Civil Engineering and Architecture 9(2): 493-499, 2021
DOI: 10.13189/cea.2021.090221 http://www.hrpub.org

Water Level Prediction Using Different Numbers of Time Series Data Based on Chaos Approach

Adib Mashuri[1], Nur Hamiza Adenan[2,*], Nor Suriya Abd Karim[2], Nor Zila Abd Hamid[2]

[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

Received January 7, 2021; Revised February 26, 2021; Accepted March 12, 2021

Cite This Paper in the following Citation Styles

(a): *[1] Adib Mashuri, Nur Hamiza Adenan, Nor Suriya Abd Karim, Nor Zila Abd 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.*

(b): *Adib Mashuri, Nur Hamiza Adenan, Nor Suriya Abd Karim, Nor Zila Abd Hamid (2021). Water Level Prediction Using Different Numbers of Time Series Data Based on Chaos Approach. Civil Engineering and Architecture, 9(2), 493-499. DOI: 10.13189/cea.2021.090221.*

Abstract The prediction of water level in floodplain area is important for early signals and flood control. A total of 6350 hourly water level time series data located at Sungai Dungun were used in this study. The data were divided into training set and testing set. The training set consisted of the first 6000 data which were used to predict the last 350 data. A total of six set data consisting of different amount of training set of data were involved in this study. Consequently, it was used to determine the influence of different amount of data on predicting accuracy by using chaos approach. Those sets of data required a combination of parameters for prediction. In this study, the different amount of data had impacts on the combination of parameter for prediction. In addition, the correlation coefficient showed different values for all sets of data and excellent prediction when they were all used in testing the data. Hence, the different total amount of data will give impact on different combination of parameters and prediction accuracy for water level prediction based on chaos approach in floodplain area.

Keywords Amount of Data, Prediction, Chaos Approach, Water Level

1. Introduction

Cao, Tao, Dong and Li [5] asserted that floods occur when excessive water level rises in river areas, whether in natural or man-made conditions. From a scientific vocabulary point of view, floods are caused by the existence of excessive heavy rain that cannot be supported by the river basin and thus the water overflows to the riverbanks or floodplains [18]. Flood disaster can cause damage to people and nature where it can affect the land structure, agriculture, livestock as well as residential areas [7]. Therefore, prediction of water level in floodplain area is important for early signals and flood control.

The dynamics of a time series data can be divided into two parts; deterministic and random. In 1963, Lorenz [15] discovered the dynamic chaotic where the knowledge was used for research in the field of science and engineering in a thorough matter while the term chaos was first introduced by Li and Yorke [14]. According to Abarbanel [1], chaotic dynamic is between the deterministic and random dynamic. Chaotic time series can be used in prediction only in a short term due to the sensitive dependence on initial conditions [19].

This chaos approach is an important discovery for predicting phenomena in scientific research. The application of chaos approach is widely used in many types of time series data such as river flow [3], ozone [9] and sea level [4]. Nowadays, the research in the application of this method on water level time series data is growing and being conducted in several countries such as in China [10], Iran [21] and Malaysia [16]. In addition, a lot of research also emphasised on time scale of water level data such as hourly scale [13], daily [12] and weekly

[12].

Grouping a number of data will affect the value of the parameter used in prediction and hence, choosing the right amount of data is important. The research in predicting water level is extended to research in prediction using different number of time series data such that the influence of data with several numbers in giving the accuracy in prediction performance for water level. Thus, this study is conducted to perform the prediction of water level with several number of data sets using chaos approach in floodplain area where flood often hits.

2. Data

The unit meter (m) is used to measure the time series data for the river water level. Data used in this study were based on hourly time series. Hourly data are suitable for applying the study of hydrology, in particular flood prediction [13]. This study was conducted at Kampung Surau Station in Sungai Dungun, Terengganu. Data used in this study started from July 2009 to March 2010 since this area was affected by flood at this range of time [11]. Referring to Figure 1, a total of 6350 hourly time series data were used in this study and they were divided into training data set and testing data set. The training data set consisted of the first 6000 data while the rest of the data were used as testing data set that were as much as 350 data.

In conducting this study, as much as 6000 data in training data set were divided into several parts as this study focused on testing the prediction performance using 1000 data set (coded as SD1000), 2000 data set (SD20000), 3000 data set (SD3000), 4000 data set (SD4000), 5000 data set (SD5000) and 6000 data set (SD6000). Since prediction was tested from data 6001 until 6350, the data used in constructing the prediction were counted backwards, in which 1000 data set (SD1000) was taken from 5001 until 6000, 2000 data set (SD2000) was taken from 4001 until 6000 and so on. The amount of data, their ranges and percentage used out of the total training data set are presented in Table 1.

Table 1. Data description

Research Code	Data Amount (Hourly)	Data taken from graph (hour)	Percentage of data used
SD1000	1000	5001 – 6000	17%
SD2000	2000	4001 – 6000	33%
SD3000	3000	3001 – 6000	50%
SD4000	4000	2001 – 6000	67%
SD5000	5000	1001 – 6000	83%
SD6000	6000	0001 – 6000	100%

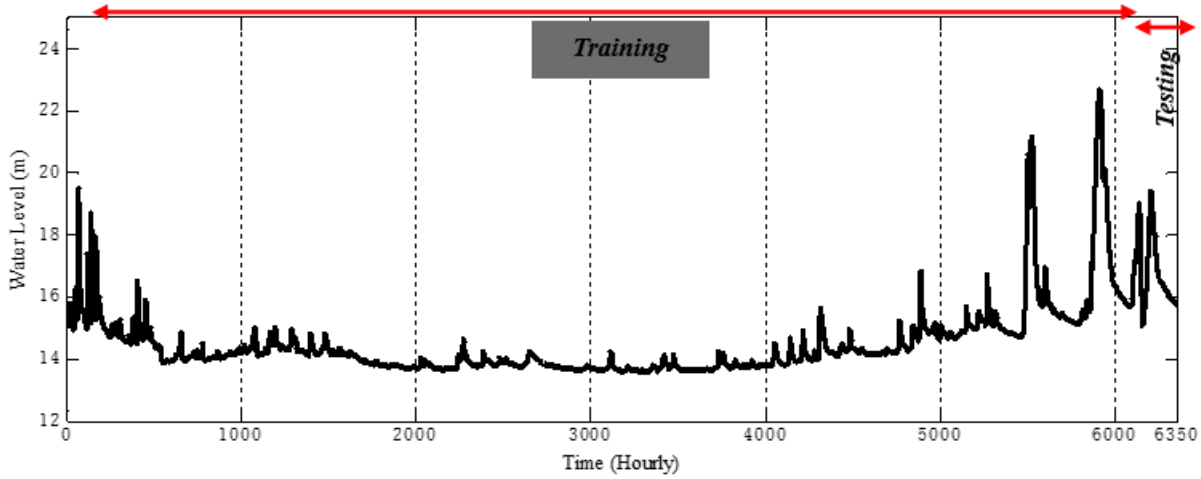


Figure 1. Training and testing of hourly water level time series data at station Kampung Sungai Station in Sungai Dungun

3. Methodology

The *X* time series is recorded by hourly as follows:

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

The *X* time series is referred to as the training set of data and N is referred to as total of data involved such that x_1 is value of data at the first hour. An example for SD6000, the total data involved are $N = 6000$. According to Takens [20], the phase space reconstruction can be developed. The phase space involves a single variable that is referred to as the training set that needs to be reconstructed to 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 such that i = 1, 2,…, $N-(m-1)\tau$. In order to obtain the value of τ, the Average Mutual Information (AMI) was used as follow:

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

where $p(x_i)$ and $p(x_{i+T})$ is 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.

In order to determine the value of *m*, Cao method was used in this study. This method does not only can provide the value of *m* but can also help to determine the chaotic dynamics of data [6]. Furthermore, this method does not depend on specific number of data and therefore it suits for this study which involves different amount of time series data to see the impact of parameters values to the prediction accuracy. Therefore, Cao method is more relevant to be used in determining the chaotic dynamics of a water level time series data compared to the other methods such as Lyapunov exponent method [17], phase space plot [8] and correlation dimension. 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_n^m refers to the neighbouring value for Y_i^m.

If $E1(m)$ is saturated with a value that is larger than m_0, then $m_0 + 1$ is the optimum dimension value [22]. 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 $E1(m)$ does not reach saturated with the increasing of *m*, hence the time series is random. Cao [6] 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 conducted using 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 value of *k* is determined by $k = 2m$, where m is the embedding dimension.

4. Result and Discussion

Table 2 shows the value of τ that was obtained using AMI method with different amount of time series data. The different values of τ obtained indicated the different amount of time series data used to influence the value of τ. The saturated value for $E1(m)$ was chosen between 0.9 until 1.0 [9]. Table 3 shows the value of $E1(m)$ for each set of data and the bold numbers were the number where $E1(m)$ started saturated for each data set. For SD1000, $E1(m)$ values started to saturate at $m = 4$ and therefore the embedding dimension, *m* for SD1000 was 4. Thus, the values for *m* were 6, 6, 7, 6, 5 for SD2000, SD3000, SD4000, SD5000 and SD6000, respectively. Hence, the different amount of data set induced different values of *m*.

Note that the value of $E1(m)$ values saturated with the increase of *m* for all sets of data. Hence, the time series was chaotic. Moreover, the existence of $E2(m) \neq 1$ for each value of *m* assured the presence of chaotic dynamics for each set of data as referred to Table 4. Therefore, the prediction using chaos approach can be conducted in all solved hourly time series data.

Table 2. Value of time delay based, τ based on AMI method

Research code	SD1000	SD2000	SD3000	SD4000	SD5000	SD6000
Time delay, τ	4	11	11	24	22	18

Table 3. The selection of m value using $E1(m)$ for every solved hourly time series data

m	1	2	3	4	5	6	7	8	9	10
SD1000	0.4066	0.8203	0.8855	0.9277	0.9616	0.9627	0.9731	0.9733	0.9662	0.9773
SD2000	0.1942	0.6100	0.7333	0.8839	0.8856	0.9315	0.9606	0.9632	0.9669	0.9747
SD3000	0.1610	0.5910	0.7739	0.8785	0.8842	0.9143	0.9458	0.9671	0.9594	0.9743
SD4000	0.1089	0.5773	0.6949	0.8169	0.8751	0.8810	0.9128	0.9113	0.9425	0.9593
SD5000	1.0928	0.5039	0.7077	0.8491	0.8920	0.9038	0.9036	0.9335	0.9412	0.9603
SD6000	0.0952	0.5097	0.7151	0.809	0.9025	0.9231	0.9321	0.9409	0.9497	0.9666

Table 4. The analysis of the chaotic dynamics using $E2(m)$ for every solved hourly time series data

m	1	2	3	4	5	6	7	8	9	10
SD1000	1.0930	1.1170	1.0900	1.0650	1.0560	1.0230	1.0140	0.9920	0.9800	0.9930
SD2000	0.9730	1.0190	0.9870	1.0210	1.0150	1.0270	1.0250	1.0050	0.9950	0.9870
SD3000	0.8560	1.0050	0.9900	1.0230	0.9900	1.0260	1.0130	1.0170	0.9940	0.9920
SD4000	0.6400	0.9520	1.0230	0.9790	1.0740	1.0180	1.0480	0.9610	1.0140	1.0030
SD5000	0.6190	0.9540	0.9870	1.0030	1.0640	1.0610	0.9860	1.0220	0.9580	0.9970
SD6000	0.6130	0.9480	0.9720	1.0160	1.0160	1.0130	1.0150	1.0010	0.9770	0.9970

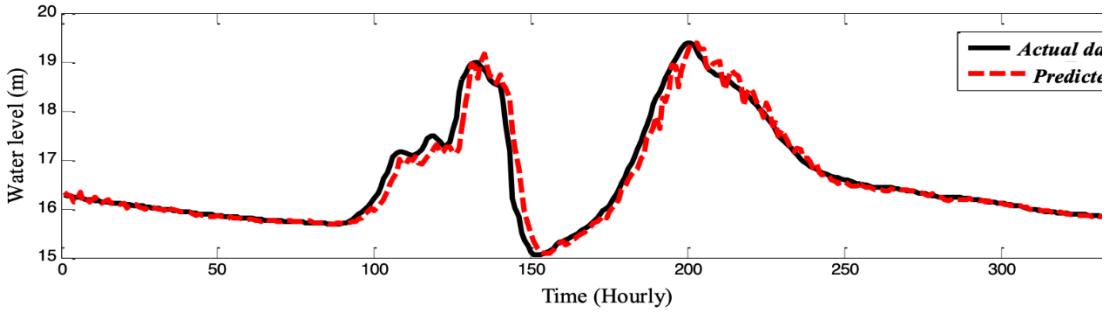
Table 5. Prediction accuracy

Research code	SD1000	SD2000	SD3000	SD4000	SD5000	SD6000
Correlation coefficient, *cc*	0.9763	0.9638	0.9638	0.9479	0.9638	0.9813

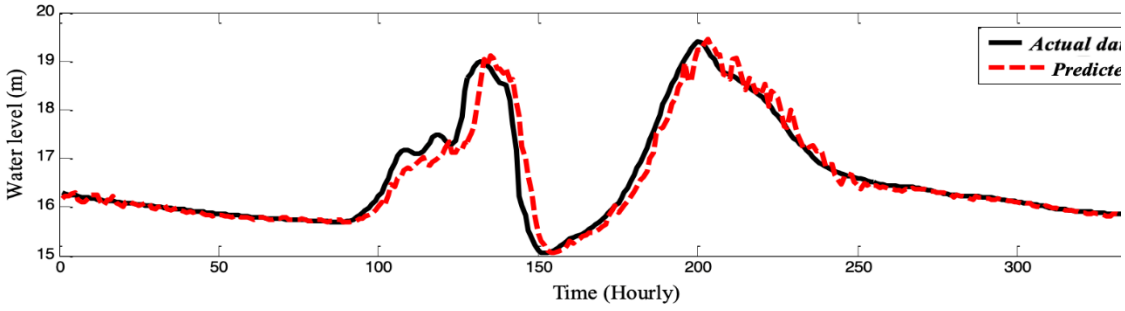
Since chaos exists in the water level time series data in Sungai Dungun, hence prediction using chaos approach can be conducted. Local linear approximation method was used. This method requires the combination values of (τ, m) from the previous calculation of using AMI method in Table 2 and $E1(m)$ in Table 3. As much as 350 data taken were used to test the prediction performed by each set of data as presented in Fig. 2. The prediction performances were compared to the actual data for each SD1000, SD2000, SD3000, SD4000, SD5000 and SD6000.

Table 5 shows the prediction performances for each data set that was represented by the correlation coefficient (CC). It can be seen that at SD1000, the prediction conducted with parameter combination of (4,4) gave correlation coefficient (*CC*) value of 0.9763. For SD2000 and SD3000, the combination of parameters was the same with (11,6) that generated *CC* values of 0.9683. This clearly shows that the same values of parameter τ and m used in prediction may contribute to the same value of *CC*. Meanwhile, the value of *CC* exceeded 0.9400 for SD4000, SD5000 and SD6000 with the parameter combinations of (24,7), (22,6) and (18,5), respectively. This shows that the different amount of data used gives different prediction values. However, the best prediction in this research can be obtained by using 6000 hours data which has the highest number of data set used.

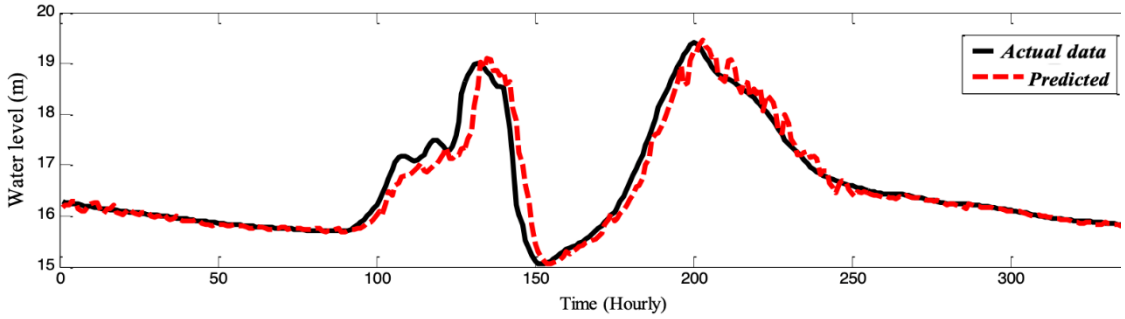
(a)


Water level (m)
Time (Hourly)
Actual da
Predicte

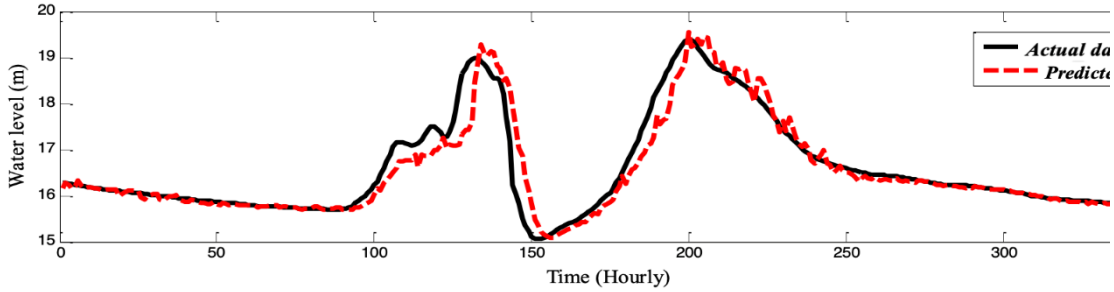
(b)


Water level (m)
Time (Hourly)
Actual da
Predicte

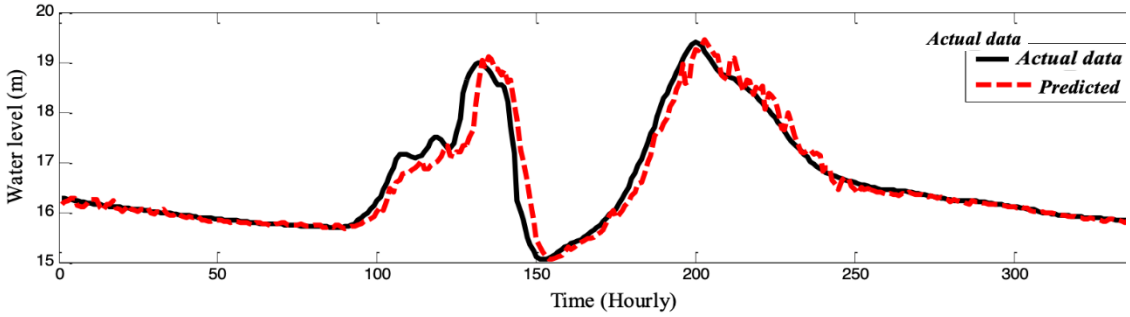
(c)


Water level (m)
Time (Hourly)
Actual data
Predicted

(d)


Water level (m)
Time (Hourly)
Actual da
Predicte

(e)


Water level (m)
Time (Hourly)
Actual data
Actual data
Predicted

(f)

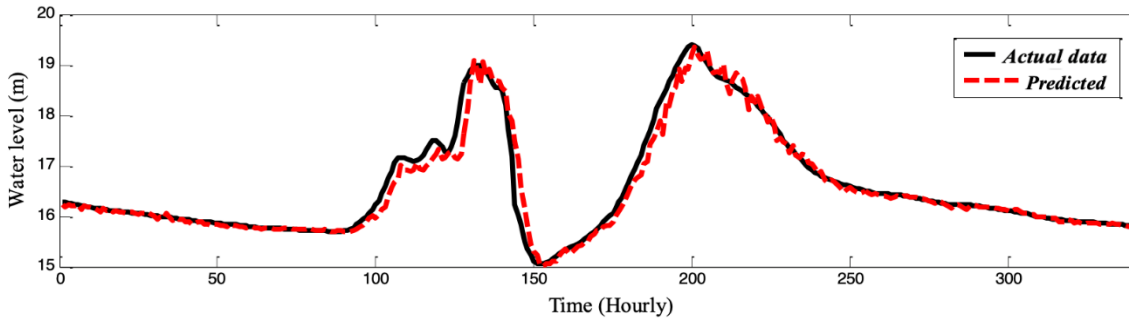


Figure 2. Prediction of hourly water level time series data with different amount of data set (a) SD1000, (b) SD2000, (c)SD3000, (d) SD40000, (e) SD5000 and (f) SD6000

5. Conclusions

This study observes the influence of different amount of data used with the combination of parameter τ and m for prediction purpose using chaos approach. In this study, the different amount of data gives impact on the combination of parameter for prediction. As such, this research suggests using different amount of data set in order to have a good combination for prediction. Furthermore, six sets of different amount time series data that consist of up to 6000 has been used in this study to determine the impact of data amount to prediction accuracy in floodplain area. As a conclusion, different amount of time series data influences the accuracy of the prediction. The purpose amount of data set that gives excellent prediction is when it uses all testing data which is 6000 data in order to predict 350 hours of data ahead. Therefore, a large amount of data needed in order to get an excellent prediction accuracy based on chaos approach prediction.

Acknowledgments

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] Abarbanel, H. D. I. (1996). *Analysis of Observed Chaotic Data*. https://doi.org/10.1007/978-1-4612-0763-4

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

[3] Adenan, N. H., & Noorani, M. S. M. (2016). Multiple Time-Scales Nonlinear Prediction of River Flow Using Chaos Approach. *Jurnal Teknologi*, *78*(7), 1–7. https://doi.org/10.11113/jt.v78.3561

[4] Ali, N. M., Hamid, N. Z. A., & Ali, N. M. (2020). Environmental Modelling through Chaotic Approach for Malaysian West Coast Sea Level. *Journal of Physics: Conference Series*, *1529*(3), 32092. https://doi.org/10.1088/1742-6596/1529/3/032092

[5] Cao, F., Tao, Q., Dong, S., & Li, X. (2020). Influence of rain pattern on flood control in mountain creek areas: a case study of northern Zhejiang. *Applied Water Science*, *10*(10), 224. https://doi.org/10.1007/s13201-020-01308-x

[6] Cao, L. (1997). Practical Method for Determining the Minimum Embedding Dimension of a Scalar time series. *Physica D: Nonlinear Phenomena*, *110*(1–2), 43–50. https://doi.org/10.1016/S0167-2789(97)00118-8

[7] Echendu, A. J. (2020). The impact of flooding on Nigeria's sustainable development goals (SDGs). *Ecosystem Health and Sustainability*, *6*(1). https://doi.org/10.1080/20964129.2020.1791735

[8] Hamid, N. Z. A. (2018). To cite this article: Nor Zila Abd Hamid. *IOP Conf. Ser.: Earth Environ. Sci*, *169*, 12107. https://doi.org/10.1088/1755-1315/169/1/012107

[9] Hamid, N. Z. A., & Noorani, M. S. M. (2017). Aplikasi Model Baharu Penambahbaikan Pendekatan Kalut ke atas Peramalan Siri Masa Kepekatan Ozon. *Sains Malaysiana*, *46*(8), 1333–1339. https://doi.org/10.17576/jsm-2017-4608-20 Aplikasi

[10] Huang, F., Huang, J., Jiang, S.-H., & Zhou, C. (2017). Prediction of Groundwater Levels using Evidence of Chaos and Support Vector Machine. *Journal of Hydroinformatics*, *19*(4), 586–606. https://doi.org/10.2166/hydro.2017.102

[11] JPS Negeri Terengganu. (2020). Laporan Banjir 2010 - 2011. Retrieved December 20, 2020, from JPS Negeri Terengganu website: http://jpsweb.terengganu.gov.my/index.php/ms/laporan-banjir-2010-2011

[12] Khatami, S. (2013). *Nonlinear Chaotic and Trend Analyses of Water Level at Urmia Lake, Iran Does Climate Variability Explain Urmia Lake Depletion*. Lund University.

[13] Khatibi, R., Ghorbani, M. A., Aalami, M. T., Kocak, K., Makarynskyy, O., Makarynska, D., & Aalinezhad, M. (2011). Dynamics of hourly sea level at Hillarys Boat Harbour, Western Australia: a chaos theory perspective. *Ocean Dynamics*, *61*(11), 1797–1807. https://doi.org/10.10

07/s10236-011-0466-8

[14] Li, T.-Y., & Yorke, J. A. (1975). Period Three Implies Chaos. *The American Mathematical Monthly*, *82*(10), 985–992.

[15] Lorenz, E. N. (1963). Deterministic Nonperiodic Flow. *Journal of the Atmospheric Sciences*, *20*(2), 130–141. https://doi.org/10.1175/1520-0469(1963)020<0130:DNF>2.0.CO;2

[16] Mashuri, A., Adenan, N. H., & Hamid, N. Z. A. (2019). Determining the Chaotic Dynamics of Hydrological Data in Flood-prone Area. *Civil Engineering and Architecture*, *7*(6), 71–76. https://doi.org/10.13189/cea.2019.071408

[17] Mihailović, D. T., Nikolić-Đorić, E., Arsenić, I., Malinović-Milićević, S., Singh, V. P., Stošić, T., & Stošić, B. (2018). *Analysis of Daily Streamflow Complexity by Kolmogorov Measures and Lyapunov Exponent*.

[18] Shakti P. C., Hirano, K., & Iizuka, S. (2020). Flood Inundation Mapping of the Hitachi Region in the Kuji River Basin, Japan, During the October 11–13, 2019 Extreme Rain Event. *Journal of Disaster Research*, *15*(6), 712–725. https://doi.org/10.20965/jdr.2020.p0712

[19] Sprott, J. C. (2003). *Chaos and Time-series Analysis*. Oxford University Press.

[20] Takens, F. (1981). *Detecting Strange Attractors in Turbulence*. Dynamical Systems and Turbulence, Warwick.

[21] Vaheddoost, B., Aksoy, H., & Abghari, H. (2016). Prediction of Water Level using Monthly Lagged Data in Lake Urmia, Iran. *Water Resources Management*, *30*(13), 4951–4967. https://doi.org/10.1007/s11269-016-1463-y

[22] Zaim, W. N. A. W. M., Hamid, N. Z. A., & Noorani, M. S. M. (2018). Peramalan Bahan Pencemar Ozon (O3) di Universiti Pendidikan Sultan Idris, Tanjung Malim Perak, Malaysia Mengikut Monsun dengan Menggunakan Pendekatan Kalut. *Sains Malaysiana*, *46*(12), 2523–2528. https://doi.org/10.17576/jsm-2017-4612-30