Sains Malaysiana 51(7)(2022): 2211-2222
http://doi.org/10.17576/jsm-2022-5107-22

Streamflow Data Analysis for Flood Detection using Persistent Homology

(Analisis Data Aliran Sungai bagi Pengesanan Banjir menggunakan Homologi Gigih)

SYED MOHAMAD SADIQ SYED MUSA*, MOHD SALMI MD NOORANI, FATIMAH ABDUL RAZAK, MUNIRA ISMAIL & MOHD ALMIE ALIAS

Department of Mathematical Sciences, Faculty of Science and Technology, Universiti Kebangsaan Malaysia, 43600 UKM Bangi, Selangor Darul Ehsan, Malaysia

Received: 20 May 2021/Accepted: 1 December 2021

ABSTRACT

Flooding is an environmental hazard that occurs almost everywhere around the world. Analysis of streamflow data can give us important climatic information for flooding events. Persistent homology (PH), a new analysis tool in topological data analysis (TDA) offers a new way to look at the information in a data set using qualitative approach. PH uses topology to extract topological features such as connected components and cycles that exist in the data set. In this paper, we present a new approach for streamflow data analysis for flood detection by using PH. An analysis was conducted at Sungai Kelantan, Malaysia. The result shows that PH gives different pattern of topological features for dry and wet periods. In particular, there are more persistent topological features in the form of connected components and cycles in the wet periods compared to the dry periods. We observed that the time series of the distance measure corresponding to the evolution of the components is consistent with the time series of the streamflow data. As a conclusion, this study suggests that the time series of the distance measure corresponding to the evolution of the components can be used for flood detection at Sungai Kelantan, Malaysia.

Keywords: Flood; persistent homology; streamflow; time delay embedding; topological data analysis

ABSTRAK

Banjir merupakan bencana alam yang berlaku hampir di seluruh dunia. Analisis data aliran sungai mampu memberikan maklumat iklim yang penting bagi kejadian banjir. Homologi gigih (HG), suatu alat analisis baharu dalam bidang analisis data bertopologi (ADB) menawarkan pendekatan baharu bagi mendapatkan maklumat dalam suatu set data menggunakan pendekatan kualitatif. HG menggunakan konsep topologi untuk mendapatkan maklumat berkaitan ciri topologi seperti komponen berkait, lubang dan lompong yang hadir dalam set data tersebut. Kajian ini membentangkan pendekatan baharu bagi analisis data aliran sungai bagi pengesanan banjir menggunakan kaedah HG. Suatu analisis telah dijalankan di Sungai Kelantan, Malaysia. Hasil kajian menunjukkan bahawa HG memberikan corak ciri-ciri topologi data aliran sungai yang berbeza bagi musim kering dan banjir. Secara khususnya, terdapat lebih banyak ciri topologi yang gigih dalam bentuk komponen berkait and lubang pada data musim banjir berbanding musim kering. Hasil kajian juga menunjukkan bahawa data siri masa ukuran jarak berkaitan perubahan komponen berkait adalah konsisten dengan data siri masa aliran sungai. Kesimpulannya, kajian ini mencadangkan data siri masa ukuran jarak berkaitan perubahan komponen berkait boleh digunakan sebagai ukuran bagi pengesanan banjir di Sungai Kelantan, Malaysia.

Kata kunci: Analisis data bertopologi; arus sungai; banjir; homologi gigih; pembenaman masa penangguhan

INTRODUCTION

Flooding is an environmental hazard that occurs almost everywhere around the world. Floods are generally defined as an overflow of water onto land that is usually dry. Flooding is one of the most common natural disaster that contributes to high number of deaths and loss of properties. Historical records of floods have shown that flooding has an imminent impact on people's livelihoods, and it is unavoidable (Jonkman & Kelman 2005). Floods occur almost every year in tropical countries including

Malaysia. In Malaysia, the major type of flooding that seriously plagued human life and the environment is monsoon flooding (Chan & Parker 1996).

Flooding has been recorded in Malaysia since the 1800s. The first recorded major flood event occurred in 1886 and had caused extensive damage in Kelantan, Malaysia (Chan & Parker 1996). Kelantan is one of the largest states in Malaysia and is affected by monsoon flooding every year (Awadalla & Noor 1991). Based on the report by Department of Drainage and Irrigation (DID 2010) on the flood events at Kelantan, starting the year 2000, the first severe flood that hit Kelantan was reported on December 2001 due to the unusual tropical cyclone Vamei. Afterwards, in the year of 2007 and 2009 heavy rainfall again had triggered major floods in Kelantan. To date, the worst flood reported in Kelantan was at the end of 2014, commonly known as Kelantan Big Yellow Flood 2014 (Alias et al. 2016).

In general, understanding of observational and historical hydro-climatological data such as streamflow data is important because they provide climate indicators for environmental risks such as flooding (Chang 2007). In monsoon areas, which can be associated with annual flooding due to high intensity of rainfall, knowledge about changes in streamflow data is very important to understand flooding risk and to allow preparation for mitigation. The answer will determine future flood management policy and decisions.

In previous research, one of the most commonly used tools to analyze and forecast extreme hydrological events is through frequency analysis of hydrograph (graph of discharged over time) (Belmar et al. 2011; Hannah et al. 2000). This approach usually defined hydrograph by a limited number of features and does not use of the hydrograph's complete information content. This may lead to negative impacts, particularly in terms of information loss and substantial simplification of the overall hydrological phenomenon. In 2012, new analysis tool known as functional data analysis (FDA) was proposed to the hydrological context to explore hydrograph (Chebana et al. 2012). FDA directly uses the whole time series of streamflow data which includes the information on shape, peak and timing. By taking into account all of the available information of the streamflow data, then, FDA treat the entire hydrograph as a functional observation (function or curve). Through looking at the whole hydrograph as a single observation, they suggest that it is more descriptive of the real phenomenon and makes better use of the whole time series. FDA had also been used to describe dry and wet periods of Malaysia through spatial and temporal variability of rainfall data (Suhaila & Yusop 2016). However, as FDA relies on smoothing step of curves, a little lack of accuracy on the function estimation may cause an increase in uncertainty of the model.

Some other previous research focused on conceptual hydrological models to get a better understanding of extreme events (Adnan 2010; Faizah 2015; Modaresi et al. 2018). In another direction, a chaotic approach has also been used to study hydrological process (Adenan & Noorani 2016; Fuwape et al. 2016). Note that the modelling approach maybe prone to some errors due to lack of information and this may lead to failure in understanding physical phenomenon. Also, even though the chaotic approach uses observational streamflow data, this approach is essentially quantitative in nature since they do not provide qualitative or topological information of the evolution of the data.

In the world of data science, topological data analysis (TDA) has recently provided a new approach for data analysis (Carlsson 2009; Ghrist 2008). TDA is an area in which data analysis, algebraic topology, statistics, and other related fields converge. TDA's main objective is to used geometry and topology ideas and findings to develop tools to study qualitative features or structures of data. To achieve this goal, accurate descriptions of qualitative features, tools to calculate them in practice and some assurance of the robustness of these features are required. One way to address all the issues is an approach in TDA called persistent homology (PH). Since this framework is based on algebraic topology, which offers a well-understood theoretical framework to explore qualitative features of data with a complex structure, it is therefore appealing for applications. It is also stable in terms of small perturbations in input data (Cohen-Steiner et al. 2007). This is an important key of PH since all the available data are used and all their features would be analyzed without necessarily increasing the uncertainty.

The idea of PH is mainly the same as in the classical analysis, e.g., representing and visualizing the data, studying the variability and trends, comparing different data set, as well as modelling and predicting. PH has been shown to be a powerful tool for analyzing complex data sets. For an excellent review of the current status of PH including its background theory, application and software (Otter et al. 2017). PH techniques have been applied to diverse problems including spatial data clustering (Pereira & de Mello 2015), complex dynamical systems (Khushboo & Shalabh 2017), financial systems (Gidea & Katz 2018), air quality research (Zulkepli et al. 2020a, 2020b, 2019) and hydrological field (Musa et al. 2020, 2019). Of recent interest is the exploration

and application of TDA to time-delay embedding of time series for the modelling and classification of dynamical system and time varying events (Musa et al. 2020; Pereira & de Mello 2015; Zulkepli 2020a, 2020b). To the best of our knowledge, PH has not been applied to streamflow data. As PH provide new information on the datasets based on the qualitative information, we believe that this approach could be used as an alternative framework, or it can also be employed parallel complement to bring additional insight on streamflow data analysis.

Motivated by this new technique of data analysis and the research gap on the application of PH on streamflow data, in this paper we apply PH to streamflow data. The main objective of this paper was to attract attention to the topological properties of streamflow data for hydrological applications through the PH framework. Also, is to illustrates the potential of this new technique for flood detection. Therefore, the paper concludes with considerations of the potential of the method for streamflow data analysis applications using PH for flood detection. This flood detection is a preliminary step of streamflow data analysis using PH as it tells us that this new information on topological properties of streamflow data contain information of the flood events that later can be used for further analysis.

In particular, in this paper, we analyze the time series data of streamflow at the Guillemard Bridge station, Sungai Kelantan, Malaysia to investigate the relation between streamflow data and flooding events. Our data consists of 15 years from 2000 to 2014. Using the PH processing pipeline for time series, first we apply time-delay embedding to obtain point cloud data from the time series, followed by computing homology groups to determine persistent homology which will then be presented in topological summaries known as barcodes and persistence diagrams. By employing Wasserstein distance measure on the evolution of the topological features, we obtain a time series of the monthly changes associated with this evolution.

In the next section, we provide a concise and informal review of the PH methodology for time series processing. In subsequent section, we introduce our streamflow time series data. We presents our analysis and results in the next section and last section concludes the paper. For the computational part on this paper, we employ the R-package 'TDA' (Fasy et al. 2021).

DATA

Malaysian climate is governed by two regimes that are the southwest and northeast monsoons (Chan & Parker 1996). The southwest monsoon which usually commences in May and ends in August is responsible for the dry period for the whole country. The northeast monsoon which usually takes place between November to February is responsible for the wet period (heavy rains) in the east coast of the Peninsular Malaysia and frequently cause monsoon flooding.

Kelantan is one of the largest states in Malaysia and is affected annually by monsoon flooding. Sungai Kelantan which is one of the main rivers in Kelantan is situated in northeast of the Peninsular Malaysia between the 4^0 40' and 6^0 12' North, and longitudes 101^0 20' and 102^0 20' East. It is the longest river in Kelantan at 248 km and drains an area of 13,100 km^2. The total area of Kelantan is 15,022 km^2 and about 68.5% of the population lives in Sungai Kelantan Basin.

Due to the northeast monsoon which brings along heavy rains, Sungai Kelantan often overflows in the period, causing an almost annual recurrence of monsoon flooding (Awadalla & Noor 1991). Since the year 2000, based on Kelantan flood report (DID 2010), the first severe flood that hit Kelantan was reported on December 2001 which was due to the unusual tropical cyclone Vamei that hits South China Sea. Heavy rainfall in the year 2007 and 2009 had also triggered major floods at Kelantan. To date, the worst flood reported in Kelantan was at the end of 2014 and is commonly known as Kelantan Big Yellow Flood 2014 (Alias et al. 2016). Based on these events, analysis for streamflow data of Sungai Kelantan is important as it can be a crucial climatic indicator for flood events in Kelantan. Therefore, in this research we focus our streamflow analysis at Sungai Kelantan.

Daily data of Sungai Kelantan flow at Guillemard Bridge station (measured in m^3s^{-1}) were obtained from the Earth Observation Centre, Institute of Climate Change, Universiti Kebangsaan Malaysia. The data involves in this analysis have 0.0549 missing data that were filled using the results from the computation of the linear interpolation method. Figure 1 shows the time series plot of the streamflow data for 15 years from year 2000 until 2014. The highest magnitude of streamflow is at the end of year 2014, followed by 2009, 2007 and 2001. Some important statistical parameters of the time series are shown in Table 1.

RECONSTRUCTION OF PHASE SPACE

In this study, PH processing pipeline for time series are implemented. Since PH is a method that extract the topological features of a data set, which is not readily available in time series in its standard form, therefore we use Takens' embedding theorem (Taken 1981) to

prepare the data. Takens' embedding theorem states that time series can be used to reconstruct the phase space of the associated dynamical system, resulting in point cloud data.

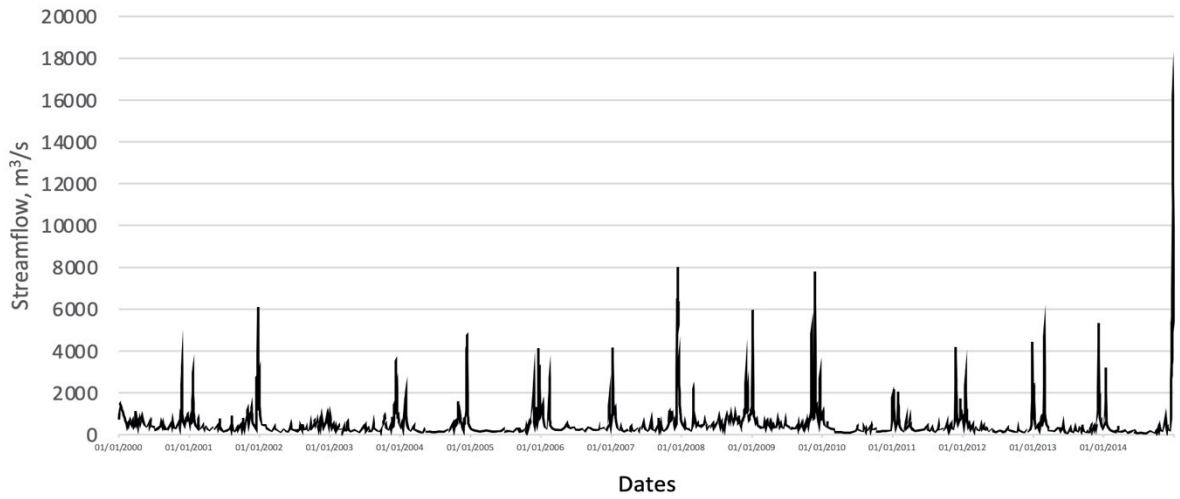

FIGURE 1. Time series plot of daily mean streamflow data of Sungai Kelantan flow at Guillemard Bridge station from 01/01/2000 until 31/12/2014

TABLE 1. Statistics of streamflow time series of Sungai Kelantan flow at Guillemard Bridge station from 01/01/2000 until 31/12/2014

Statistics	Daily
Number of data	5478
Average	454.03
Max	18339.4
Min	49.9
Standard deviation	795.08
Skew	10.06
Kurtosis	154.21

Given a time series $x_1, x_2, \ldots, x_N$, Takens' embedding theorem (Taken 1981) states that the constructed phase space consists of vectors $x_n\ (m,\tau) = (x_n, x_{n+\tau}, \ldots, x_{n+(m-1)\tau})$ where m is the embedding dimension and τ is the time delay which have to be chosen appropriately. The value of m and τ can be found using method of average mutual information and Cao method, respectively (Hamid & Noorani 2017; Zaim & Hamid 2017). However, in this research we fixed the value $\tau = 1$ and $m = 2$ as our previous research on PH on water level data of Sungai Kelantan (Musa et al. 2020, 2019) using these values shows good analysis. As we obtained 2-dimensional point cloud data from the reconstruction of phase space, therefore, we can then extract the topological features associated with the constructed phase space and compute the homology groups to determine the persistent homology.

PERSISTENT HOMOLOGY

The core idea in PH is to analyze topological features in the data set. From the point cloud data obtained via

the reconstruction of phase space, we then construct simplicial complexes. The building blocks for a simplicial complex are k-simplex which are 0-simplex (vertex), 1-simplex (edge), 2-simplex (triangle), illustrated in Figure 2.

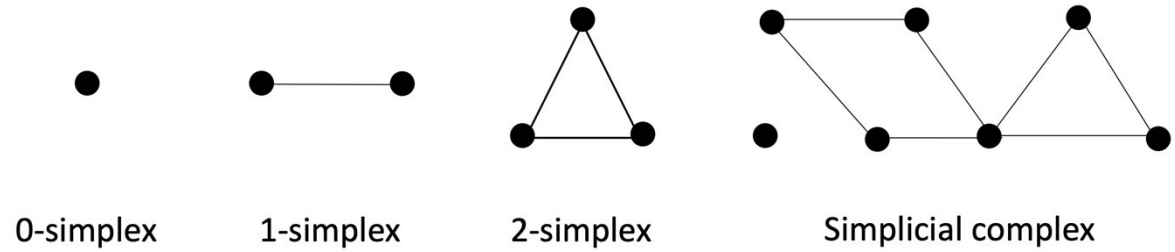

FIGURE 2. Buildup of simplicial complex from simplices

To analyze how topological features appears and disappears, the idea of filtered simplicial complex is used. For example, consider the set of points in $\mathbb{R}^2$ shown in Figure 3 (top). Let ε be a nonnegative real number which is interpreted as a distance parameter. By considering ε ball at each data point, we build an edge (1-simplex) between two points a and b if and only if the distance between them is less than ε. Similarly, we build a triangle (2-simplex) if and only if the pairwise distances between three points are each less than ε. This will produce filtered simplicial complexes as illustrated in Figure 3 (top). This construction of simplicial complexes is known as filtered Vietoris-Rips simplicial complexes or Rips complexes (Edelsbrunner & Harer 2010).

Based on the Rips complexes, one usually interested to understand their basic topological features such as the number of components, holes, and voids. Algebraic topology captures these topological features by counting the rank of each homology group of the simplicial complex. For each simplicial complex, X, algebraic topology can compute its k-dimensional homology H_k (X) for each natural number $k \in \{0,1,2,\ldots\}$. The rank of the 0-dimensional homology group H_0 (X) counts the number of connected components, the rank of the 1-dimensional homology group H_1 (X) count of number of holes, the rank of the 2-dimensional homology group H_2 (X) count of number of voids and so on. These ranks of homology group also known as Betti number. Therefore, the pth Betti number counts the number of p-dimensional holes of the simplicial complex. As the value of distance parameter ε increases, simplices are added to the simplicial complex and therefore the Betti number also changes, complexes. Persistent homology then captures which topological features that are persist across the scale.

Precisely, Figure 3 (top) shows an example of filtered simplicial complexes for point cloud data consisting of four points. In the beginning, at filtration value ε_0, we can see that there are 4 components, H_0 (X) = 4. The components survive through filtration value ε_1 and ε_2. At filtration value ε_3, edges or 1-simplexes are formed and connect all the points together into a single connected component and hence changes the Betti number of 0-dimensional holes to H_0 (X) = 1. The component never varnishes as the filtration value is further increased. Also, at filtration value ε_3, a 1-dimensional hole in the data is born as the edges form a rectangle, H_1 (X) = 1. The 1-dimensional hole dies out at filtration value ε_4 when the 2-simplex or triangle appears.

TOPOLOGICAL SUMMARIES

By constructing filtered simplicial complex, PH can show us the evolution of topological features that exist in the data set. Now we need tools to summarize every topological feature that have been captured by PH. These topological summaries will provide a concise description of the topological changes over all scales of the data. These topological summaries store information of the growth, birth and death of different topological features across dimensions.

The first topological summary that has been introduced is commonly known as a barcode. Barcode is a finite collection of intervals that represent the lifetime of topological features (connected components, holes). The left endpoint of the interval represents the filtration value ε_i at which the topological feature is born, and its right endpoint represents the filtration value ε_j, with $j > i$, at which the topological feature dies.

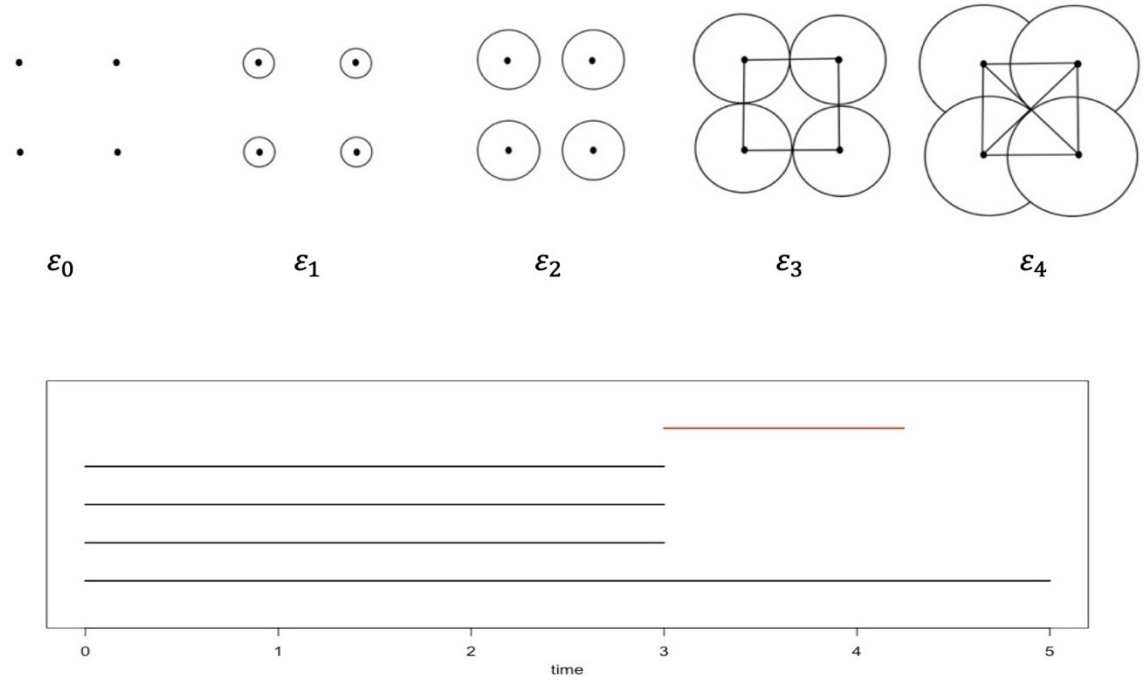

FIGURE 3. Filtered simplicial complex (top) and the corresponding barcode (bottom)

If the topological feature live forever, we represent its lifetime by the interval $[\varepsilon_i, \varepsilon_\infty)$. Figure 3 (bottom) shows the corresponding barcode for point cloud data in Figure 3 (top).

Based on Figure 3 (bottom), from the barcode we can see that for topological features of dimension 0 (components), there are four components that are born at filtration value $\varepsilon_0 = 0$ and three dies at filtration value $\varepsilon_3 = 3$ (given by the first three black intervals $[\varepsilon_0, \varepsilon_3)$). These three components that vanishes merged with the fourth component which lives forever and is given by the interval $[\varepsilon_0, \varepsilon_\infty)$. For topological features of dimension 1 (holes) the interval is colored as red in the barcode. The hole appears at filtration value $\varepsilon_3 = 3$ and dies at filtration value $\varepsilon_4 = 4.25$ given by the red interval $[\varepsilon_3, \varepsilon_4)$. In the barcode, the lifetime of a topological feature is the difference between death point and birth point of the topological feature. The longer the lifetime of the topological features the more persistence the topological features are.

An alternative way to visualize topological features captured by PH from the data set is through persistence diagram. Persistence diagram is a finite multiset of all birth-death pairs of topological features points in the extended $\overline{\mathbb{R}^2}$ plane, where $\overline{\mathbb{R}} = \mathbb{R} \cup \{\infty\}$. The line in the barcode corresponding to the interval $(\varepsilon_i, \varepsilon_j)$ is represented as the point $(\varepsilon_i, \varepsilon_j)$ in the persistence diagram while the points on the diagonal line represent topological features that are born and dies at the same time (each of the points on the diagonal has infinite multiplicity). This diagonal line helps us see which topological features that are persistent. Point that lies close to the diagonal line indicates that the topological feature is not persistent, while points that stand far from the diagonal line corresponds to the persistent topological features. Figure 4 shows the point cloud data from Figure 3 with the respective barcode and persistence diagram.

DISTANCE MEASURES

As topological summaries such as barcodes and persistence diagrams only give us a visualization of the topological features that exist in the data that have been captured by PH, we need another mathematical tool to extract information from these topological summaries. Since barcodes and persistence diagrams contain the same information on the topological features that exist in the data, therefore for further analysis we will only look at persistence diagrams instead of barcodes.

Given a set of persistence diagrams, it is only natural for us to compare them with respect to their topological similarity. This can be done by using a distance metric.

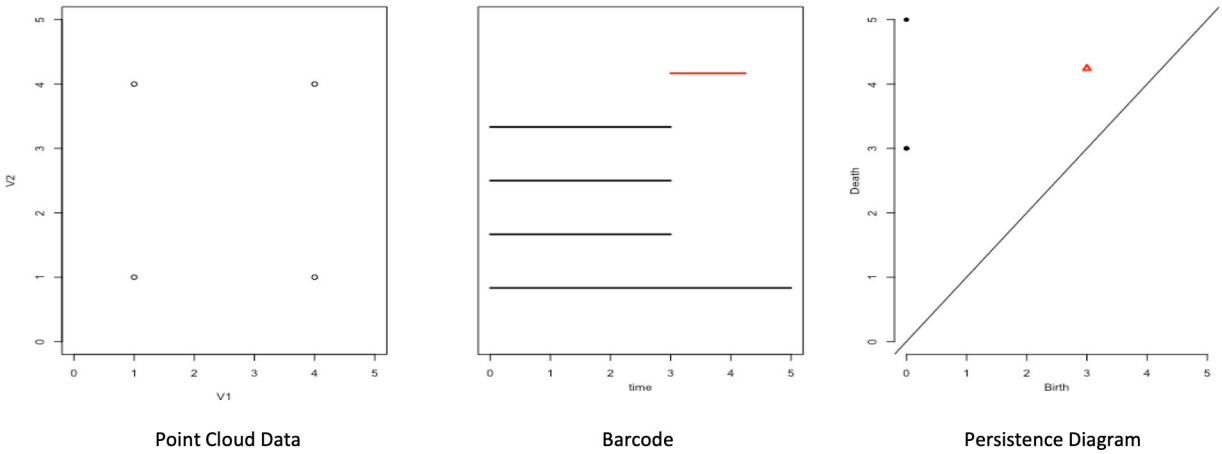

FIGURE 4. Point cloud data, barcode and persistence diagram

One of the common distance metrics is Wasserstein metric (Edelsbrunner & Harer 2010), which will be used in this study. The definition of p^{th} Wasserstein metric is as follows:

$$D_p(X,Y) = \inf_{\phi}\left[\sum_{q\in X}\|q-\phi(q)\|_{\infty}^{p}\right]^{\frac{1}{p}} \quad (1)$$

where X and Y are persistence diagrams and the summation is over all bijections ϕ: $X \rightarrow Y$. Here $\|\cdot\|_{\infty}$ is the sup norm. Since the diagonal set is by default part of all persistent diagrams, the pairing points between X and Y via ϕ can include pairing between off-diagonal points and diagonal points.

Note that different values of degree p yield different types of measurement of the distances between persistent diagrams. Using $p = \infty$, the corresponding distance only measures the distance between the most significant features in the diagrams. Using $p \geq 1$, the corresponding distance D_p puts more weight on the significant features than on the least significant ones. One of the remarkable properties of persistence diagrams is their robustness (Cohen-Steiner et al. 2007), meaning that small changes in the input data produce persistence diagrams that are close to one another relative to the Wasserstein metric. In this study we use Wasserstein metric degree $p = 1$.

RESULTS AND DISCUSSION

In this section, we apply our PH processing pipeline to the daily streamflow data of Sungai Kelantan at Guillemard Bridge station and discuss the findings. The discussion of the findings is divided into two parts. The first part shows some examples of the differences in topological features for dry and wet period based on their reconstructed phase space, barcodes and persistence diagrams. The purpose of this part is to investigate the pattern of topological features for dry and wet periods. In the second part we employ a distance measure on the evolution of the topological features and obtain a time series of the monthly changes associated with this evolution.

For the first part of the results, here we show the results that we obtained by applying PH to selected streamflow data for dry and wet periods. Since the dry period usually commences in May and up to August (Chan & Parker 1996) therefore for the analysis here we select the data during dry period for year 2000, May until August 2000. Here we have chosen the year 2000 to illustrate our findings for the dry periods because we observed that the pattern of topological features for dry periods are consistent for each year throughout the 15 years. For the wet periods, we choose our data set based on the historical severe flood events, December 2001, December 2007 and November 2009, lastly December 2014.

The reconstructed phase space, barcodes and persistence diagrams for the dry periods are shown in Figure 5 while for the wet periods are shown in Figure 6. For the dry period, the points in the reconstructed phase space (left column in Figure 5) are densely packed together near the origin which indicate that the values of streamflow are low and in the same range. However, for the wet period (Figure 6), the points in the reconstructed phase space are spread out, which indicate there are low

and high values of streamflow data with a wide range of values. For comparison purpose, we fixed the upper limit of the axes in the reconstructed phase spaces to be 10000. An exception was made for the wet period December 2014 (Kelantan Big Yellow Flood 2014) where we had to increase the limit up to 20000 to cover all the points.

The topological features that PH can extract from two-dimensional point cloud data are connected components (0-dimensional topological features) and holes (1-dimensional topological features). In a barcode, each black interval corresponds to a connected component while red interval corresponds to a hole. The left endpoint of the interval gives us the filtration value at which a topological feature is born while the right endpoint tells us when the topological feature dies. In Figures 5 and 6, note that the left endpoint for all black intervals is 0, since all the points or components are born at filtration value 0. The longest interval which starts at the filtration value 0 and dies at the filtration value 5000 (maximum filtration value) indicates that once the graph is fully connected into a single connected component, it

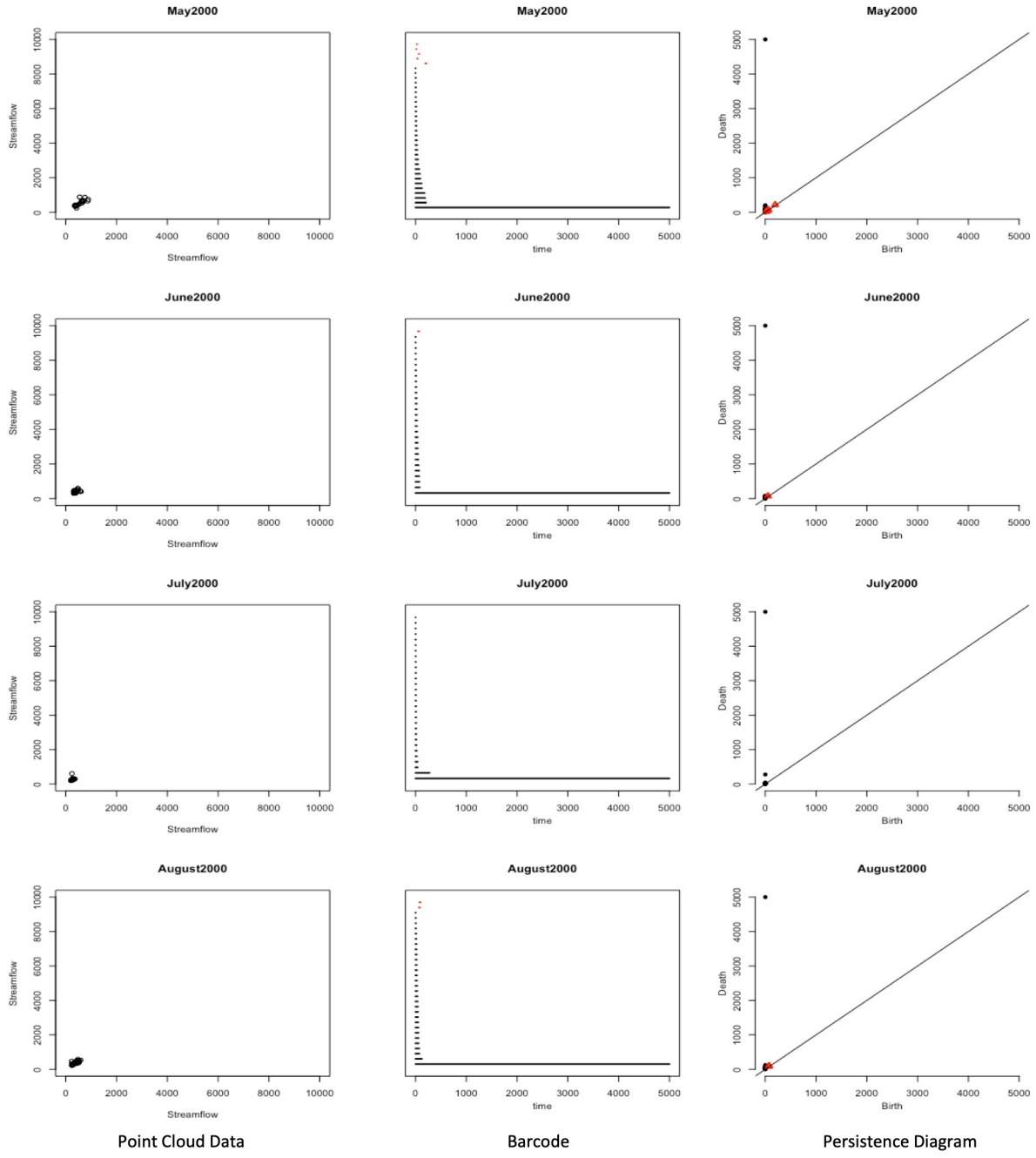

FIGURE 5. Persistent homology during dry period May to August 2000

remains fully connected (hence the component never dies) as the filtration value is further increased.

Persistency of topological features can be measured by finding the difference between the death and birth of the topological features. The larger the difference between death and birth of the topological feature the more persistent the topological feature is. Therefore, based on the barcodes for dry periods in Figure 5 (middle row), we can see that there are only short black and red intervals that indicate short-lived (non-persistent) connected components and holes. For the wet periods in Figure 6 (middle row), some of the black intervals passed through filtration value 3000 which are significantly longer compared to the dry periods. This indicates that there are long-lived (persistent) connected components for the wet period data set. Barcodes for the wet periods also contain longer red intervals compared to the barcodes for the dry periods. This tells us that there exist long-lived (persistent) holes in wet period data.

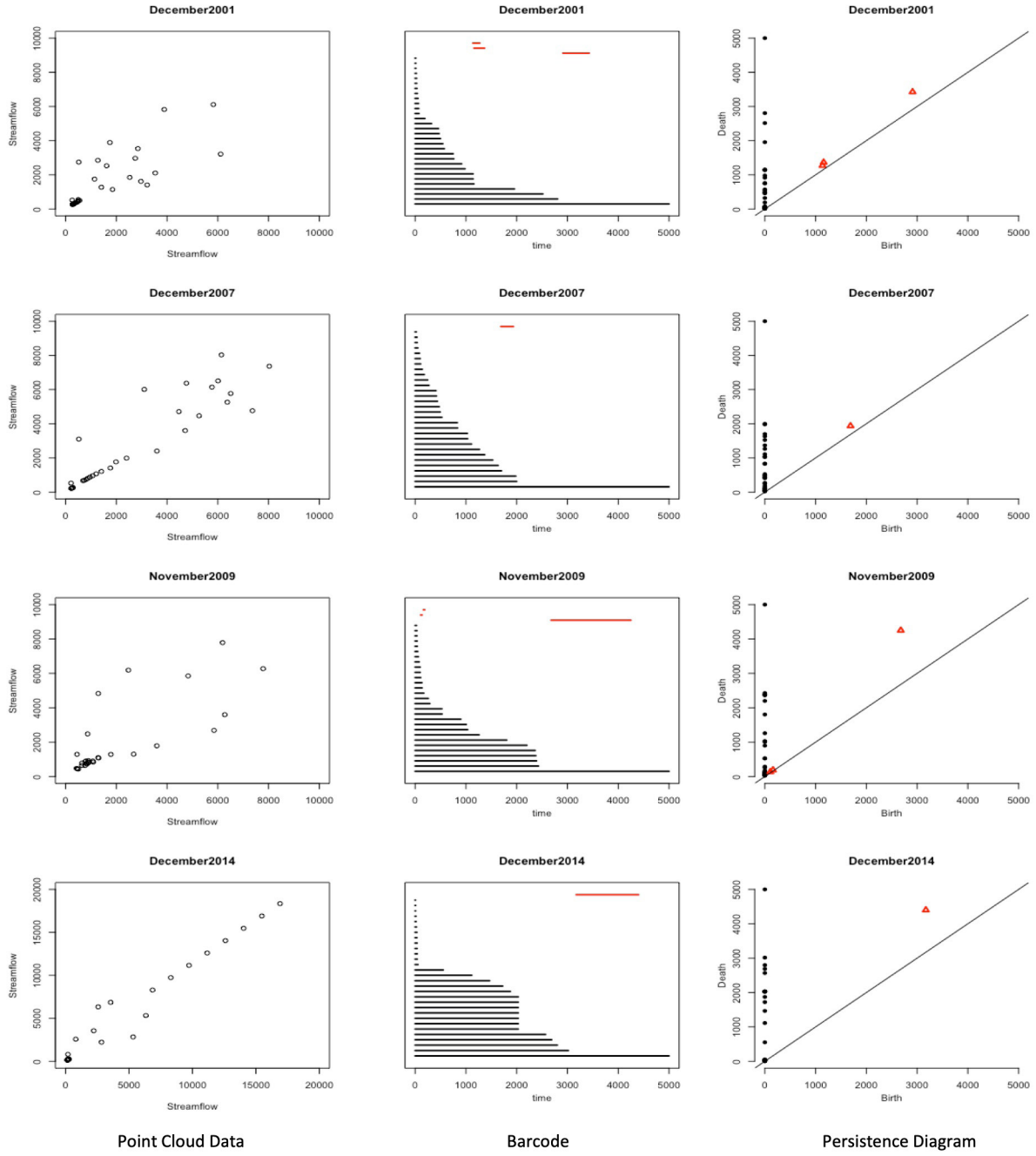

FIGURE 6. Persistent homology for severe flood events December 2001, December 2007, November 2009 and December 2014

An alternative way to summarize the information on topological features is through persistence diagrams. In a persistence diagram, each black dot corresponds to a connected component (0-dimensional topological features), while red dot corresponds to a hole (1-dimensional topological features). The x-coordinate of a dot gives us filtration value at which a topological feature is born while the y-coordinate tell us when the topological feature dies. In Figures 5 and 6 (right columns), the x-coordinate for all black dots is 0, since all the points or components are born at filtration value 0. The black dot with the highest y-coordinate, that is 5000 (maximum filtration value), corresponds to the fully connected graph.

The concentration of the black and red dots near the diagonal line of Figure 5 (right column) indicates that there are only short-lived (non-persistent) connected components and holes in the dry period data. Nevertheless, Figure 6 (right column) shows that there are black and red dots located far from the diagonal line. This indicates that there exists long-lived (persistent) connected components and holes in the wet period data. These observations can further be quantified by computing the time series of the Wasserstein distance of the persistence diagrams.

For the second part of the results, we employ a distance measure on the evolution of the topological features in the time-ordered persistence diagrams and obtain a time series of the monthly changes associated with this evolution. The Wasserstein metric provide a means for comparing the topological similarity between persistent diagrams. As we want to see the changes associated with the evolution of the topological features so here, we compare the topological similarity for each persistence diagram relative to an 'origin' persistence diagram. Here an origin persistence diagram corresponds to an empty persistence diagram which only consists of the diagonal line. Each comparison of persistence diagrams with the origin will result in one distance measure, so that we will have a time series of the distance measures providing an evolution of the monthly changes of the topological features. Our aim is to assess how various topological features of the data affect the growth of the distance measures of the persistence diagrams.

Overall, based on the time series of the Wasserstein metric for the connected components (0-dimensional topological features) in Figure 7 (top), we can see that there are 15 peaks in this time series which corresponds

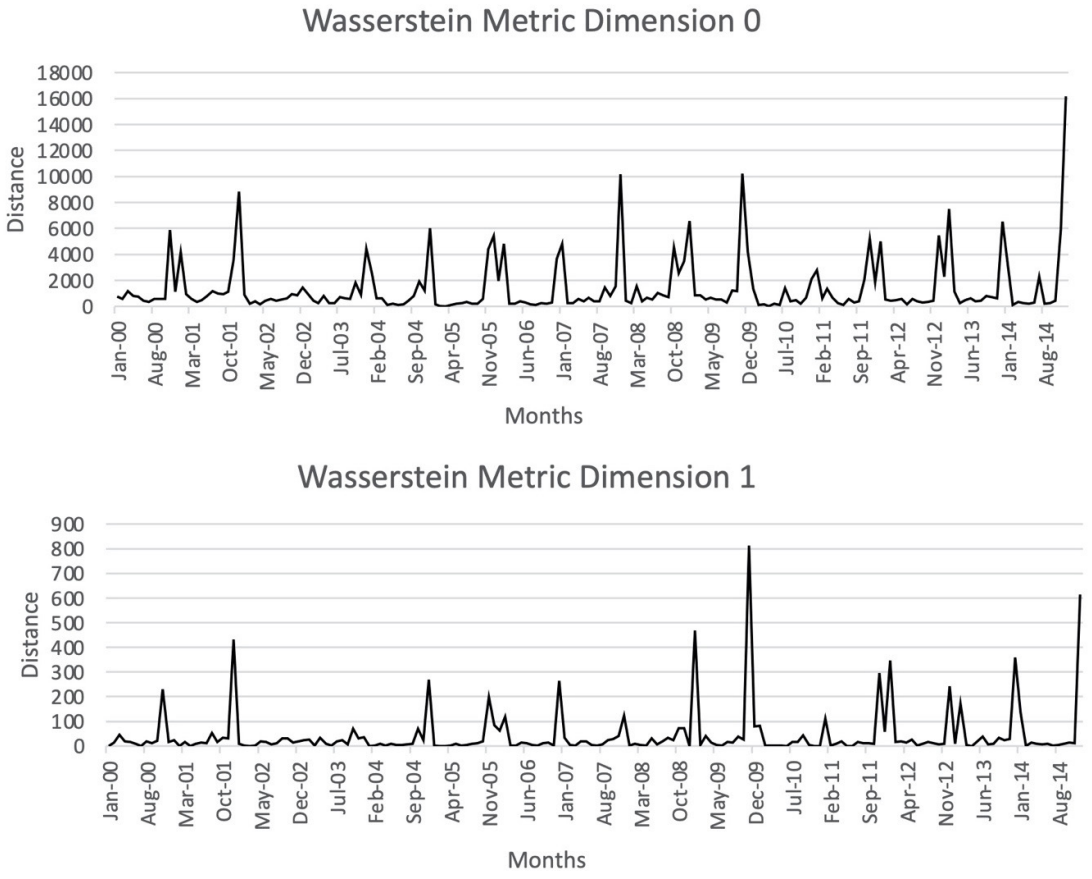

FIGURE 7. Wasserstein metric for dimension 0 (top) and dimension 1 (bottom)

to the 15 wet periods of the 15 years of the data set. The highest magnitude is during December 2014, followed by November 2009, December 2007, and December 2001. All these peaks in the time series of the distance measures correspond to the flood events in the original streamflow data, implying the consistency of the Wasserstein distance of the 0-dimensional topological features (connected components). Therefore, by calculating the distance measures of persistence diagrams which corresponds to the connected components (0-dimensional topological features) can give us information on detecting floods at Sungai Kelantan, Malaysia.

For the time series of the Wasserstein distance for holes (1-dimensional topological features) in Figure 7 (bottom), there are wet periods which show no clear peaks. The highest magnitude is found during December 2009, followed by December 2014, January 2009, and December 2001. This result shows that the time series of the distance measures corresponding to the evolution of the holes is not consistent with the original time series of the streamflow data. Thus, the distance measures corresponding to the holes (1-dimensional topological features) does not seems to give us significant information on the flooding events at Sungai Kelantan, Malaysia.

Conclusion

The primary aim of this paper is to introduce the PH framework to hydrological applications based on the topological properties of streamflow time series. This paper has presented an objective procedure for flood detection using PH as an aid to explore the streamflow time series. This study involved the application of TDA, specifically PH on detecting floods based on streamflow data. PH provides qualitative information of the data set or more precisely topological features such as connected components and cycles that exists in the data set. This paper presents a new approach for streamflow data analysis by using PH to the daily data of Sungai Kelantan streamflow from 2000 to 2014.

From the results, it is clear that PH can characterize the dry and wet periods through producing different patterns of topological features for both situations. In this regard, there are more persistent connected components and holes during the wet periods compared to the dry periods. By employing a distance measure on the evolution of the topological features we obtained a time series of the monthly changes associated with this evolution. We observed that the time series corresponding to the evolution of the components is consistent with the time series of the streamflow data. In conclusion, this study suggests that the time series of the distance measure corresponding to the evolution of the components can be used for flood detection at Sungai Kelantan, Malaysia.

Acknowledgements

This work was supported by the UKM Grant: GUP-2020-032; and the Ministry of Education Malaysia Grant: FRGS/1/2019/STG06/UKM/01/3. The authors also acknowledge Earth Observation Centre, Institute of Climate Change, Universiti Kebangsaan Malaysia for providing the streamflow data.

References

Adnan, N.A. 2010. Quantifying the impacts of climate and land use changes on the hydrological response of a monsoonal catchment. Dissertation, University of Southampton, Southampton, England (Unpublished).

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.

Alias, N.E., Mohamad, H., Chin, W.Y. & Yusop, Z. 2016. Rainfall analysis of the Kelantan big yellow flood 2014. *Jurnal Teknologi* 78(9-4): 83-90.

Awadalla, S. & Noor, I.M. 1991. Induced climate change on surface runoff in Kelantan Malaysia. *International Journal of Water Resources Development* 7(1): 53-59.

Belmar, O., Velasco, J. & Martinez-Capel, F. 2011. Hydrological classification of natural flow regimes to support environmental flow assessments in the intensively regulated Mediterranean rivers, Segura river basin (Spain). *Environmental Management* 47(5): 992-1004.

Carlsson, G. 2009. Topology and data. *Bulletin of the American Mathematical Society (N.S)* 46(2): 255-308.

Chan, N.W. & Parker, D.J. 1996. Response to dynamic flood hazard factors in Peninsular Malaysia. *The Geographical Journal* 162(3): 313-325.

Chang, H. 2007. Comparative streamflow characteristics for urbanizing basins in the Portland metropolitan area, Oregon, USA. *Hydrological Processes* 21(2): 211-222.

Chebana, F., Dabo-Niang, S. & Ouarda, T.B.M.J. 2012. Exploratory functional flood frequency analysis and outlier detection. *Water Resources Research* 48(4): W04514.

Cohen-Steiner, D., Edelsbrunner, H. & Harer, J. 2007. Stability of persistence diagrams. *Discrete and Computational Geometry* 37(1): 103-120.

Drainage and Irrigation Department (DID). 2010. *Updating of Condition of Flooding and Flood Damage Assessment in Malaysia: State Report for Kelantan.* Unpublished report. Kuala Lumpur: DID.

Edelsbrunner, H. & Harer, J. 2010. *Computational Topology: An Introduction.* Applied Math Textbook.

Faizah, C.R. 2015. Study on early forecasting of flood through

historical hydrologic data analysis and numerical simulation in Kelantan Watershed, Malaysia. Dissertation, University of Tokyo, Tokyo, Japan (Unpublished).

Fasy, B.T., Kim, J., Lecci, F., Maria, C. & Rouvreau, V. 2021. *Statistical Tools for the Topological Data Analysis.* https://cran.r-project.org/web/packages/TDA/TDA.pdf.

Fuwape, I.A., Ogunjo, S.T., Oluyamo, S.S. & Rabiu, A.B. 2016. Spatial variation of deterministic chaos in mean daily temperature and rainfall over Nigeria. *Theoretical and Applied Climatology* 130(1-2): 119-132.

Ghrist, R. 2008. Barcodes: The persistent topology of data. *Bulletin of the American Mathematical Society (N.S.)* 45(1): 61-75.

Gidea, M. & Katz, Y. 2018. Topological data analysis of financial time series: Landscapes of crashes. *Physica A: Statistical Mechanics and its Applications* 491: 820-834.

Hamid, N.Z. & Noorani, M.S.M. 2017. Aplikasi model baharu penambahbaikan pendekatan kalut ke atas peramalan siri masa kepekatan ozon. *Sains Malaysiana* 46(8): 1333-1339.

Hannah, D.M., Smith, B.P.G., Gurnell, A.M. & McGregor, G.R. 2000. An approach to hydrograph classification. *Hydrological Processes* 14(2): 317-338.

Jonkman, S.N. & Kelman, I. 2005. An analysis of the causes and circumstances of flood disaster deaths. *Disasters* 29(1): 75-97.

Khushboo, M. & Shalabh, G. 2017. Topological characterization and early detection of burifications and chaos in complex system using persistent homology. *Chaos: An Interdisciplinary Journal of Nonlinear Science* 27(5): 051102.

Modaresi, F., Araghinejad, S. & Ebrahimi, K. 2018. Selected model fusion: An approach for improving the accuracy of monthly streamflow forecasting. *Journal of Hydroinformatics* 20(4): 917-933.

Musa, S.M.S., Noorani, M.S.M., Razak, F.A., Ismail, M., Alias, M.A. & Hussain, S.I. 2020. Using persistent homology as preprocessing of early warning signals for critical transition in flood. *Scientific Reports* 11: 7234.

Musa, S.M.S., Noorani, M.S.M., Razak, F.A., Ismail, M. & Alias, M.A. 2019. Streamflow data analysis using persistent homology. *AIP Conference Proceedings* 2111(1): 020021.

Otter, N., Potter, M.A., Tillmann, U., Grindrod, P. & Harrington, H.A. 2017. A roadmap for the computation of persistent homology. *EPJ Data Science* 6: 17.

Pereira, C.M.M. & de Mello, R.F. 2015. Persistent homology for time series and spatial data clustering. *Expert Systems with Applications* 42(15-16): 6026-6038.

Suhaila, J. & Yusop, Z. 2016. Spatial and temporal variabilities of rainfall data using functional data analysis. *Theoretical and Applied Climatology* 129(1-2): 229-242.

Takens, F. 1981. Detecting strange attractors in turbulence. In *Dynamical Systems and Turbulence*, edited by Rand, D. & Young, LS. Lecture Notes in Mathematics. Vol. 898, Springer, Berlin, Heildelberg, New York. pp. 336-381.

Zaim, W.N.A.B.W.M. & Hamid, N.Z. 2017. Peramalan bahan pencemar ozon (0_3) di Universiti Pendidikan Sultan Idris, Tanjung Malim, Perak, Malaysia mengikut monsun dengan menggunakan pendekatan kalut. *Sains Malaysiana* 46(12): 2523-2528.

Zulkepli, N.F.S., Noorani, M.S.M., Razak, F.A., Ismail, M. & Alias, M.A. 2020a. Cluster analysis of haze episodes based on topological features. *Sustainability* 12(10): 3985.

Zulkepli, N.F.S., Noorani, M.S.M., Razak, F.A., Ismail, M. & Alias, M.A. 2020b. Pendekatan baharu untuk mengelompok stesen pengawasan kualiti udara menggunakan homologi gigih. *Sains Malaysiana* 49(4): 963-970.

Zulkepli, N.F.S., Noorani, M.S.M., Razak, F.A., Ismail, M. & Alias, M.A. 2019. Topological characterization of haze episodes using persistent homology. *Aerosol and Air Quality Research* 19: 1614-1624.

*Corresponding author; email: syedmohdsadiq1992@yahoo.com