (*CC*). Table 3 also shows the value of *CC* for SD50, SD60, SD70, SD80 and SD90. For SD50, the value of *CC* was the highest among other data sets with 0.9878. The values of *CC* for SD60, SD70, SD80 and SD90 were 0.9788, 0.9743 0. 9676 and 0.9581, respectively. This result showed the same combination of τ and *m* in the prediction result of different *CC* values according to different ratios of the data set.

As you can see, Fig.1 shows training data within 0 to 5400 hours were below 15m and some testing data were above 15m. Fig. 2 shows that the prediction accuracy was still above 95% even the trend of training data used slightly different from the trend of testing data used. This is the uniqueness of chaos method in predicting uncertain time series data [28].

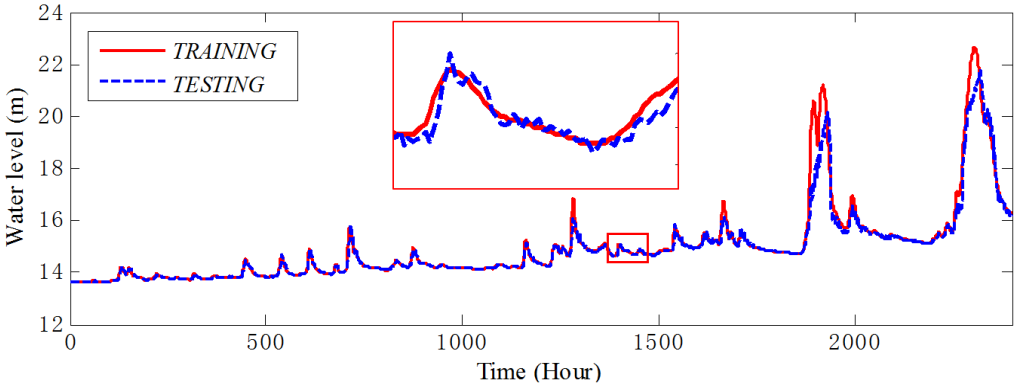
The values of *CC* portray a substantial prediction value for the value of 0.9 until 0.99, while the perfect prediction for the *CC* value is 1.0 [29, 30]. This study found that the prediction using all data sets was in the range of significant prediction. However, the prediction using the ratio of training and testing with the ratio 50:50 had the best performance since the value of *CC* for this data set was the highest. Refer to Fig. 2, testing time-series datasets, and prediction results were plotted against time (hour). The prediction result gave excellent accuracy from the beginning to the end by comparing it to testing data where all datasets showed over 95% accuracy.

Table 3. The value of τ, *m* and performance accuracy for corresponding data set

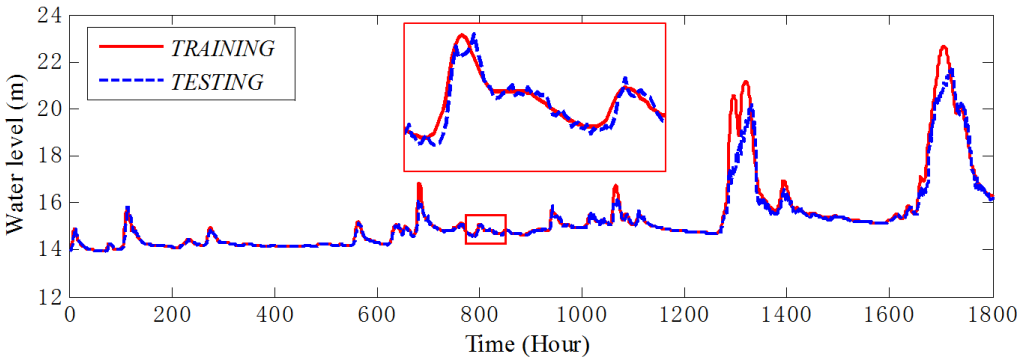
Data Set	SD50	SD60	SD70	SD80	SD90
(τ, *m*)	(11,5)	(15,6)	(15,6)	(15,6)	(32,6)
Correlation coefficient, *CC*	0.9878	0.9788	0.9743	0.9676	0.9581



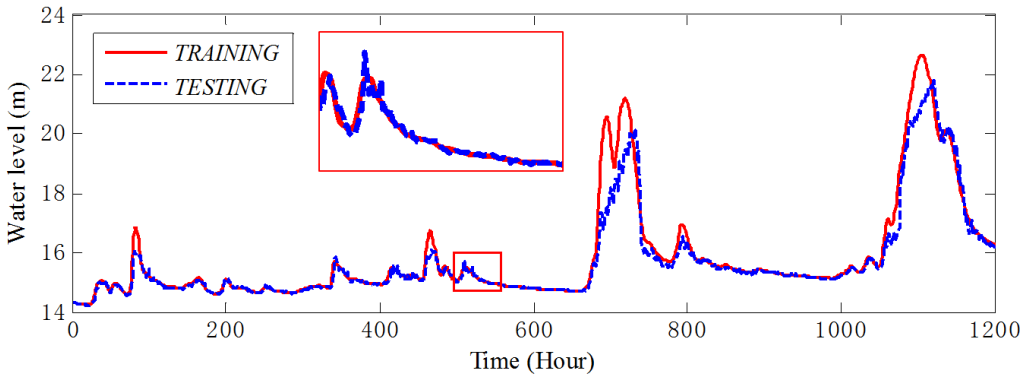
A



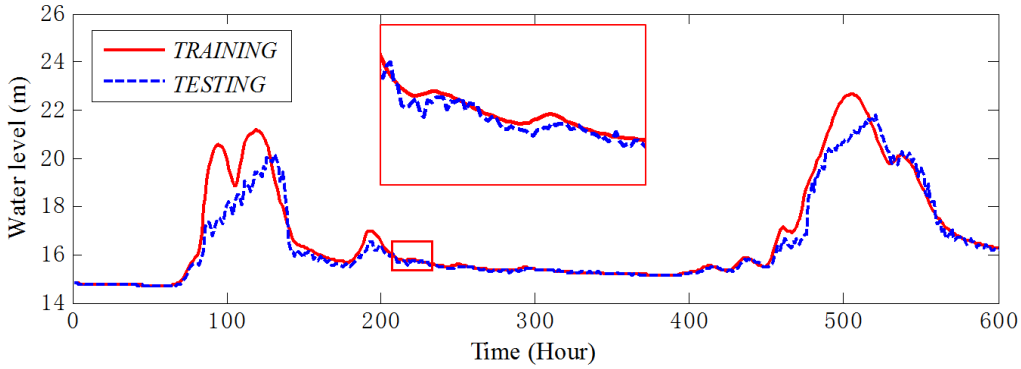
B



C



D



E

Figure 2. Hourly water level time series data forecasting with different data sets (a)SD50, (b)SD60, (c)SD70, (d)SD80 and (e)SD90

5. Conclusions

This study applied data splitting into hourly water level time series data to forecast the water level at Sungai Dungun using the local linear approximation method. This data splitting was done by separating the data into five data sets with different ratios of training and testing data which included 50:50 (SD50), 60:40 (SD60), 70:30 (SD70), 80:20 (SD80) and 90:10 (SD90). The result shows a chaotic dynamic in water level time series data [13].

The implication of data splitting shows that the ratio of training data affected the value of τ. The increasing number of training data resulted in the value of τ. This analysis also identified that m = 6 for all data sets except SD50, which was 5. Next, SD60, SD70, SD80, and SD90 had the *CC* values of 0.9788, 0.9743, 0.9676, and 0.9581, respectively. This result shows that the same combination of τ and m resulted in varied *CC* values depending on the ratio of the training data set. The values of *CC* portray a strong forecast accuracy for each ratio of data splitting. However, the forecast that applied the 50:50 ratio outperformed the others since the value of *CC* for this data set was the greatest.

We can conclude that the chaos method gives excellent prediction results, which gave more than 95% accuracy, even when forecasting using different ratios of data set. Therefore, this result is essential to chaos since no research on this method applied to this objective. So, research implementing a chaos approach may have a significant reference about the number of data needed to have significant prediction accuracy. In a nutshell, this study provides essential information in Sungai Dungun hourly time-series data, giving different results based on data splitting. This indicates that differences in splitting testing and training data will produce the different results of future prediction.

Acknowledgements

The authors thankfully acknowledged the financial support provided by the Ministry of Education Malaysia (2019-0009-102-02: FRGS/1/20/2018/STG06/UPSI/02/3) as well as the Department of Irrigation and Drainage Malaysia for providing the hydrological data.

REFERENCES

[1] A. Mashuri, H. Adenan, and N. Z. A. Hamid, "Determining the Chaotic Dynamics of Hydrological Data in Flood-Prone Area," Civil Engineering and Architecture, vol. 7, no. 6A, pp. 71–76, 2019, doi: 10.13189/cea.2019.071408.

[2] A. Mashuri, N. H. Adenan, N. S. A. Karim, R. A. Tarmizi, N. Z. A. Hamid, and M. S. Adenan, "Improvement of Local Mean Approximation Method for Prediction of Water Level Time Series Data in Flooded Area," *Math. Sci. Informatics J.*, vol. 2, no. 1, pp. 40–48, 2021, Accessed: Aug. 09, 2021. [Online]. Available: http://www.mijuitmjournal.com.

[3] A. W. H. Soeharn, M. Farid, D. S. Abidah, T. R. Maitsa, Setianingsih, and N. Majidah, "Effect of extreme rain and land covering change in Jatihandap on 20 March 2018 flash flood," *MATEC Web Conf.*, vol. 270, p. 04003, Feb. 2019, doi: 10.1051/matecconf/201927004003.

[4] A. J. Echendu, "The impact of flooding on Nigeria's sustainable development goals (SDGs)," *https://doi.org/10.1080/20964129.2020.1791735*, vol. 6, no. 1, Dec. 2020, doi: 10.1080/20964129.2020.1791735.

[5] S. Dazzi, R. Vacondio, and P. Mignosa, "Flood Stage Forecasting Using Machine-Learning Methods: A Case Study on the Parma River (Italy)," *Water 2021, Vol. 13, Page 1612*, vol. 13, no. 12, p. 1612, Jun. 2021, doi: 10.3390/W13121612.

[6] Q. Lin, J. Leandro, W. Wu, P. Bhola, and M. Disse, "Prediction of Maximum Flood Inundation Extents With Resilient Backpropagation Neural Network: Case Study of Kulmbach," *Front. Earth Sci.*, vol. 0, p. 332, Aug. 2020, doi: 10.3389/FEART.2020.00332.

[7] W. M. Wong, S. K. Subramaniam, F. S. Feroz, I. D. Subramaniam, and L. A. F. Rose, "Flood prediction using ARIMA model in Sungai Melaka, Malaysia," *Int. J. Adv. Trends Comput. Sci. Eng.*, vol. 9, no. 4, pp. 5287–5295, Jul. 2020, doi: 10.30534/IJATCSE/2020/160942020.

[8] N. H. Adenan *et al.*, "Traffic Flow Prediction in Urban Area Using Inverse Approach of Chaos Theory," Civil Engineering and Architecture, vol. 9, no. 4, pp. 1277–1282, 2021, doi: 10.13189/cea.2021.090429.

[9] N. H. Adenan and M. S. Md Noorani, "Multiple Time-Scales Nonlinear Prediction of River Flow Using Chaos Approach," *J. Teknol.*, vol. 78, no. 7, Jun. 2016, doi: 10.11113/jt.v78.3561.

[10] H. Litimi, A. Bensa ïda, L. Belkacem, and O. Abdallah, "Chaotic behavior in financial market volatility," *J. Risk*, vol. 21, no. 3, pp. 27–54, Feb. 2019, doi: 10.21314/JOR.2018.400.

[11] B.-W. Shen *et al.*, "Is Weather Chaotic?: Coexistence of Chaos and Order within a Generalized Lorenz Model," *Bull. Am. Meteorol. Soc.*, vol. 102, no. 1, pp. E148–E158, Jan. 2021, doi: 10.1175/BAMS-D-19-0165.1.

[12] N. M. Ali and N. Z. A. Hamid, "Environmental Modelling through Chaotic Approach for Malaysian West Coast Sea Level.," *J. Phys. Conf. Ser.*, vol. 1529, no. 3, p. 032092, Apr. 2020, doi: 10.1088/1742-6596/1529/3/032092.

[13] A. Mashuri, N. H. Adenan, N. S. A. Karim, and N. Z. A. Hamid, "Water Level Prediction Using Different Numbers of Time Series Data Based on Chaos Approach," Civil Engineering and Architecture, vol. 9, no. 2, pp. 493–499, 2021, doi: 10.13189/cea.2021.090221.

[14] N. H. Adenan and M. S. M. Noorani, "Peramalan Data Siri Masa Aliran Sungai di Dataran Banjir dengan Menggunakan Pendekatan Kalut (Predicting Time Series Data at Floodplain Area using Chaos Approach)," *Sains Malaysiana*, vol. 44, no. 3, pp. 463–471, 2015, Accessed: Oct. 19, 2018. [Online]. Available: http://journalarticle.ukm.my/8490/1/19_Nur_Hamiza.pdf.

[15] P. Yousefi, G. Courtice, G. Naser, and H. Mohammadi, "Nonlinear Dynamic Modeling of Urban Water Consumption Using Chaotic Approach (Case Study: City of Kelowna)," *Water 2020, Vol. 12, Page 753*, vol. 12, no. 3, p. 753, Mar. 2020, doi: 10.3390/W12030753.

[16] R. R. Picard and K. N. Berk, "Data Splitting," *Am. Stat.*, vol. 44, no. 2, p. 140, May 1990, doi: 10.2307/2684155.

[17] Q. H. Nguyen *et al.*, "Influence of Data Splitting on Performance of Machine Learning Models in Prediction of Shear Strength of Soil," 2021, doi: 10.1155/2021/4832864.

[18] V. R. Joseph and A. Vakayil, "SPlit: An Optimal Method for Data Splitting," *Technometrics*, Dec. 2020, doi: 10.1080/00401706.2021.1921037.

[19] A. Rácz, D. Bajusz, and K. Héberger, "Effect of dataset size and train/test split ratios in qsar/qspr multiclass classification," *Molecules*, vol. 26, no. 4, Feb. 2021, doi: 10.3390/MOLECULES26041111.

[20] M. L. Sapini, N. S. Adam, N. Ibrahim, N. Rosmen, and N. M. Yusof, "The presence of chaos in rainfall by using 0-1 test and correlation dimension," in *AIP Conference Proceedings*, Nov. 2017, vol. 1905, no. 1, p. 050040, doi: 10.1063/1.5012259.

[21] JPS Negeri Terengganu, "Laporan Banjir 2010 - 2011," *JPS Negeri Terengganu*, 2020.

[22] F. Takens, *Detecting strange attractors in turbulence*. Dynamical Systems and Turbulence, Warwick, 1981.

[23] L. Cao, "Practical method for determining the minimum embedding dimension of a scalar time series," *Phys. D Nonlinear Phenom.*, vol. 110, no. 1–2, pp. 43–50, Dec. 1997, doi: 10.1016/S0167-2789(97)00118-8.

[24] D. T. Mihailović *et al.*, "Analysis of Daily Streamflow Complexity by Kolmogorov Measures and Lyapunov Exponent," Sep. 2018, Accessed: Jan. 06, 2019. [Online]. Available: http://arxiv.org/abs/1809.08633.

[25] N. Z. A. Hamid, "Application of Chaotic Approach in Forecasting Highland's Temperature Time Series," in *IOP Conference Series: Earth and Environmental Science*, Aug. 2018, vol. 169, no. 1, p. 012107, doi: 10.1088/1755-1315/169/1/012107.

[26] N. H. Adenan, "Analisis Dan Peramalan Data Siri Masa Aliran Sungai Dengan Menggunakan Pendekatan Kalut.," Universiti Kebangsaan Malaysia (UKM), 2015.

[27] N. H. Zakaria, N. H. Adenan, N. Suriya, A. Karim, and A. Mashuri, "Peramalan Siri Masa Aras Sungai Empangan di Selangor Menggunakan Pendekatan Kalut dan Kaedah Penghampiran Linear Setempat Prediction of Water Level Time Series Data for Dam at Selangor using Chaotic Approach and Local Linear Approximation Method berlaku s," vol. 9, pp. 10–17, 2021.

[28] A. Mashuri, N. H. Adenan, and N. S. A. Karim, "Chaotic Identification of Hourly and Daily Water Level Time Series Data in Different Areas of Elevation at Pahang River," *Turkish J. Comput. Math. Educ.* , vol. 12, no. 3, pp. 2427–2435, 2021, Accessed: Feb. 07, 2022. [Online]. Available: https://www.researchgate.net/publication/357933611_Chaotic_Identification_of_Hourly_and_Daily_Water_Level_Time_Series_Data_in_Different_Areas_of_Elevation_at_Pahang_River.

[29] H. Akoglu, "User's guide to correlation coefficients," *Turkish J. Emerg. Med.*, vol. 18, no. 3, p. 91, Sep. 2018, doi: 10.1016/J.TJEM.2018.08.001.

[30] P. Schober, C. Boer, and L. A. Schwarte, "Correlation Coefficients," *Anesth. Analg.*, vol. 126, no. 5, pp. 1763–1768, May 2018, doi: 10.1213/ANE.0000000000002864.

[31] P. Schober and L. A. Schwarte, "Correlation coefficients: Appropriate use and interpretation," *Anesth. Analg.*, vol. 126, no. 5, pp. 1763–1768, May 2018, doi: 10.1213/ANE.0000000000002864.