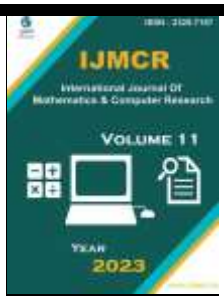
**International Journal of Mathematics and Computer Research**

**ISSN: 2320-7167**

**Volume 11 Issue 05 May 2023, Page no. – 3451-3457**

**Index Copernicus ICV: 57.55, Impact Factor: 7.362**

**DOI: 10.47191/ijmcr/v11i5.10**




# The Applied of Functional Data Analysis to Comparison 100 Times Simulated Monthly Rainfall Using Some two Parameters Distribution

**Masroh[1], Rado Yendra[2], Ari Pani Desvina[3], Muhammad Marizal[4]**

[1,2,3,4]Department of Mathematics, Faculty of Science and Technology, Universitas Islam Negeri Sultan Syarif Kasim Riau

**ARTICLE INFO**

Published Online:
30 May 2023

Corresponding Author:
**Muhammad Marizal**

**ABSTRACT**

Three probability models of mothly rainfall such as Gamma, Weibull and Log Normal distribution are evaluated in terms of their ability to reproduce the mean statistics derived from 100 times simulation of monthly rainfall in the Pekanbaru City, Indonesia. One of the important studies is to investigate and understand the simulate mean monthly rainfall patterns that occur throughout the year. To identify the pattern, it requires a rainfall curve to represent monthly observation of rainfall received during the year. Functional data analysis (FDA) methods are capable to convert discrete data into a function that can represent the rainfall curve and as a result, try to describe the hidden patterns of the rainfall. This study is focused on investigating 100 curve average monthly rainfall simulatated by three different quantile functions using the FDA. The mean and standard deviation of FDA for average monthly precipitation are obtained. Through these two statistics a the confidence interval curves of the mean curve are presented represent 95% pointwise confidence intervals. In this study, most of the monthly average rainfall from the actual data were around the FDA mean and the monthly average rainfall was within the FDA confidence interval. In this study, 100 times monthly rainfall simulations using the quantile function of the gamma and log normal distributions found that the mean FDA can capture most of the mean monthly rainfall from historical data, and within the FDA interval. The contradictory results shown by the monthly rainfall simulations using the Weibull distribution, most of the monthly average rainfall historical data cannot be captured by the FDA mean and are outside the FDA confidence interval. Based on the Mean Absolute Error (MAE) value of the average monthly rainfall of historical data and the monthly average of the FDA, it can be concluded that the gamma distribution can produce simulated rain better than the log normal distribution.

**KEYWORDS:** Probability Models, Simulation Monthly Rainfall, Functional Data Analysis, Gamma Distribution, Weibull Distribution, Log Normal Distribution

## I. INTRODUCTION

The simulation or generation of synthetic monthly rainfall data is important as it enables the generation of synthetic rainfall that has similar characteristics to the observed data. Thus, it will assist in cases where data is unavailable. Monthly rainfall data is generally needed in the simulation of water resources systems, and in the estimation of water yield from large catchments. Monthly streamflow data generation models are usually applied to generate monthly rainfall data, The probability model is very useful in simulating or generating synthetic monthly rainfall. In particular, the gamma distribution has been used many times to model rainfall totals on wet days[1]. Valuable general reviews on weather generators are published by [2] and [3]. More elaborate models have been proposed for the distribution of precipitation amounts given the occurrence of a wet day. Stern and Coe (1984)[4] used the two-parameter gamma distribution to describe the precipitation amount on wet days. An excellent review of stochastic weather models has been presented [5]. Although a large number of precipitation models have been developed, many practical applications require that weather generators produce other meteorological variables in addition to precipitation. Synthetic rain data generation for various time periods such as daily, monthly and yearly is well done in Australia, the quantile function of the two-parameter gamma probability density function plays

a very important role in this purpose [6]. Several algorithms have been developed for the purpose of generating daily, monthly and yearly synthetic rainfall. Rainfall generation algorithm (rGen) has been generated to produce annual synthetic rainfall [7] and new synthetic daily rainfall generated together [8] with a similar model to produce monthly synthetics total. Using the FDA method can transform the simulate monthly rainfall data into a curve or function Therefore, the FDA method is considered to be one of the most advanced techniques using all available data as curves [,9,10,11]. Ramsay and Silverman [12] gave a very good explanation on several functional methods such as principle component analysis, linear model, canonical correlation and discriminant analysis. It has been used in so many applications such as in environmental problem (Gao dan Niemier, [13]) and economy (Laukaitis and Rackauskas, [14]). FDA can also be used to detect the outlier in water quality (Muniz et al., [15]) and Martinez et al. [16] on outlier in air quality. Burfield et al. [17] used FDA in characterizing the chemical data and conclude that it is a powerful technique to detect the function minima and maxima even though they argued that the computational part was more complex compared to classical multivariate analysis. Sierra et al.[18] and Ruiz-bellido et al. [19] shared the same thought that functional data analysis is a promising and valuable tool in their research. There are two main focuses in this research, the first this study focuses on using several two-parameter probability models such as Weibull, Gamma and Log Normal in simulating monthly rainfall based on 100 time simulation using the quantile function of probability models. Parameter estimation using the maximum likelihood method The second, the study is focused on investigating 100 curve average monthly rainfall simulatated by three different quantile functions to choose the best distribution using the FDA. The mean and standard deviation of FDA for average monthly precipitation are obtained. Through these two statistics a the confidence interval curves of the mean curve are presented represent 95% pointwise confidence intervals.

## II. DATA

The analysis on the daily rainfall data for Pekanbaru meteorological stations, provided by the Meteorological, Climatological, and Geophysical Agency (BMKG) of Pekanbaru, Indonesia, is developed in this study. We have available records of rainfall data starting earlier than 1 January 1990. Data series continue until 31 December 2008, resulting in 6940 daily observations corresponding to 19 years. Note that in order to have clear visualisations of our analysis, in the following sections are shown just data source mean monthly precipitation over Pekanbaru city, Indonesia.

## III. METHODS

### *A. Probability Density Function (pdf) and Cumulataive Distribution Function (cdf)*

The primary tools to describe the mothly rainfall characteristics are probability density functions. Three two parameters probability density functions such as Weibull, Gamma and Log Normal will be used in this research. The pdf and cdf for each distribution that we consider are as given in Table 2, where *x* denote the observed values of the random variable representing the event of interest.

**Table I Pdf and Cdf distributions model**

| | Distribution | pdf ( *f(y)* ) dan cdf ( *F(y)* ) |
|---|---|---|
| 1 | Weibull (x;η,κ) | $f(x) = \frac{\eta}{\kappa}\left(\frac{x}{\kappa}\right)^{\eta-1} e^{\left(-\left(\frac{x}{\kappa}\right)^{\eta}\right)}, x > 0,\ \eta,\kappa > 0$<br>$F(x) = 1 - e^{\left(-\left(\frac{x}{\kappa}\right)^{\eta}\right)}$ |
| 2 | Gamma (x;α,β) | $f(x) = \frac{1}{\beta^{\alpha}\Gamma(\alpha)} x^{\alpha-1} e^{-\frac{x}{\beta}}, \quad x > 0, \alpha,\beta > 0$<br>$F(x) = \frac{1}{\beta^{\alpha}\Gamma(\alpha)} \int_{o}^{x} x^{\alpha-1} e^{-x/\beta} dx$ |
| 3 | Log Normal (x;μ,σ) | $f(x) = \frac{1}{\sqrt{2\pi\sigma x}} e^{\frac{-(\log x-\mu)}{2\sigma^2}}, x > 0, \quad \mu,\sigma > 0$<br>$F(x) = \Phi\left(\frac{\log x - \mu}{\sigma}\right)$<br>where $\Phi$ is the cumulative distribution function of the standard normal distribution (i.e., $N(0,1)$). |

Parameter estimation is the first thing that must be done in probability modeling. Like most studies that have been done, it can be concluded that the maximum likelihood (MLE) method is the most dominant used in this case. The MLE function for this model is implicit and complex and we will not discuss it in detail in this paper. The nonlinear equation generated by the maximum log likelihood function (ln L) requires a numerical method, namely Newton's rhapshon, to obtain a solution of the equation. But this method has been used in iteration systems to find the solution. Several initial values have been tested for this procedure. If the initial value used causes iteration of a certain value or the iterations converge to a value, then that value can be considered as the

selected estimation parameter. In this study the MLE method is also the main choice to be used in generating parameter estimates.

***B. Simulated Monthly Rainfall***

A large number of simulate monthly rainfall are generated using a two-parameter distribution such as gamma, weibull and log normal distribution. The two parameters, η and κ, used to describe the weibull distribution, α and β used to describe gamma distribution whereas μ and σ used to describe log normal distribution are found using maximum likelihood estimation. To simulate a sequence {$x[t]$} of synthetic monthly rainfall, we first generate realisations {$r[t]$} of a sequence {$R[t]$} of independent random numbers, each one uniformly distributed on the unit interval [0, 1], and then use quantile functions to solve the equation

$$F^{-1}[\alpha[t], \beta[t]](x) = r[t]$$

to find the corresponding monthly rainfall denoted by $x = x[t]$. $\alpha[t]$ and $\beta[t]$ are defined by the maximum likelihood estimates from the observed monthly data. using the same method can be applied to the weibull and log normal quantile function for the purpose of simulating monthly synthetic rainfall, preceded of course by parameter estimation for each probability density function.

***C. Smoothing functional data of rainfall , Mean and Standard Deviation Functional***

This analysis starts with smoothing raw data of rainfall using a technique of fitting models to data by minimizing the sum of squared errors. This approach consist on fitting the discrete observations $r_j, j = 1, \cdots, n$ for rainfall, using the following models:

$$r_j = x(l_j) + \epsilon_j$$

and a basis function expansion for $x(l)$ of the form

$$x(l) = \sum_{k=1}^{K} c_k \phi_k(l) = c'\phi$$

where vectors c of length K contain the coefficients $c_k$ and dm and assume that the residuals $\epsilon_j$ about the true curve are independently and identically distributed with mean zero and constant variance $\sigma^2$. Let define the $n$ x $K$ matrix $\mathbf{\Phi}$ as containing the values $\phi_k(l_j)$. Then, a simple linear smoother is obtained if the coefficients of the expansions $c_k$ are determined by minimizing the least squares criterions

$$LSE(r, c) = \sum_{j=1}^{n} \left( r_j - \sum_{k=1}^{K} c_k \phi_k(l_j) \right)^2$$

which in matrix form are expressed as:

$$LSE(r, c) = (r - \mathbf{\Phi}c)'(r - \mathbf{\Phi}c)$$

Taking the derivative of criterions $LSE(r, c)$ with respect to $c$ yield the equations

$$2\mathbf{\Phi}\mathbf{\Phi}'c - \mathbf{\Phi}'r = 0$$

and solving this for $c$ provides the estimators $\hat{c}$ that minimizes the least squares solution,

$$\hat{c} = (\mathbf{\Phi}'\mathbf{\Phi})^{-1}\mathbf{\Phi}'r$$

The vector $\hat{r}$ of fitted values is

$$\hat{r} = \mathbf{\Phi}(\mathbf{\Phi}'\mathbf{\Phi})^{-1}\mathbf{\Phi}'r$$

he functional observation for rainfall is expressed by

$$x(l) = \sum_{k=1}^{K} c_k \phi_k(l)$$

The smoothness of the fit can be controlled by the choice of K, which indicates the number of basis functions. The smaller the number of basis functions, the smoother the fit, and the larger the number of basis functions, the closer the fit will be to the data. The basis functions employed in this analysis, are Fourier basis functions since they perfectly represent periodic data. The set of basis functions for Fourier series includes one constant function and then pairs of sine and cosine functions to capture the variation in phase (the number of basis must always be odd):

$$\phi_1(l) = 1, \phi_2(l) = sin(l\omega), \phi_3(l) = cos(l\omega), \cdots, \phi_k(l) = sin\left(\frac{k}{2}l\omega\right), \phi_{k+1}(l) = cos\left(\frac{k}{2}l\omega\right)$$

where $\phi_k$ is the $k$th basis functions and $\omega = 2\pi/T$ where T is the period of the function.The traditional statistics for multivariate data are conventional to functional data. The mean function of curves is given as [20,21,22]:

$$\mu(l) = \frac{1}{N}\sum_{i=1}^{N} x_i(l)$$

while The standard deviation (SD) function of curves is given as

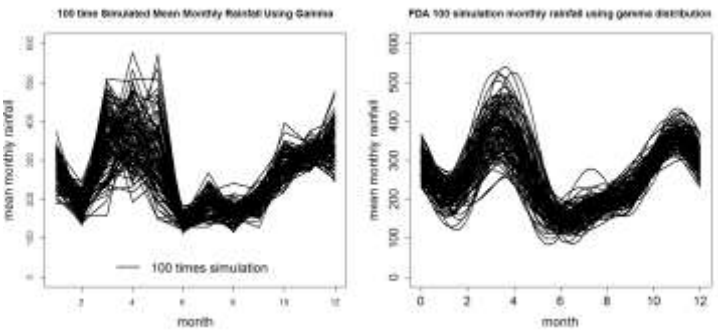
$$SD = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(x_i(l) - \mu(l)\right)^2}$$

## IV. RESULT

The parameters of Gamma, Weibull and Log Normal distribution are estimated using the maximum likelihood . Table 2 presents the estimated parameters of all distribution . Based on the parameters, simulate rainfall will be generated every month for 100 times . The quantile function or invers of the ditribution function will be used for this purpose. Statistics such as mean every month for 100 times simulate will be displayed using graphs. The mean of simulate monthly rainfall using gamma distribution represented using 100 black graph. This situation is explained from Figure 1 (a). The first step is to transform the discrete rainfall details into continuous functions or curves. Then, Fourier basis functions are preferred. The choice of k = 7 can be justified to capture the rainfall variation within a month. the discrete rainfall data were converted into functional data objects by using Fourier bases. All of the obtained functional data simulate of the mean monthly rainfall are represented in 100 black curve, as shown in Figure 1 (b); this plot gives an idea that data are periodic. Therefore, this periodicity can justify the use of the Fourier basis functions, in order to mainly reach the peaks and circles.

**Table II. Estimated Parameters for Some Two Parameters Distributions**

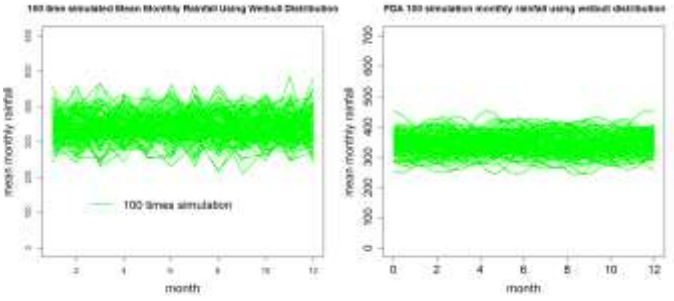
| | Weibull | | Gamma | | Log Normal | |
|---|---|---|---|---|---|---|
| | η | κ | α | β | μ | σ |
| January | 1.795 | 310.41 | 2.62 | 104.53 | 5.45 | 0.589 |
| February | 1.799 | 209.20 | 2.64 | 69.93 | 5.06 | 0.593 |
| March | 1.318 | 395.16 | 1.04 | 343.22 | 5.64 | 0.626 |
| April | 1.445 | 407.91 | 1.49 | 245.42 | 5.67 | 0.661 |
| May | 1.158 | 324.94 | 0.65 | 465.72 | 5.41 | 0.662 |
| June | 2.060 | 163.35 | 3.47 | 41.43 | 4.82 | 0.577 |
| July | 1.496 | 207.36 | 2.07 | 90.62 | 4.95 | 0.888 |
| August | 2.015 | 182.16 | 3.36 | 47.72 | 4.96 | 0.501 |
| September | 1.919 | 228.77 | 3.16 | 63.97 | 5.13 | 0.664 |
| October | 2.392 | 325.83 | 4.79 | 60.08 | 5.57 | 0.431 |
| November | 3.726 | 348.71 | 9.93 | 31.59 | 5.69 | 0.351 |
| December | 1.739 | 384.68 | 2.35 | 144.31 | 5.67 | 0.543 |




(a) (b)

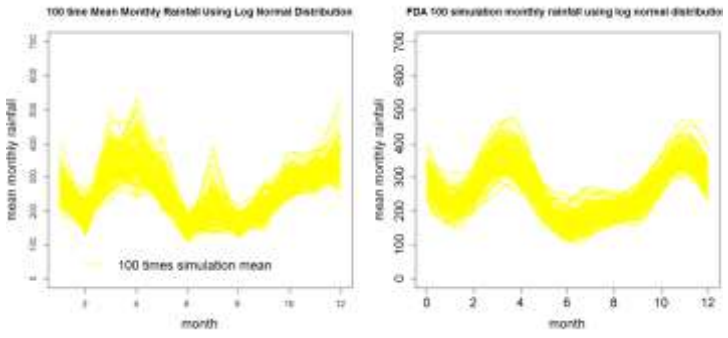
**Figure 1 (a) 100 times simulation mean monthly rainfall using gamma distribution (b)100 times functional data analysis**

The mean of simulate monthly rainfall using weibull distribution represented using 100 green graph, as shown in Figure 2(c). Using k = 7, then Fourier basis functions are conducted. The choice of k = 7 can be justified to capture the rainfall variation within a month. the discrete rainfall data were converted into functional data objects by using Fourier bases. All of the obtained functional data simulate of the mean monthly rainfall are represented in 100 green curve, as shown in Figure. 2 (d). Figure 3(e) shows the 100 mean simulate month rainfall obtained by using log normal distribution as displayed by 100 yellow graphs. The mean of simulate monthly rainfall curves obtain by using functional data analysis for 100 times simulation as shown in Figure 3(f). The pattern for mean monthly simulate rainfall curves show the bimodal shaped with many fluctuations. This pattern is believed to be the result of Pekanbaru's climate which is influenced by the two main monsoon seasons in Indonesia.




(c) (d)

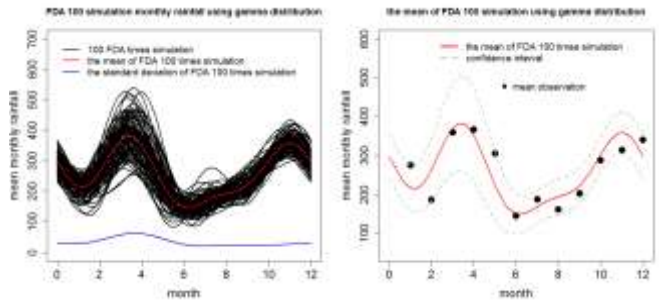
**Figure 2 (c) 100 times simulation mean monthly rainfall using weibull distribution (d)100 times functional data analysis**




(e) (f)

**Figure 3 (e) 100 times simulation mean monthly rainfall using log normal distribution (f) 100 times functional data analysis**

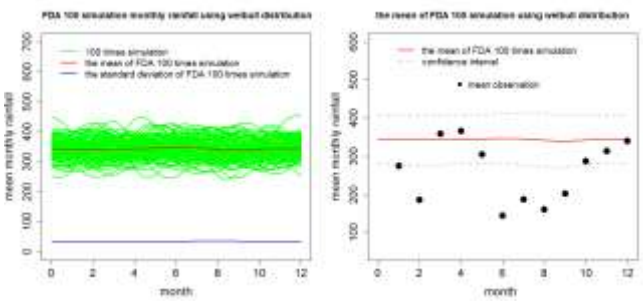
The mean and standard deviation of functional data for average simulate monthly rainfall using gamma distribution are presented in Figure 4(g) ,the solid curve is ploted in red is the mean function of average simulate monthly rainfall derived from the least squre error and standard deviation curve is plotted in blue line. The confidence interval curves of the mean curve are presented in Figure 4(h), and the grey dashed lines represent 95% pointwise confidence intervals on the mean curve based on the least square error smoothing estimate of measurement plotted in Figure 1(b) and standard deviation function plotted in Figure 4(g). From Figure 4 (h) it can be seen that the average monthly rainfall of historical data (black dots) can mostly be captured by the mean of FDA curve and all of these historical data are within the confidence intervals.




(g) (h)

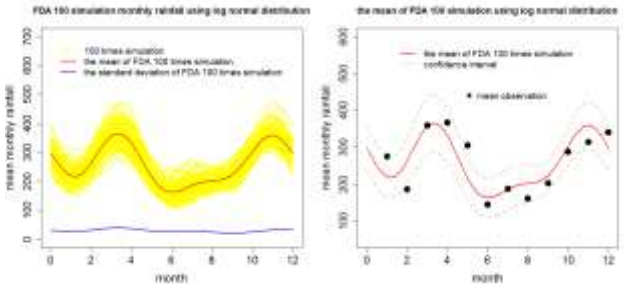
**Figure 4 (g) Plot of 100 times FDA mean simulation using gamma distribution (black lines), the mean FDA (red lines) and standard deviation (blue lines) (h) Plot of confidence interval curves of the mean FDA and mean monthly rainfall observation (black points)**

The mean and standard deviation of FDA for mean simulate monthly rainfall using weibull distribution are presented in Figure 5(i) ,the solid curve is ploted in red is the mean function of average simulate monthly rainfall derived from the least squre error and standard deviation curve is plotted in blue line. The confidence interval curves of the mean curve are presented in Figure 5(j), and the grey dashed lines represent 95% pointwise confidence intervals on the mean curve based on the least square error smoothing estimate of measurement plotted in Figure 2(d) and standard deviation function plotted in Figure 5(i). From Figure 5 (j) it can be seen that the average monthly rainfall of historical data (black dots) can not mostly be captured by the mean of FDA curve and mostly these historical data are outside the confidence intervals. Furthermore, the FDA will also be used to analyze the ability of the normal log distribution to produce monthly simulated rain. For this reason, a graph of the FDA average and FDA standard deviation for the average monthly rainfall of 100 simulations has been produced as shown in Figure 6(k). The confidence interval curves of the mean curve are presented in Figure 6(j), and the grey dashed lines represent 95% pointwise confidence intervals on the mean curve based on the least square error smoothing estimate of measurement plotted in Figure 3(b) and standard deviation function plotted in Figure 6(k). From the figure it can be seen that almost all of means monthly rainfall historical data (black dots) are within the confidence inteval, although the average rainfall for months 2 and 5 is still outside the confidence interval.




(i) (j)

**Figure 5 (i) Plot of 100 times FDA mean simulation using weibull distribution (black lines), the mean FDA (red lines) and standard deviation (blue lines) (j) Plot of confidence interval curves of the mean FDA and mean monthly rainfall observation (black points)**




(k) (l)

**Figure 6 (k) Plot of 100 times FDA mean simulation using log normal distribution (black lines), the mean FDA (red lines) and standard deviation (blue lines) (l) Plot of confidence interval curves of the mean FDA and mean monthly rainfall observation (black points)**

The Mean Absolute Error value obtained from the difference between mean monthly rainfall historical data and FDA is also displayed in ensuring the best distribution in the simulation. For this reason, Figure 7 is included for this purpose. From Figure 7 it can be seen that the gamma distribution is the best used to produce monthly rainfall simulations.

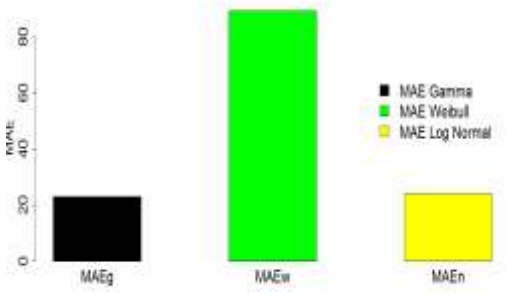



**Figure 7 The comparison Mean Absolut Error of Mean Monthly Rainfall FDA Using Gamma, Weibull, and Log Normal Distribution**

## V. CONCLUSION

This paper has performed a different practice in analyzing data by using functional data analysis. A functional data analysis has been applied in 100 types of mean the simulate monthly rainfall data using gamma, weibull, and log normal ditribution. This research analyzes and visualizes the average monthly rainfall throughout the last two decades for Pekanbaru City. FDA approaches with emphasis on smoothing and visualization were modified and applied for the rainfall measurements as an important step in a full FDA. Based on the results, the following main conclusions can be drawn from this work:

1. The entire rainfall observations were treated by FDA techniques like function data shown by curves which represent the actual phenomena better and display many graphic displays of rainfall data.
2. The least square error smoothing made it easy to choose the best smoothing parameter so that the average monthly rainfall functional data were determined. Therefore, the noise was decreased, and errors were eliminated.
3. The Mean Absolute Error value obtained from the difference between mean monthly rainfall historical data and FDA is also displayed in ensuring the best distribution in the simulation. Thr gamma distribution is the best used to produce monthly rainfall simulations

## REFERENCES


1. K. Rosenberg, J. Boland & P. G. Howlett. (2004). Simulati n of monthly rainfall totals, *ANZIAM J*. 46 (E) ppE85–E104,
2. Srikanthan, R., & McMahon, T. A. (2001). Stochastic generation of annual, monthly and daily climate data: A review. *Hydrology and Earth System Sciences, 5*(4), 633–670.
3. Wilks, D. S., & Wilby, R. L. (1999). The weather generation game: a review of stochastic weather models. *Progress in Physical Geography, 23*(3), 329–357.
4. Stern, R. D., & Coe, R. (1984). A model fitting analysis of daily rainfall. *Journal of the Royal Statistical Society. Series A., 147*, Part 1, 1–34.
5. Wilks, D. S., & Wilby, R. L. (1999). The weather generation game: a review of stochastic weather models. *Progress in Physical Geography, 23*(3), 329–357.
6. Julia Piantadosi, John Boland & Phil Howlett. (2008). Generating Synthetic Rainfall on Various Timescales—Daily, Monthly and Yearly, *Environ Model Assess*. 14:431–438. doi 10.1007/s10666-008-9157-3
7. Dick, N. P.,&Bowden, D. C. (1973).Maximum likelihood estimation for mixtures of two normal distributions. *Biometrics, 29*, 781–790.
8. Piantadosi, J., Howlett, P. G., & Boland, J. W. (2008). *A new model for correlated daily rainfall* (in press).
9. Sierra C, Flor-blanco G, Ordoñez C, Flor G, Gallego JR (2017) Analyzing coastal environments by means of functional data analysis. Sediment Geol 357:99–108
10. [10]. Bur R, Neumann C, Saunders CP (2015) Review and application of functional data analysis to chemical data—the example of the comparison, classification, and database search of forensic ink chromatograms. Chemom Intell Lab Syst 149:47–106
11. Müller H, Sen R, Stadtmüller U (2011) Functional data analysis for volatility. J Econom 165(2):233–245
12. Ramsay J. O. and Silverman B. W. Applied Functional Data Analysis: Methods and Case Studies. Springer. 2007.
13. Gao H. O., Niemeier D. A. Using functional data analysis of diurnal ozone and NOx cycles to inform transportation emissions control. Transportation Research Part D: Transport and Environment. 2008 June 1. 13(4): 221-38.
14. Laukaitis A, Rakauskas A. Functional data analysis for client's segmentation tasks. European Journal of Operational Research. 2005 May 16. 163(1): 210-6.
15. Mu˜niz C. D., Nieto P. G., Fern´andez J. A., Torres J. M., Taboada J. Detection of outliers in water quality monitoring samples using functional data analysis in San Esteban estuary (Northern Spain).

Science of the Total Environment. 2012 Nov 15. 439: 54-61.

16. Mart´ınez J., Saavedra ´A ., Garc´ıa-Nieto P. J., Pi˜neiro J. I., Iglesias C., Taboada J., Sancho J., Pastor J. Air quality parameters outlier's detection using functional data analysis in the Langreo urban area (Northern Spain). Applied Mathematics and Computation. 2014 Aug 15. 241: 1-0.
17. Burfield R., Neumann C., Saunders C. P. Review and application of functional data analysis to chemical data—The example of the comparison, classification, and database search of forensic ink chromatograms. Chemometrics and Intelligent Laboratory Systems. 2015 Dec 15. 149: 97 106.
18. Sierra C., Flor-Blanco G., Ordo˜nez C., Flor G., Gallego J. R. Analyzing coastalm environments by means of functional data analysis. Sedimentary Geology. 2017 July 15. 357: 99-108.
19. Ruiz-Bellido M. A., Romero-Gil V., Garc´ıa-Garc´ıa P., Rodr´ıguez-G´omez F., Arroyo-L´opez F. N., Garrido-Fern´andez A. Assessment of table olive fermentation by functional data analysis. International Journal of Food Microbiology. 2016 Dec. 5. 238: 1-6.
20. Chebana F, Dabo-Niang S, Ouarda TBMJ (2012) Exploratory functional flood frequency analysis and outlier detection. *Water Resour Res* 48(4):1–20
21. Ramsay JO, Silverman B (2005) Functional data analysis. Springer, New York. https ://doi.org/10.1007/b9888 8
22. Suhaila J, Yusop Z (2017) Spatial and temporal variabilities of rainfall data using functional data analysis. *Theor Appl Climatol* 129(1–2):229–242