Journal of Quality Measurement and Analysis *JQMA* 21(3) 2025, 121-138
e-ISSN: 2600-8602 *http://www.ukm.my/jqma*
https://doi.org/10.17576/jqma.2103.2025.07

SCALAR-ON-FUNCTION REGRESSION FOR MODELING CLIMATE VARIABLE ASSOCIATIONS

(*Regresi Skalar-Atas-Fungsi untuk Pemodelan Hubungan Pembolehubah Iklim*)

YAP SOOK MEI & JAMALUDIN SUHAILA*

ABSTRACT

This study aims to investigate the scalar-on-functional relationship through functional regression, focusing on scalar responses, total annual rainfall, and functional predictors represented by daily temperature curves within the context of Malaysia's climate patterns. The distinctiveness of Malaysia's tropical climate, marked by high humidity and consistent temperatures throughout the year, underscores the importance of understanding this relationship for accurate climate modeling and forecasting. A key component of this study is the application of a roughness penalty in constructing Fourier basis functions for temperature regression coefficients, which ensures a smooth and flexible representation of the temperature curves. The study will compare models with a limited number of basis functions against those with a larger number of basis functions supplemented by an additional roughness penalty. The analysis utilises climate data from twelve stations across Peninsular Malaysia, spanning 2010 to 2017. The findings indicate that models incorporating a roughness penalty demonstrate superior performance, as the penalty helps mitigate overfitting by controlling excessive complexity in the estimated functions. Moreover, the results underscore the significant interactions between rainfall and temperature over time, offering critical insights into the dynamics of the Malaysian climate. These insights potentially enhance the region's water resource management, agricultural planning, and climate adaptation strategies.

Keywords: functional regression; scalar-on-function; roughness penalty

ABSTRAK

Kajian ini bertujuan untuk menyiasat hubungan skalar-atas-fungsi melalui regresi fungsi, memfokuskan pada pembolehubah bersandar skalar, seperti jumlah hujan tahunan, dan fungsi pemboleh ubah penerang yang diwakili oleh lengkung suhu harian dalam konteks corak iklim Malaysia. Keistimewaan iklim tropika Malaysia, yang dicirikan dengan kelembapan yang tinggi dan suhu yang konsisten sepanjang tahun, menekankan kepentingan memahami hubungan ini untuk pemodelan dan ramalan iklim yang tepat. Komponen utama kajian ini ialah penggunaan dendaan kasar dalam membina fungsi asas Fourier untuk pekali regresi suhu, bagi memastikan perwakilan licin dan fleksibel bagi lengkung suhu. Kajian ini akan membandingkan model dengan bilangan fungsi asas yang terhad dengan model yang mempunyai bilangan fungsi asas yang lebih besar ditambah dengan dendaan kasar. Analisis menggunakan data iklim daripada dua belas stesen di seluruh Semenanjung Malaysia, merangkumi 2010 hingga 2017. Keputusan kajian menunjukkan bahawa model yang menggabungkan dendaan kasar mempamirkan prestasi yang lebih baik, kerana ia membantu mengurangkan isu terlebih padanan dengan mengawal kerumitan yang berlebihan dalam fungsi yang dianggarkan. Selain itu, keputusan kajian menekankan hubungan yang ketara antara hujan dan suhu dari masa ke masa, memberikan pandangan kritis dalam dinamik iklim Malaysia. Kajian ini berpotensi meningkatkan pengurusan sumber air, perancangan pertanian dan strategi penyesuaian iklim di rantau ini.

Kata kunci: fungsi regresi; regresi scalar atas fungsi; dendaan kasar

1. Introduction

In the context of climate change, rising global temperatures drive substantial changes in rainfall patterns, consequently impacting local ecosystems, agriculture, water resources, and environmental sustainability. The observed increase in global temperatures has resulted in shifts in the distribution and intensity of rainfall, leading to more frequent and severe weather events such as floods and droughts (Lindwall 2022; Tabari 2020). According to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change (IPCC AR6), the average global temperature during the first two decades of the 21st century (2001–2020) has increased by approximately 0.99°C (with a range of 0.84°C to 1.10°C) above preindustrial levels (1850–1900) (Intergovernmental Panel on Climate Change 2021). This accelerated warming presents significant risks to society, the economy, and natural ecosystems, resulting in heightened public awareness of extreme weather events in recent years. For example, heatwaves are becoming more frequent and severe across many regions, posing substantial threats to human health, agriculture, and energy systems. The elevated evaporation rates associated with higher temperatures intensify drought conditions, exacerbating water scarcity issues in various parts of the world (Lai 2022). These alterations in rainfall patterns and the heightened occurrence of extreme weather events have profound implications for global food security, economic stability, and social well-being.

Extensive research has been undertaken to analyze historical climate data trends to understand the influence of climate on the frequency and intensity of extreme weather events (Iraqi and Abdallah 2022; Felix *et al*. 2021; Quan *et al*. 2021; Kunkel *et al.* 2020). Several Asian countries, including India, Malaysia, Bangladesh, China, Thailand, and Vietnam, have experienced exceptionally high temperatures due to a severe heatwave that began in April 2023. The Meteorological Department of Malaysia (MetMalaysia) has issued heatwave alerts for multiple states, forecasting that hot weather will persist until June, with above average temperatures and below average rainfall across the country (Saieed 2023). Adding to these concerns, Malaysia experienced a weak to moderate El Niño event starting in June 2023, projected to reduce rainfall by 20% to 40% and increase temperatures by 0.5ºC to 1.0ºC (Chu 2023). These projections align with the findings of Muhammad *et al*. (2024), who observed that Peninsular Malaysia has experienced a higher frequency of heatwaves since 2000. Their study assesses population exposure to heatwaves, indicating a gradual increase in spatial and temporal dimensions. The central south region of Peninsular Malaysia has the highest population affected by heat waves. The potential for more prolonged heatwaves, exacerbated by the El Niño effect, has led to warnings from health experts. While the study focuses on heat waves and their characteristics, it does not delve deeply into the statistical relationships between multiple climate variables beyond temperature.

To effectively address and mitigate the impacts of climate change, it is crucial to possess a comprehensive understanding of climatic trends and their relationship. Understanding the intricate relationship between climate variables is crucial for effective climate adaptation and mitigation strategies. Traditional statistical methods often fail to capture these climatic variables' continuous and dynamic nature, leading to suboptimal predictions and analyses. Hence, applying Functional Data Analysis (FDA) to climate data significantly advances the understanding and modeling of complex climate systems. Climate data, characterized by its high dimensionality and temporal dependencies, poses substantial challenges for traditional statistical methods. FDA offers a robust framework to handle these challenges by treating data as continuous functions rather than discrete observations, allowing for a deeper analysis of temporal and spatial patterns (Suhaila & Yusop 2017). The versatility of the FDA will enable it to handle irregularly spaced, unevenly sampled, and varying length data effectively. Studies

have demonstrated its effectiveness in modeling and predicting various climatic phenomena, such as temperature and precipitation patterns, sea level changes, and extreme weather events (Suhaila & Yusop 2017; Suhaila 2021; Kutrolli & Benth 2019; Alaya *et al.* 2020; Ghumman *et al.* 2020; Hael *et al.* 2020).

Suhaila and Yusop (2017) constructed functional data objects from discrete rainfall observations to analyze spatial and temporal patterns across Peninsular Malaysia. Their methodology employed smoothed curves to extract information from both the function and its derivatives. The researchers opted against using roughness penalties in their analysis due to the relatively small number of basis functions employed. Subsequently, Suhaila (2021) expanded this work by emphasizing FDA exploratory tools to identify functional outliers, which they found to be associated with El Niño and La Niña phenomena. This later study incorporated roughness penalties with optimized smoothing parameters to generate more refined curves. Despite these methodological advances, their findings remained primarily confined to visualization aspects rather than extending to deeper analytical insights.

Functional regression, a powerful technique within functional data analysis, is a functional version of regression analysis when responses or covariates include functional data. This method profoundly explains how these variables interact over time, capturing the underlying patterns and periodicities that drive weather phenomena. Functional regression has received considerable attention in various scientific fields because of its observed high-dimensional and complex data structures (Beyaztas & Shang 2020; Acal *et al.* 2021). The ability of functional regression to analyze dynamic dependencies between climatic variables is a significant strength. For example, by modeling temperature as a function and rainfall as a scalar, researchers can explore how temperature variations influence rainfall patterns over time. This approach provides a detailed understanding of the relationships between climatic variables, which is essential for accurate climate modeling and prediction.

Despite its potential, the application of functional regression in climatology, particularly in the context of the Malaysian climate, remains underexplored. This gap highlights the need for comprehensive research to develop functional regression models tailored to the unique climatic conditions of Malaysia. Addressing this gap is essential for improving our predictive capabilities and developing robust strategies for managing the impacts of climate change in this region. This study seeks to investigate the influence of the comprehensive temperature profile, including its specific temporal characteristics, on the total amount of rainfall instead of relying solely on discrete temperature values at specific time points. The functional regression model employed in this analysis is designed to identify which aspects or features of the temperature curve are most strongly associated with observed rainfall, thereby determining the relevant periods or patterns within the temperature data that serve as predictors of rainfall. A key issue in this analysis is the application of a roughness penalty in estimating the temperature regression coefficients $\beta(t)$ since the estimated $\beta(t)$ might be too wiggly, capturing noise rather than true temperature effects. This penalty is crucial for controlling overfitting and ensuring smoothness in the estimated relationships. Therefore, this study will compare the performance between the low dimensional $\beta(t)$ model with the high dimensional $\beta(t)$ that implements a roughness penalty.

2. Material and Methods

2.1. *Study area*

Peninsular Malaysia is strategically located within the tropical region, lying between latitudes 1° and 7°N and longitudes 100° to 103°E. This geographical positioning places it close to the equator, resulting in a hot and humid climate characteristic of tropical regions. The climate of Peninsular Malaysia is heavily influenced by its proximity to large bodies of water, including the South China Sea and the Straits of Malacca, as well as the monsoonal wind patterns that dominate the region. The region experiences relatively stable temperatures throughout the year, with minimal seasonal variation. However, the climate is profoundly shaped by the annual monsoons. The southwest monsoon (SWM), occurring from May to August, typically brings drier weather to the Peninsula, while the northeast monsoon (NEM), from November to February, is associated with heavier rainfall, particularly along the eastern coast (Suhaila and Yusop 2017). This seasonal variability in rainfall is further influenced by the inter-monsoon periods, during which the region experiences increased rainfall in the transitional months of March to April and September to October. This study utilizes daily rainfall and temperature data from 12 meteorological stations across Peninsular Malaysia, covering eight years from 2010 to 2017, provided by the Malaysian Meteorological Service. Figure 1 displays the physical map of Peninsular Malaysia, showing the locations of the stations selected for analysis.

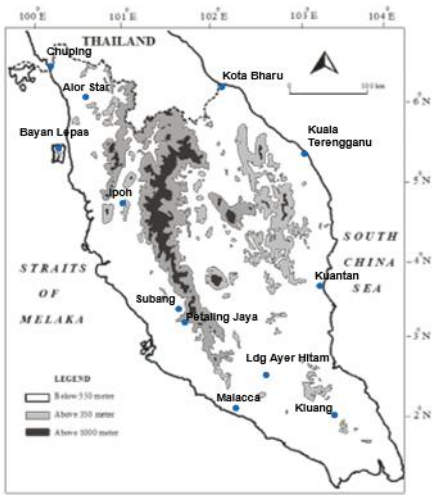


Figure 1: Map of the geographical locations of the studied stations.

2.2. *Functional data representation*

Suppose we have a data set $y_i(t_j)$ such as $\mathbf{Y}_i = \left(y_i(t_1), y_i(t_2), ..., y_i(t_T)\right)'$ for i = 1, 2, . . . , n discrete observations at time j=1, 2, . . . , T. These discrete observations are transformed into smoothing curves $X_i(t)$ as temporal functions with a base period of T = 365 days. The first step in the FDA is to build a set of basis functions that best represent the functional data, which can be written as

$$X_i(t) = \sum_{k=1}^{K} d_{ik}\phi_k(t) = \mathbf{d}'\phi(t), \tag{1}$$

where d_{ik} refers to the basis coefficient of each observation *i*, ϕ_k is the basis function and *K* is the maximum number of basis functions. Fourier basis functions are a set of functions used to represent functional data, particularly when dealing with functions that exhibit periodic or oscillatory behavior. It is suitable for examining annual trends with seasonal variation (Kutrolli and Benth 2019). It can be written in the form of sine and cosine functions as $X(t) = d_0 + d_1\sin(\omega t) + d_2\cos(\omega t) + ...$ defined by the basis function $\phi_0(t) = 1, \phi_{2k-1}(t) = \sin k\omega t, \phi_{2k}(t) = \cos k\omega t$, $t = t_1, t_2, ..., t_T$. The basis is periodic and the constant ω is related to the period *T* by the relation $\omega = \frac{2\pi}{T}$.

The coefficients of the expansion d_{ik} are determined by minimizing the least squares criterion;

$$SSE_i = \sum_{j=1}^{T}\left(y_i(t_j) - \sum_{k=1}^{K} d_{ik}\phi_k(t_j)\right)^2. \tag{2}$$

The model equation can be written in a matrix form as $\mathbf{y} = \mathbf{\Phi d} + \boldsymbol{\varepsilon}$ where **y** is the vector of observations at time *t*, **d** refers to the vector of coefficients d_{ik} while $\mathbf{\Phi}$ representing the value of the basis function ϕ_k, and epsilon $\boldsymbol{\varepsilon}$ is the error vector. The estimated coefficients can be solved using the following Eq. (3),

$$\mathbf{d} = \left(\mathbf{\Phi}^T\mathbf{\Phi}\right)^{-1}\left(\mathbf{\Phi}^T\mathbf{y}\right). \tag{3}$$

However, the least squares fitting criterion in Eq. (2) can be modified to avoid overfitting by adding the roughness penalty in the second term. The penalized residual sum of squares is given as

$$\mathbf{d} = \left(\mathbf{\Phi}^T\mathbf{\Phi} + \lambda\mathbf{R}\right)^{-1}\left(\mathbf{\Phi}^T\mathbf{y}\right) \tag{4}$$

with λ is a smoothing parameter, and **R** is a penalty matrix related to the second derivative of the function. A detailed explanation can be found in Suhaila *et al.* (2011), Suhaila and Yusop (2017), Ramsay and Silverman (2005).

2.3. *Functional descriptive statistics*

Functional descriptive statistics extend traditional descriptive statistics to the context of functional data, where observations are curves or functions. Let $x_i(t)$, $i = 1, 2, ..., n$ be a sample of curves or functions that fit the data. The mean function, $\bar{x}(t)$ is computed as $\bar{x}(t) = \frac{1}{n}\sum_{i=1}^{n} x_i(t)$ to measure the central tendency of the functional data. The variance

function $Var_X(t) = \frac{1}{n-1}\sum_{i=1}^{n}\left(x_i(t) - \bar{x}(t)\right)^2$ measures how much an individual function deviates from the mean function.

Suppose that we have another function, $z_i(s),\ i = 1,2,...,n$, then the covariance function of two smoothing functions $x_i(t)$ and $z_i(s)$ at time *t* and *s* is given as

$$Cov_{X,Z}(t,s) = \frac{1}{n-1}\sum_{i=1}^{n}\left(x_i(t) - \bar{x}(t)\right)\left(z_i(s) - \bar{z}(s)\right) \tag{5}$$

with $\bar{x}(t)$ and $\bar{z}(s)$ are the sample means of $x_i(t)$ and $z_i(s)$, respectively. Similarly, the cross-correlation function can be written as

$$Corr_{X,Z}(t,s) = \frac{Cov_{X,Z}(t,s)}{\sqrt{Var_X(t)Var_Z(s)}}. \tag{6}$$

2.4. *Scalar-on-functional regression*

A classical linear regression is often written in the form of

$$Y_i = \alpha + \sum_{p=1}^{P} X_{ip}\beta_p + \varepsilon_i, \text{ for } i = 1,2,...,n \tag{7}$$

where Y_i is the scalar response variable, α refers to the intercept term, X_{ip} is the covariate or predictor variables and β_p represents the regression coefficient while ε_i is the random error with the assumptions that $\varepsilon_i \sim N\left(0,\sigma^2\right)$. A functional linear regression is introduced to model the relationship between a scalar response Y and a functional predictor $X(t)$, called scalar-on-function regression. Here, the functional predictor is a curve or function observed over time, while the response is a single scalar value.

The functional linear model is an extension of the model linear regression in Eq. (7) which can be expressed as

$$Y_i = \alpha + \int_T X_i(t)\beta(t)dt + \varepsilon_i \tag{8}$$

with $X_i(t)$ is the functional predictor observed over domain *T,* while $\beta(t)$ is the coefficient function that describes how the functional predictor relates to the response variable. Since $\beta(t)$ is a function, then the model will yield an infinite-dimensional beta coefficient despite being based on a finite number of observations *n* (Ramsay *et al*. 2009). Hence, it is high possibility that we can identify many functions that perfectly satisfies the model, yielding zero errors. To address this issue, Cardot and Sarda (2006) recommended using basis expansion for $\beta(t)$.

The expansion of the basis function $\beta(t)$ is given as

$$\beta(t) = \sum_{l}^{L} b_l \psi_l(t) \tag{9}$$

where b_l refers to the basis coefficient and ψ_l is the basis function and L is the maximum number of basis functions. Eq. (9) can then be rewritten as

$$Y_i = \alpha + \int_T \sum_{k=1}^{K} d_{ik} \phi_k(t) \sum_{l}^{L} b_l \psi_l(t) dt + \varepsilon_i$$

which is equivalent to

$$Y_i = \alpha + \sum_{l=1}^{L} b_l \left(\sum_{k=1}^{K} d_{ik} \int_T \phi_k(t) \psi_l(t) dt \right) + \varepsilon_i . \tag{10}$$

Suppose that $z_{il} = \sum_{k=1}^{K} d_{ik} \int_T \phi_k(t) \psi_l(t) dt$ then Eq. (10) can be written as $Y_i = \alpha + \sum_{l=1}^{L} b_l z_{il} + \varepsilon_i$. This is now a standard linear regression model with z_{il} as the predictors. The coefficients b_l then can be estimated by minimizing the least squares criterion, which is given as

$$\sum_{i=1}^{n} \left(Y_i - \alpha - \sum_{l=1}^{L} b_l z_{il} \right)^2 . \tag{11}$$

After estimating the coefficients $\hat{b}_l$, the estimated coefficient function $\hat{\beta}(t)$ can be reconstructed as $\hat{\beta}(t) = \sum_{l}^{L} \hat{b}_l \psi_l(t)$.

The roughness penalty approach can be employed to exert greater control over the smoothing process, particularly in conjunction with a high-dimensional basis. The method effectively reduces the risk of overlooking significant features or incorporating irrelevant ones (Ramsay *et al.* 2009). However, in some cases, satisfactory results can be achieved without resorting to this technique if the number of basis functions is smaller than the number of individuals in the sample.

The roughness penalty typically involves the derivative of the coefficient function and a common form is based on the integral of the squared second derivative. The estimation $\beta(t)$ involves minimizing a penalized least squares criterion, which combines the residual sum of squares with the roughness penalty. The penalized residual sum of squares is defined as

$$PENSSE(\beta(t)) = \arg\min \left(\sum_{i=1}^{n} \left(Y_i - \alpha - \int_T X_i(t) \beta(t) dt \right)^2 + \lambda \int_T \left(\frac{d^2 \beta(t)}{dt^2} \right)^2 dt \right) \tag{12}$$

with λ is the smoothing parameter that represents a compromise between the fit of the model to the data and the smoothness of $\beta(t)$. Detail explanation of the smoothing parameter for functional regression can be found in Ramsay *et al*. (2009) and López *et al.* (2022).

Estimating the regression coefficient function $\beta(t)$ in scalar-on-function regression involves expressing both the functional predictor $X(t)$ and $\beta(t)$ in terms of basis functions, transforming the problem into a linear regression form, and then estimating the coefficients using standard regression techniques. The resulting estimated coefficient function reveals how the functional predictor $X(t)$, influences the scalar response Y across different values of t. Specifically, positive coefficient values $\hat{\beta}(t)$ suggest a positive influence on Y, whereas negative values $\hat{\beta}(t)$ indicate a negative effect.

2.5. *Measure the adequacy of the functional linear model*

The coefficient of determination R^2, represents the proportion of variance in the scalar response variable that is accounted for the predictor variables in the model. A higher R^2 indicates a better fit of the model to the data, suggesting that the functional predictors are more effective in explaining the variability for the scalar response. The calculation of R^2 follows the standard formulation as

$$R^2 = 1 - \frac{SSE}{SST} \tag{13}$$

in which the residuals sum of squares (SSE) and total sum of squares (SST) are defined as follows:

$$SSE = \sum_{i=1}^{n} \left(y_i - \hat{y}_i \right)^2 , \qquad SST = \sum_{i=1}^{n} \left(y_i - \bar{y} \right)^2 \tag{14}$$

with y_i is the observed scalar response, $\hat{y}_i$ is the fitted values and $\bar{y}$ is the average of the response variable. On the other hand, the statistical test, F-test, is conducted to determine the overall effect of the functional predictor $X(t)$ on the model. The testing is computed as

$$F_{Test} = \frac{SST - SSE/df(\text{model})}{SSE/df(\text{error})} \tag{15}$$

with df(model) representing the degrees of freedom associated with the model where it represents the number of basis function l use for $\beta(t)$. On the other hand, df(error) representing the degrees of freedom associated with the error with n-l-1. However, in a penalized regression framework, the effective degrees of freedom (EDF) replace the traditional degrees of freedom. EDF accounts for the influence of the penalty on the estimation of $\beta(t)$. Normally in computer algorithm, the p-value is used to determine the significance of the model. A small p-value indicates that the null hypothesis can be rejected, suggesting that the functional predictor statistically affects the scalar response.

In a functional setup, the test statistic distribution under the null hypothesis is complex and challenging to derive (Ramsay *et al.* 2009; Cuevas *et al*. 2004; Suhaila and Yusop 2017). Consequently, the permutation test is employed to approximate the null distribution of the test statistic and to determine an estimated critical value for the test. This procedure involves rearranging the vector of responses while maintaining the order of the covariates and then

refitting the model. The *p*-value for the test is calculated as the proportion of permuted F_{perm} greater than the F_{Test} of the observed data.

3. Results and Discussion

This section is divided into two main subsections. The first part focuses on the exploratory tools of functional data analysis. These include smoothing processes for rainfall and temperature data, summary statistics of functional data, like the mean, standard deviation, cross-covariance, and cross-correlation functions. The second subsection will focus on the inferential aspect of functional data, highlighting functional regression with a scalar response and predictor functions.

3.1. *Functional smoothing of rainfall and temperature*

This study aims to analyze the fluctuations in rainfall and temperature at each station throughout the year. The mean daily rainfall and mean daily temperature were computed for each station at each time t with T = 365 days. Since the climate in Malaysia is strongly influenced by the monsoon seasons, we employ the Fourier basis as the basis function to convert discrete climate data into a smoothing function. The *R* package "fda.usc" and a specific algorithm called optim.basis is used to find the optimal number of basis functions and their smoothing parameter for roughness penalty. Generalized cross-validation (GCV) was applied in the analysis to choose the best smoothing parameter λ with the optimal basis function. The lowest value of GCV will give the optimal basis functions and smoothing parameter. A detailed explanation has already been discussed in Suhaila *et al.* (2011), Suhaila and Yusop (2017), Ramsay and Silverman (2005).

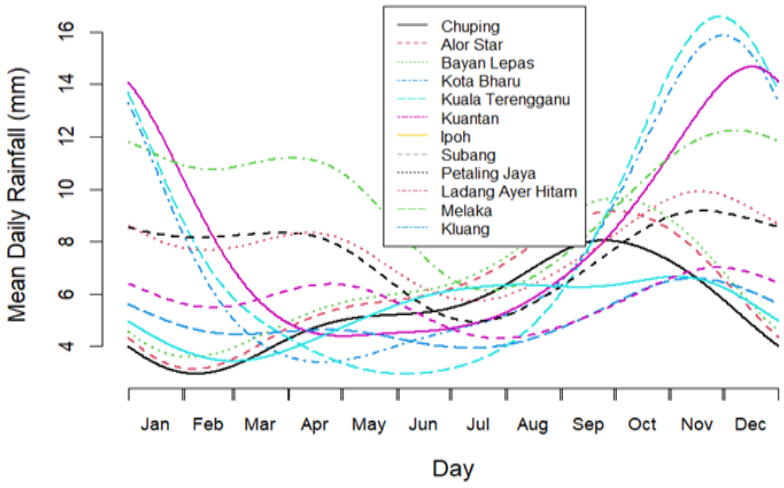


Figure 2: Smoothing rainfall curve using 73 Fourier basis functions with penalized parameter, $\lambda=10^6$

The analysis begins by applying a Fourier basis function to smooth the raw rainfall and temperature data. Figure 2 displays the smoothed rainfall curves for all 12 stations under study. Based on the lowest value of the GCV, the optimal model utilizes 73 basis functions with a corresponding smoothing parameter, $\lambda = 10^6$. The high number of basis functions provides flexibility in representing station-specific patterns. In contrast, the large smoothing parameter successfully prevented excessive wiggliness without over-smoothing important

features. The curves follow seasonal patterns while filtering out what would likely be noise from day-to-day weather fluctuations. The smoothing level effectively captures the general trends in rainfall patterns while preserving meaningful differences between stations. Overall, the result suggests this parameterization struck a good balance.

Rainfall patterns across the Peninsula demonstrate significant spatial variability. East coast stations, namely Kota Bharu, Kuantan, and Kuala Terengganu, exhibit a unimodal rainfall pattern, with prominent peaks at the beginning and end of the year corresponding to the NEM period. In contrast, other stations observe additional rainfall peaks during the inter-monsoon periods in April and October. Northwestern stations, including Chuping, Alor Star, and Bayan Lepas, manifest a more substantial rainfall peak during the second inter-monsoon period (October) relative to the first inter-monsoon phase. Western Peninsula stations, encompassing Subang, Petaling Jaya, Ipoh, and Ladang Ayer Hitam, display a distinct bimodal rainfall regime, with initial maxima occurring in the April-May interval and secondary maxima manifesting between November and early December. As evidenced by the data presented in Figure 2, February through July constitutes a comparatively dry period for the majority of monitoring stations throughout the region.

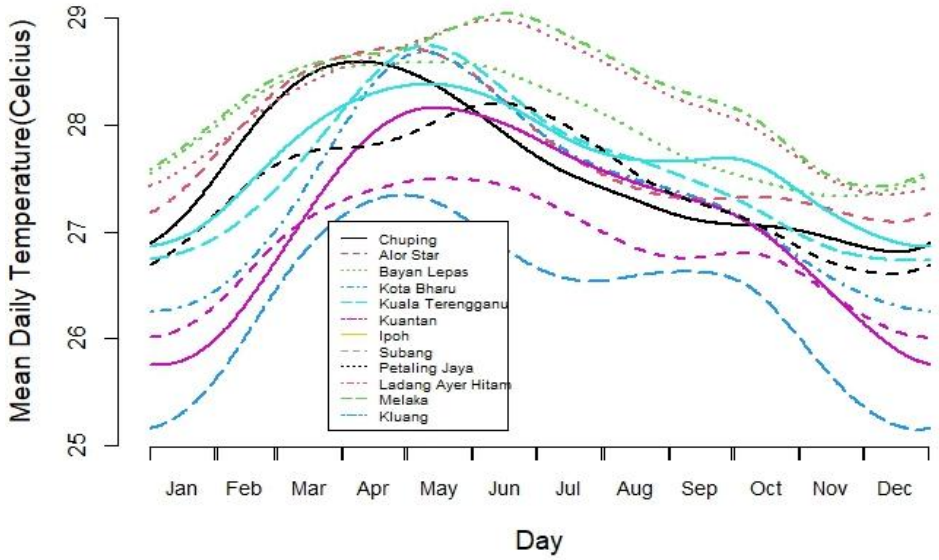


Figure 3: Smoothing temperature curve using 65 Fourier basis functions with penalized parameter, $\lambda=10^5$

A similar procedure was employed to smooth the temperature data, using a Fourier basis function with 65 basis functions and the corresponding smoothing parameter. The smoothed temperature curves are presented in Figure 3. The curves are smooth enough to filter out noise and random fluctuations. In addition, the curves do not show the excessive wiggliness that would indicate noise. The smoothing parameter appears appropriate for temperature data, allowing for meaningful interpretation of seasonal temperature patterns across different stations.

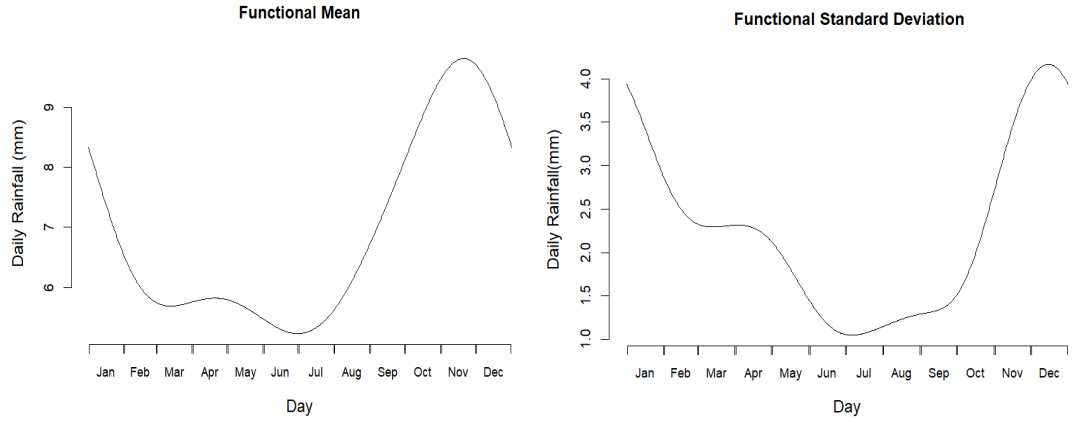
Temperatures range approximately between 25°C and 29°C across all locations. The seasonal patterns reveal slight variations in temperature across the studied stations over the years. Most stations recorded their highest temperatures between April and May and June and July. There is a consistent cooling trend moving toward December for most stations. Kluang, Ladang Ayer Hitam, and Kuantan recorded relatively low mean temperatures. In contrast, stations in the northwest, including Chuping, Alor Star, Bayan Lepas, Subang, and Petaling

Jaya in the west Peninsula, experienced higher mean temperatures, indicating that these stations are warmer than other stations.

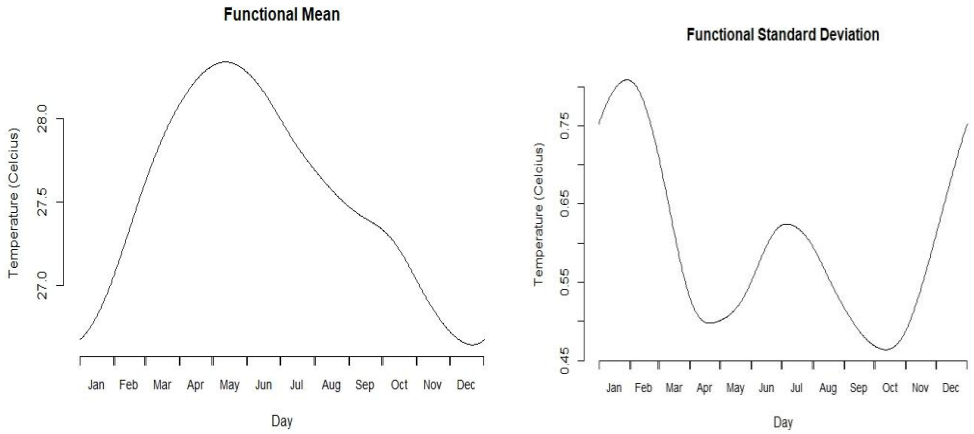
3.2. *Functional descriptive statistics*

Figure 4 presents summary statistics of rainfall and temperature data, consisting of their mean and standard deviation, across all studied regions. Dry periods are identified from February to July, as depicted in Figure 4(a). The rainfall patterns are mainly characterized by a low functional mean rainfall during those months. The analysis reveals two annual rainfall peaks, the first occurring during the inter-monsoon period (April-May) and the second in November, coinciding with the NEM season. Notably, there is substantial variability in rainfall across the studied stations during the NEM months, whereas lower variability is observed between June and July.

Figure 4(b) highlights a high functional mean temperature during April, May, and June, contrasting with the low functional mean rainfall observed during these months, as depicted in Figure 4(a). The cooling period, which exhibits low mean temperature, is observed during the NEM period. Additionally, the standard deviation plot for temperature indicates significant variations in mean temperature across the stations during the NEM period.



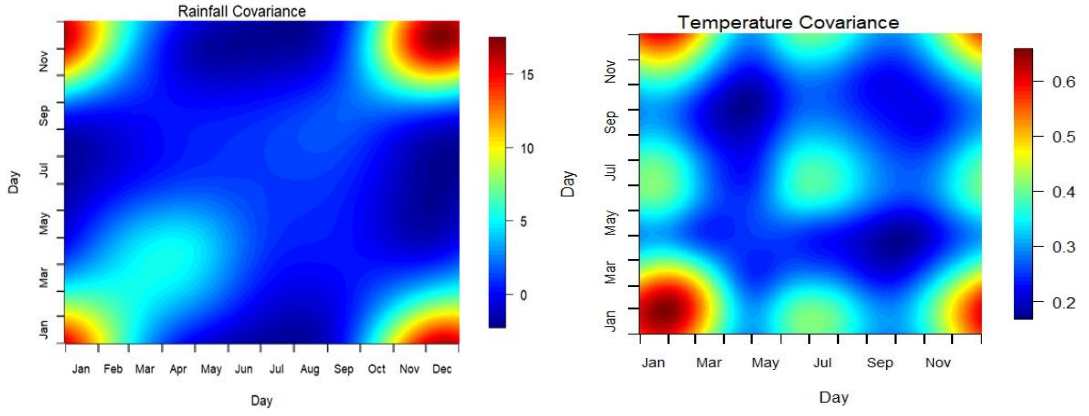
(a)



(b)

Figure 4: The functional mean and standard deviation for (a) rainfall and (b) temperature curves

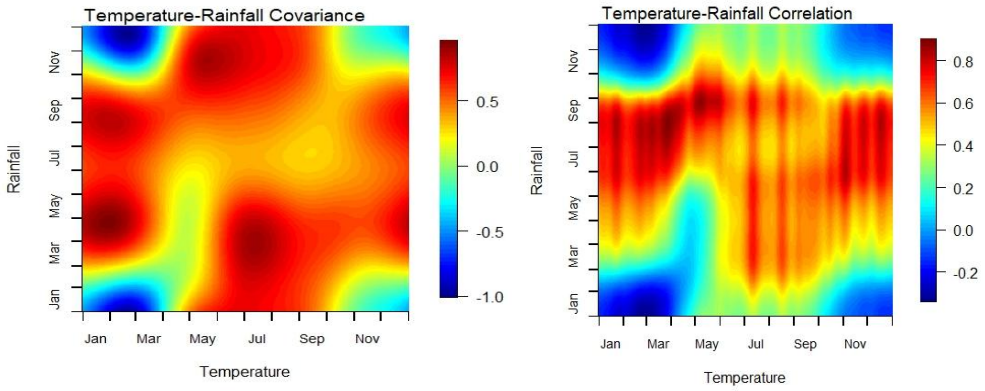
The variance-covariance function defines the covariance between rainfall and temperature curves at t and s. Figure 5 presents heat maps illustrating these relationships for both rainfall and temperature. The highest variability in rainfall among the studied stations is observed at the beginning and end of the year, as shown in Figure 5(a), corresponding to the NEM period. In contrast, lower rainfall variability is seen for other months. In Figure 5(b), temperature variability is most pronounced in the early part of the year, with smaller fluctuations occurring during the remaining months.



(a) (b)

Figure 5: Variance-covariance bivariate function for (a) rainfall and (b) temperature curves

Figure 6 presents the cross-covariance and cross-correlation values between rainfall and temperature. The functional covariance analysis offers insights into the temporal relationships between these variables across different periods. As depicted in Figure 6(a), there is a strong positive variability observed for rainfall during the SWM associated with temperature during the NEM seasons, and vice versa. Conversely, negative covariances are observed between rainfall and temperature at the beginning and end of the year, which suggests an inverse relationship, where periods of high rainfall are linked to lower temperatures. Figure 6(b) illustrates that the correlation values between rainfall and temperature range from -0.4 to 0.8. Low negative correlations between rainfall and temperature during the NEM season (early and end of the year) suggest that cooler periods are associated with increased rainfall. However, strong positive correlations between rainfall during the SWM season and temperature in the NEM season indicate that warmer temperatures are linked with higher rainfall, possibly due to increased evaporation and convection. In summary, the applications of functional descriptive statistics provide a deeper understanding of climate relationships than traditional univariate and multivariate methods, offering more comprehensive information beyond a single value or matrix. Understanding this can provide insights into the climate system's temporal dynamics and help predict future conditions.



(a) (b)

Figure 6: (a) Cross-covariance and (b) cross-correlation functions between rainfall and temperature

3.3. *Predicting scalar response based on functional predictor*

This section analyzes the rainfall data as a scalar response, while the temperature data is treated as a functional predictor. To predict the total annual rainfall for the 12 studied stations based on the temperature profile, let AR_i represent the logarithm of the total annual rainfall at station i, and $Temp_i(t)$ denote the daily temperature function at time t. Similar to Eq. (8), the model equation can be written as

$$AR_i = \alpha + \int_T Temp_i(t)\beta(t)dt + \varepsilon_i \tag{16}$$

where $\beta(t)$ refer to regression coefficient. The first step illustrates the use of a low-dimensional basis function, while the second step includes a large number of basis functions with a roughness penalty.

3.3.1. *Low-dimensional basis regression coefficient β(t)*

The most straightforward regression analysis approach employs a low-dimensional basis for $\beta(t)$ to achieve a smooth fit. The analysis will commence with 65 Fourier basis functions to represent the temperature curves while five basis functions for the regression coefficient and use a constant function as the intercept coefficient. The choice of five basis functions is deliberate, ensuring that their number is less than the number of observations, which is 12 (corresponding to the number of stations). Estimating $\beta(t)$ and the intercept term is done using the least squares regression method. The intercept coefficient is 1.5269, and the estimated regression coefficients for predicting the logarithm of annual rainfall with five Fourier basis functions are illustrated in Figure 7.

Figure 7 depicts the estimated regression coefficient values $\beta(t)$ for temperature data, providing crucial insights into the temporal relationship between temperature and rainfall. The smoothing with a small number of basis function appears reasonably balanced while capturing meaningful seasonal variation without introducing complex patterns. The dashed lines represent the 95% confidence limits for the $\beta(t)$ values. Positive $\beta(t)$ values observed during May through August (Southwest Monsoon period) indicate a strong positive correlation between temperature and rainfall, suggesting elevated temperatures correspond with increased rainfall during this interval. The confidence intervals (lower bound in green, upper bound in

red) exclude zero during this period, confirming temperature's substantial influence on rainfall patterns. This finding supports the cross-correlation analysis results discussed previously.

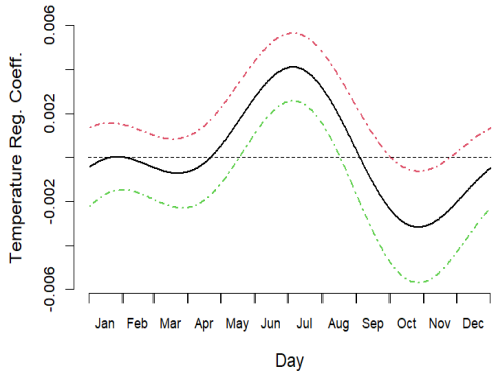


Figure 7: Estimated regression coefficients for predicting log annual rainfall with five basis regression coefficient $\beta(t)$

Conversely, negative regression coefficients emerge from February through April, potentially indicating an inverse relationship where lower temperatures correlate with enhanced rainfall. However, the confidence intervals during early months encompass zero, suggesting temperature exerts minimal influence on rainfall during this period. In contrast, the negative regression coefficients observed between October and November and confidence intervals that exclude zero demonstrate temperature's significant influence on rainfall patterns during these months despite the inverse relationship.

The squared multiple correlation coefficient (R^2) was employed to evaluate model fitness. As shown in Eq. (13), the calculated R^2 value of 0.898 demonstrates a strong relationship between rainfall and temperature variables, indicating that approximately 90% of the variance in the response variable can be attributed to temperature fluctuations. Consequently, the model generates predictions that closely align with the empirical observations presented in Figure 8. Mean square error (MSE) provided an additional metric for assessing model performance, with a computed value of 0.001491. Furthermore, the F-statistic of 10.531 (df = 5, 6) yielded a p-value below the 0.05 significance threshold, confirming that the temperature function exerts a statistically significant effect on rainfall patterns. These combined statistical indicators substantiate the model's validity and predictive capacity.

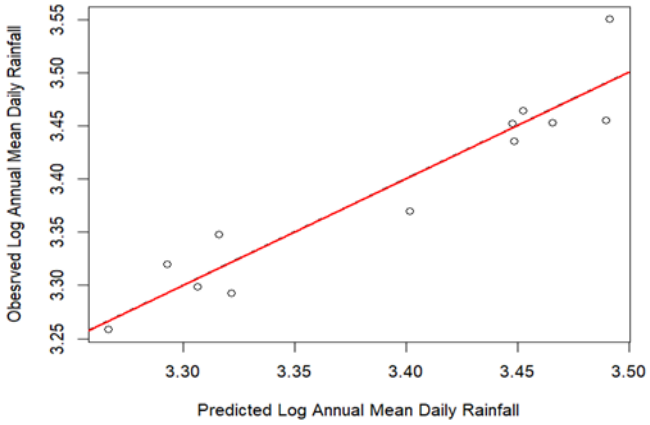


Figure 8: Predicted and observe value of log annual mean daily rainfall

3.3.2. *High-dimensional basis regression coefficient $\beta(t)$ with roughness penalty*

The best way to control the smoothness while determining the regression coefficient $\beta(t)$ is by introducing the roughness penalty term. In this section, 65 Fourier basis functions are used to represent the temperature curves while 11 Fourier basis functions are applied for $\beta(t)$. Figure 9 plots the cross-validation score against the logarithms of various values of λ. We choose $\lambda = 10^{10}$ for the final fit, corresponding to the lower minimum in the plot.

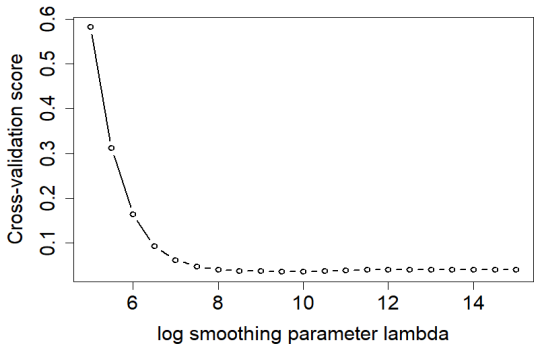


Figure 9: The cross-validation score function for fitting log annual rainfall by daily temperature variation, with a penalty on the size of harmonic acceleration

Figure 10 presents the estimated regression function obtained using the roughness penalty approach and pointwise 95% confidence intervals. This curve shows less wiggliness than the curve in Figure 7, suggesting a more controlled smoothing process. It appears to have a good balance in the smoothing process since it captures the essential seasonal pattern while removing potential noise. This suggests the effective use of roughness penalties to achieve a smoothing level. The dashed lines represent the 95% confidence limits for the $\beta(t)$ values. The intercept value is calculated as 0.00232. During the early months of the year, the confidence intervals encompass zero, indicating that temperature has negligible influence on rainfall during this period. In contrast, a pronounced temperature effect emerges from May to August, corresponding to the Southwest Monsoon (SWM) season, when rainfall peaks significantly. This trend gradually declines from October through December. The influence of temperature on rainfall proves highly significant during these periods, as evidenced by confidence intervals that do not include zero. Overall, the observed pattern effectively highlights a contrast between temperatures during the SWM and Northeast Monsoon (NEM) periods, with the model favoring stations that experience relatively warmer conditions during the SWM and cooler conditions during the NEM.

The squared multiple correlation coefficient was calculated as 0.887, with an *F*-statistic of 12.161 (effective degrees of freedom (EDF) = 5.29, denominator *df* = 6.72). The mean square error was computed as 0.001487. Compared to previous *F*-test results, the model with the roughness penalty showed greater statistical significance based on the lower mean square errors, indicating improved performance.

A permutation approach was implemented to further validate the model by randomly rearranging the response vector while maintaining the original predictor order. This procedure was iterated 1,000 times, yielding a *p*-value of 0.000. This value, falling well below the 0.05 significance threshold, provides compelling evidence that the temperature function exerts a statistically significant influence on precipitation patterns. The combination of these statistical

indicators strongly supports the validity of the proposed model with roughness penalty and confirms the substantial impact of temperature variation on rainfall dynamics.

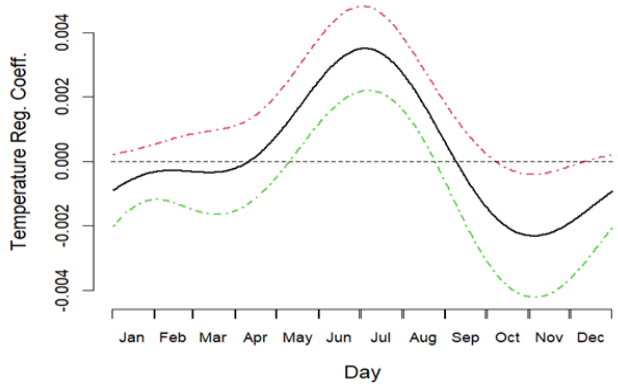


Figure 10: Estimate $\beta(t)$ for predicting log annual rainfall from average daily temperature with a harmonic acceleration penalty and smoothing parameter set to 10^{10}

Figure 11 displays the normal Q-Q plot to assess the normality of the residuals from our regression model. This normality check confirms that the assumption of normally distributed residuals is met. It verifies the accuracy of the regression model results and ensures the reliability of the analysis.

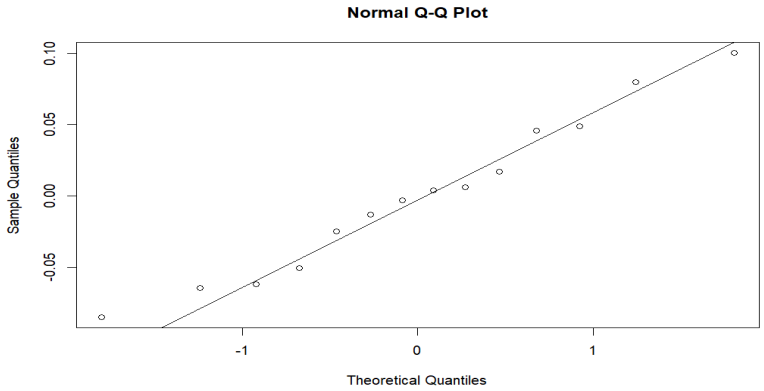


Figure 11: Normal Q-Q plot of residuals

4. Conclusion

In conclusion, this study demonstrates the utility of functional descriptive statistics and functional regression models in analyzing climate data, specifically the temporal patterns of rainfall and temperature across twelve stations in Malaysia. By leveraging the entire function of the data rather than discrete points, these methods allow for more accurate and interpretable predictions, uncovering significant insights into the relationship between rainfall and temperature. The findings highlight the strong influence of the Northeast Monsoon and Southwest Monsoon seasons on the variations and associations between these climate variables. Functional covariance and correlation have proven effective in revealing the temporal interdependence between rainfall and temperature, identifying periods of strong or

weak associations, and detecting any lagged or synchronous relationships. Moreover, the application of scalar response functional regression, including a roughness penalty, has shown to be particularly effective in capturing the intricate relationships between rainfall and temperature. The roughness penalty helps control overfitting, ensuring that the model balances complexity and accuracy, thereby providing a robust framework for climate analysis within the context of Malaysia's climate.

The findings suggest that scalar-on-functional regression can be a powerful tool in understanding and predicting climate-related phenomena in Malaysia, helping to inform better decision-making in the face of climate variability and change. The model seeks to improve the accuracy of rainfall predictions by leveraging the detailed information contained within temperature functions, offering a detailed analysis rather than models relying solely on scalar temperature summaries.

Future research directions could include (i) incorporating additional functional predictors such as humidity, atmospheric pressure, and wind velocity to develop more comprehensive climate models; (ii) examining sub-regional variations within Malaysia to account for localized climate effects; (iii) extending the temporal scope to investigate long-term climate trends; and (iv) developing spatiotemporal functional models that simultaneously address both spatial and temporal dimensions of climate variability. Such advancements would further enhance the precision of climate prediction systems and strengthen adaptive capacity in response to emerging climate challenges in Malaysia and comparable tropical regions.

Acknowledgment

The authors would like to express their gratitude to Universiti Teknologi Malaysia for the funding given under the Research University Grant Scheme Q.J130000.3854.23H36.

References

Acal C., Escabias M., Aguilera A.M. & Valderrama M.J. 2021. COVID-19 data imputation by multiple function-on-function principal component regression. *Mathematics* **9**(11): 1237.

Alaya M.A.B, Ternynck C., Dabo-Niang S., Chebana F. & Ouarda T.B.M.J. 2020. Change point detection of flood events using a functional data framework. *Advances in Water Resources* **137**: 103522.

Beyaztas U. & Shang H.L. 2020. On function-on-function regression: partial least squares approach. *Environmental and Ecological Statistics* **27**:95–114.

Cardot H. & Sarda P. 2006. Linear regression models for functional data. In Sperlich S., Härdle W. & Aydınlı G. (eds.). *The Art of Semiparametrics*: 49–66. Contributions to statistics. Physica-Verlag HD.

Cuevas A., Febrero M. & Fraiman R. 2004. An anova test for functional data. *Computational Statistics & Data Analysis* **47**(1): 111–122.

Felix M.L., Kim Y.-K., Choi M., Kim J.-C., Do X.K., Nguyen T.H. & Jung K. 2021. Detailed trend analysis of extreme climate indices in the Upper Geum River Basin. *Water* **13**(22): 3171.

Ghumman A.R., Rauf A.-U., Haider H. & Shafiquzamman M. 2020. Functional data analysis of models for predicting temperature and precipitation under climate change scenarios. *Journal of Water and Climate Change* **11**(4): 1748–1765.

Hael M.A., Yongsheng Y. & Saleh B.I. 2020. Visualization of rainfall data using functional data analysis. *SN Applied Sciences* **2**(3): 461.

Intergovernmental Panel on Climate Change. 2021. Summary for policymakers. In Masson-Delmotte V., Zhai P., Pirani A., Connors S.L., Péan C., Berger S., Caud N., Chen Y., Goldfarb L., Gomis M.I., Huang M., Leitzell K., Lonnoy E., Matthews J.B.R., Maycock T.K., Waterfield T., Yelekçi O., Yu R. & Zhou B. (eds.). *Climate change 2021: The physical science basis. Contribution of Working Group I to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change*: 3-32. New York, NY: Cambridge University Press.

Iraqi A.A. & Abdallah A.M. 2022. Analysis of long-term climatic changes at Al-Hodeidah-Yemen during the period between 1985 and 2019. *Theoretical and Applied Climatology* **150**(3): 1067–1081.

Kunkel K.E., Karl T.R., Squires M.F., Yin X., Stegall S.T. & Easterling D.R. 2020. Precipitation extremes: Trends and relationships with average precipitation and precipitable water in the contiguous United States. *Journal of Applied Meteorology and Climatology* **59**(1): 125–142.

Kutrolli G. & Benth F.E. 2019. An application of functional data analysis to forecast weather variables. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.3766459.
Lai O. 2022. Water Shortage: Causes and Effects. *Earth.Org.* https://earth.org/causes-and-effects-of-water-shortage/ (26 June 2022).
Lindwall C. 2022. What Are the Effects of Climate Change? *NRDC*. https://www.nrdc.org/stories/what-are-effects-climate-change (24 October 2022).
López O.A.M., López A.M. & Crossa J. 2022. *Multivariate Statistical Machine Learning Methods for Genomic Prediction*. Cham, Switzerland: Springer Nature.
Muhammad M.K.I., Hamed M.M., Harun, S., Sa'adi Z., Sammen S.S., Al-Ansari N., Shahid S. & Scholz M. 2024. Heatwaves in Peninsular Malaysia: a spatiotemporal analysis. *Scientific Reports* **14**: 4255.
Quan N.T., Khoi D.N., Hoan N.X., Phung N.K. & Dang T.D. 2021. Spatiotemporal trend analysis of precipitation extremes in Ho Chi Minh City, Vietnam during 1980-2017. *International Journal of Disaster Risk Science* **12**:131–146.
Ramsay J.O., Hooker G. & Graves S. 2009. *Functional Data Analysis with R and MATLAB*. New York, NY: Springer.
Ramsay J.O. & Silverman B. 2005. *Functional Data Analysis*. 2nd Ed. New York, NY: Springer.
Chu M.M. 2023. Malaysia forecast to face weak to moderate El Niño from June onwards. *Reuters*. https://www.reuters.com/business/environment/malaysia-forecast-face-weak-moderate-el-nino-june-onwards-2023-06-07/ (7 June 2023).
Suhaila J. 2021. Functional data visualization and outlier detection on the anomaly of El Niño southern oscillation. *Climate* **9**(7): 118.
Suhaila J., Jemain A.A., Hamdan M.F. & Zin W.W.Z. 2011. Comparing rainfall patterns between regions in Peninsular Malaysia via functional data analysis techniques. *Journal of Hydrology* **411**(3-4): 197–206.
Suhaila J. & Yusop Z. 2017. Spatial and temporal variabilities of rainfall data using functional data analysis. *Theoretical and Applied Climatology* **129**: 229-242.
Tabari H. 2020. Climate change impact on flood and extreme precipitation increases with water availability. *Scientific Report* **10:** 13768.
Saieed Z. 2023. Malaysia's heatwave expected to last until June, with haze likely to follow. *The Straits Times*. https://www.straitstimes.com/asia/se-asia/malaysia-s-heatwave-expected-to-last-until-june-with-haze-likely-after-that (5 May 2023).

Department of Mathematical Sciences
Faculty of Science
Universiti Teknologi Malaysia
81310 Johor Bahru
Johor, MALAYSIA
E-mail: suhailasj@utm.my[*]

Received: 27 January 2025
Accepted: 10 April 2025

[*]Corresponding author