Exploring rainfall variabilities using statistical functional data analysis

N A Mazelan[1] and J Suhaila[1,2*]

[1] Department of Mathematical Sciences, Faculty of Science, Universiti Teknologi Malaysia, 81310, Johor Bahru, Johor, Malaysia

[2] UTM Centre for Industrial and Applied Mathematics, Ibnu Sina Institute for Scientific and Industrial Research, Universiti Teknologi Malaysia, 81310, Johor Bahru, Johor, Malaysia

*Corresponding E-mail: suhailasj@utm.my

Abstract. Functional data analysis (FDA) has been widely applied in various scientific fields, including climatological, hydrological, environmental, and biomedical. The flexibility of the FDA in incorporating temporal elements into the statistical analysis makes the method highly demanded compared to the conventional statistical approach. This study introduces FDA methods to investigate the variations and patterns of rainfall throughout Peninsular Malaysia, which includes 16 rain gauge stations in Peninsular Malaysia from 1999 to 2019. A descriptive statistic of the functional data depicted the mean and variation of the rainfall curve over time, while the functional principal component analysis measured the temporal variability of the rainfall curve. According to the findings, the first and second principal components accounted for 87.4% of all variations. The first principal component was highly characterised by the stations over the eastern region during the northeast monsoon since the highest variability was observed from November to January. On the other hand, the stations impacted by the inter-monsoon season were best described by the second principal component. Based on the factor scores derived from the functional principal component, those rain gauge stations with comparable features were then clustered. Overall, the results showed that the rainfall pattern is strongly influenced by their geographical and topographical features and the seasonal monsoon effect.

Keywords: Functional data analysis; Functional principal component analysis; Rainfall variability.

1. Introduction

Rainfall events are typically defined by a finite number of discrete values that comprise multivariate data that may or may not contain all of the information accessible in a smoothing curve. A modern statistical method such as functional data analysis (FDA) has been used to capture the intriguing patterns that can be found in rainfall data. The FDA technique is used to convert discrete data into a smoothing

curve or function. The FDA builds up a smoothing curve that can be examined at any time interval from a series of observations at a discrete time interval. The smoothness and differentiability of the smoothing curve of functional data distinguish it from conventional multivariate data. The conventional multivariate approach fails to capture the underlying process in rainfall data. If a functional concept is ignored, then additional information related to the functional nature of the rainfall dataset is not properly utilized.

Functional data analysis techniques have been successfully carried out by several researchers in various fields [1–6]. Suhaila and Yusop [7] used a functional analysis of variance to explain the geographical and temporal variability in rainfall. Their study brought attention to the variations and distinctions in rainfall profiles. On the other hand, Hael et al. [8] used FDA on projected data in the Taiz region to analyse the temporal variations which help planners manage rainwater conservation. Functional principal component analysis (FPCA) is the most popular tool in FDA. It is used to reduce the dimensionality of functional data, capturing the largest amount of variance and exhibiting various aspects of the underlying data. The FPCA has been used to interpret lactate curves [9], analyse kinematic data [10], gene classification [11], explore variations in glomerular filtration rate curves [12], and model aircraft degradation [13]. Suhaila [14] recently applied the FPCA techniques to determine the variations of the multivariate El Niño Southern Oscillation Index (MEI). Their findings suggested that the two principal components likely represented the variations of El Niño and La Niña episodes.

The conventional statistical method allows the expression of data variability as either a single value or as multivariate values that indicate the standard deviation or the coefficient of variation. However, there are no explanations for the temporal aspects. Therefore, this study intends to evaluate the changes and variabilities of rainfall data in terms of smoothing curves via functional data analysis. The time when the high or low variability occurrences took place will be analysed based on the smoothing curves. The FDA method allows for the assessment of temporal aspects. One could argue that the FDA is more willing to provide information that isn't accessible through conventional statistical techniques. Additionally, it is simpler to comprehend rainfall behaviours and their temporal variation when rainfall variabilities are visualised in smoothing curves. FDA is a relatively new technique in climate studies. The results of this study should give a thorough analysis of the rainfall pattern and aid in future forecasting.

2. Data and Study Area

The daily rainfall data at 16 rain gauge stations was obtained from the Malaysian Meteorological Department from 1999 to 2019. Based on their geographical locations, all rain gauge stations were divided into the Northwest, Eastern, Central, Southwest, and Western regions. The climate of Peninsular Malaysia is significantly influenced by the monsoons, particularly the Southwest Monsoon (SWM), which runs from May to August, and the Northeast Monsoon (NEM), which runs from November to February. The two inter-monsoons, from March to April and September to October, are also predicted to bring heavy rains. During the NEM season, the eastern region frequently receives more rain than those sheltered by the Main Range Titiwangsa. The Main Range Titiwangsa divides Peninsular Malaysia's eastern and western parts over 483 km. The locations of the stations under study are shown in Figure 1.

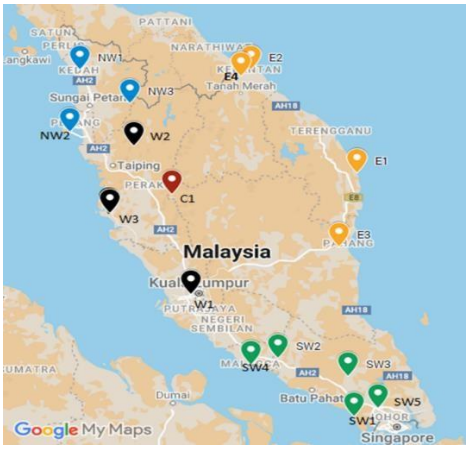


Figure 1. The location of the studied stations.

3. Methodology

In this section, the idea of constructing a series of basis functions to transform discrete rainfall data into a smoothing curve will be presented. The variation in the rainfall data will next be investigated using a functional principal component analysis and a functional summary of functional data with curve visualization.

3.1. Smoothing Rainfall Data

Consider observed discrete rainfall data, $\mathbf{X}_i = \left(x_1(t_1), x_2(t_2), ..., x_i(t_T)\right)'$ for i = 1,2,…,n; j=1,2,…,T where T =365 days, n is the number of rain gauge stations and $x_i(t_j)$ is the amount of rain at day t_j of the i-th stations. These values are converted into smoothing curves $y_i(t)$ which can be expressed linearly as a combination of the basis function;

$$y_i(t) = \sum_{k=1}^{K} d_k \beta_k(t) \tag{1}$$

where d_k is the basis coefficient, $B_k(.)$ is the known basis function, and K denotes the maximum size of basis functions. The periodic function is represented as a Fourier basis which can be written as

$$\hat{y}_i(t) = d_o + d_1 \sin(\omega t) + d_2 \cos(\omega t) + d_3 \sin(2\omega t) + d_4 \cos(2\omega t) + ... \tag{2}$$

which is defined by the basis $\beta_0(t) = 1, \beta_{2r-1}(t) = \sin(r\omega t)$, $\beta_{2r}(t) = \cos(r\omega t)$. The constant ω is related to the period T by the relation $\omega = 2\pi / T$. By minimizing the sum of squared residuals, the expansion of the coefficients d_k are calculated using the least squares approach, as in

$$SSE_i = \sum_{j=1} \left(x_j - y(t_j)\right)^2, i = 1,2,...,n \tag{3}$$

The errors of the curves can be minimized by using the penalized sum square of error (PSSE), which is given as

$$PSSE_\lambda = \sum_j \left[x_j - y(t_j)\right]^2 + \lambda \int \left[D^4 y(t)\right]^2 dt \tag{4}$$

where λ is the parameter for smoothing while $\left[D^4 y(t)\right]^2$ measuring the curvature in y at t.

The smoothing parameter λ will be chosen based on the minimum generalized cross-validation (GCV),

$$GCV(\lambda) = \left(\frac{n}{n - df(\lambda)}\right)\left(\frac{SSE}{n - df(\lambda)}\right) \tag{5}$$

where $df(\lambda)$ are the equivalent degrees of freedom of the smoothing operator.

3.2 Summary of Functional Data

The mean and variance function of curves are given, respectively as

$$\mu_y(t) = \frac{1}{N}\sum_{i=1}^{N} y_i(t) \tag{6}$$

$$VAR_y(t) = \frac{1}{N}\sum_{i=1}^{n}\left[y_i(t) - \underline{y}(t)\right]^2 \tag{7}$$

where $y_i(t)$ is the sample curves or functions and $\bar{y}(t)$ is the sample mean.

The dependence of measurements across various argument values is summarised by the functional covariance, which is computed for all *s* and *t* by

$$\sigma(s,t) = \frac{1}{N}\sum_{i=1}^{N}\left(y_i(s) - \bar{y}(s)\right)\left(y_i(t) - \bar{y}(t)\right). \tag{8}$$

3.3. Functional Principal Component Analysis (FPCA)

The FPCA is used to reduce the dimensionality of the variables by establishing a set of new variables as linear combinations of the original ones. Let $z_i(t)$ be the center for functional datasets such as

$$z_i(t) = y_i(t) - \bar{y}(t), i = 1, 2, ..., n \tag{9}$$

where $\bar{y}(t)$ is the mean function of $\left(y_i(t), ..., y_n(t)\right)$. A FPCA is applied to $z_i(t), i = 1, 2, ..., n,$ produce a small group of harmonic functions which draw attention to the principal source of variation in the data. The first principal component $w_1(t)$ describes a weight function, $z_i(t)$ that exists over the same range. The first component gives the maximum variation in the functional principal component scores as

$$f_{i1} = \int w_1(t) z_i(t) dt \tag{10}$$

subject to the normalisation constraint $\int w_1(t)^2 dt = 1$.

Repeat for the following components $f_{ik} = \int w_k(t) z_i(t) dt$ by maximising the variance of the corresponding scores under the constraints $\int w_k(s) w_j(s) ds = 0, k \geq 2, k \neq j.$

4. Results and Discussion

This section gives a summary statistic of the rainfall data at the stations under study. The process of rainfall smoothing and its variations will be discussed based on the functional principal component analysis.

4.1 Descriptive analysis of annual rainfall data

Table 1 displays the summary statistics of annual rainfall data for stations over the northwest, east, central, southwest, and west regions of Peninsular Malaysia. Petaling Jaya, which is located in the western region of Peninsular Malaysia, has the greatest annual mean rainfall with nearly 3384.41 mm but the lowest variability. It appears that the annual rainfall values at Petaling Jaya did not exhibit significant year-to-year variation. Sitiawan, at the high north of the western area, recorded the lowest annual rainfall. The eastern stations recorded higher annual mean rainfall and a higher coefficient of variation than other regions, as indicated in Table 1. Throughout the NEM season, the eastern region is known for having more frequent and intense downpours than other areas. As a result, stations in the eastern region of Peninsular Malaysia have higher annual mean rainfall than stations in other parts of the country.

Table 1. Summary statistics of annual rainfall data for each studied station.

Location		Coding	Latitude	Longitude	Annual Mean	CV
Northwest Region	Alor Setar	NW1	6° 12' N	102° 15' E	2176.04	13.66
	Bayan Lepas	NW2	5° 18' N	100° 16' E	2367.96	17.49
	Baling	NW3	5° 68' N	100° 91' E	2232.14	16.89
	Dungun	E1	4° 46' N	103° 25' E	2872.19	21.16
Eastern Region	Kota Bharu	E2	6° 10' N	102° 18' E	2726.14	21.98
	Kuantan	E3	3° 46' N	103° 13' E	3015.48	19.31
	Pasir Mas	E4	06° 02' N	102° 07' E	2855.36	25.37
Central Region	Tanah Rata	C1	4° 28' N	101° 23' E	2604.46	15.48
	Pontian	SW1	01° 29' N	103° 23' E	1921.30	16.34
Southwest Region	Tangkak	SW2	02° 16' N	102° 32' E	1947.81	16.76
	Kluang	SW3	02° 01 'N	103° 19' E	2147.85	17.54
	Melaka	SW4	2° 16 'N	102° 15' E	1978.54	12.79
	Senai	SW5	1° 38' N	103° 40' E	2554.30	15.57
Western Region	Petaling Jaya	W1	3° 06' U	101°39' T	3384.41	11.01
	Lenggong	W2	5° 06' N	100° 58' E	2105.07	15.92
	Sitiawan	W3	4° 13' N	101° 42' E	1825.10	15.08

4.2 Smoothing rainfall data using Fourier Basis function

A Fourier basis function was employed to describe the rainfall pattern since the climate in Malaysia is influenced by the monsoon season. The rainfall data were smoothed using a Fourier basis function that consists of sine and cosine functions, which can be expressed in terms of harmonics. Seven basis functions were used to describe the rain patterns in the central and southwest regions, but only five were needed for the northwest and west. On the other hand, the rainfall pattern in the eastern region required the use of nine basis functions. The deviations for the three, five, and seven basis functions were still significantly higher. When nine basis functions were applied, however, the deviation was significantly decreased, and no more basis functions were needed. Consequently, nine basis functions, including four harmonics, could accurately depict the rainfall distribution over the eastern stations. Due to the high unpredictability and fluctuations in rainfall patterns observed at these rainfall stations, an increase in the number of basis functions is essential.

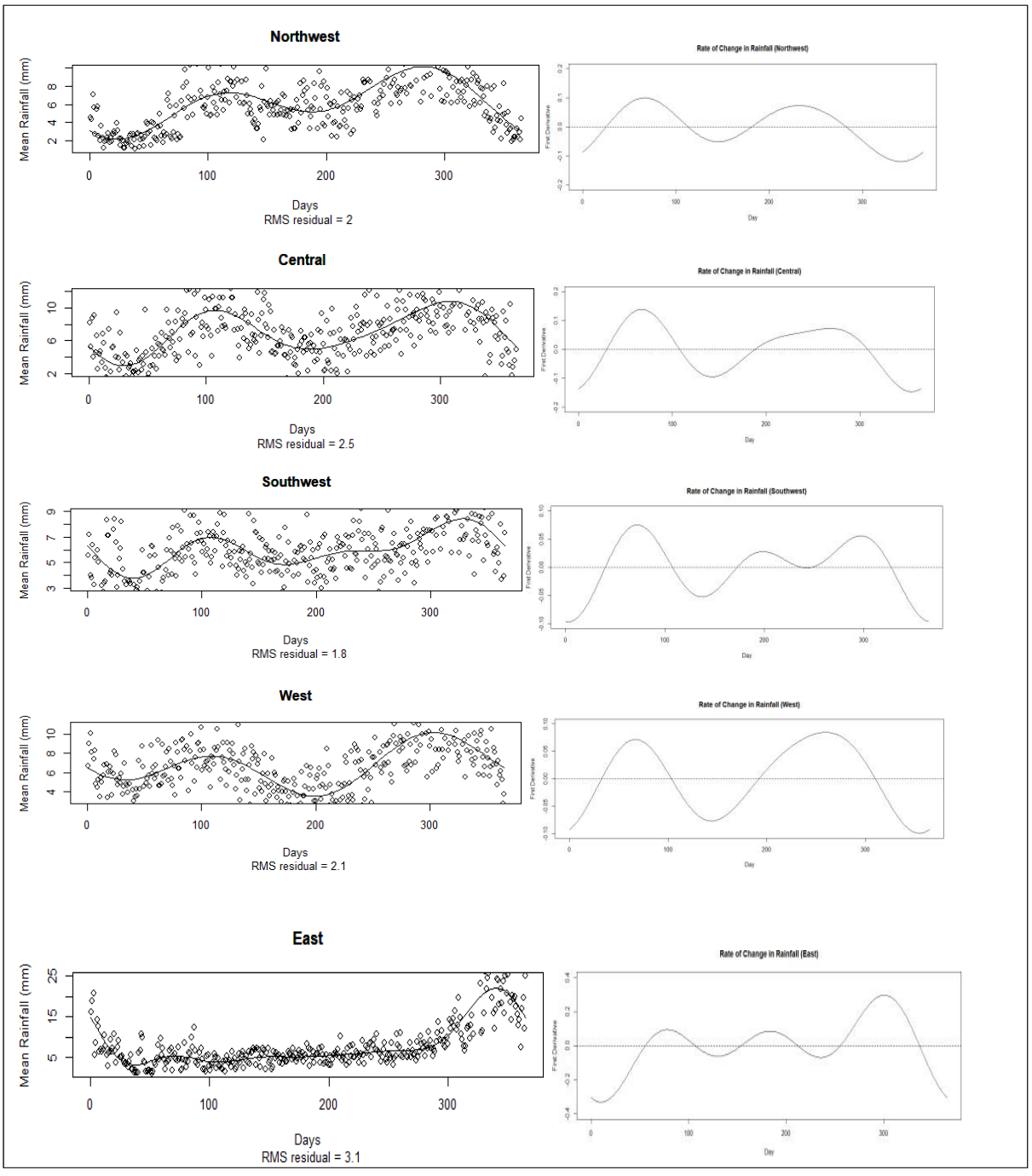


Figure 2. The right panel shows the observed mean daily rainfall with its corresponding smoothing curves for the regions and their rate of change in the left panel.

Figure 2 illustrates the smoothing curves for each region based on the required number of basis functions. The right panel of Figure 2 shows a bimodal rainfall pattern for the northwest, southwest, west, and central regions. During the inter-monsoon, the first peaks in these four regions occurred between April and May. It was demonstrated that the first peak was lower than the second peak. The temporal delay between the northwest and the other three regions was very different during the second peak, though. While rainfall peaked in November for the other three regions, it did so in September and

October for the northwest region. Rainfall decreased after the second peak and hit its lowest level in early February. But for the western regions, the lowest rainfall occurs in July. The left panel of Figure 2 shows the rate of change in rainfall. Positive changes in rainfall were observed from February to April and August to September for the west, southwest, northwest, and central regions, where the season changes from SWM and NEM to inter-monsoon, respectively. On the other hand, negative changes can be seen at the end and the beginning of the year.

Figure 2 demonstrates how the pattern of the rainfall stations in the eastern area differs from those in the other four regions. The range of the average daily rainfall is 5 to 25 mm. The rainfall has been seen to follow a unimodal trend, with November and December seeing the highest peak. Rainfall levels are low from mid-May to mid-September, increasing in October until it reaches its peak and drastically decreasing after mid-February. Based on the rate of change, a significant change in rainfall occurred during the NEM flow from October to early February, but only small tweaks occurred from mid-February to October.

4.3 Summary of functional data

Figure 3 illustrates the descriptive statistics of functional mean and functional variation for 16 stations from 1999 to 2019. It reveals that the highest mean rainfall was observed in November during the NEM season. Conversely, the lowest mean rainfall was recorded between February and March and July and August. The NEM brings heavy rain and stormy seas, and it occasionally causes flooding in the eastern part of Peninsular Malaysia. Rainfall variability is high between the end of November and early January, while other months have moderate fluctuations. The high rainfall variability reflects the influence of the northeast monsoon. During this period, the cold surge from northern Asia impacted most of the southern part of the South China Sea, including Malaysia. In addition, the Intertropical Convergence Zone (ITCZ), a cloud band that circles the globe around the equator and is frequently wetter than the southwest monsoon, impacts the area. Hence, a high rainfall peak was observed from November to January.

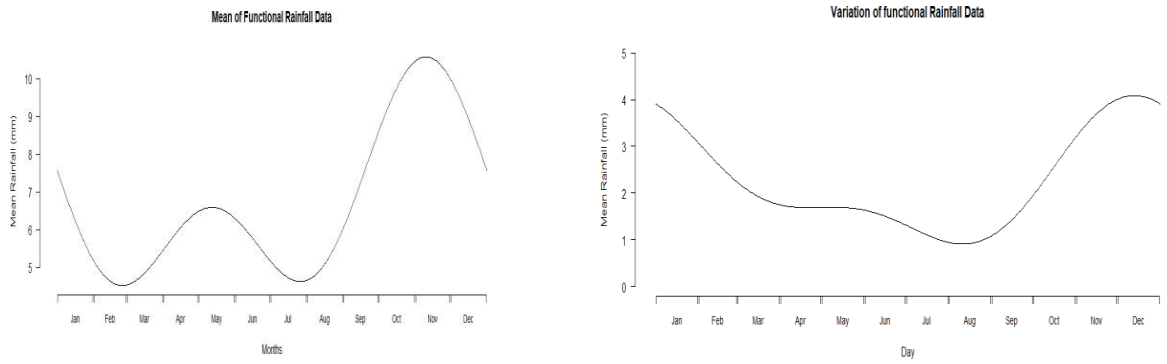


Figure 3. The smoothing mean and variance curves for all regions.

4.4 Estimating the variation of functional data via functional principal component analysis

Functional principal component analysis was carried out to identify the main sources of variation in the rainfall data. For the first two principal components, the variance rates were 70.1% and 17.3%, respectively. The overall variation in the rainfall data was explained by these two components, which accumulated to 87.4%. The first principal component (PC1) that showed the largest variation was seen in November, as shown in Figure 4. Contrarily, the second principal component (PC2) revealed that April had the greatest fluctuating rainfall values.

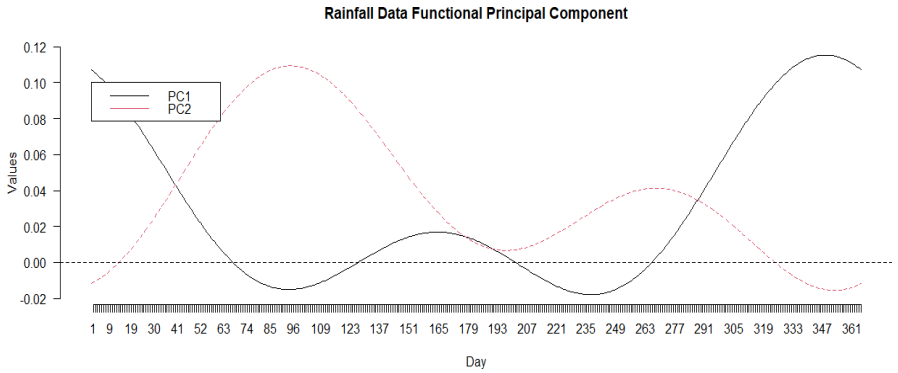
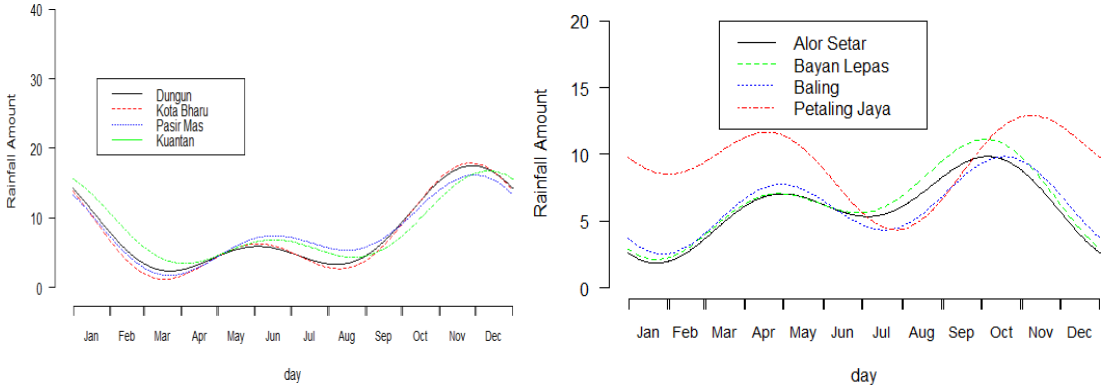


Figure 4. Centered Harmonic Functions.

A high correlation between the principal component and the given variable was demonstrated by the high loading score for that variable. Based on the score values of PC1, the highest score values were dominated by the eastern stations, as depicted in Figure 5(a). It shows that PC1 was associated with the eastern region of Peninsular Malaysia during the NEM. Rains during the inter-monsoon season from April to May and September to October played a significant role in forming the second component, PC2. Stations in the northwesterly and western areas dominated it, as shown in Figure 5(b). The interval between monsoons, or the "grey period," is when the southwest and northeast monsoons alternately change, which is referred to here as the inter-monsoon. During the southwest and northeast monsoon shifts, the weather plays havoc and suddenly switches from a hot, sunny day to a violent thunderstorm. The weather is cloudy and wet most of the time during the period.



(a) (b)

Figure 5. Smoothing curves with the highest loading for the (a) first principal component and (b) second principal component.

The scores of the first two principal components were then mapped onto Figure 6. The figure shows some exact groupings that reveal the most important type of variation in the rainfall curve. These stations are grouped based on their high PC1 scores: Kuantan, Dungun, Pasir Mas, and Kota Bharu. On the other hand, based on the shape of their curves, Alor Setar, Bayan Lepas, and Baling could be grouped together as a single cluster. In Figure 6, Petaling Jaya stands out as a distinct cluster since it has the highest PC2 score, suggesting that the area may have a distinct rainfall shape from other stations. Also, Tanah Rata in the central part of Peninsular Malaysia and Senai in the southwest could each have their cluster.

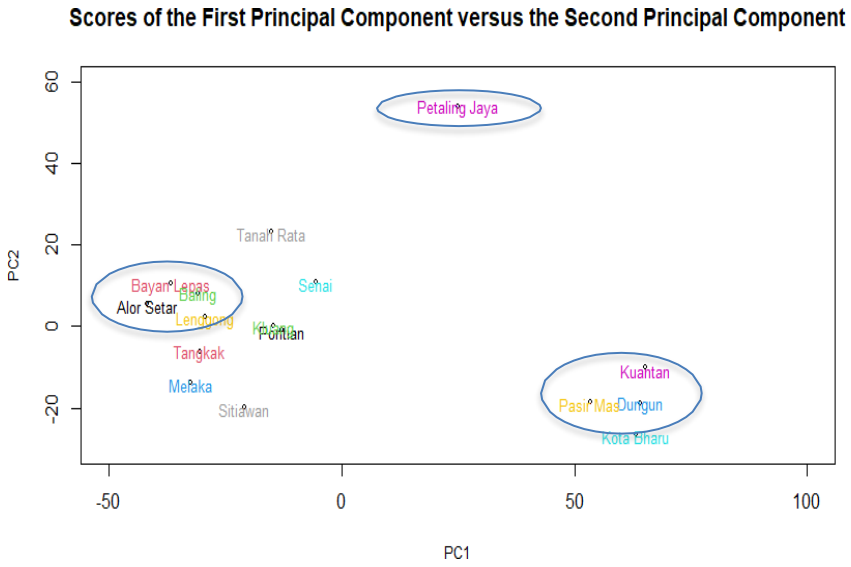


Figure 6. Map of scores for PC1 and PC2.

4. Conclusion

In this study, functional data analysis was used to convert rainfall data from each region into smoothing curves that can be analysed at any time interval. Additional information, such as the changes in rainfall, could be estimated based on the derivative of the smoothing Fourier function. A positive or negative change could be observed from the smoothing curve over the time interval. The functional findings of this study showed that more basis functions are needed than in the other regions to fully represent the variation in the pattern of rainfall over the eastern region. The highest rainfall is recorded during the NEM season, while the lowest rainfall occurs between mid-May and September during the southwest monsoon. In contrast, the northwest regions got less rain during the months when the northeast monsoon flowed, while the second inter-monsoon was the wettest time.

The rainfall patterns over Peninsular Malaysia could be adequately described by the first two principal components, which represented 87.4% of the total variation. According to the first principal component, the northeast monsoon season occurs most frequently in the east. In contrast, the first inter-monsoon season occurs most commonly in the regions represented by the second principal component. Additionally, the topographical and geographical implications will significantly impact Peninsular Malaysia's rainfall patterns. The current work is limited to exploring and profiling rainfall patterns. However, the findings from this study can provide additional information for those who work with the water management system as an alternative way to describe rainfall events, such as in terms of rainfall variability, rate of change, and potential anomalies. In addition, further research should consider inferential aspects in the analysis, such as comparing rainfall patterns using the functional analysis of variance, building up the relationship using the functional linear model, and forecasting rainfall values using functional time series.

References

[1] Alaya MAB, Ternynck C, Dabo-Niang S, Chebana F and Ouarda TBMJ 2020 Change point detection of flood events usinga functional data framework *Adv Water Resour* **137** 103522

[2] de Pinedo AR, Couplet M, Iooss B, Marie N, Marrel A, Merle E and Sueur R 2021 Functional Outlier Detection by Means of h-Mode Depth and Dynamic Time Warping *Appl. Sci.* **11**, 11475

[3] Dai W, Mrkvička T, Sun Y and Genton MG Functional outlier detection and taxonomy by sequential transformationsComput. Stat. Data Anal **149** 106960
[4] Ghumman AR, Rauf AU, Husnain Haider H and Shafiquzamman M 2020 Functional data analysis of models for predictingtemperature and precipitation under climate change scenarios J. Water Clim. Chang **11. 4** 1748-1765
[5] Wang D, Zhong Z, Bai K and He L 2019 Spatial and Temporal Variabilities of PM2.5 Concentrations in China UsingFunctional Data Analysis Sustainability **11** 1620
[6] Chebana F, Dabo-Niang S and Ouarda TBMJ 2012 Exploratory functional flood frequency analysis and outlier detectionWater Resour. Res. **48** W04514
[7] Suhaila J and Yusop Z 2017 Spatial and temporal variabilities of rainfall data using functional data analysis *Theor. Appl. Climatol.* **129** 229–242
[8] Hael MA 2021 Modeling of rainfall variability using functional principal component method: a case study of Taiz region, Yemen *Model. Earth Syst. Environ.* **7**, 17–27
[9] Newell J, McMillan K, Grant S and McCabe G 2006 Using functional data analysis to summarise and interpret lactatecurves *Comput. Biol. Med.* **36**(3) 262-275
[10] Ryan W, Harrison A and Hayes K 2006 Functional data analysis of knee joint kinematics in the vertical jump *Sports Biomech.* **5**(1) 121-138
[11] Song JJ, Deng W, Lee H-J and Kwon D 2008 Optimal classification for time-course gene expression data using functional data analysis *Comput Biol Chem.* **32**(6) 426-432
[12] Dong JJ, Wang L, Gill J and Cao J 2018 Functional principal component analysis of glomerular filtration rate curves after kidney transplant *Stat. Methods. Med. Res*. **27**(12) 3785-3796
[13] Zhang B, Zheng K, Huang Q, Feng S, Zhou S and Zhang Y 2020 Aircraft engine prognostics based on informative sensor selection and adaptive degradation modeling with functional principal component analysis *Sensors* **20**(3) 920
[14] Suhaila J 2021 Functional Data Visualization and Outlier Detection on the Anomaly of El Niño Southern Oscillation *Climate* **9** (118)

Acknowledgments

The authors would like to acknowledge the financial support from the Ministry of Education Malaysia under the Fundamental Research Grant Scheme (FRGS/1/2020/STG06/UTM/02/3) and the Universiti Teknologi Malaysia for the funding under UTMER (QJ130000.3854.19J58).