


## Tweedie models for Malaysia rainfall simulations with seasonal variabilities

Jamaludin Suhaila <sup>a,b</sup>

<sup>a</sup> Department of Mathematical Sciences, Faculty of Science, Universiti Teknologi Malaysia, 81310, Johor Bahru, Johor Malaysia

<sup>b</sup> UTM Centre for Industrial and Applied Mathematics, Ibnu Sina Institute for Scientific and Industrial Research, Universiti Teknologi Malaysia, 81310, Johor Bahru, Johor Malaysia

E-mail: suhailasj@utm.my

 JS, 0000-0001-8609-3807

### ABSTRACT

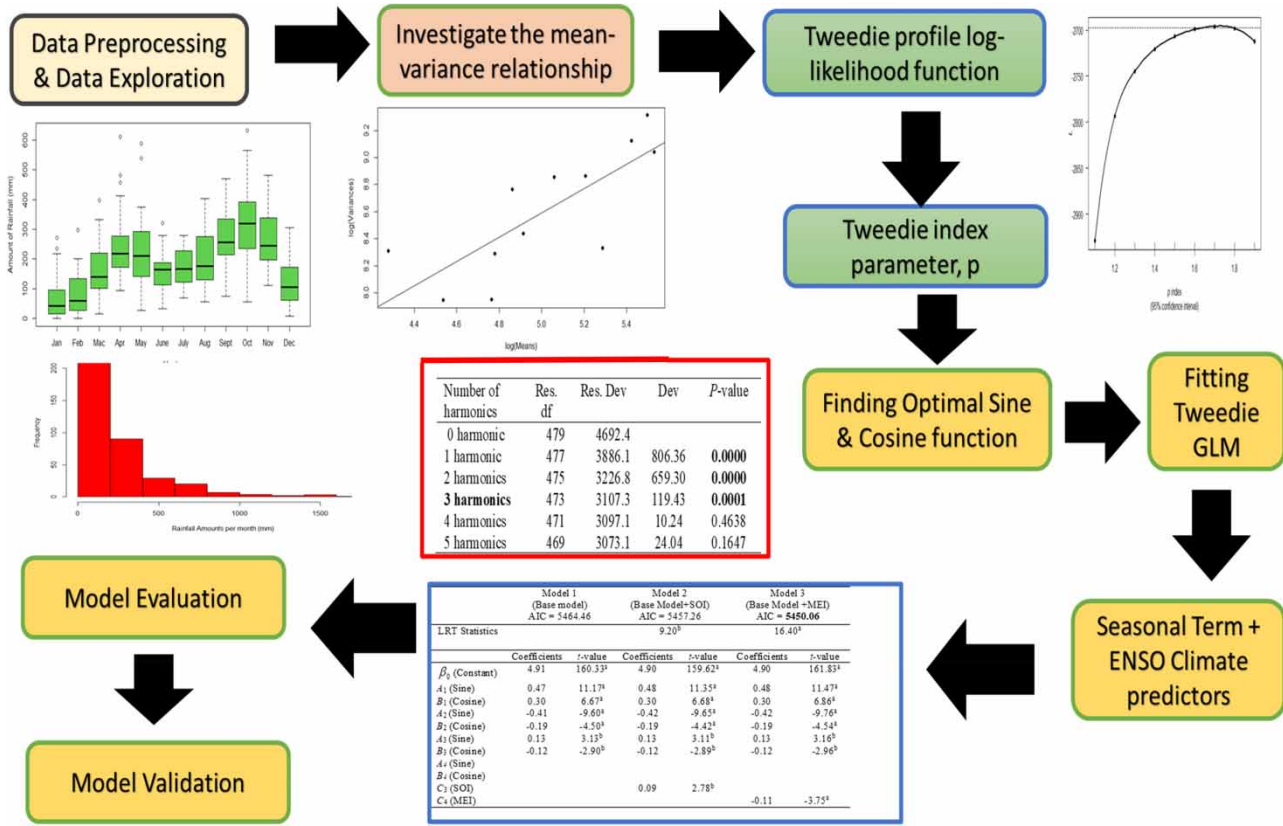
This study aims to evaluate the suitability of the Tweedie generalised linear model for characterising monthly rainfall patterns across 18 meteorological stations in Peninsular Malaysia. It incorporates harmonic functions consisting of sine and cosine functions as seasonal predictors and El Niño Southern Oscillation (ENSO) indices as climatic predictors. Results indicate that three harmonic functions are essential to accurately portray rainfall dynamics in the southwestern and northwestern regions, while two suffice for the inland and western regions. However, incorporating four harmonic functions is the most optimal representation of the eastern region. An additional 1-month lag in ENSO indices is introduced to the optimal seasonal predictor model. Based on the findings, the southern oscillation index notably impacts monthly rainfall significantly in eastern and inland areas, while meteorological stations in the western and northwestern areas fit better with the multivariate ENSO index. Strikingly, no substantial impact of climate predictors is observed on the monthly rainfall within the southwestern region. Thus, the influence of climate indices is very much influenced by the geographical locations of the regions. Importantly, generating simulated data through the Tweedie model contributes to a more accurate representation of the statistical properties inherent in rainfall analysis.

**Key words:** El Niño Southern oscillation, exponential dispersion model, generalised linear models, Poisson–Gamma, seasonal rainfall, tweedie index parameter

### HIGHLIGHTS

- Incorporating seasonal terms using harmonic functions enhanced the model's ability to describe rainfall patterns.
- The effect of seasonal and ENSO climate predictors on monthly rainfall amounts varies depending on the geographical locations and monsoon influence.
- Simulated data based on the Tweedie model able to capture the distributional properties, and excess zeros observed in rainfall analysis.

GRAPHICAL ABSTRACT



INTRODUCTION

Due to the enormous global population growth and economic development, understanding rainfall behaviour is essential for its significance to water resource management and climate change. Consequently, climatologists and meteorologists are consistently engaged in exploring the causative factors and underlying mechanisms governing seasonal fluctuations in rainfall. An illustrative example lies in the discrete wavelet transform, as *Ruwangika et al. (2020)* applied it to identify complex rainfall patterns. Furthermore, integrating the Fourier series into the generalised linear model help in accounting for the nuances of seasonal rainfall variations. Leveraging machine learning techniques, as demonstrated by *Barrera-Animas et al. (2022)*, *Das et al. (2022)*, and *Liyew & Melese (2021)*, have proven effective in enhancing the analytical capacity for rainfall forecasting. The main characteristics of rainfall include its amount, frequency, length of dry and wet spells, and intensity. These values may vary from place to place, year to year, month to month, day to day, and even hour to hour. Scientists have always faced challenges in modelling the chaotic behaviour of rainfall due to its spatial and temporal variabilities, coupled with constraints imposed by data availability within the study region. Therefore, finding a suitable rainfall model that can account for both issues is quite promising.

Rainfall processes may be categorised into two: one is concerned with the amount of rain on wet days, while the other deals with the sequence of dry and wet days. A combination of the Markov chain and the gamma distribution function is recognised as a simple approach for modelling. It has effectively generated daily rainfall data in many applications (*Gabriel & Neumann 1962; Stern & Coe 1982; Geng et al. 1986*). Because of their simplicity, those two-part models have been successfully used to simulate rainfall data. However, the effectiveness of the best order of Markov chains has been questioned, which led to other studies by several researchers, such as higher-order (*Deni et al. 2009*), hybrid (*Wilks 1999*), and hidden Markov chains (*Robertson et al. 2003*). Although Markov chain models have been extensively studied, these models only monitor the occurrence of rain and are thus limited in their efficacy in describing the amount of rain.

Furthermore, the shape of the rainfall distribution is often skewed to the right; hence, several positive skewed distributions other than Gamma are used in fitting the distribution, such as the lognormal distribution (Chebana & Ouarda 2021), the Weibull distribution (Sharda & Das 2005; Ximenes *et al.* 2021), and the skewed normal distribution (Chapman 1998). The two-part models are known to have potential uses in predicting and simulating rainfall. Several studies have been done to extend the two-part models in the framework of generalised linear models (GLMs) (Coe & Stern 1982; Chandler & Wheater 2002; Yang *et al.* 2005; Suhaila & Jemain 2009a; Serinaldi & Kilsby 2014; Pomee *et al.* 2020).

GLMs were originally used to model daily rainfall time series data at a single site (Coe & Stern 1982; Stern & Coe 1984). These models were the first to be implemented worldwide and are the most basic stochastic weather models. GLMs are parametric models that have the structure to condition the outcome variable, which is daily rainfall, on observed covariates, such as sine and cosine functions of time, to account for seasonality and other climatological variables. Chandler & Wheater (2002) developed a GLM-based framework for comprehending the spatiotemporal rainfall pattern. The division of the rainfall process into occurrences, describing wet and dry states and the amount of rainfall measured during wet spells, is a feature of the GLM models (Wilks 1998; Chandler & Wheater 2002). Yang *et al.* (2005) used the GLM approach to simulate multisite daily rainfall sequences incorporating intersite dependence structures.

However, when two separate models are employed, there is a possibility that the results fail to describe the overall characteristics of the rainfall. Therefore, Serinaldi (2009) developed a new model structure in modelling to deal with the issue of the construction of two separate models. He incorporated the concept of copula-based Markov chain models. The rainfall values at day  $t$ , given the rainfall value on the previous day ( $t - 1$ ), are conditioned using the probability distribution based on the bivariate copula model. The conditional distribution could describe the discrete-continuous nature of the rainfall without splitting the occurrence and amount of processes. The same approach was followed by Serinaldi & Kilsby (2014). They merged meta-Gaussian random fields and a generalised additive model to simulate daily rainfall over large areas without splitting the processes of rainfall occurrence and rainfall amount. As a result, the combined models become more parsimonious compared to other models for multisite rainfall simulation and, at the same time, preserve the rainfall characteristics.

The Tweedie family of distributions is a viable alternative for modelling datasets that exhibit discrete (zero) and continuous (non-zero) values. It is a special case of an exponential dispersion model (EDM), which is often used as a distribution for generalised linear models. The Tweedie family is used to fit within the GLM framework. Within this framework, the Tweedie family accommodates incorporating covariates, allowing response variables to adapt to these predictors. The Poisson–Gamma distribution in the Tweedie family was sufficient to represent both rainfall components with a single complete rainfall model (Dunn 2004; Hasan & Dunn 2010, 2011; Dzapire *et al.* 2018; Hasan *et al.* 2019). Hasan & Dunn (2010) used Tweedie GLM with a sine and cosine term (base model) as predictors for modelling monthly rainfall by fitting a single model for each station. The model fits well with the monthly rainfall data for most Australian stations. In another study, Hasan & Dunn (2012) improved the base model by introducing climatological covariates. Three separate covariates were successively applied to the base model: the southern oscillation index (SOI), the SOI phase, and the NINO 3.4 index. Their results indicated that there had been a significant shift in the amount expected and the possibility of no rain due to the impact of these climate variables. A later study by Hasan *et al.* (2019) investigated the relationship between the estimated parameters from the fitted model of daily, monthly, and seasonal rainfall totals across Australia. They discussed the possibility of estimating the parameters of daily data using the fitted parameters for monthly rainfall. They concluded that the established relationship between the parameters in the monthly and daily models might aid in comprehending the features and generating data on a daily timescale.

A Tweedie GLM has also been applied in a statistical downscaling approach (Hertig & Jacobeit 2015; Hertig *et al.* 2017; Pomee *et al.* 2020). Hertig & Jacobeit (2015) used the Tweedie EDM to downscale weather-dependent daily precipitation data in the Mediterranean region. The probability of occurrence and rainfall distribution corresponds well with the Poisson–Gamma distribution but not in situations with extreme values. In a recent publication, Pomee *et al.* (2020) employed a GLM approach to model the relationship between atmospheric variables and precipitation in the Indus River Basin of Pakistan. GLMs based on a gamma-distributed model were used in the framework. However, for the case of having exact zeros in the observed time series, a Tweedie EDM within a GLM framework was applied because of its ability to model discrete and continuous precipitation features simultaneously.

Seasonal variability and cyclical patterns are likely to be apparent in Malaysian rainfall, with some areas having dry and rainy periods regularly from year to year. It is well known that the monsoons and geographical locations have a significant impact on the Malaysian climate (Suhaila & Jemain 2009b; Suhaila *et al.* 2011). In their scholarly contributions, Suhaila &

Jemain (2009a) introduced a Fourier series to enhance the representation of model parameters. This particular technique holds the advantage of succinctly elucidating the intricate patterns and temporal variations in rainfall. Incorporating sine and cosine functions within this series framework facilitated the systematic quantification of harmonics, thereby providing alternative ways of characterising the complex rainfall patterns prevalent in Peninsular Malaysia. It is notable, however, that their investigation concentrated solely on modelling rainfall amounts, with rainfall occurrences treated independently.

On the other hand, Yunus *et al.* (2017) used Tweedie GLM to investigate several possible climate covariates on daily rainfall at four studied stations over Peninsular Malaysia. The covariates in the model served various purposes, such as sinusoidal functions for capturing seasonal variability, temporal autocorrelation accounting for the influence of rainfall on the previous, and climate predictors. They found that adding climate indices such as the SOI and NINO 3.4 index improved the fit and substantially changed the predicted daily rainfall outcomes.

The concurrent modelling of discrete and continuous distributions poses a significant challenge for most models, primarily due to the dissimilarity in their marginal distributions. As a result, addressing this mixture of distributions in a unified framework can introduce considerable complexity. Therefore, the Poisson–Gamma Tweedie model emerges as one of the alternatives for the concurrent modelling of rainfall amount and occurrence, particularly within the context of Malaysia. To the best of the author’s knowledge, the application of the Tweedie distribution in the domain of rainfall modelling remains new within the Malaysian context. Hence, the primary goal of this study is twofold: firstly, to examine the adaptability and versatility of the Tweedie GLM in characterising the intricate attributes of Malaysian rainfall data, and secondly, to determine the impact exerted by El Niño Southern Oscillation (ENSO) indices on the distribution of rainfall patterns.

This study introduces a dynamic approach by incorporating a linear function comprising several harmonic components and climatological predictors to enhance the flexibility of the Tweedie model. In contrast, the prior investigation by Yunus *et al.* (2017) exclusively focused on a single harmonic function across four meteorological stations situated in Malaysia. Since Malaysia has a distinctive seasonal rainfall pattern, relying solely on a single harmonic function proves inadequate in effectively describing the rainfall distributions across all stations. Consequently, the present study engages in a comparative analysis of various harmonic functions employed as seasonal predictors, subsequently selecting the optimal number that best describes the rainfall pattern in Malaysia. In addition, this study introduces the monthly SOI and Multivariate El Niño southern oscillation index (MEI) as climate predictors to improve the model accuracy. The predictors serve to assess the significant effect on Malaysian rainfall patterns. The relationship between the ENSO indices and rainfall could be established through the application of fitted Tweedie generalised linear models. Furthermore, the simulated data derived from the fitted model could be used for diverse applications within rainfall modelling, particularly in domains like water resource management and agricultural planning.

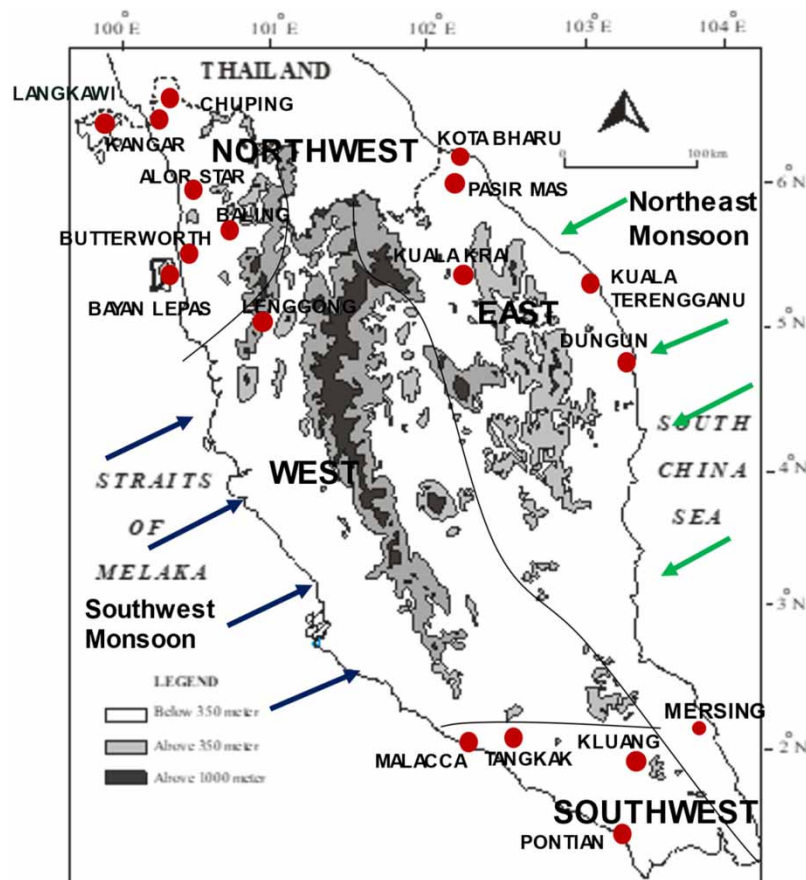
## STUDY AREA

Peninsular Malaysia is situated in the tropics between the latitudes of 1°N and 7°N and the longitudes of 100°E and 103°E. It has an equatorial climate with uniformly high temperatures, high humidity, and relatively light winds. Peninsular Malaysia receives rainfall that varies from season to season. The country has a uniform temperature of 25.5–32 °C throughout the year, with an annual rainfall normally between 2,000 and 4,000 mm, while the annual number of wet days ranges from 150 to 200 (Suhaila *et al.* 2010). The climate of Peninsular Malaysia is primarily dominated by the Northeast Monsoon (NEM) (November–February), the Southwest Monsoon (SWM) (May–August), and two inter-monsoons (March–April and September–October) (Suhaila & Yusop 2017; Hui-Mean *et al.* 2019; Annual Report Malaysian Meteorological Department 2020). The NEM brings heavy rainfall to the east coast of the Peninsula, while the northwest region and the inland areas sheltered by the mountain ranges (Titiwangsa range) are relatively free from its influence. The coasts that are exposed to the NEM appear to be wetter than those exposed to the SWM during the NEM season (Suhaila *et al.* 2011; Sa’adi *et al.* 2017; Liang *et al.* 2023). Figure 1 shows the map of the studied stations over Peninsular Malaysia along with the direction of the southwest and NEM flows.

## DATA AND METHODOLOGY

### Rainfall: data source and data quality checking

This study selected 18 representative meteorological stations that are well distributed across Peninsular Malaysia based on the availability of the data. The datasets were obtained from the Malaysian Meteorological Department, and most of the studied stations (67%) have data available from January 1980 to December 2019. While the rainfall data at the remaining stations



**Figure 1** | Map of the studied stations and the flows of the SWM and NEM.

started in 1985 and 1988, some ended in 2012 and 2015. Detailed information regarding the meteorological stations is presented in [Table 1](#). This table includes summary statistics highlighting the months when no rainfall was recorded. It was found that Kangar exhibited the highest percentage, with 42.4% of months having no recorded rainfall. Following Kangar, Chuping and Alor Star are in the second and third positions, with percentages of 32.5 and 27.5%, respectively, signifying notable occurrences of months without rainfall. Conversely, the remaining stations experienced months without rainfall at least once within the study period.

Complete rainfall data are observed for all stations except for the Pasir Mas station, which had 0.6% of its values recorded as missing. A simple inverse distance weighting method was used to estimate the missing values at the Pasir Mas station using the method of the nearest neighbouring stations. The data were grouped into monthly rainfall series and then tested for homogeneity. In the Malaysian context, there are no specific guidelines for choosing the best method for the homogenisation procedure. In this study, the homogeneity of each data series was conducted using the approach introduced by [Wijngaard et al. \(2003\)](#). This study employed four distinct homogeneity tests: the standard normal homogeneity, the Buishand range test, the Pettitt, and the Von Neumann Ratio. All these tests applied the information from the single station series since the meteorological stations used are sparsely distributed. The classification of the monthly station rainfall series was then based on their performance in the homogeneity tests, which were classified as ‘useful (U)’ (when at least three tests indicated homogeneity), ‘doubtful (D)’ (when a minimum of two tests indicated homogeneity), and ‘suspect (S)’ (none or only one test indicated homogeneity). For a more comprehensive understanding of the methods employed, readers are directed to the studies by [Wijngaard et al. \(2003\)](#) and [Alexandersson \(1986\)](#). The results of the homogeneity tests are presented in [Table 2](#). Inhomogeneous rainfall series were detected in January for Baling, Chuping, Kota Bharu, and Pasir Mas. In contrast, the remaining months of these stations exhibited a homogeneous rainfall series. Conversely, it is notable that all other stations subjected to investigation displayed consistent homogeneity in their respective monthly rainfall series.

**Table 1** | The geographical coordinates of the studied stations with the percentage of months with no rains

Stations	Latitude	Longitude	Elevation (m)	Months with no rains (%)	Period
Alor Star	6°12' N	100°24'E	3.9	11 (27.5)	1980–2019
Baling	5°41' N	100° 55'E	51.9	4 (10)	1980–2019
Bayan Lepas	5°18'N	100°16'E	3.0	2 (5)	1980–2019
Butterworth	5°27' N	100°23'E	3.3	1 (2.86)	1985–2019
Chuping	6°29' N	100°16'E	22	13 (32.5)	1980–2019
Dungun	4°46' N	103°25'E	3.1	2 (5.6)	1980–2015
Kangar	6° 26' N	100°12'E	3.1	14 (42.4)	1980–2012
Kluang	2°01' N	103°19'E	88.1	1 (2.5)	1980–2019
Kota Bharu	6°10' N	102°18'E	4.4	7 (17.5)	1980–2019
Kuala Krai	5°32' N	102°12'E	68.3	2 (5.89)	1985–2019
KualaTerengganu	5°23' N	103°06'E	5.2	4 (11.43)	1985–2019
Langkawi	6°20'N	99°44'E	6.4	8 (25)	1988–2019
Lenggong	5°06' N	100°58'E	100.7	1 (2.5)	1980–2019
Melaka	2°16' N	102°15'E	8.5	1 (2.5)	1980–2019
Mersing	2°27' N	103°50'E	43.6	1 (2.5)	1980–2019
Pasir Mas	6°02' N	102°07'E	9.1	5 (12.5)	1980–2019
Pontian	1°29' N	103°23'E	4.6	1 (2.5)	1980–2019
Tangkak	2°16' N	102°32'E	30.5	1 (2.5)	1980–2019

**Table 2** | Categorising homogeneity test in monthly rainfall series across 18 meteorological stations

Stations	J	F	M	A	M	J	J	A	S	O	N	D
Alor Star	D	U	U	U	U	U	U	U	U	U	U	U
Baling	S	U	U	U	U	U	U	U	U	U	U	U
Bayan Lepas	U	U	U	U	U	U	U	U	U	U	U	U
Butterworth	U	U	U	U	U	U	U	U	U	U	U	U
Chuping	S	U	U	U	U	U	U	U	U	U	U	U
Dungun	U	U	U	U	U	U	U	U	U	U	U	U
Kangar	U	U	U	U	U	U	U	U	U	U	U	U
Kluang	U	U	U	U	U	U	U	U	U	U	U	U
Kota Bharu	S	U	U	U	U	U	U	U	U	U	U	U
Kuala Krai	U	U	U	U	U	U	U	U	U	U	U	U
Kuala Trengganu	U	U	U	U	U	U	U	U	U	U	U	U
Langkawi	U	U	U	U	U	U	U	U	U	U	U	U
Lenggong	U	U	U	U	U	U	U	U	U	U	U	U
Melaka	U	U	U	U	U	U	U	U	U	U	U	U
Mersing	U	U	U	U	U	U	U	U	U	U	U	U
Pasir Mas	S	U	U	U	U	U	U	U	U	U	U	U
Pontian	U	U	U	U	U	U	U	U	U	U	U	U
Tangkak	U	U	U	U	U	U	U	U	U	U	U	U

Note: S, suspect; D, doubtful; U, useful.

The ENSO index influences the Malaysian climate, which can vary depending on the season and location, with drier-than-normal conditions during El Niño and wetter-than-normal conditions during La Niña (Tangang *et al.* 2017). According to studies by Tangang & Juneng (2004), a strong ENSO index is only apparent in East Malaysia during the season, and the inter-annual variability of Malaysia rainfall associated with ENSO events peaks during the NEM period (Juneng & Tangang 2005). Gomyo & Koichiro (2009) also affirm a strong connection between ENSO and rainfall patterns in East Malaysia. Historical analysis reveals that strong El Niño occurrences, such as those in 1997–1998 and 2015–2016, have led to extended droughts and water crises across multiple regions within Malaysia (Tan *et al.* 2022). Consequently, this study employed the SOI as the initial ENSO indicator, alongside the recent multivariate ENSO index (MEI), to serve as climate predictors. The purpose is to explore the relationship between rainfall distribution and the ENSO index phenomenon using the Tweedie GLM. Within this study, the MEI values can be comprehended through the description provided and accessed via the website: <https://psl.noaa.gov/enso/mei/>. Similarly, the SOI values required for this research were acquired from the following source: <https://www.cpc.ncep.noaa.gov/data/indices/soi/>.

### Exponential dispersion model

The EDM is a statistical framework that incorporates a wide range of probability distributions in various fields of statistics and data analysis. The model proves to be rather useful, particularly when working with count data or continuous data whose distributions exhibit variability and skewness in their distributions. The probability density function of the EDM has the following general form:

$$f(y; \theta, \phi) = a(y, \phi) \exp\left\{\frac{1}{\phi}[y\theta - \kappa(\theta)]\right\} \quad (1)$$

where  $\phi > 0$  is the dispersion parameter,  $\theta$  is the canonical parameter, and  $\kappa(\theta)$  is the cumulant function. In particular, the mean of the response variables  $Y$  is  $E(Y) = \mu = \kappa'(\theta)$ , and the variance of  $Y$  is  $\text{Var}(Y) = \phi\kappa''(\theta)$ . It is important to note that  $\kappa'(\theta)$  is a function of the mean. Thus,  $\kappa''(\theta)$  is also dependent on the mean. Therefore,  $\kappa''(\theta)$  is often described as a variance function  $V(\mu)$ , so that  $V(\mu) = \kappa''(\theta)$ . Hence, the variance of  $Y$  can be written as  $\text{Var}(Y) = \phi V(\mu)$ , where  $Y$  follows an EDM with mean  $\mu$ , variance function  $V(\mu)$ , and dispersion parameter  $\phi$ .

### Tweedie family of distributions

Let  $N$  be the number of rainfall events in any month. Assume that  $N$  has a Poisson distribution with mean,  $\lambda$  such that  $N \sim \text{Poisson}(\lambda)$ . Suppose any rainfall event  $i$  produces an amount of rainfall  $S_i$ , where each  $S_i$  is distributed as a Gamma distribution  $(\alpha, \gamma)$  with mean  $\alpha\gamma$  and variance  $\alpha\gamma^2$ . The total monthly rainfall  $Y$ ,  $Y = S_1 + S_2 + \dots + S_N$  is a sum of independent and identically Gamma random variables, which results in a continuous distribution for the positive rainfall amount. If there is no recorded rainfall ( $Y = 0$ ) in a particular month, then the probability of recording no rainfall events  $N = 0$  can be expressed as follows:

$$P(Y = 0) = \frac{e^{-\lambda}\lambda^0}{0!} = e^{-\lambda}. \quad (2)$$

The process has a point mass at zero, which implies that it is not an entirely continuous random variable. The Tweedie family of distributions belongs to the class of the EDM (Jørgensen 1997), which can be used to fit into the GLM framework. For those cases having exact zeroes in rainfall time series, a Tweedie GLM was employed due to its ability for simultaneous modelling of discrete and continuous rainfall features (Hasan & Dunn 2010, 2012; Hertig & Jacobeit 2015; Hertig *et al.* 2017; Pomee *et al.* 2020).

### Fitting Tweedie distributions

The Tweedie distributions are those EDMs with a variance function of the form  $V(\mu) = \phi\mu^p$  where  $p \notin (0, 1)$  (Jørgensen 1997). The Tweedie family of distributions has three parameters: mean  $\mu$ , dispersion parameter  $\phi$ , and index parameter  $p$ , which is denoted as  $TW_p(\mu, \phi)$ . If  $p = 0$ , the Tweedie distributions correspond to normal distributions, but when  $p = 1$ ,  $\phi = 1$ , it becomes the Poisson distribution. In addition, when  $p = 2$ , the Tweedie distributions resemble a Gamma

distribution, while if  $p = 3$ , it is an inverse Gaussian distribution. Apart from these four special cases, the densities of the Tweedie distribution cannot be written in closed form (Dunn & Smyth 2005).

A special case of EDM class is a Tweedie distribution with the index parameter  $p$  between 1 and 2. It can be explained as a mixture of Poisson and Gamma distributions. The probability density function is given as follows:

$$f(y; \mu, \phi, p) = a(y, \phi) \exp\left\{\frac{1}{\phi} \left[ y \frac{\mu^{1-p}}{1-p} - \frac{\mu^{2-p}}{2-p} \right]\right\} \quad (3)$$

$$\text{where } a(y, \phi) = \frac{1}{y} \sum_{j=1}^{\infty} \frac{y^{-j\alpha} (p-1)^{\alpha j}}{\phi^{j(1-\alpha)} (2-p)^j j! \Gamma(-j\alpha)} \text{ with } \alpha = \frac{2-p}{1-p}.$$

The corresponding log-likelihood function of Equation (3) can be written as follows:

$$l(\mu, \phi, p|y) = \sum_{i=1}^n \log(a(y_i, \phi)) + \sum_{i=1}^n \left( \frac{1}{\phi} \left[ y_i \frac{\mu_i^{1-p}}{1-p} - \frac{\mu_i^{2-p}}{2-p} \right] \right) \quad (4)$$

In the context of GLM,  $\mu_i$  corresponds to predicted values that can be estimated directly from the model using a standard algorithm. However, the maximum likelihood estimation of  $\phi$  is quite complicated (Dunn 2004). Similarly, the estimation procedure to compute the index parameter,  $p$ , is also difficult and requires high computational resources (Dunn & Smyth 2005; Hasan & Dunn 2010, 2012; Hasan *et al.* 2019). With modern and advanced computer technologies, the algorithm to compute the parameter estimations for Tweedie distribution has been successfully developed. Dunn (2022) created the algorithm for the Tweedie profile function in the R package by considering a set of values for the index parameter  $p$  and computing the log-likelihood. This can be done by directly maximising the profile likelihood function. The Tweedie package in R has a function to estimate  $p$  using *tweedie.profile* functions. Once the maximum likelihood value of  $p$  is found, the distribution within the class of the Tweedie family is identified. In addition, the algorithms that were used to estimate  $p$  can also be used to compute the Maximum Likelihood Estimation (MLE) of  $\phi$  (Dunn & Smyth 2005). Hence, this study will employ the algorithm developed by Dunn (2022), which is available in the R package.

The relationship between the parameters of the Poisson and Gamma distributions with the EDM notation (Hasan & Dunn 2010) is given as follows:

$$\lambda = \frac{1}{\phi} \frac{\mu^{2-p}}{2-p}, \alpha = \frac{2-p}{1-p}, \gamma = \phi(p-1)\mu^{p-1} \quad (5)$$

### Generalized linear models

After the Tweedie Poisson–Gamma distribution is specified, the GLM framework, as discussed in McCullagh & Nelder (1989), will be used to fit the dataset. To model the response  $\mu_i$  with a linear component  $X_i\beta$ , the link function needs to be chosen correctly. Since the model fits GLM to the mean rainfall, the logarithm link function is applied for  $\mu_i$ . Based on the Tweedie GLM, models in the form of  $\log \mu_i = X_i\beta$  are fitted, where  $\mu_i$  is the mean rainfall in month  $i$ ,  $X_i$  is a vector of predictors, and  $\beta$  represents a vector of parameter coefficients.

### Modelling the predictors in a Tweedie GLM

Sine and cosine functions are included in the Tweedie model to describe seasonal rainfall patterns, which can be expressed as follows:

#### Model 1 (base model)

$$\log \mu_i = \beta_0 + \sum_{j=1}^k \left\{ A_j \sin\left(\frac{j\pi(m-6)}{6}\right) + B_j \cos\left(\frac{j\pi(m-6)}{6}\right) \right\} \quad (6)$$

where  $\mu_i \sim Tw_p(\mu, \phi)$ ;  $j$  is the harmonic,  $k$  is the maximum number of harmonics required,  $A_j$  and  $B_j$  are the parameter coefficients of the series, which are often denoted as  $\beta_j$  (starting at  $j = 1$ ) parameters, and  $m$  represents the month of the year (1 = January, 2 = February, and so on). Notice that  $i = 12(t-1) + m$ ,  $t = 1, 2, \dots, T$  starting at  $t = 1$  for the year 1980 and  $t = 40$  for



the year 2019. Monthly mean rainfall  $\mu_i$  is allowed to depend on the time of the year, in which the mean  $\mu_i$  varies continuously with time. The first model (Model 1) is called the base model in this analysis.

Instead of considering only a sine and cosine terms as applied in the previous analyses (e.g., Hasan & Dunn 2010; Yunus *et al.* 2017), the present study added several sine and cosine terms to determine the optimum number of harmonic functions that best described the rainfall data of the stations since it is known that Malaysian rainfall patterns are heavily influenced by the monsoon seasons. After the optimum number of harmonic functions has been determined, the lagged 1 month SOI and MEI were added to the base model of Equation (6), which are presented as Model 2 and 3 as follows:

### Model 2 (base model + SOI)

$$\log \mu_i = \beta_0 + \sum_{j=1}^k \left\{ A_j \sin\left(\frac{j\pi(m-6)}{6}\right) + B_j \cos\left(\frac{j\pi(m-6)}{6}\right) \right\} + C_3(\text{SOI})_{i-1} \quad (7)$$

### Model 3 (base model + MEI)

$$\log \mu_i = \beta_0 + \sum_{j=1}^k \left\{ A_j \sin\left(\frac{j\pi(m-6)}{6}\right) + B_j \cos\left(\frac{j\pi(m-6)}{6}\right) \right\} + C_4(\text{MEI})_{i-1} \quad (8)$$

For model comparison, the best-fit model is determined based on the Akaike information criterion (AIC), which is defined as follows:

$$\text{AIC} = -2 \log L + 2k \quad (9)$$

where  $L$  is the maximised likelihood function and  $k$  represents the number of parameters. However, since those tested models have different parameters, a question may arise regarding the statistical significance of the AIC values associated with the most suitable model. A likelihood ratio test (LRT) concept has been introduced and employed to address this concern.

Suppose we have two models with the AICs given as follows:

$$\text{AIC}_1 = -2 \log L_1 + 2k_1$$

$$\text{AIC}_2 = -2 \log L_2 + 2k_2$$

Then, the log-LRT can be written as follows:

$$-2(\log(L_1) - \log(L_2)) = \text{AIC}_1 - \text{AIC}_2 + 2(k_2 - k_1) \sim \chi_{k_2 - k_1}^2 \quad (10)$$

The LRT will be used to compare the fitness of Models 2 and 3 with the base model (Model 1).

### Model validation

Two approaches are proposed to assess the effectiveness of the best-fitting model. The first approach uses the  $k$ -fold cross-validation method of the generalised linear model. The entire dataset is divided into  $k$  pieces (equal sample size) for  $k$ -fold cross-validation. Following that, the remaining  $(k-1)$  folds act as the new training set, with each fold acting as a validation set. This process is carried out iteratively  $k$  times, with a different group of observations being used as a validation set each time.

The mean square error (MSE) of each fold is computed as  $\text{MSE} = 1/n \sum_{i=1}^n (y_i - \hat{y}_i)^2$ , where  $y_i$  and  $\hat{y}_i$  are the observed and predicted values, respectively. The generalisation error estimate can be determined by averaging the MSEs over  $k$  folds of validation. This is referred to as the cross-validation error and can be expressed as  $\text{CV}_{(k)} = 1/k \sum_{i=1}^k \text{MSE}_i$ . The  $k$ -fold cross-validation method reduces the computational time compared to leave-one-out cross-validation (LOOCV). However,

it may introduce some bias in the performance estimate (James *et al.* 2013). This study uses an algorithm **cv.glm** in the library (**boot**) to evaluate the cross-validated prediction error. The adjusted cross-validation estimate is chosen since it is designed to compensate for the bias introduced by not using LOOCV. A  $k$ -fold cross-validation is a great option for most applications since it strikes a good balance between computational efficiency and reliability.

The second model validation approach uses a graphical display via a Taylor diagram. It is a method that can be used to compare the performance of different models or simulations against observations by plotting the standard deviation, centred root mean square difference, and correlation coefficient of the model output relative to the observations. Karl Taylor introduced the method in 2001, and the diagram was named after him. However, Taylor diagrams do not offer a formal statistical test of the model's accuracy or generalisation. Instead, they can provide a visual and qualitative assessment of the model's performance. Therefore, Taylor diagrams should be used in combination with other validation methods to obtain a more comprehensive and reliable assessment of the model's performance. A detailed explanation is given by Taylor (2001).

## RESULTS AND DISCUSSION

### Parameter estimates of tweedie distribution

The first essential step in the Tweedie GLM framework is an exploration of the mean–variance relationship through a log-log plot of the variance against the mean. The relationship between mean and variance is considered relevant since the variance is proportional to the power of the mean. This relationship will then be used to find a suitable distribution for monthly rainfall totals. An appropriate value of the variance power  $p$  can be determined statistically using the profile log-likelihood function. The choice of  $p$  will determine which member of the Tweedie family of distributions will be used in the analysis. The analysis starts by fitting the simplest model of sine and cosine functions with a single harmonic function to describe the seasonal rainfall patterns. The model can be expressed as follows:

$$\log \mu_i = \beta_0 + A_1 \sin\left(\frac{\pi(m-6)}{6}\right) + B_1 \cos\left(\frac{\pi(m-6)}{6}\right) \quad (11)$$

where  $\mu_i \sim Tw_p(\mu, \phi)$ ;  $A_1$  and  $B_1$  are the parameter coefficients of the first harmonic function, and  $m$  represents the month of the year (1 = January, 2 = February, and so on). Notice that  $i = 12(t-1) + m$ ,  $t = 1, 2, \dots, T$  starting at  $t = 1$  (first year of studied data), 2 (second year of studied data), ..., and so on. The parameters of Tweedie distribution were then estimated using *tweedie.profile* functions in R programming. The value  $\hat{p}$  is estimated by maximising the profile log-likelihood function.

Table 3 lists the estimated index parameter  $p$ , the confidence interval of  $p$ , and the dispersion parameter  $\phi$  for all studied stations. As shown in Table 3, the estimated values of  $p$  for the stations under study that maximise the log-likelihood function fall between 1.32 and 1.72. Dungun demonstrates the highest value of all the estimated values, with  $p = 1.720$ . The values of the estimated  $p$  for stations situated on the east coast of Peninsular Malaysia are higher than other stations in other parts of the Peninsula. The distinction in index parameter  $p$  could be attributed to heavy rains on the east coast of Peninsular Malaysia that may be influencing these outcomes. However, the MLE of  $p$  for all studied stations remains in the interval of  $1 < p < 2$ . This finding implies that the Tweedie Poisson–Gamma distribution can be effectively employed for modelling purposes. The estimation of the Tweedie index parameter is the starting point for the formulation of the Tweedie model. The present study uses monthly rainfall as the basis. However, the Tweedie index parameter is computed for each station (not each month), producing much simpler models and reducing the computational burden. After performing the profile likelihood estimate, the desired Tweedie family distribution could be used to fit a GLM.

### Determining the optimum number of harmonic functions for seasonal rainfall distribution

For simplicity, the fitting process for additional sine and cosine functions will be carried out using the same Tweedie index parameter  $p$  obtained for each station using Equation (11). As mentioned in the study by Hasan & Dunn (2012), the value of the estimated  $p$  makes no difference to the estimated GLM coefficients, but it has a small impact on the monthly rainfall variation of the stations. All 18 meteorological stations were analysed in this study. However, to facilitate the discussion, the findings for the following five stations will be used as case studies based on their geographical locations in Peninsular

**Table 3** | Tweedie parameters of each studied station

Stations	$p$ -Index	Confidence interval of $p$	Dispersion parameter, $\phi$
Alor Star	1.443	(1.391,1.510)	7.027
Baling	1.410	(1.337,1.502)	7.454
Bayan Lepas	1.486	(1.388, 1.592)	4.986
Butterworth	1.492	(1.380,1.636)	4.522
Chuping	1.443	(1.386, 1.510)	6.978
Dungun	1.720	(1.633, 1.792)	2.387
Kangar	1.378	(1.312, 1.449)	9.059
Kluang	1.557	(1.397, 1.708)	3.281
Kota Bharu	1.655	(1.580, 1.727)	3.760
Kuala Krai	1.606	(1.503, 1.700)	3.497
Kuala Trengganu	1.704	(1.622, 1.772)	2.935
Langkawi	1.476	(1.443, 1.543)	6.068
Lenggong	1.508	(1.364, 1.662)	3.948
Melaka	1.378	(1.233, 1.542)	6.570
Mersing	1.655	(1.559, 1.761)	2.575
Pasir Mas	1.622	(1.553, 1.706)	4.105
Pontian	1.329	(1.194, 1.505)	8.209
Tangkak	1.443	(1.346, 1.579)	4.987

Malaysia. Chuping represents the studied stations located in the northern region; Bayan Lepas in the inland region; Lenggong represents the stations in the western region; Kota Bharu represents the stations in the eastern region of the Peninsula; and Pontian represents the stations in the southwest region.

The model fitting starts with a constant term and then proceeds with a sine and cosine term, representing the first harmonic. Then, the second and third harmonics were added until no significant effect was observed in the model. Next, model deviance is used to describe the adequacy of the fitted model. The term residual deviance (Res.dev) is used to measure the unexplained variation that remains after fitting a model to the data, while deviance (dev) here represents the reduction in the residual deviance between the  $n$ th and  $(n + 1)$ th harmonics. Models with different numbers of harmonics are then compared by considering the reductions in deviance. In addition, residual degrees of freedom (Res.df) represent the number of observations or data points minus the number of estimated parameters. The degree of freedom indicates the number of observations that are free to vary after accounting for the parameters in the model. The probability value, or  $p$ -value, of the test, is then computed using a distribution with degrees of freedom equal to the difference in the number of estimated parameters. A smaller  $p$ -value will lead to rejecting the null hypothesis, which suggests stronger evidence against the null hypothesis, implying that the factor significantly affects the response variable. Finding the optimal number of harmonic functions will proceed until no significant effect is observed in the reduction of deviance.

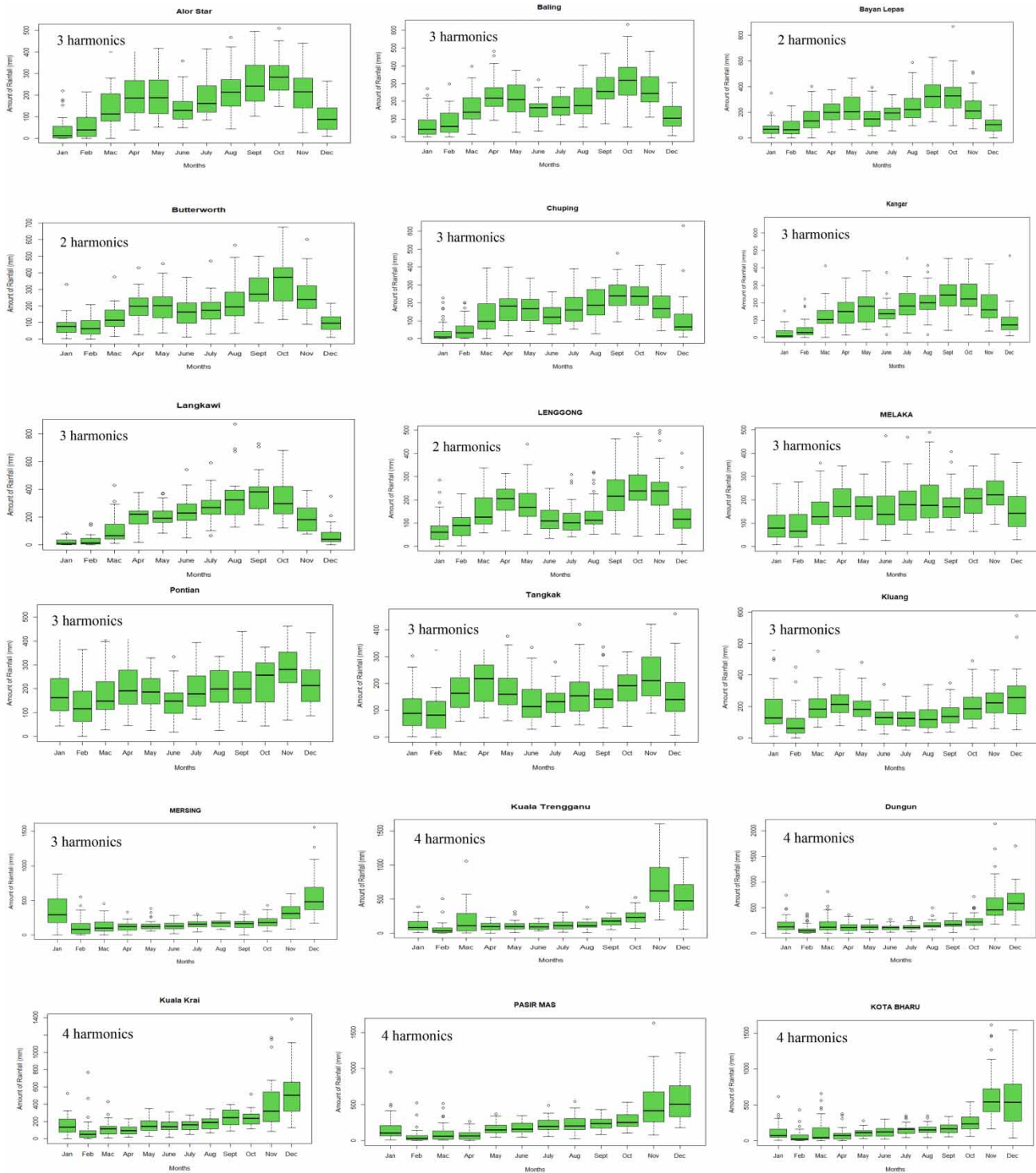
The deviance analyses for Chuping and Bayan Lepas stations are presented in Tables 4(a) and 4(b). The performance of the model is measured based on the reduction in deviance. Every time a harmonic is added, it will remove two degrees of freedom corresponding to the sine and cosine terms. The chi-square test statistics for a model with one harmonic yield a  $p$ -value of 0.000. This  $p$ -value is less than the significance level of 0.05, indicating a significant effect of adding one harmonic to the model with no harmonics. Note that further increases in the number of harmonics yield a  $p$ -value less than the significance level of 0.05, which means that based on the reduction in deviance, additional harmonics are needed in the model. The second and third harmonics still produce a significant result until the fourth harmonics are added to the model; they do not significantly reduce the deviance. Thus, three harmonics are adequate to fit the mean monthly rainfall at the Chuping station.

**Table 4** | The effect of increasing the number of harmonics fitted to the mean monthly rainfall for (a) Chuping, (b) Bayan Lepas, (c) Lenggong, (d) Pontian, and (e) Kota Bharu stations

Number of harmonics	Res. df	Res. Dev	Dev	P-value
(a)				
0 harmonic	479	4,692.4		
1 harmonic	477	3,886.1	806.36	<b>0.0000</b>
2 harmonics	475	3,226.8	659.30	<b>0.0000</b>
<b>3 harmonics</b>	473	3,107.3	119.43	<b>0.0001</b>
4 harmonics	471	3,097.1	10.24	0.4638
5 harmonics	469	3,073.1	24.04	0.1647
(b)				
0 harmonic	479	3,408.0		
1 harmonic	477	2,628.5	779.55	<b>0.0000</b>
<b>2 harmonics</b>	475	2,031.1	597.39	<b>0.0000</b>
3 harmonics	473	2,024.8	6.27	0.4632
4 harmonics	471	2,021.8	3.02	0.6904
5 harmonics	469	2,018.7	3.10	0.6831
(c)				
0 harmonic	479	2,181.9		
1 harmonic	477	2,051.1	130.88	<b>0.0000</b>
<b>2 harmonics</b>	475	1,453.0	598.03	<b>0.0000</b>
3 harmonics	473	1,435.9	17.08	0.0616
4 harmonics	471	1,429.4	6.55	0.3434
5 harmonics	469	1,422.0	7.41	0.2986
(d)				
0 harmonic	479	2,481.3		
1 harmonic	477	2,338.6	142.675	<b>0.0000</b>
2 harmonics	475	2,250.0	88.556	<b>0.0000</b>
<b>3 harmonics</b>	473	2,179.4	70.654	<b>0.0000</b>
4 harmonics	471	2,170.3	9.093	0.3474
5 harmonics	469	2,167.3	3.030	0.7031
(e)				
0 harmonic	479	3,070.6		
1 harmonic	477	2,083.6	986.98	<b>0.0000</b>
2 harmonics	475	1,760.0	323.60	<b>0.0000</b>
3 harmonics	473	1,559.9	200.13	<b>0.0000</b>
<b>4 harmonics</b>	471	1,503.7	56.18	<b>0.0002</b>
5 harmonics	469	1,493.0	10.70	0.2018

The bold values show the optimal number of harmonics required with the  $p$ -value  $< 0.05$ .

The seasonal rainfall peaks for Chuping are observed in April and September during the intermonsoon months, as shown in the box plots in Figure 2. The second peak in September is higher than the first, while the lowest rainfall peak is recorded during the NEM months (December–February). Similar patterns during the NEM season have been observed for Alor Star, Baling, Kangar, and Langkawi stations. Compared to the stations in the northern region, rainfall patterns at Bayan Lepas station in the northern inland areas are very well described by two harmonic functions. Two rainfall peaks are observed. The first rainfall peak is seen in May during the SWM season, while the highest rainfall peak is observed during the



**Figure 2 |** Boxplots of the observed monthly rainfall distributions.

second intermonsoon season, as shown in [Figure 2](#). As mentioned in the earlier section, these northwestern and inland areas are sheltered by the Titiwangsa ranges and have less influence on the NEM season. The difference in the number of harmonics between those areas may be due to the fluctuation in rainfall values throughout the year based on their geographical locations.

Table 4(c) shows that two harmonics are the best way to characterise the rainfall pattern for Lenggong station in the western Peninsula. The results presented in this table indicate that with only one harmonic, the deviations are still much larger, suggesting that the model still did not suit very well. When two harmonics are added to the model, the deviation is substantially reduced, suggesting that this term is still needed, but no more harmonics are required. As a result, two harmonics are ideal for describing the bimodal rainfall pattern in Lenggong. As shown in Figure 2, Lenggong displayed a bimodal rainfall pattern, representing two seasonal rainfall peaks in April and October during the intermonsoon season. This result is consistent with the findings of Suhaila & Jemain (2009a), in which they concluded that heavy rainfall was observed during the intermonsoon in the western Peninsula.

The seasonal rainfall pattern observed in Pontian is characterised by the utilisation of three harmonics, as depicted in Table 4(d). Figure 2 shows three rainfall peaks that could be considered: April, August, and November, corresponding to the SWM, NEM, and intermonsoon seasons. The result suggests that Pontian, situated in the southwestern region of the Peninsula, receives rainfall consistently throughout the year. A similar pattern could be seen for Tangkak and Melaka stations. Upon comparing these three peaks, it becomes apparent that the rainfall distribution on the southwestern Peninsula is more likely to be influenced by the NEM season due to the highest rainfall peak compared to the other seasons. Despite the northwest and southwest Peninsula yielding three harmonics, the rainfall patterns between those studied stations differ. These distinctions could be due to their geographical locations, which impact the local rainfall patterns with the monsoon circulations. Conversely, the eastern region required a large number of harmonics to depict its rainfall pattern effectively. Four harmonics are needed to describe the rainfall pattern in Kota Bharu due to fluctuating rainfall values throughout the year. The highest peak presented in Figure 2 is observed either in November or December. Suhaila & Jemain (2009a) affirmed that rainfall distribution in the eastern region is highly associated with the NEM flow.

The present study introduces a distinct model configuration compared to the approach adopted by Hasan & Dunn (2010, 2012). The number of sine and cosine terms was first examined, and the optimal number was chosen based on the deviance analysis. The selection of an optimal number of seasonal terms yielded the most accurate depiction of the rainfall distributions across the observed stations within Peninsular Malaysia. This choice is particularly relevant due to the profound impact of monsoon patterns on the climatic conditions prevailing in Malaysia.

### The fitted GLM Tweedie models

Table 5 presents the findings from the analysis of five selected case studies chosen as illustrative examples. The influence of the seasonal component, characterised by sine and cosine functions, in conjunction with the climatic predictors, ENSO indices, specifically SOI and MEI, has a significant impact on the monthly rainfall amount observed at the Chuping station, as shown in Table 5(a). The LRT is employed to assess the significance of climatic predictors compared to the initial base model centred on seasonal terms. Based on the lowest AIC, it was found that the presence of MEI values within the model has a statistically significant effect on the measured rainfall amount at the Chuping station. In contrast, when considering the model incorporating SOI and two harmonic functions, it is noted that this configuration yields a better fit for the rainfall distribution in the inland region than the model with MEI, as presented in Table 5(b).

Table 5(c) shows that all seasonal terms with two harmonics, except for the first cosine term, significantly impact the monthly rainfall amounts for the Lenggong station. In addition, based on the LRT statistics, models featuring climate predictors offer a significantly improved fit compared to the base model. Among these, the model with MEI demonstrates the most optimal fit. Contrasting results are observed at the southwestern station, as exemplified by the Pontian station, where none of the climate predictors substantially impact the monthly rainfall patterns, as detailed in Table 5(d). Consequently, adding climate predictors did not significantly influence the model for stations in the southwestern region. In other words, the monthly rainfall distribution for stations in the southwestern region is completely characterised by the model, consisting of seasonal terms. For the Kota Bharu station in the eastern region, both SOI and MEI together with the seasonal terms (excluding the sine term for the fourth harmonic) show a significant impact on the monthly rainfall amounts. Through an evaluation of the AIC, the model featuring the SOI emerges as the preferred choice for accurately modelling the monthly rainfall at the Kota Bharu station.

According to Richard & Walsh (2018), a La Niña event can be identified by a positive value of the SOI, typically accompanied by an increase in the total amount of rain. In addition, as outlined in the work of Wong *et al.* (2016), negative MEI values reflect the cold ENSO phase associated with the La Niña events, whereas positive MEI values correspond to the

**Table 5** | Estimated values of the regression coefficients for (a) the Chuping station (northwest region), (b) the Bayan Lepas station (inland region), (c) the Lenggong station (west region), (d) the Pontian station (southwest region), and (e) the Kota Bharu station (east region)

<b>(a)</b>						
<b>LRT Statistics</b>	<b>Model 1 (Base model)</b> <b>AIC = 5464.46</b>		<b>Model 2 (Base Model + SOI)</b> <b>AIC = 5457.26</b>		<b>Model 3 (Base Model +MEI)</b> <b>AIC = 5450.06</b>	
			<b>9.20<sup>a</sup></b>		<b>16.40<sup>b</sup></b>	
	<b>Coefficients</b>	<b>t-value</b>	<b>Coefficients</b>	<b>t-value</b>	<b>Coefficients</b>	<b>t-value</b>
$\beta_0$ (constant)	4.91	160.33 <sup>b</sup>	4.90	159.62 <sup>b</sup>	4.90	161.83 <sup>b</sup>
$A_1$ (sine)	0.47	11.17 <sup>b</sup>	0.48	11.35 <sup>b</sup>	0.48	11.47 <sup>b</sup>
$B_1$ (cosine)	0.30	6.67 <sup>b</sup>	0.30	6.68 <sup>b</sup>	0.30	6.86 <sup>b</sup>
$A_2$ (sine)	-0.41	-9.60 <sup>b</sup>	-0.42	-9.65 <sup>b</sup>	-0.42	-9.76 <sup>b</sup>
$B_2$ (cosine)	-0.19	-4.50 <sup>b</sup>	-0.19	-4.42 <sup>b</sup>	-0.19	-4.54 <sup>b</sup>
$A_3$ (sine)	0.13	3.13 <sup>a</sup>	0.13	3.11 <sup>a</sup>	0.13	3.16 <sup>a</sup>
$B_3$ (cosine)	-0.12	-2.90 <sup>a</sup>	-0.12	-2.89 <sup>a</sup>	-0.12	-2.96 <sup>a</sup>
$A_4$ (sine)						
$B_4$ (cosine)						
$C_3$ (SOI)			0.09	2.78 <sup>a</sup>		
$C_4$ (MEI)					-0.11	-3.75 <sup>b</sup>
<b>(b)</b>						
<b>LRT Statistics</b>	<b>Model 1 (Base model) AIC = 5647.83</b>		<b>Model 2 (Base Model + SOI)</b> <b>AIC = 5643.09<sup>c</sup></b>		<b>Model 3 (Base Model +MEI)</b> <b>AIC = 5646.10</b>	
			<b>6.75<sup>a</sup></b>		<b>3.73</b>	
	<b>Coefficients</b>	<b>t-value</b>	<b>Coefficients</b>	<b>t-value</b>	<b>Coefficients</b>	<b>t-value</b>
$\beta_0$ (constant)	5.17	210.84 <sup>b</sup>	5.17	211.62 <sup>b</sup>	5.17	211.42 <sup>b</sup>
$A_1$ (sine)	0.44	13.20 <sup>b</sup>	0.45	13.44 <sup>b</sup>	0.46	13.31 <sup>b</sup>
$B_1$ (cosine)	0.25	7.09 <sup>b</sup>	0.25	7.12 <sup>b</sup>	0.25	7.15 <sup>b</sup>
$A_2$ (sine)	-0.35	-10.31 <sup>b</sup>	-0.35	-10.42 <sup>b</sup>	-0.35	-10.36 <sup>b</sup>
$B_2$ (cosine)	-0.22	-6.54 <sup>b</sup>	-0.22	-6.51 <sup>b</sup>		
$A_3$ (sine)						
$B_3$ (cosine)						
$A_4$ (sine)						
$B_4$ (cosine)						
$C_5$ (SOI)			0.07	2.62 <sup>a</sup>		
$C_4$ (MEI)					-0.05	-1.91
<b>(c)</b>						
<b>LRT Statistics</b>	<b>Model 1 (Base model)</b> <b>AIC = 5451.101</b>		<b>Model 2 (Base Model + SOI)</b> <b>AIC = 5448.981</b>		<b>Model 3 (Base Model +MEI)</b> <b>AIC = 5444.948</b>	
			<b>4.12<sup>c</sup></b>		<b>8.15<sup>a</sup></b>	
	<b>Coefficients</b>	<b>t-value</b>	<b>Coefficients</b>	<b>t-value</b>	<b>Coefficients</b>	<b>t-value</b>
$\beta_0$ (constant)	5.02	214.63 <sup>b</sup>	5.01	215.66 <sup>b</sup>	5.01	217.45 <sup>b</sup>
$A_1$ (sine)	0.21	6.52 <sup>b</sup>	0.21	6.67 <sup>b</sup>	0.21	6.64 <sup>b</sup>
$B_1$ (cosine)	0.02	0.50	0.02	0.49	0.02	0.54
$A_2$ (sine)	-0.43	-13.04 <sup>b</sup>	-0.43	-13.20 <sup>b</sup>	-0.43	-13.30 <sup>b</sup>
$B_2$ (cosine)	-0.18	-5.39 <sup>b</sup>	-0.17	-5.38 <sup>b</sup>	-0.17	-5.41 <sup>b</sup>
$A_3$ (sine)						

(Continued.)

Table 5 | Continued

(c)

LRT statistics	Model 1 (base model + SOI) AIC = 5451.101		Model 2 (base model + SOI) AIC = 5448.981		Model 3 (base model + MEI) AIC = 5444.948	
	Coefficients	t-value	Coefficients	t-value	Coefficients	t-value
			4.12 <sup>c</sup>		8.15 <sup>a</sup>	
$B_3$ (cosine)						
$A_4$ (sine)						
$B_4$ (cosine)						
$C_3$ (SOI)			0.05	2.00 <sup>c</sup>		
$C_4$ (MEI)					-0.07	-2.78 <sup>a</sup>

(d)

LRT Statistics	Model 1 (Base model) AIC = 5679.23		Model 2 (Base Model + SOI) AIC = 5678.86		Model 3 (Base Model +MEI) AIC = 5678.83	
	Coefficients	t-value	Coefficients	t-value	Coefficients	t-value
			2.37		2.41	
$\beta_0$ (constant)	5.27	241.10 <sup>b</sup>	5.27	241.02 <sup>b</sup>	5.27	241.13 <sup>b</sup>
$A_1$ (sine)	0.14	4.61 <sup>b</sup>	0.14	4.61 <sup>b</sup>	0.14	4.61 <sup>b</sup>
$B_1$ (cosine)	-0.01	-3.23 <sup>a</sup>	-0.10	-3.22 <sup>a</sup>	-0.10	-3.22 <sup>a</sup>
$A_2$ (sine)	-0.14	-4.59 <sup>b</sup>	-0.14	-4.60	-0.14	-4.60 <sup>b</sup>
$B_2$ (cosine)	0.01	0.24	0.01	0.23	0.01	0.23
$A_3$ (sine)	0.05	1.48	0.05	1.50	0.05	1.50
$B_3$ (cosine)	-0.12	-3.80 <sup>a</sup>	-0.12	-3.78 <sup>b</sup>	-0.12	-3.78 <sup>b</sup>
$A_4$ (sine)						
$B_4$ (cosine)						
$C_3$ (SOI)			0.01	0.22		
$C_4$ (MEI)					0.01	0.28

(e)

LRT Statistics	Model 1 (Base model) AIC = 5784.20		Model 2 (Base Model + SOI) AIC = 5757.89		Model 3 (Base Model +MEI) AIC = 5766.55	
	Coefficients	t-value	Coefficients	t-value	Coefficients	t-value
			28.31 <sup>b</sup>		19.65 <sup>b</sup>	
$\beta_0$ (constant)	5.12	146.72 <sup>b</sup>	5.10	153.31 <sup>b</sup>	5.11	150.75 <sup>b</sup>
$A_1$ (sine)	0.50	9.80 <sup>b</sup>	0.53	10.99 <sup>b</sup>	0.51	10.45 <sup>b</sup>
$B_1$ (cosine)	-0.48	-9.91 <sup>b</sup>	-0.47	-10.24 <sup>b</sup>	-0.47	-10.03 <sup>b</sup>
$A_2$ (sine)	-0.31	-6.26 <sup>b</sup>	-0.33	-6.84 <sup>b</sup>	-0.32	-6.58 <sup>b</sup>
$B_2$ (cosine)	0.34	6.95 <sup>b</sup>	0.35	7.63 <sup>b</sup>	0.34	7.34 <sup>b</sup>
$A_3$ (sine)	0.28	5.71 <sup>b</sup>	0.28	5.92 <sup>b</sup>	0.28	5.84 <sup>b</sup>
$B_3$ (cosine)	-0.24	-4.94 <sup>b</sup>	-0.24	-5.23 <sup>b</sup>	-0.24	-5.11 <sup>b</sup>
$A_4$ (sine)	-0.06	-1.19	-0.06	-1.27	-0.06	-1.26
$B_4$ (cosine)	0.19	3.94 <sup>b</sup>	0.19	4.13 <sup>b</sup>	0.19	4.03 <sup>b</sup>
$C_3$ (SOI)			0.17	5.07 <sup>b</sup>		
$C_4$ (MEI)					-0.14	-4.18 <sup>b</sup>

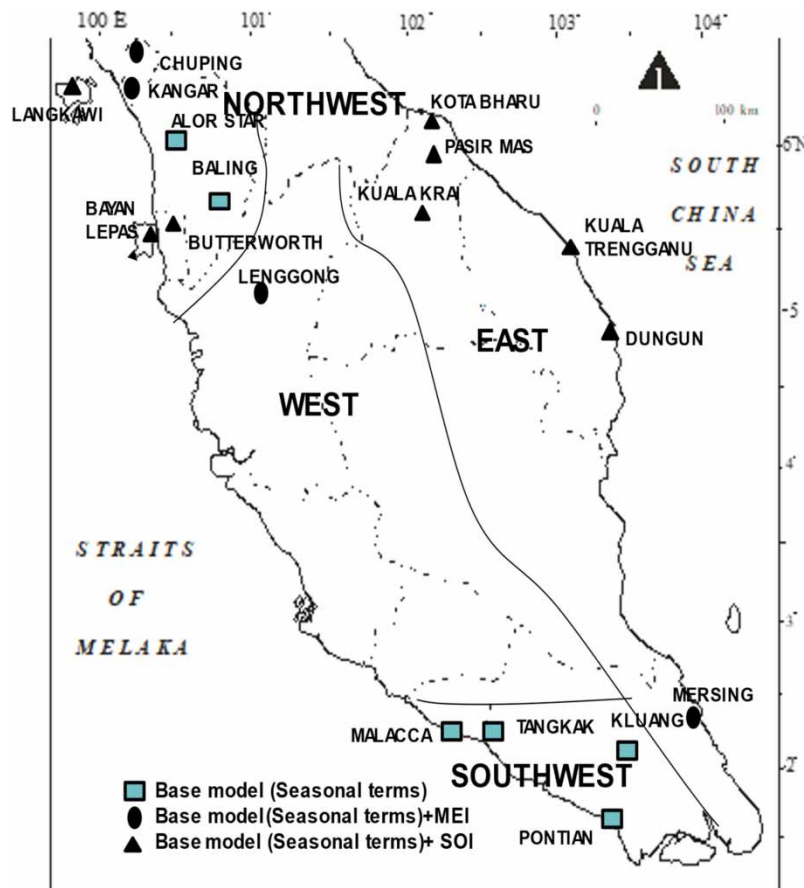
<sup>a</sup> $p < 0.001$ .<sup>b</sup> $0.001 \leq p < 0.01$ .<sup>c</sup> $0.001 \leq p < 0.05$ .



warm ENSO phase associated with the El Niño episodes. As can be seen in Tables 5(a), 5(c), and 5(e), there were cases of regression coefficients that exhibited significant positive SOI as well as significant negative MEI. These combined positive values of SOI and negative values of MEI suggest the prevalence of the La Niña episodes.

The SOI shows positive regression coefficients for all case-studied stations, with the largest impact at the Kota Bharu station in the eastern areas, as shown in Table 5(e). These findings can be explained in such a way that the increase in the positive coefficients of SOI will lead to an increase in the amount of rainfall. On the other hand, a converse relationship is observed between the MEI and the rainfall amount. The more severely negative MEI values will lead to an increase in the amount of rain. This phenomenon is particularly evident at the Chuping and Lenggong stations, which represent the northwestern and western regions of Peninsular Malaysia in the study. These stations exhibit notably larger and more significant negative MEI values. On the contrary, neither SOI nor MEI has a substantial impact on the amount of rainfall recorded at the Pontian station.

Figure 3 illustrates the preferred model determined by LRT statistics across all 18 meteorological stations. The findings concluded that the model with climatic predictors significantly enhances the accuracy of modelling monthly rainfall amounts for most stations. The base model with SOI values was preferred by those stations located in the inland and eastern regions of Peninsular Malaysia. The fitted model with SOI is noticeably better at these stations than the base model with MEI. Conversely, the base model with MEI is preferred for the Chuping, Kangar, Lenggong, and Mersing stations. Interesting findings are observed for Alor Star and Baling stations in the northwestern and the whole stations in the southwestern regions where there are no significant improvements of the models with the climatological predictors. Therefore, a base model comprising seasonal terms emerges as the most suitable choice for these stations. The findings from the LRT statistics could be summarised such that eight stations preferred the base model with SOI, four chose the base model with MEI, and the remaining six preferred the base model without any climate predictors to be the optimal fit.



**Figure 3** | The preferred model of each studied station based on the lowest AIC and LRT statistics.

### Model interpretation

Consider the model fitted to the monthly rainfall amounts for the Kota Bharu station with the index Tweedie parameter  $\hat{p} = 1.655$  (Table 2). Based on the deviance analysis, the fitted model with four harmonics is the best for the Kota Bharu station. Note that the model uses the logarithmic link function.

#### Model 1 (base model)

$$\begin{aligned} \mu_i = & \exp(5.12) \exp\left\{0.50 \sin\left(\frac{\pi(m-6)}{6}\right) - 0.48 \cos\left(\frac{\pi(m-6)}{6}\right)\right\} \exp\left\{-0.31 \sin\left(\frac{2\pi(m-6)}{6}\right) + 0.34 \cos\left(\frac{2\pi(m-6)}{6}\right)\right\} \\ & \exp\left\{0.28 \sin\left(\frac{3\pi(m-6)}{6}\right) - 0.24 \cos\left(\frac{3\pi(m-6)}{6}\right)\right\} \exp\left\{-0.06 \sin\left(\frac{4\pi(m-6)}{6}\right) + 0.19 \cos\left(\frac{4\pi(m-6)}{6}\right)\right\} \end{aligned}$$

#### Model 2 (base model + SOI)

$$\begin{aligned} \mu_i = & \exp(5.10) \exp\left\{0.53 \sin\left(\frac{\pi(m-6)}{6}\right) - 0.47 \cos\left(\frac{\pi(m-6)}{6}\right)\right\} \exp\left\{-0.33 \sin\left(\frac{2\pi(m-6)}{6}\right) + 0.35 \cos\left(\frac{2\pi(m-6)}{6}\right)\right\} \\ & \exp\left\{0.28 \sin\left(\frac{3\pi(m-6)}{6}\right) - 0.24 \cos\left(\frac{3\pi(m-6)}{6}\right)\right\} \exp\left\{-0.06 \sin\left(\frac{4\pi(m-6)}{6}\right) + 0.19 \cos\left(\frac{4\pi(m-6)}{6}\right)\right\} \\ & \exp\{0.18\text{SOI}_{i-1}\} \end{aligned}$$

#### Model 3 (base model + MEI)

$$\begin{aligned} \mu_i = & \exp(5.11) \exp\left\{0.51 \sin\left(\frac{\pi(m-6)}{6}\right) - 0.47 \cos\left(\frac{\pi(m-6)}{6}\right)\right\} \exp\left\{-0.32 \sin\left(\frac{2\pi(m-6)}{6}\right) + 0.35 \cos\left(\frac{2\pi(m-6)}{6}\right)\right\} \\ & \exp\left\{0.28 \sin\left(\frac{3\pi(m-6)}{6}\right) - 0.24 \cos\left(\frac{3\pi(m-6)}{6}\right)\right\} \exp\left\{-0.06 \sin\left(\frac{4\pi(m-6)}{6}\right) + 0.19 \cos\left(\frac{4\pi(m-6)}{6}\right)\right\} \\ & \exp\{-0.14\text{MEI}_{i-1}\} \end{aligned}$$

Suppose that  $m = 3$ ,  $t = 10$  (tenth year of study 1989), then  $i = 12(t - 1) + m = 12(9) + 3 = 113$ . The values of SOI and MEI for February 1989 are 1.2 and  $-1.06$ , respectively. The mean predicted rainfall amounts in March 1989 using model 1, model 2, and model 3 are 115.58, 135.10, and 130.11 mm, respectively. The mean predicted rainfall amounts for the rest of the months can be found by substituting the appropriate  $m$  and  $i$ .

On the other hand, the Tweedie model can also provide useful information on certain rainfall events. The parameters of the Tweedie distribution  $(\mu, p, \phi)$  can be reparameterized to the Poisson with a parameter  $\lambda$  that represents the mean number of rainfall events and Gamma distribution with two parameters  $(\alpha, \gamma)$ . The mean amount of rains per rainfall event can be computed as  $\alpha\gamma$ , where  $\alpha$  and  $\gamma$  represent the shape and scale parameters, respectively. Consider the Kota Bharu station as an example. The estimated parameters of the Tweedie model, as presented earlier in Table 2, are  $\hat{p} = 1.655$ , and  $\hat{\phi} = 3.760$ . The parameters of  $(\lambda, \gamma, \alpha)$  for Model 1 can be estimated as follows:

$$\hat{\lambda} = \frac{\mu^{2-p}}{\phi(2-p)} = \frac{115.58^{2-1.655}}{3.760(2-1.655)} = 3.90$$

$$\hat{\gamma} = \phi(p-1)\mu^{p-1} = 3.760(1.665-1)115.58^{1.665-1} = 58.86$$

$$\hat{\alpha} = \frac{p-2}{1-p} = \frac{1.665-2}{1-1.665} = 0.504.$$

Based on these parameters, the predicted mean number of rainfall events in March 1989 is 3.90 and the mean amounts of rainfall per event is  $\hat{\alpha}\hat{\gamma} = (0.504)(58.86) = 29.67\text{mm}$ . The estimations are repeated for Model 2 and 3. The predicted mean numbers of rainfall events are given as 4.10 and 4.06 mm, and the mean amounts of rainfall per event are 32.91, and 32.10 mm, respectively. In addition, the probability of obtaining no rain in that month for Model 1 can be predicted as  $P(Y = 0) = \exp(-\lambda) = \exp(-3.90) = 0.02$ , which is 2%. For months with no rain, the fitted Tweedie model predicts a small value of the probability of getting no rain even though the observed probability of no rain in March is zero, but the model predicts a probability of 0.02. It shows that while historical data do not have zero rainfall months, the probability of months with no rainfall in the future is still being considered in the model.

### Cross-validation and model selection

The methods of  $k$ -fold cross-validation and the Taylor diagram have been employed for all three tested models to assess the effectiveness of the best-fitting model. Table 6 lists the adjusted cross-validation prediction error for five stations taken as a case study. This analysis encompasses the utilisation of both 10-fold and 20-fold cross-validation to ensure the consistency of the results. Chuping and Lenggong stations are well fitted with the base model with the addition of the MEI, while Bayan Lepas and Kota Bharu stations are well suited with the addition of the SOI and the base model. For the Pontian station, no addition of climate predictors best described the monthly rainfall data of the station.

Figure 4 presents the Taylor diagram illustrating the statistics of the predicted rainfall of each tested model against observations at five case study stations. A low correlation (0.3) is observed for the Pontian station. In contrast, Chuping, Lenggong, and Bayan Lepas stations exhibit moderate correlations of around 0.6, reflecting a more substantial alignment between predictions and actual rainfall observations. The Kota Bharu station has a high correlation of nearly 0.75, signifying a high degree of association between predicted and observed rainfall.

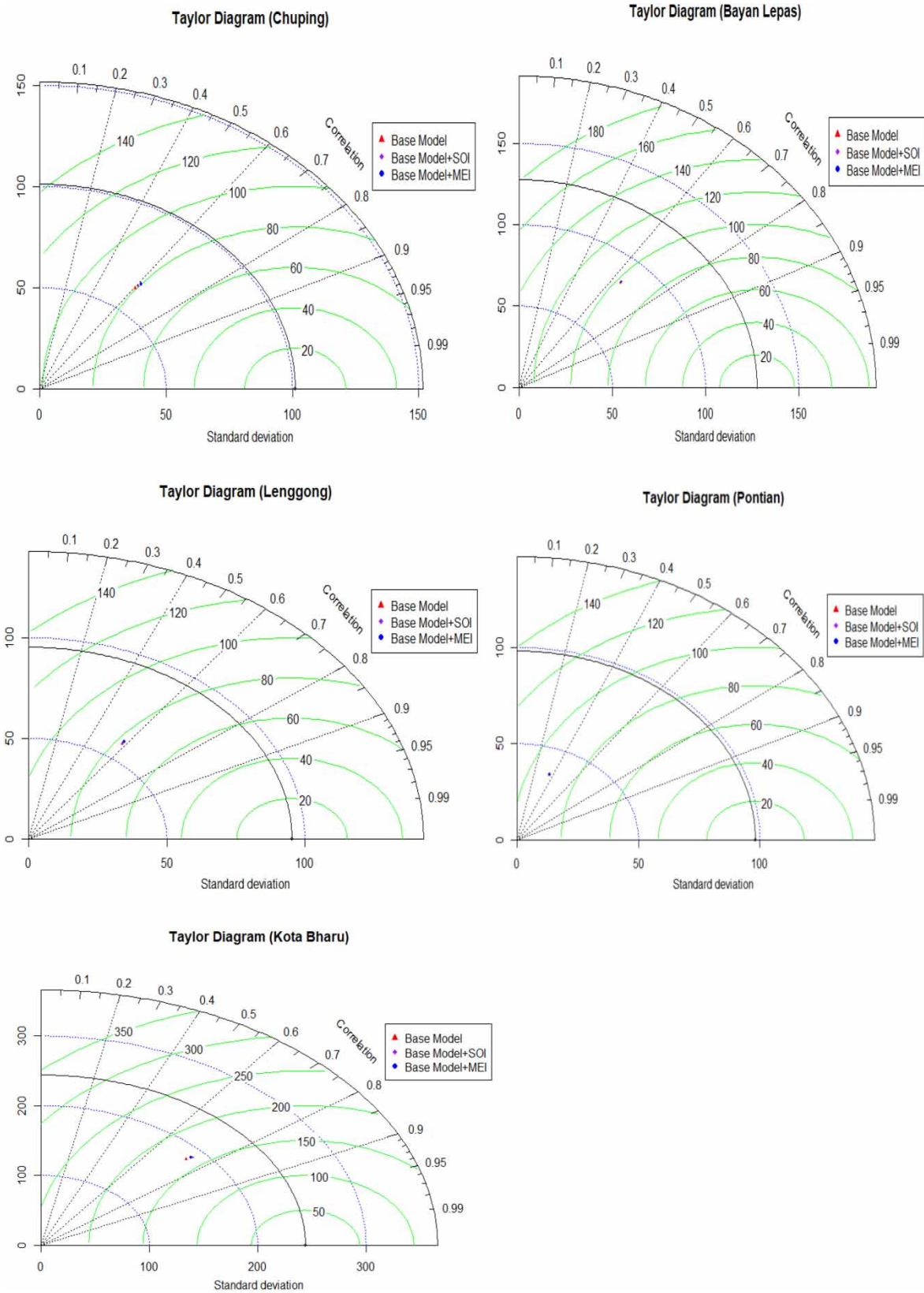
In comparing all three tested models, a base model with the SOI displays the closest agreement to the observed monthly rainfall values for the Kota Bharu station in the eastern region. On the other hand, the base model enhanced with the MEI index emerges as the optimal choice for the Chuping station in the northwest region, surpassing the performance of the other two models. For the rest of the stations, the Taylor diagram suggests a minimal difference in statistical metrics among the three models.

### Limitations of study and future work

This study aims to identify an appropriate set of harmonic functions, which include sine and cosine functions, that can effectively represent the monthly rainfall data in Peninsular Malaysia. The main concern revolves around the availability of

**Table 6** | Tweedie GLM cross-validation prediction error of each station

Stations	Models	Cross-validation prediction error	
		K = 10	K = 20
Chuping (northwest)	Base model	6,704.72	6,670.65
	Base model + SOI	6,705.90	6,656.03
	Base model + MEI	<b>6,623.24</b>	<b>6,639.70</b>
Bayan Lepas (inland)	Base model	9,785.80	9,771.71
	Base model + SOI	<b>9,725.98</b>	<b>9,754.93</b>
	Base model + MEI	9,866.33	9,819.34
Lenggong (west)	Base model	6,162.00	6,134.14
	Base model + SOI	6,241.50	6,190.96
	Base model + MEI	<b>6,152.54</b>	<b>6,125.03</b>
Pontian (southwest)	Base model	<b>8,540.16</b>	<b>8,566.46</b>
	Base model + SOI	8,693.56	8,611.66
	Base model + MEI	8,604.62	8,594.16
Kota Bharu (east)	Base model	28,571.92	28,404.75
	Base model + SOI	<b>27,559.41</b>	<b>27,886.04</b>
	Base model + MEI	28,319.96	28,052.38



**Figure 4** | Taylor diagram showing the correlation and standard deviation ratio between the predicted and observed monthly rainfall at five case study stations. The legend shows the base model, the base model with SOI, and the base model with MEI.

datasets, as the study focuses solely on stations with an extensive study period and a percentage of missing daily rainfall data below 1%. Moreover, to achieve the objective of the study, only stations that recorded both zero and non-zero monthly rainfall during the study period were taken into account.

The subsequent challenge relates to the computational time required to calculate the regression coefficients for each tested harmonic function, potentially impacting the model's efficiency. Nevertheless, the outcomes yielded the optimal number of sine and cosine functions for accurately describing the monthly rainfall data, considering the substantial influence of monsoons on Malaysia's climate. These results exhibited higher accuracy than models that utilised a single harmonic function composed exclusively of a sine and cosine component.

The third limitation focuses on the Tweedie index parameter  $p$ . In this study, the Tweedie index parameter  $p$  was computed for each station but not on a monthly basis. According to Hasan & Dunn (2010), the estimated  $p$  values do not significantly affect the regression coefficient estimates but can have a slight impact on monthly variation. For the sake of simplicity, the same index parameter is applied to all three tested models: the base model (seasonal terms), the base model with the SOI, and the base model with the MEI. Cross-validation was conducted to assess the model's fitness using  $k$ -fold and Taylor diagrams presented in a graphical form. Despite these limitations, this study presents an optimal approach for selecting the correct number of sine and cosine functions to capture the seasonal rainfall pattern precisely. Moreover, this method could be employed to determine the influence of climate predictors on the monthly rainfall of the stations.

For future research, it is recommended to develop separate models for modelling rainfall data and comparing their performance against the Tweedie Poisson–Gamma distribution. Furthermore, it is advisable to compute the Tweedie index parameter for each tested model, enhancing its accuracy in describing monthly rainfall variation. The Tweedie family of distributions applies not only to monthly timescales but also to daily timescales. Finally, considering additional ENSO climate predictors in the model could provide insight into their influence on monthly rainfall.

## CONCLUSION

Modelling rainfall processes separately pose many challenges. Direct modelling of rainfall processes is often problematic because rainfall data consists of zeroes and non-zeroes. Many fitting models encounter problems when handling this kind of dataset. In Malaysia, the variation in rainfall patterns is strongly affected by monsoonal winds that bring heavy rainfall during certain months in different regions. Since the studied stations have some months with zero rainfall values within the continuous rainfall total, the Tweedie family of distributions is the best model to be applied to the Malaysian rainfall series. The present study focused on modelling the monthly rainfall distribution at 18 rain gauge stations across Peninsular Malaysia using a Tweedie Poisson–Gamma distribution under the GLM framework.

This study successfully demonstrated the practical applications of the Tweedie distribution to address situations where continuous rainfall data contain exact zero values. The choice of Tweedie distribution was justified within the range of Tweedie index parameter values ( $1 < p < 2$ ). The discrepancy in Tweedie parameter values observed among stations was attributed to their geographical locations and monsoon-induced variations. In contrast to the previous literature (Hasan & Dunn 2010, 2012; Yunus *et al.* 2017), incorporating seasonal terms using harmonic functions enhanced the model's ability to describe rainfall patterns and reveal distinct patterns across different regions. Furthermore, the influence of climatological predictors on monthly rainfall was examined, highlighting the significance of the SOI and MEI indices in explaining rainfall variations. The analysis showed that the base model with SOI provided the best fit for most stations, particularly in eastern and inland areas, while the base model with MEI demonstrated a significant impact on rainfall at specific northern stations.

The main contribution of this study lies in developing a Tweedie GLM that handles both zero and non-zero rainfall values and its extension to incorporate climatological predictors and the optimal number of seasonal terms. By re-parameterizing Tweedie parameters to the Poisson–Gamma distribution, simulations can be conducted to predict statistical aspects of monthly rainfall, aiding in agricultural planning, disaster mitigation, and water resource management.

In summary, the Tweedie GLM emerges as a suitable alternative for modelling rainfall distributions, effectively addressing the challenges posed by zero-inflated data. The insights gained from this study offer practical modelling solutions and contribute to a deeper understanding of the complex associations between climatological variables and rainfall distribution. This knowledge holds potential implications for climate forecasting and resource management within the region, emphasising the relevance of the findings for future applications and research.

## ACKNOWLEDGEMENTS

The authors are indebted to the Malaysian Meteorological Department for providing the daily rainfall data used in the study. The authors would like to express their gratitude to the Ministry of Higher Education (MOHE) for the funding under the Fundamental Research Grant Scheme (FRGS/1/2020/STG06/UTM/02/3) under vote 5F311. We are also grateful to Universiti Teknologi Malaysia for supporting this project.

## DATA AVAILABILITY STATEMENT

Data cannot be made publicly available; readers should contact the corresponding author for details.

## CONFLICT OF INTEREST

The authors declare there is no conflict.

## REFERENCES

- Alexandersson, H. 1986 A homogeneity test applied to precipitation data. *Int. J. Climatol.* **6**, 661–675.
- Annual Report 2020 Available from: <https://www.met.gov.my/en/penerbitan/laporan-tahunan/>
- Barrera-Animas, A. Y., Oyedele, L. O., Bilal, M., Akinosho, T. D., Delgado, J. M. D. & Akanbi, L. A. 2022 Rainfall prediction: a comparative analysis of modern machine learning algorithms for time-series forecasting. *Mach. Learn. Appl.* <https://doi.org/10.1016/j.mlwa.2021.100204>.
- Chandler, R. E. & Wheater, H. S. 2002 Analysis of rainfall variability using generalized linear models: a case study from the west of Ireland. *Water Resour. Res.* **38** (10), 1192.
- Chapman, T. G. 1998 Stochastic modelling of daily rainfall: the impact of adjoining wet days on the distribution of rainfall amounts. *Environ. Model. Softw.* **13**, 317–324.
- Chebana, F. & Ouarda, T. B. 2021 Multivariate non-stationary hydrological frequency analysis. *J. Hydrol.* **593**, 125907. <https://doi.org/10.1016/j.jhydrol.2020.125907>.
- Coe, R. & Stern, R. D. 1982 Fitting models to daily rainfall. *J. Appl. Meteorol.* **21**, 1024–1031.
- Das, P., Sachindra, D. A. & Chanda, K. 2022 Machine learning-based rainfall forecasting with multiple non-linear feature selection algorithms. *Water Resour. Manage.* **36**, 6043–6071. <https://doi.org/10.1007/s11269-022-03341-8>.
- Deni, S. M., Jemain, A. A. & Ibrahim, K. 2009 Fitting optimum order Markov chain models for daily rainfall occurrences in peninsular Malaysia. *Theor. Appl. Climatol.* **97**, 109–121.
- Dunn, P. K. 2004 Occurrence and quantity of precipitation can be modeled simultaneously. *Int. J. Climatol.* **24**, 1231–1239.
- Dunn, P. K. 2022 Evaluation of Tweedie Exponential Family Models. Package 'Tweedie'. <https://cran.r-project.org/web/packages/tweedie/tweedie.pdf>.
- Dunn, P. K. & Smyth, G. K. 2005 Series evaluation of tweedie exponential dispersion model densities. *Stat. Comput.* **15**, 267–280.
- Dzupire, N. C., Ngare, P. & Odongo, L. 2018 A Poisson-Gamma model for Zero Inflated Rainfall Data. *J. Probab. Stat.* **2018**, 1–12. <https://doi.org/10.1155/2018/1012647>.
- Gabriel, K. R. & Neumann, J. 1962 A Markov chain model for daily rainfall occurrence at Tel Aviv. *Q. J. R. Meteorol. Soc.* **88**, 90–95.
- Geng, S., Penning de Vries, F. W. T. & Supit, I. 1986 A simple method for generating daily rainfall data. *Agric. For. Meteorol.* **36**, 363–376.
- Gomyo, M. & Koichiro, K. 2009 Spatial and temporal variations in rainfall and the ENSO-rainfall relationship over Sarawak, Malaysian Borneo. *SOLA* **5**, 041–044. doi:10.2151/sola.2009–011.
- Hasan, M. & Dunn, P. K. 2010 A simple Poisson-Gamma model for modeling rainfall occurrence and amount simultaneously. *Agric. For. Meteorol.* **150**, 1319–1330.
- Hasan, M. & Dunn, P. K. 2011 Two Tweedie distributions that are near-optimal for modeling monthly rainfall in Australia. *Int. J. Climatol.* **31** (9), 1389–1397.
- Hasan, M. & Dunn, P. K. 2012 Understanding the effect of climatology on monthly rainfall amounts in Australia using Tweedie GLMs. *Int. J. Climatol.* **32**, 1006–1017.
- Hasan, M., Croke, B. F. W. & Karim, F. 2019 Spatial and seasonal variations and inter-relationship in fitted model parameters for rainfall totals across Australia at various timescales. *Climate* **7** (4), 1–11. <http://dx.doi.org/10.3390/cli7010004>.
- Hertig, E. & Jacobeit, J. 2015 Considering observed and future nonstationarities in statistical downscaling of Mediterranean precipitation. *Theor. Appl. Climatol.* **122**, 667–683. <https://doi.org/10.1007/s00704-014-1314-9>.
- Hertig, E., Merckenschlager, C. & Jacobeit, J. 2017 Change points in predictors–predictand relationships within the scope of statistical downscaling. *Int. J. Climatol.* **37**, 1619–1633. <https://doi.org/10.1002/joc.4801>.
- Hui-Mean, F., Yusof, F., Yusop, Z. & Suhaila, J. 2019 Trivariate copula in drought analysis: a case study in peninsular Malaysia. *Theor. Appl. Climatol.* **138**, 657–671. <https://doi.org/10.1007/s00704-019-02847-3>.
- James, G., Witten, D., Hastie, T. & Tibshirani, R. 2013 *An Introduction to Statistical Learning with Applications in R*, 1st edn. Springer Texts in Statistics. Springer, New York.

- Jørgensen, B. 1997 *The Theory of Dispersion Models*. Chapman and Hall, London.
- Juneng, L. & Tangang, F. T. 2005 Evolution of ENSO-related rainfall anomalies in Southeast Asia region and its relationship with atmosphere-ocean variations in Indo-Pacific sector. *Clim. Dyn.* **25** (4), 337–350.
- Liang, J., Tan, M. L., Catto, J. L., Hawcroft, M. K., Hodges, K. I. & Haywood, J. M. 2023 Projected near-term changes in monsoon precipitation over peninsular Malaysia in the high Res MIP multi-model ensembles. *Clim. Dyn.* **60**, 1151–1171. <https://doi.org/10.1007/s00382-022-06363-5>.
- Liyew, C. M. & Melese, H. A. 2021 Machine learning techniques to predict daily rainfall amount. *J. Big Data* **8**, 153. <https://doi.org/10.1186/s40537-021-00545-4>.
- McCullagh, P. & Nelder, J. A. 1989 *Generalized Linear Models*, 2nd edn. Chapman and Hall, London.
- Pomee, M. S., Ashfaq, M., Ahmad, B. & Hertig, E. 2020 Modeling regional precipitation over the Indus River basin of Pakistan using statistical downscaling. *Theor. Appl. Climatol.* **142**, 29–57. <https://doi.org/10.1007/s00704-020-03246-9>.
- Richard, S. & Walsh, K. J. E. 2018 The influence of El Niño–Southern oscillation on boreal winter rainfall over Peninsular Malaysia. *Theor. Appl. Climatol.* **134**, 121–138. <https://doi.org/10.1007/s00704-017-2262-y>.
- Robertson, A. W., Kirshner, S. & Smyth, P. 2003 *Hidden Markov Models for Modelling Daily Rainfall Occurrence over Brazil*. Technical report. University of California, CA.
- Ruwangika, A. M., Perera, A. & Rathnayake, U. 2020 Comparison of statistical, graphical, and wavelet transform analyses for rainfall trends and patterns in Badulu Oya Catchment, Sri Lanka. *Complexity*. **2020**, Article ID 7146593, 1–13. <https://doi.org/10.1155/2020/7146593>.
- Sa'adi, Z., Shahid, S., Tarmizi, I., Chung, E.-S. & Wang, X.-J. 2017 Trends analysis of rainfall and rainfall extremes in Sarawak, Malaysia using modified Mann–Kendall test. *Meteorol. Atmos. Phys.* **131**, 2. <https://doi.org/10.1007/s00703-017-0564-3>.
- Serinaldi, F. 2009 A multisite daily rainfall generator driven by bivariate copula-based mixed distributions. *J. Geophys. Res.* **114**, D10103. doi:10.1029/2008JD011258.
- Serinaldi, F. & Kilsby, C. G. 2014 Simulating daily rainfall fields over large areas for collective risk estimation. *J. Hydrol.* **512**, 285–302.
- Sharda, V. N. & Das, P. K. 2005 Modelling weekly rainfall data for crop planning in a sub-humid climate of India. *Agric. Water Manage.* **76**, 120–138.
- Stern, R. D. & Coe, R. 1982 The use of rainfall models in agricultural planning. *Agric. Meteorol.* **26**, 35–50.
- Stern, R. D. & Coe, R. 1984 A model fitting analysis of daily rainfall data. *J. R. Stat. Soc. Ser. A* **147**, 1–34.
- Suhaila, J. & Jemain, A. A. 2009a A comparison of the rainfall patterns between stations on the East and the West coasts of Peninsular Malaysia using the smoothing model of rainfall amounts. *Meteorol. Appl.* **16** (3), 391–401.
- Suhaila, J. & Jemain, A. A. 2009b Investigating the impacts of adjoining wet days on the distribution of daily rainfall amounts in Peninsular Malaysia. *J. Hydrol.* **368** (1–4), 17–25.
- Suhaila, J. & Yusop, Z. 2017 Trend analysis and change point detection of annual and seasonal temperature series in Peninsular Malaysia. *Meteorol. Atmos. Phys.*, 1–17. doi:10.1007/s00703-017-0537-6.
- Suhaila, J., Deni, S. M., Zin, W. Z. W. & Jemain, A. A. 2010 Spatial patterns and trends of daily rainfall regime in Peninsular Malaysia during the southwest and northeast monsoons:1975–2004. *Meteorol Atmos Phys* **110**, 1–18. <https://doi.org/10.1007/s00703-010-0108-6>
- Suhaila, J., Jemain, A. A., Hamdan, M. F. & Zin, W. Z. W. 2011 Comparing rainfall patterns between regions in Peninsular Malaysia via a functional data analysis technique. *J. Hydrol.* **411**, 197–206.
- Tan, M. L., Zhang, F., Chieh Derek, C. J., Yu, K. H., Shaharudin, S. M., Chan, N. W. & Asyirah, A. R. 2022 Spatio-temporal analysis of precipitation, temperature and drought from 1985 to 2020 in Penang, Malaysia. *Water Supply* **22** (5), 4757. doi:10.2166/ws.2022.140.
- Tangang, F. T. & Juneng, L. 2004 Mechanism of Malaysian rainfall anomalies. *J. Clim.* **17** (18), 3615–3621.
- Tangang, F., Farzanmanesh, R., Mirzaei, A., Supari., Salimun, E., Jamaluddin, A. F. & Juneng, L. 2017 Characteristics of precipitation extremes in Malaysia associated with El Niño and La Niña events. *International Journal of Climatology* **37** (S1), 696–716. <https://doi.org/10.1002/joc.5032>.
- Taylor, K. E. 2001 Summarizing multiple aspects of model performance in a single diagram. *J. Geophys. Res.* **106**, 7183–7192. doi:10.1029/2000JD900719.
- Wijngaard, J. B., Klein, T. A. M. G. & Können, G. P. 2003 Homogeneity of 20th century European daily temperature and precipitation series. *Int. J. Climatol.* **23**, 679–692.
- Wilks, D. S. 1998 Multisite generalization of a daily stochastic precipitation generation model. *J. Hydrol.* **210**, 178–191.
- Wilks, D. S. 1999 Interannual variability and extreme-value characteristics of several stochastic daily precipitation models. *Agric. For. Meteorol.* **93**, 153–169.
- Wong, C. L., Liew, J., Yusop, Z., Ismail, T., Venneker, R. & Uhlenbrook, S. 2016 Rainfall characteristics and regionalization in peninsular Malaysia based on a high resolution gridded data set. *Water* **8**, 500. doi:10.3390/w8110500.
- Ximenes, P. S. M. P., da Silva, A. S. A., Ashkar, F. & Stosic, T. 2021 Best-fit probability distribution models for monthly rainfall of Northeastern Brazil. *Water Sci. Technol.* **84** (6), 1541–1556. <https://doi.org/10.2166/wst.2021.304>.
- Yang, C., Chandler, R. E., Isham, V. S. & Wheeler, H. S. 2005 Spatial-temporal rainfall simulation using generalized linear models. *Water Resour. Res.* **41**, W11415. doi:10.1029/2004WR003739.
- Yunus, R. M., Hasan, M. M., Razak, N. A., Zubairi, Y. Z. & Dunn, P. K. 2017 Modeling daily rainfall with climatological predictors: Poisson-gamma generalized linear modeling approach. *Int. J. Climatol.* **37** (3), 1391–1399.