Hindawi
Journal of Probability and Statistics
Volume 2018, Article ID 1012647, 12 pages
https://doi.org/10.1155/2018/1012647

Research Article

A Poisson-Gamma Model for Zero Inflated Rainfall Data

Nelson Christopher Dzupire,[1] **Philip Ngare**,[1,2] **and Leo Odongo**[1,3]

[1]*Pan African University Institute of Basic Sciences, Technology and Innovation, Juja, Kenya*
[2]*University of Nairobi, Nairobi, Kenya*
[3]*Kenyatta University, Nairobi, Kenya*

Correspondence should be addressed to Nelson Christopher Dzupire; ndzupire@cc.ac.mw

Received 7 November 2017; Accepted 20 February 2018; Published 4 April 2018

Academic Editor: Steve Su

Rainfall modeling is significant for prediction and forecasting purposes in agriculture, weather derivatives, hydrology, and risk and disaster preparedness. Normally two models are used to model the rainfall process as a chain dependent process representing the occurrence and intensity of rainfall. Such two models help in understanding the physical features and dynamics of rainfall process. However rainfall data is zero inflated and exhibits overdispersion which is always underestimated by such models. In this study we have modeled the two processes simultaneously as a compound Poisson process. The rainfall events are modeled as a Poisson process while the intensity of each rainfall event is Gamma distributed. We minimize overdispersion by introducing the dispersion parameter in the model implemented through Tweedie distributions. Simulated rainfall data from the model shows a resemblance of the actual rainfall data in terms of seasonal variation, means, variance, and magnitude. The model also provides mechanisms for small but important properties of the rainfall process. The model developed can be used in forecasting and predicting rainfall amounts and occurrences which is important in weather derivatives, agriculture, hydrology, and prediction of drought and flood occurrences.

1. Introduction

Climate variables, in particular, rainfall occurrence and intensity, hugely impact human and physical environment. Knowledge of the frequency of the occurrence and intensity of rainfall events is essential for planning, designing, and management of various water resources system [1]. Specifically rain-fed agriculture is a sensitive sector to weather and crop production is directly dependent on the amount of rainfall and its occurrence. Rainfall modeling has a great impact on crop growth, weather derivatives, hydrological systems, drought, and flood management and crop simulated studies.

Rainfall modeling is also important in pricing of weather derivatives which are financial instruments that are used as a tool for risk management to reduce risk associated with adverse or unexpected weather conditions.

Further as climate change greatly affects the environment there is an urgent need for predicting the variability of rainfall for future periods for different climate change scenarios in order to provide necessary information for high quality climate related impact studies [1].

However modeling precipitation poses a lot of challenges, namely, accurate measurement of precipitation since rainfall data consists of sequences of values which are either zero or some positive numbers (intensity) depending on the depth of accumulation over discrete intervals. In addition factors like wind can affect collection accuracy. Rainfall is localized unlike temperature which is highly correlated across regions; therefore a derivative holder based on rainfall may suffer geographical basis risk in case of pricing weather derivatives. The final challenge is the choice of a proper probability distribution function to describe precipitation data. The statistical property of precipitation is far more complex and a more sophisticated distribution is required [2].

Rainfall has been modeled as a chain dependent process where a two-state Markov chain model represents the occurrence of rainfall and the intensity of rainfall is modeled by fitting a suitable distribution like Gamma [3], exponential, and mixed exponential [1, 4]. These models are easy to

understand and interpret and use maximum likelihood to find the parameters. However models involve many parameters to fully describe the dynamics of rainfall as well as making several assumptions for the process.

Wilks [5] proposed a multisite model for daily precipitation using a combination of two-state Markov process (for the rainfall occurrence) and a mixed exponential distribution (for the precipitation amount). He found that the mixture of exponential distributions offered a much better fit than the commonly used Gamma distribution.

In study of Leobacher and Ngare [3] the precipitation is modeled on a monthly basis by constructing a suitable Markov-Gamma process to take into account seasonal changes of precipitation. It is assumed that rainfall data for different years of the same month is independent and identically distributed. It is assumed that precipitation can be forecast with sufficient accuracy for a month.

Another approach of modeling rainfall is based on the Poisson cluster model where two of the most recognized cluster based models in the stochastic modeling of rainfall are the Newman-Scott Rectangular Pulses model and the Bartlett-Lewis Rectangular Pulse model. These models represent rainfall sequences in time and rainfall fields in space where both the occurrence and depth processes are combined. The difficulty in Poisson cluster models as observed by Onof et al. [6] is the challenge of how many features should be addressed so that the model is still mathematically tractable. In addition the models are best fitted by the method of moments and so requires matching analytic expressions for the statistical properties such as mean and variance.

Carmona and Diko [7] developed a time-homogeneous jump Markov process to describe rainfall dynamics. The rainfall process was assumed to be in form of storms which consists of cells themselves. At a cell arrival time the rainfall process jumps up by a random amount and at extinction time it jumps down by a random amount, both modeled as Poisson process. Each time the rain intensity changes, an exponential increase occurs either upwards or downwards. To preserve nonnegative intensity, the downward jump size is truncated to the current jump size. The Markov jump process also allows for a jump directly to zero corresponding to the state of no rain [8].

In this study the rainfall process is modeled as a single model where the occurrence and intensity of rainfall are simultaneously modeled. The Poisson process models the daily occurrence of rainfall while the intensity is modeled using Gamma distribution as the magnitude of the jumps of the Poisson process. Hence we have a compound Poisson process which is Poisson-Gamma model. The contribution of this study is twofold: a Poisson-Gamma model that simultaneously describes the rainfall occurrence and intensity at once and a suitable model for zero inflated data which reduces overdispersion.

This paper is structured as follows. In Section 2 the Poisson-Gamma model is described and then formulated mathematically while Section 3 presents methods of estimating the parameters of the model. In Section 4 the model is fitted to the data and goodness of fit of the model is evaluated by mean deviance whereas quantile residuals perform the diagnostics check of the model. Simulation and forecasting are carried out in Section 5 and the study concludes in Section 6.

2. Model Formulation

2.1. Model Description. Rainfall comprises discrete and continuous components in that if it does not rain the amount of rainfall is discrete whereas if it rains the amount is continuous. In most research works [3, 4, 9] the rainfall process is presented by use of two separate models: one is for the occurrence and conditioned on the occurrence and another model is developed for the amount of rainfall. Rainfall occurrence is basically modeled as first or higher order Markov chain process and conditioned on this process a distribution is used to fit the precipitation amount. Commonly used distributions are Gamma, exponential, mixture of exponential, Weibull, and so on. These models work based on several assumptions and inclusion of several parameters to capture the observed temporal dependence of the rainfall process. However rainfall data exhibit overdispersion [10] which is caused by various factors like clustering, unaccounted temporal correlation, or the fact that the data is a product of Bernoulli trials with unequal probability of events. The stochastic models developed in this way underestimate the overdispersion of rainfall data which may result in underestimating the risk of low or high seasonal rainfall.

Our interest in this research is to simultaneously model the occurrence and intensity of rainfall in one model. We would model the rainfall process by using a Poisson-Gamma probability distribution which is flexible to model the exact zeros and the amount of rainfall together.

Rainfall is modeled as a compound Poisson process which is a Lévy process with Gamma distributed jumps. This is motivated by the sudden changes of rainfall amount from zero to a large positive value following each rainfall event which are modeled as pure jumps of the compound Poisson process.

We assume rainfall arrives in forms of storms following a Poisson process, and at each arrival time the current intensity increases by a random amount based on Gamma distribution. The jumps of the driving process represent the arrival of the storm events generating a jump size of random size. Each storm comprises cells that also arrive following another Poisson process.

The Poisson cluster processes gives an appropriate tool as rainfall data indicating presence of clusters of rainfall cells. As observed by Onof et al. [6] use of Gamma distributed variables for cell depth improves the reproduction of extreme values.

Lord [11] used the Poisson-Gamma compound process to model the motor vehicle crashes where they examined the effects of low sample mean values and small sample size on the estimation of the fixed dispersion parameter. Wang [12] proposed a Poisson-Gamma compound approach for species richness estimation.

2.2. Mathematical Formulation. Let N_t be total number of rainfall event per day following a Poisson process such that

$$P\left(N_t = n\right) = e^{-\lambda}\frac{\lambda^n}{n!}, \quad \forall n \in \mathbb{N},$$
$$N_t = \sum_{t\geq 1} 1_{[t,\infty)}\left(t\right). \tag{1}$$

The amount of rainfall is the total sum of the jumps of each rainfall event, say $(y_i)_{i\geq 1}$, assumed to be identically and independently Gamma distributed and independent of the times of the occurrence of rainfall:

$$L\left(t\right) = \begin{cases} \sum_{i=1}^{N_t} y_i & N_t = 1, 2, 3, \ldots \\ 0 & N_t = 0, \end{cases} \tag{2}$$

such that $y_i \sim \text{Gamma}(\alpha, P)$ is with probability density function

$$f\left(y\right) = \begin{cases} \dfrac{\alpha^P y^{P-1} e^{-\alpha y}}{\Gamma\left(P\right)} & y > 0, \\ 0 & y \leq 0. \end{cases} \tag{3}$$

Lemma 1. *The compound Poisson process (2) has a cumulant function*

$$\psi\left(s, t, x\right) = \lambda t\left(e^{M_Y(x)} - 1\right), \tag{4}$$

for $0 \leq s < t$ and $x \in \mathbb{R}$, where $M_Y(x)$ is the moment generating function of the Gamma distribution.

Proof. The moment generating function $\Phi(s)$ of $L(s)$ is given by

$$\begin{aligned} M_L\left(s\right) &= \mathbf{E}\left(e^{sL(t)}\right) \\ &= \sum_{j=0}^{\infty} \mathbf{E}\left(e^{sL(t)} \mid N\left(t\right) = j\right) P\left(N\left(t\right) = j\right) \\ &= \sum_{j=0}^{\infty} \mathbf{E}\left(e^{s(L(1)+L(2)+\cdots+L(j))} \mid N\left(t\right) = j\right) P\left(N\left(t\right) = j\right) \\ &= \sum_{j=0}^{\infty} \mathbf{E}\left(e^{s(L(1)+L(2)+\cdots+L(j))}\right) P\left(N\left(t\right) = j\right) \\ &\qquad \text{because of independence of } L \text{ and } N\left(t\right) \\ &= \sum_{j=0}^{\infty} \left(M_Y\left(s\right)\right)^j e^{-\lambda t}\frac{\left(\lambda t\right)^j}{j!} = e^{-\lambda t}\sum_{j=0}^{\infty}\left(M_Y\left(s\right)\right)^j \frac{\left(\lambda t\right)^j}{j!} \\ &= e^{-\lambda t + M_Y(s)\lambda t}. \end{aligned} \tag{5}$$

So the cumulant of L is

$$\ln M_L\left(s\right) = \lambda\left(M_Y\left(s\right) - 1\right) = \lambda\left[\left(1 - \alpha x\right)^{-P} - 1\right]. \tag{6}$$

□

If we observe the occurrence of rainfall for n periods, then we have the sequence $\{L_i\}_{i=1}^n$ which is independent and identically distributed.

If on a particular day there is no rainfall that occurred, then

$$P\left(L = 0\right) = \exp\left(-\lambda\right)\frac{\left(\lambda\right)^0}{0!} = \exp\left(-\lambda\right) = p_0. \tag{7}$$

Therefore the process has a point mass at 0 which implies that it is not entirely continuous random variable.

Lemma 2. *The probability density function of L in (2) is*

$$f_\theta\left(L\right) = \exp\left(-\lambda\right)\delta\left(L\right) + \exp\left(-\lambda - \alpha L\right)L^{-1} r_P\left(\nu L^P\right), \tag{8}$$

where $\delta_0(L)$ is a dirac function at zero.

Proof. Let $q_0 = 1 - p_0$ be the probability that it rained. Hence for $L_i > 0$ we have

$$\begin{aligned} f_\theta^+\left(L\right) &= \sum_{i=1}^{\infty}\frac{p_i}{q_0}\left(\frac{\alpha^{iP}L^{iP-1}\exp\left(-\alpha L\right)}{\Gamma\left(ip\right)}\right) \\ &\qquad \text{where } p_i = \exp\left(-\lambda\right)\frac{\left(\lambda\right)^i}{i!} \\ &= \frac{1}{q_0}\left[\sum_{i=1}^{\infty} p_i \exp\left(-\alpha L\right)\frac{\alpha^{iP}L^{iP-1}}{\Gamma\left(iP\right)}\right] \\ &= \frac{1}{q_0}\left[\exp\left(-\alpha L\right)\sum_{i=1}^{\infty} p_i \frac{\alpha^{iP}L^{iP-1}}{\Gamma\left(iP\right)}\right] \\ &= \frac{1}{q_0}\left[\exp\left(-\alpha L\right)\sum_{i=1}^{\infty}\left(\exp\left(-\lambda\right)\frac{\left(\lambda\right)^i}{i!}\right)\frac{\alpha^{iP}L^{iP-1}}{\Gamma\left(iP\right)}\right] \\ &= \frac{\exp\left(-\lambda\right)}{q_0}\left[\exp\left(-\alpha L\right)\sum_{i=1}^{\infty}\left(\frac{\left(\lambda\right)^i}{i!}\right)\frac{\alpha^{ip}L^{iP-1}}{\Gamma\left(ip\right)}\right] \\ &= \frac{\exp\left(-\lambda\right)}{q_0}\exp\left(-\alpha L\right)\left[\sum_{i=1}^{\infty}\frac{\left(\lambda\right)^i\left(\alpha L\right)^{ip}}{Li!\Gamma\left(iP\right)}\right] \\ &= \frac{L^{-1}\exp\left(-\alpha L\right)}{\left(\exp\left(\lambda\right) - 1\right)}\sum_{i=1}^{\infty}\frac{\lambda\alpha^P L^P}{i!\Gamma\left(iP\right)}. \end{aligned} \tag{9}$$

If we let $\nu = \lambda\alpha^P$ and $r_p(\nu L^P) = \sum_{i=1}^{\infty}(\nu L^P/i!\Gamma(iP))$, then we have

$$f_\theta^+\left(L\right) = \frac{L^{-1}\exp\left(-\alpha L\right)}{\left(\exp\left(\lambda\right) - 1\right)} r_P\left(\nu L^P\right). \tag{10}$$

We can express the probability density function $f_\theta(L)$ in terms of a Dirac function as

$$\begin{aligned} f_\theta(L) &= p_0\delta_0(L) + q_0 f_\theta^+(L) \\ &= \exp(-\lambda)\,\delta_0(L) \\ &\quad + \left[\frac{q_0}{(\exp(\lambda)-1)}\right] L^{-1}\exp(-\alpha L)\, r_P\left(vL^P\right) \\ &= \exp(-\lambda)\,\delta_0(L) \\ &\quad + \exp(-\lambda-\alpha L)\, L^{-1} r_P\left(vL^P\right). \end{aligned} \tag{11}$$

□

Consider a random sample of size n of L_i with the probability density function

$$f_\theta(L) = \exp(-\lambda)\,\delta(L) + \exp(-\lambda-\alpha L)\, L^{-1} r_p\left(vL^p\right). \tag{12}$$

If we assume that there are m positive values $L_1, L_2, \ldots, L_m$, then there are $M = n - m$ zeros where $m > 0$.

We observe that $m \sim Bi(n, 1-\exp(-\lambda))$ and $p(m=0) = \exp(-n\lambda)$; hence the likelihood function is

$$L = \binom{n}{m} p_0^{n-m} q_0^m \prod_{i=1}^{m} f_\theta^+(L_i) \tag{13}$$

and the log-likelihood for $\theta = (\lambda, \alpha, p)$ is

$$\begin{aligned} &\log L(\theta; L_1, L_2, \ldots, L_n) \\ &= \log\left(\binom{n}{m} p_0^{n-m} q_0^m \prod_{i=1}^{m} f_\theta^+(L_i)\right) \\ &= \log\left(\binom{n}{m} e^{-\lambda n+\lambda m}\left(1-e^{-\lambda}\right)^m \prod_{i=1}^{m} e^{-\lambda-\alpha L_i}\frac{1}{L_i}\right. \\ &\left.\cdot \sum_{j=1}^{\infty}\frac{\left(\lambda\alpha^p L_{ij}^p\right)^j}{j!\Gamma(jp)}\right) = \log\binom{n}{m} + \lambda(m-n) + m \\ &\cdot \log\left(1-e^{-\lambda}\right) + \sum_{i=1}^{m} -\lambda-\alpha L_i - \log L_i \\ &+ \log\sum_{i=1}^{m}\sum_{j=1}^{\infty}\frac{\left(\lambda\alpha^p L_{ij}^p\right)^j}{j!\Gamma(jp)}. \end{aligned} \tag{14}$$

Now for $\widehat{\lambda}$ we have

$$\begin{aligned} &\frac{\partial \log L(\theta; L_1, L_2, \ldots, L_n)}{\partial\lambda} \\ &= m - n + \frac{m}{1-e^{-\lambda}} + (-1)^m \\ &\quad + \frac{1}{\lambda}\sum_{i=1}^{m}\sum_{j=1}^{\infty} i\frac{\partial \log L(\theta; L_1, L_2, \ldots, L_n)}{\partial\lambda} = 0 \Longrightarrow \\ &m - n + \frac{m}{1-e^{-\lambda}} + (-1)^m + \frac{1}{\lambda}\sum_{i=1}^{m}\sum_{j=1}^{\infty} i = 0. \end{aligned} \tag{15}$$

We can observe from the above evaluation that λ can not be expressed in closed form; similar derivation also shows that α as well can not be expressed in closed form. Therefore we can only estimate λ and α using numerical methods. Withers and Nadarajah [13] also observed that the probability density function can not be expressed in closed form and therefore it is difficult to find the analytic form of the estimators. So we will express the probability density function in terms of exponential dispersion models as described below.

Definition 3 (see [14]). A probability density function of the form

$$f(y;\theta,\Theta) = a(y,\Theta)\exp\left\{\frac{1}{\Theta}\left[y\theta - k(\theta)\right]\right\} \tag{16}$$

for suitable functions $k()$ and $a()$ is called an exponential dispersion model.

$\Theta > 0$ is the dispersion parameter. The function $k(\theta)$ is the cumulant of the exponential dispersion model; since $\Theta = 1$, then $k'()$ are the successive cumulants of the distribution [15]. The exponential dispersion models were first introduced by Fisher in 1922.

If we let $L_i = \log f(y_i;\theta_i,\Theta)$ as a contribution of y_i to the likelihood function $L = \sum_i L_i$, then

$$\begin{aligned} L_i &= \frac{1}{\Theta}\left[y_i\theta - k(\theta_i)\right] + \log a(y,\Theta), \\ \frac{\partial L_i}{\partial\theta_i} &= \frac{1}{\Theta}\left(y_i - k'(\theta_i)\right), \\ \frac{\partial^2 L_i}{\partial\theta_i^2} &= -\frac{1}{\Theta}k''(\theta_i). \end{aligned} \tag{17}$$

However we expect that $\mathbf{E}(\partial L_i/\partial\theta_i) = 0$ and $-\mathbf{E}(\partial^2 L_i/\partial\theta_i^2) = \mathbf{E}(\partial L_i/\partial\theta_i)^2$ so that

$$\begin{aligned} \mathbf{E}\left(\frac{1}{\Theta}\left(y_i - k'(\theta_i)\right)\right) &= 0, \\ \frac{1}{\Theta}\left(\mathbf{E}(y_i) - k'(\theta_i)\right) &= 0, \\ \mathbf{E}(y_i) &= k'(\theta_i). \end{aligned} \tag{18}$$

Furthermore

$$\begin{aligned} -\mathbf{E}\left(\frac{\partial^2 L_i}{\partial\theta_i^2}\right) &= \mathbf{E}\left(\frac{\partial L_i}{\partial\theta_i}\right)^2, \\ -\mathbf{E}\left(-\frac{1}{\Theta}k''(\theta_i)\right) &= \mathbf{E}\left(\frac{1}{\Theta}\left(y_i - k'(\theta_i)\right)\right)^2, \\ \frac{k''(\theta_i)}{\Theta} &= \frac{\mathrm{Var}(y_i)}{\Theta^2}, \\ \mathrm{Var}(y_i) &= \Theta k''(\theta_i). \end{aligned} \tag{19}$$

Therefore the mean of the distribution is $\mathbf{E}[Y] = \mu = dk(\theta)/d\theta$ and the variance is $\mathrm{Var}(Y) = \Theta(d^2k(\theta)/d\theta^2)$.

The relationship $\mu = dk(\theta)/d\theta$ is invertible so that θ can be expressed as a function of μ; as such we have $\text{Var}(Y) = \Theta V(\mu)$, where $V(\mu)$ is called a variance function.

Definition 4. The family of exponential dispersion models, whose variance functions are of the form $V(\mu) = \mu^p$ for $p \in (-\infty, 0] \cup [1, \infty)$, are called Tweedie family distributions.

Examples are as follows: for $p = 0$ then we have a normal distribution, $p = 1$, and $\Theta = 1$; it is a Poisson distribution, and Gamma distribution for $p = 2$, while when $p = 3$ it is Gaussian inverse distribution. Tweedie densities can not be expressed in closed form (apart from the examples above) but can instead be identified by their cumulants generating functions.

From $\text{Var}(Y) = \Theta(d^2k(\theta)/d\theta^2)$, then for Tweedie family distribution we have

$$\text{Var}(Y) = \Theta \frac{d^2k(\theta)}{d\theta^2} = \Theta V(\mu) = \Theta\mu^p. \tag{20}$$

Hence we can solve for μ and $k(\theta)$ as follows:

$$\begin{aligned} \mu &= \frac{dk(\theta)}{d\theta}, \\ \frac{d\mu}{d\theta} &= \mu^p \Longrightarrow \\ \int \frac{d\mu}{\mu^p} &= \int d\theta, \\ \theta &= \begin{cases} \dfrac{\mu^{1-p}}{1-p} & p \neq 1, \\ \log\mu & p = 1 \end{cases} \end{aligned} \tag{21}$$

by equating the constants of integration above to zero.

For $p \neq 1$ we have $\mu = [(1-p)\theta]^{1/(1-p)}$ so that

$$\begin{aligned} \int dk(\theta) &= \int [(1-p)\theta]^{1/(1-p)} d\theta, \\ k(\theta) &= \frac{[(1-p)\theta]^{(2-p)/(1-p)}}{2-p} = \frac{\mu^{(2-p)/(1-p)}}{2-p}, \\ & \qquad p \neq 2. \end{aligned} \tag{22}$$

Proposition 5. *The cumulant generating function of a Tweedie distribution for* $1 < p < 2$ *is*

$$\log M_Y(t) = \frac{1}{\Theta}\frac{\mu^{2-p}}{p-1}\left[\left(1 + t\Theta(1-p)\mu^{p-1}\right)^{(2-p)/(1-p)} - 1\right]. \tag{23}$$

Proof. From (16) the moment generating function is given by

$$\begin{aligned} M_Y(t) &= \int \exp(ty)\, a(y,\Theta) \exp\left\{\frac{1}{\Theta}[y\theta - k(\theta)]\right\} dy \\ &= \int a(y,\Theta) \exp\left\{\frac{1}{\Theta}[y(\theta + t\Theta) - k(\theta)]\right\} dy \\ &= \int a(y,\Theta) \exp\left(\frac{y(\theta + t\Theta) - k(\theta)}{\Theta}\right. \\ &\quad \left. + \frac{k(\theta + t\Theta) - k(\theta)}{\Theta}\right) dy = \int a(y,\Theta) \\ &\quad \cdot \exp\left(\frac{y(\theta + t\Theta) - k(\theta + t\Theta)}{\Theta}\right. \\ &\quad \left. + \frac{k(\theta + t\Theta) - k(\theta + t\Theta)}{\Theta}\right) dy = \int a(y,\Theta) \\ &\quad \cdot \exp\left(\frac{y(\theta + t\Theta) - k(\theta + t\Theta)}{\Theta}\right) \\ &\quad \cdot \exp\left(\frac{k(\theta + t\Theta) - k(\theta + t\Theta)}{\Theta}\right) dy \\ &= \exp\left(\frac{k(\theta + t\Theta) - k(\theta + t\Theta)}{\Theta}\right) \int a(y,\Theta) \\ &\quad \cdot \exp\left(\frac{y(\theta + t\Theta) - k(\theta + t\Theta)}{\Theta}\right) dy \\ &= \exp\left\{\frac{1}{\Theta}[k(\theta + t\Theta) - k(\theta)]\right\}. \end{aligned} \tag{24}$$

Hence cumulant generating function is

$$\log M_Y(t) = \frac{1}{\Theta}[k(\theta + t\Theta) - k(\theta)]. \tag{25}$$

For $1 < p < 2$ we substitute θ and $k(\theta)$ to have

$$\log M_Y(t) = \frac{1}{\Theta}\frac{\mu^{2-p}}{p-1}\left[\left(1 + t\Theta(1-p)\mu^{p-1}\right)^{(2-p)/(1-p)} - 1\right]. \tag{26}$$

□

By comparing the cumulant generating functions in Lemma 1 and Proposition 5 the compound Poisson process can be thought of as Tweedie distribution with parameters (λ, α, P) expressed as follows:

$$\begin{aligned} \lambda &= \frac{\mu^{2-p}}{\Theta(2-p)}, \\ \alpha &= \Theta(p-1)\mu^{p-1}, \\ P &= \frac{2-p}{p-1}. \end{aligned} \tag{27}$$

The requirement that the Gamma shape parameter P be positive implies that only Tweedie distributions between $1 < p < 2$ can represent the Poisson-Gamma compound process. In addition, for $\lambda > 0$, $\alpha > 0$ implies $\mu > 0$ and $\Theta > 0$.

Proposition 6. *Based on Tweedie distribution, the probability of receiving no rainfall at all is*

$$P(L=0)=\exp\left[-\frac{\mu^{2-p}}{\Theta(2-p)}\right] \tag{28}$$

and the probability of having a rainfall event is

$$\begin{aligned}&P(L>0)\\&\quad=W(\lambda,\alpha,L,P)\exp\left[\frac{L}{(1-p)\mu^{p-1}}-\frac{\mu^{2-p}}{2-p}\right],\end{aligned} \tag{29}$$

where

$$W(\lambda,\alpha,L,P)=\sum_{j=1}^{\infty}\frac{\lambda^j(\alpha L)^{jP}e^{-\lambda}}{j!\Gamma(jP)}. \tag{30}$$

Proof. This follows by directly substituting the values of λ and $\theta, k(\theta)$ into (16). □

The function $W(\lambda,\alpha,L,P)$ is an example of Wright's generalized Bessel function; however it can not be expressed in terms of the more common Bessel function. To evaluate it the value of j is determined for which the function W_j reaches the maximum [15].

3. Parameter Estimation

We approximate the function $W(\lambda,\alpha,L,P) = \sum_{j=1}^{\infty}(\lambda^j(\alpha L)^{jP}e^{-\lambda}/j!\Gamma(jP)) = \sum_{j=1}^{\infty}W_j$ following the procedure by [15] where the value of j is determined for which W_j reaches maximum. We treat j as continuous so that W_j is differentiated with respect to j and set the derivative to zero. So for $L>0$ we have the following.

Lemma 7 (see [15]). *The log maximum approximation of W_j is given by*

$$\begin{aligned}\log W_{\max}&=\frac{L^{2-p}}{(2-p)\Theta}\left[\log\frac{L^P(p-1)^P}{\Theta^{(1-P)}(2-p)}+(1+P)\right.\\&\quad\left.-P\log P-(1-P)\log\frac{L^{2-p}}{(2-p)\Theta}\right]-\log(2\pi)-\frac{1}{2}\\&\quad\cdot\log P-\log\frac{L^{2-p}}{(2-p)\Theta},\end{aligned} \tag{31}$$

where $j_{\max}=L^{2-p}/(2-p)\Theta$.

Proof.

$$\begin{aligned}W(\lambda,\alpha,L,P)&=\sum_{j=1}^{\infty}\frac{\lambda^j(\alpha L)^{jP-1}e^{-\lambda}}{j!\Gamma(jP)}\\&=\sum_{j=1}^{\infty}\frac{\lambda^jL^{jP-1}e^{-L/\tau}e^{-\lambda}}{j!\tau^{Pj}\Gamma(Pj)}\quad\text{where }\tau=\frac{1}{\alpha}.\end{aligned} \tag{32}$$

Substituting the values of λ,α in the above equation we have

$$\begin{aligned}&W(\lambda,\alpha,L,P)\\&=\sum_{j=1}^{\infty}\frac{\left(\mu^{2-p}/\Theta(2-p)\right)^jL^{jP-1}\left[\Theta(1-p)\mu^{p-1}\right]^{jP}e^{-L/\tau}e^{-\lambda}}{j!\Gamma(Pj)}\\&=e^{-L/\tau-\lambda}L^{-1}\sum_{j=1}^{\infty}\frac{\mu^{(2-p)j}\left(\Theta(p-1)\mu^{p-1}\right)^{jP}L^{jP}}{\Theta^j(2-p)^j\,j!\Gamma(jP)}\\&=e^{-L/\tau-\lambda}L^{-1}\sum_{j=1}^{\infty}\frac{L^{jP}(p-1)^{jP}\mu^{(2-p)j+(p-1)jP}}{\Theta^{j(1-P)}(2-p)^j\,j!\Gamma(jP)}.\end{aligned} \tag{33}$$

The term $\mu^{(2-p)j+(p-1)jP}$ depends on the L,p,P,Θ values so we maximize the summation

$$\begin{aligned}W(L,\Theta,P)&=\sum_{j=1}^{\infty}\frac{L^{jP}(p-1)^{jP}}{\Theta^{j(1-P)}(2-p)^j\,j!\Gamma(jP)}\\&=\sum_{j=1}^{\infty}\frac{z^j}{j!\Gamma(jP)}\qquad\text{where }z=\frac{L^P(p-1)^P}{\Theta^{(1-P)}(2-p)}\\&=W_j.\end{aligned} \tag{34}$$

Considering W_j we have

$$\begin{aligned}\log W_j&=j\log z-\log j!-\log(Pj)\\&=j\log z-\log\Gamma(j+1)-\log(Pj).\end{aligned} \tag{35}$$

Using Stirling's approximation of Gamma functions we have

$$\begin{aligned}\log\Gamma(1+j)&\approx(1+j)\log(1+j)-(1+j)\\&\quad+\frac{1}{2}\log\left(\frac{2\pi}{1+j}\right),\\\log\Gamma(Pj)&\approx Pj\log(Pj)-Pj+\frac{1}{2}\log\left(\frac{2\pi}{Pj}\right).\end{aligned} \tag{36}$$

And hence we have

$$\begin{aligned}W_j&\approx j\left[\log z+(1+P)-P\log P-(1-P)\log j\right]\\&\quad-\log(2\pi)-\frac{1}{2}\log P-\log j.\end{aligned} \tag{37}$$

For $1<p<2$ we have $P=(2-p)/(p-1)>0$; hence the logarithms have positive arguments. Differentiating with respect to j we have

$$\begin{aligned}\frac{\partial\log W_j}{\partial j}&\approx\log z-\frac{1}{j}-\log j-P\log(Pj)\\&\approx\log z-\log j-P\log(Pj),\end{aligned} \tag{38}$$

where $1/j$ is ignored for large j. Solving for $(\partial \log Wj)/\partial j = 0$ we have

$$j_{\max} = \frac{L^{2-p}}{(2-p)\,\Theta}. \tag{39}$$

Substituting $j_{\max}$ in $\log W_j$ to find the maximum approximation of W_j we have

$$\log W_{\max} = \frac{L^{2-p}}{(2-p)\,\Theta}\left[\log \frac{L^{P}\,(p-1)^{P}}{\Theta^{(1-P)}\,(2-p)} + (1+P)\right.$$
$$\left. - P\log P - (1-P)\log \frac{L^{2-p}}{(2-p)\,\Theta}\right] - \log(2\pi) - \frac{1}{2} \tag{40}$$
$$\cdot \log P - \log \frac{L^{2-p}}{(2-p)\,\Theta}.$$

Hence the result follows. □

It can be observed that $\partial W_j/\partial j$ is monotonically decreasing; hence $\log W_j$ is strictly convex as a function of j. Therefore W_j decays faster than geometrically on either side of $j_{\max}$ [15]. Therefore if we are to estimate $W(L,\Theta,P)$ by $\widehat{W}(L,\Theta,P) = \sum_{j=j_d}^{j_u} W_j$ the approximation error is bounded by geometric sum

$$\begin{aligned} &W(L,\Theta,P) - \widehat{W}(L,\Theta,P) \\ &\quad < W_{j_d-1}\frac{1-r_l^{j_d-1}}{1-r_l} + W_{j_u+1}\frac{1}{1-r_u}, \\ &r_l = \exp\left(\frac{\partial W_j}{\partial j}\right)\Big|_j = j_d - 1, \\ &r_u = \exp\left(\frac{\partial W_j}{\partial j}\right)\Big|_j = j_u + 1. \end{aligned} \tag{41}$$

For quick and accurate evaluation of $W(\lambda,\alpha,L,P)$, the series is summed for only those terms in the series which contribute significantly to the sum.

Generalized linear models extend the standard linear regression models to incorporate nonnormal response distributions and possibly nonlinear functions of the mean. The advantage of GLMs is that the fitting process maximizes the likelihood for the choice of the distribution for a random variable y and the choice is not restricted to normality unlike linear regression [16].

The exponential dispersion models are the response distributions for the generalized linear models. Tweedie distributions are members of the exponential dispersion models upon which the generalized linear models are based. Consequently fitting a Tweedie distribution follows the framework of fitting a generalized linear model.

Lemma 8. *In case of a canonical link function, the sufficient statistics for $\{\beta_j\}$ are $\{\sum_{i=1}^{n} y_i x_{ij}\}$.*

Proof. For n independent observations y_i of the exponential dispersion model (16) the log-likelihood function is

$$\begin{aligned} L(\beta) &= \sum_{i=1}^{n} L_i = \sum_{i}^{n} \log f(y_i,\theta_i,\Theta) \\ &= \sum_{i=1}^{n} \frac{y_i\theta_i - k(\theta_i)}{\Theta} + \sum_{i}^{n} \log a(y_i,\Theta). \end{aligned} \tag{42}$$

But $\theta_i = \sum_j^p \beta_j x_{ij}$; hence

$$\sum_{i}^{n} y_i\theta_i = \sum_{i=1}^{n} y_i \sum_{j}^{p} \beta_j x_{ij} = \sum_{j}^{p} \beta_j \sum_{i=1}^{n} y_i x_{ij}. \tag{43}$$

□

Proposition 9. *Given that y_i is distributed as (16) then its distribution depends only on its first two moments, namely, μ_i and $\mathrm{Var}(y_i)$.*

Proof. Let $g(\mu_i)$ be the link function of the GLM such that $\eta_i = \sum_{j=1}^{p} \beta_j x_{ij} = g(\mu_i)$. The likelihood equations are

$$\frac{\partial L(\beta)}{\partial \beta} = \sum_{i=1}^{n} \frac{\partial L_i}{\partial \beta_j} \quad \forall j. \tag{44}$$

Using chain rule we have

$$\frac{\partial L_i}{\partial \beta_j} = \frac{\partial L_i}{\partial \theta_i}\frac{\partial \theta_i}{\partial \mu_i}\frac{\partial \mu_i}{\partial \eta_i}\frac{\partial \eta_i}{\partial \beta_j} = \frac{y_i - \mu_i}{\mathrm{Var}(y_i)} x_{ij}\frac{\partial \mu_i}{\partial \eta_i}. \tag{45}$$

Hence

$$\frac{\partial L(\beta)}{\partial \beta} = \frac{y_i - \mu_i}{\mathrm{Var}(y_i)} x_{ij}\frac{\partial \mu_i}{\partial \eta_i} = \frac{y_i - \mu_i}{\Theta \mu_i^{p}} x_{ij}\frac{\partial \mu_i}{\partial \eta_i}. \tag{46}$$

Since $\mathrm{Var}(y_i) = V(\mu_i)$, the relationship between the mean and variance characterizes the distribution. □

Clearly a GLM only requires the first two moments of the response y_i; hence despite the difficulty of full likelihood analysis of Tweedie distribution as it can not be expressed in closed form for $1 < p < 2$ we can still fit a Tweedie distribution family. The likelihood is only required to estimate p and Θ as well as diagnostic check of the model.

Proposition 10. *Under the standard regularity conditions, for large n, the maximum likelihood estimator $\widehat{\beta}$ of β for generalized linear model is efficient and has an approximate normal distribution.*

Proof. From the log-likelihood, the covariance matrix of the distribution is the inverse of the information matrix $\mathbf{J} = \mathbf{E}(-\partial^2 L(\beta)/\partial\beta_h\partial\beta_j)$.

So

$$\begin{aligned}\mathbf{J} &= \mathbf{E}\left(-\frac{\partial^2 L(\beta)}{\partial\beta_h\partial\beta_j}\right) = \mathbf{E}\left[\left(\frac{\partial^2 L_i}{\partial\beta_h}\right)\left(\frac{\partial^2 L_i}{\partial\beta_j}\right)\right] \\ &= \left[\left(\frac{y_i-\mu_i}{\mathrm{Var}(y_i)}x_{ih}\frac{\partial\mu_i}{\partial\eta_i}\right)\left(\frac{y_i-\mu_i}{\mathrm{Var}(y_i)}x_{ij}\frac{\partial\mu_i}{\partial\eta_i}\right)\right] \\ &= \frac{x_{ih}x_{ij}}{\mathrm{Var}(y_i)}\left(\frac{\partial\mu_i}{\partial\eta_i}\right)^2.\end{aligned} \tag{47}$$

Hence

$$\mathbf{E}\left(-\frac{\partial^2 L(\beta)}{\partial\beta_h\partial\beta_j}\right) = \sum_i^n \frac{x_{ih}x_{ij}}{\mathrm{Var}(y_i)}\left(\frac{\partial\mu_i}{\partial\eta_i}\right)^2 = \left(X^T W X\right), \tag{48}$$

where $W = \mathrm{diag}[(1/\mathrm{Var}(y_i))(\partial\mu_i/\partial\eta_i)^2]$.

Therefore $\widehat{\beta}$ has an approximate $N[\beta, (X^T W X)^{-1}]$ with $\mathrm{Var}(\widehat{\beta}) = (X^T\widehat{W}X)^{-1}$, where $\widehat{W}$ is evaluated at $\widehat{\beta}$. □

To compute $\widehat{\beta}$ we use the iteratively reweighted least square algorithm proposed by Dobson and Barnett [17] where the iterations use the working weights w_i:

$$\frac{w_i}{V(\mu_i)\,\dot{g}(\mu_i)^2}, \tag{49}$$

where $V(\mu_i) = \mu_i^p$.

However estimating p is more difficult than estimating β and Θ such that most researchers working with Tweedie densities have p a priori. In this study we use the procedure in [15] where the maximum likelihood estimator of p is obtained by directly maximizing the profile likelihood function. For any given value of p we find the maximum likelihood estimate of β, Θ and compute the log-likelihood function. This is repeated several times until we have a value of p which maximizes the log-likelihood function.

Given the estimated values of p and β, then the unbiased estimator of Θ is given by

$$\widehat{\Theta} = \sum_{i=1}^n \frac{\left[L_i - \mu_i\left(\widehat{\beta}\right)\right]^2}{\mu_i\left(\widehat{\beta}\right)^{\widehat{p}}}. \tag{50}$$

Since for $1 < p < 2$ the Tweedie density can not be expressed in closed form, it is recommended that the maximum likelihood estimate of Θ must be computed iteratively from full data [15].

4. Data and Model Fitting

4.1. Data Analysis. Daily rainfall data of Balaka district in Malawi covering the period 1995–2015 is used. The data was obtained from Meteorological Surveys of Malawi. Figure 1 shows a plot of the data.

In summary the minimum value is 0 mm which indicates that there were no rainfall on particular days, whereas the maximum amount is 123.7 mm. The mean rainfall for the whole period is 3.167 mm.

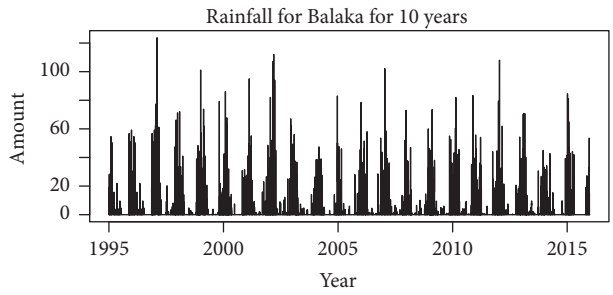

Figure 1: Daily rainfall amount for Balaka district.

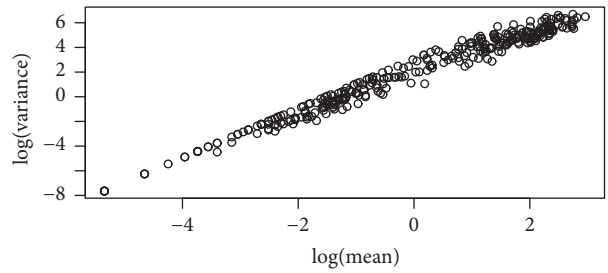

Figure 2: Variance mean relationship.

We investigated the relationship between the variance and the mean of the data by plotting the log(variance) against log(mean) as shown in Figure 2. From the figure we can observe a linear relationship between the variance and the mean which can be expressed as

$$\log(\text{Variance}) = \alpha + \beta\log(\text{mean}) \tag{51}$$

$$\text{Variace} = A * \text{mean}^{\beta}, \quad A \in \mathbb{R}. \tag{52}$$

Hence the variance can be expressed as some power $\beta \in \mathbb{R}$ of the mean agreeing with the Tweedie variance function requirement.

4.2. Fitted Model. To model the daily rainfall data we use sin and cos as predictors due to the cyclic nature and seasonality of rainfall. We have assumed that February ends on 28th for all the years to be uniform in our modeling.

The canonical link function is given by

$$\log\mu_i = a_0 + a_1\sin\left(\frac{2\pi i}{365}\right) + a_2\cos\left(\frac{2\pi i}{365}\right), \tag{53}$$

where $i = 1, 2, \ldots, 365$ corresponds to days of the year and a_0, a_1, a_2 are the coefficients of regression.

In the first place we estimate $\widehat{p}$ by maximizing the profile log-likelihood function. Figure 3 shows the graph of the profile log-likelihood function. As can be observed the value of p that maximizes the function is 1.5306.

From the results obtained after fitting the model, both the cyclic cosine and sine terms are important characteristics for daily rainfall Table 1. The covariates were determined to take into account the seasonal variations in the stochastic model.

Table 1: Estimated parameter values.

Parameter	Estimate	Std. error	t value	Pr(> $\lvert t \rvert$)
$\widehat{a}_0$	0.1653	0.0473	3.4930	0.0005^{***}
$\widehat{a}_1$	0.9049	0.0572	15.81100	$<2e-16^{***}$
$\widehat{a}_2$	2.0326	0.0622	32.6720	$<2e-16^{***}$
$\widehat{\Theta}$	14.8057	-	-	-

With *signif* code: 0 ∗ ∗ ∗.

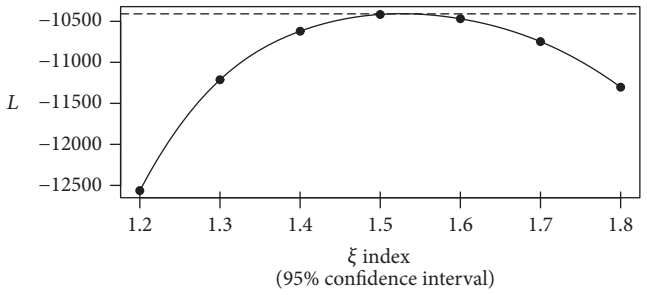

Figure 3: Profile likelihood.

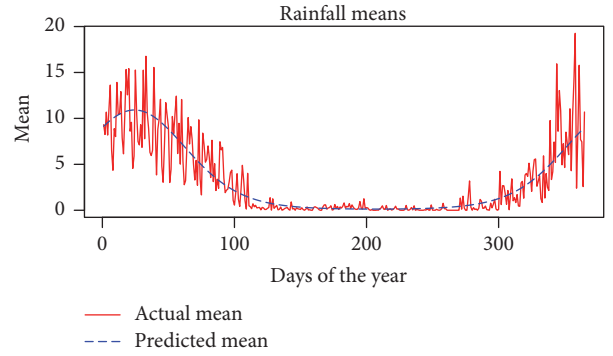

Figure 4: Actual versus predicted mean.

The predicted $\widehat{\mu}_i, \widehat{p}, \widehat{\Theta}$ for each day only depends on the day's conditions so that for each day i we have

$$\begin{aligned}\widehat{\mu}_i &= \exp\left[0.1653 + 0.9049\sin\left(\frac{2\pi i}{365}\right)\right.\\ &\quad \left. + 2.0326\cos\left(\frac{2\pi i}{365}\right)\right],\\ \widehat{p} &= 1.5306,\\ \widehat{\Theta} &= 14.8057.\end{aligned} \tag{54}$$

From these estimated values we can calculate the parameter $(\widehat{\lambda}_i, \widehat{\alpha}_i, \widehat{P})$ from the corresponding formulas above as

$$\begin{aligned}\widehat{\lambda}_i &= \frac{1}{6.5716}\left(\exp\left[0.1653 + 0.9049\sin\left(\frac{2\pi i}{365}\right)\right.\right.\\ &\quad \left.\left. + 2.03263\cos\left(\frac{2\pi i}{365}\right)\right]\right)^{0.4694},\\ \widehat{\alpha} &= 7.4284\left(\exp\left[0.1653 + 0.9049\sin\left(\frac{2\pi i}{365}\right)\right.\right.\\ &\quad \left.\left. + 2.0326\cos\left(\frac{2\pi i}{365}\right)\right]\right)^{0.5306},\\ \widehat{P} &= 0.8847.\end{aligned} \tag{55}$$

Comparing the actual means and the predicted means for 2 July we have $\widehat{\mu} = 0.3820$, whereas $\mu = 0.4333$; similarly for 31 December we have $\widehat{\mu} = 9.0065$ and $\mu = 10.6952$, respectively. Figure 4 shows the estimated mean and actual mean where the model behaves well generally.

4.3. Goodness of Fit of the Model. Let the maximum likelihood estimate of θ_i be $\widehat{\theta}_i$ for all i and $\widehat{\mu}$ as the model's mean estimate. Let $\widetilde{\theta}_i$ denote the estimate of θ_i for the saturated model with corresponding $\widetilde{\mu} = y_i$.

The goodness of fit is determined by deviance which is defined as

$$\begin{aligned}&-2\left[\frac{\text{maximum likelihood of the fitted model}}{\text{Maximum likelihood of the saturated model}}\right]\\ &= -2\left[L\left(\widehat{\mu}; y\right) - L\left(y, y\right)\right]\\ &= 2\sum_{i=1}^{n}\frac{y_i\widetilde{\theta}_i - k\left(\widetilde{\theta}_i\right)}{\Theta} - 2\sum_{i=1}^{n}\frac{y_i\widehat{\theta}_i - k\left(\widehat{\theta}_i\right)}{\Theta}\\ &= 2\sum_{i=1}^{n}\frac{y_i\left(\widetilde{\theta}_i - \widehat{\theta}_i\right) - k\left(\widetilde{\theta}_i\right) + k\left(\widehat{\theta}_i\right)}{\Theta} = \frac{\text{Dev}\left(y, \widehat{\mu}\right)}{\Theta}.\end{aligned} \tag{56}$$

$\text{Dev}(y, \widehat{\mu})$ is called the deviance of the model and the greater the deviance, the poorer the fitted model as maximizing the likelihood corresponds to minimizing the deviance.

In terms of Tweedie distributions with $1 < p < 2$, the deviance is

$$\begin{aligned}&\text{Dev}_p\\ &= 2\sum_{i=1}^{n}\left(\frac{y_i^{2-p} - (2-p)\,y_i\mu_i^{1-p} + (1-p)\,\mu_i^{2-p}}{(1-p)(2-p)}\right).\end{aligned} \tag{57}$$

Based on results from fitting the model, the residual deviance is 43144 less than the null deviance 62955 which implies that the fitted model explains the data better than a null model.

4.4. Diagnostic Check. The model diagnostic is considered as a way of residual analysis. The fitted model faces challenges

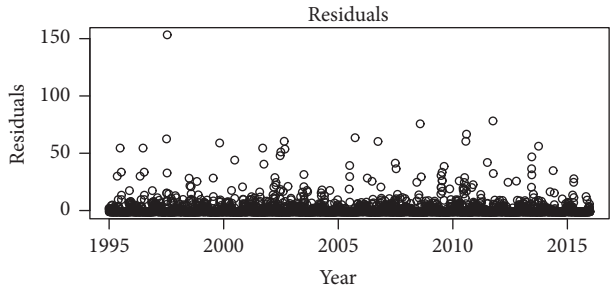

Figure 5: Residuals of the model.

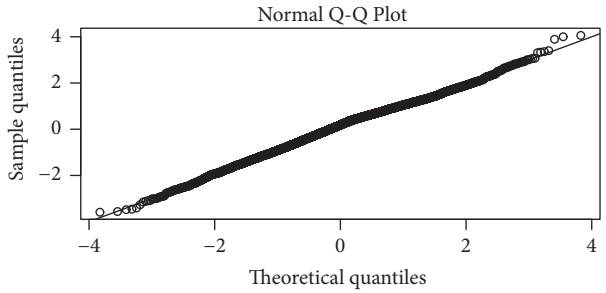

Figure 6: Q-Q plot of the quantile residuals.

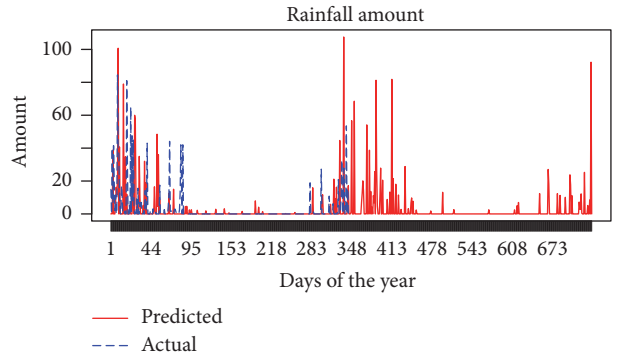

Figure 7: Simulated rainfall and observed rainfall.

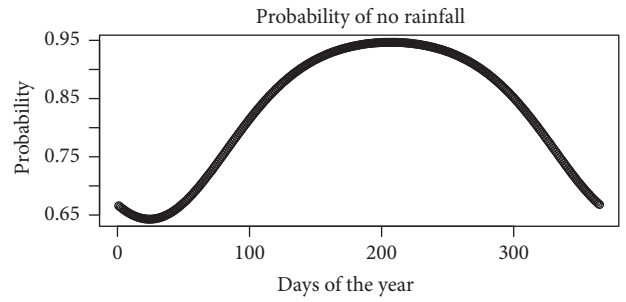

Figure 8: Probability of rainfall occurrence.

to be assessed especially for days with no rainfall at all as they produce spurious results and distracting patterns similarly as observed by [15]. Since this is a nonnormal regression, residuals are far from being normally distributed and having equal variances unlike in a normal linear regression. Here the residuals lie parallel to distinct values; hence it is difficult to make any meaningful decision about the fitted model (Figure 5).

So we assess the model based on quantile residuals which remove the pattern in discrete data by adding the smallest amount of randomization necessary on the cumulative probability scale.

The quantile residuals are obtained by inverting the distribution function for each response and finding the equivalent standard normal quantile.

Mathematically, let $a_i = \lim_{y \uparrow y_i} F(y; \widehat{\mu}_i, \widehat{\Theta})$ and $b_i = F(y_i; \widehat{\mu}_i, \widehat{\Theta})$, where F is the cumulative function of the probability density function $f(y; \mu, \Theta)$; then the randomized quantile residuals for y_i are

$$r_{q,i} = \Phi^{-1}(u_i) \tag{58}$$

with u_i being the uniform random variable on $(a_i, b_i]$. The randomized quantile residuals are distributed normally barring the variability in $\widehat{\mu}$ and $\widehat{\Theta}$.

Figure 6 shows the normalized Q-Q plot and as can be observed there are no large deviations from the straight line, only small deviations at the tail. The linearity observed indicates an acceptable fitted model.

5. Simulation

The model is simulated to test whether it produces data with similar characteristics to the actual observed rainfall. The simulation is done for a period of two years where one was the last year of the data (2015) and the other year (2016) was a future prediction. Then comparison was done with a graph for 2015 data as shown in Figure 7.

The different statistics of the simulated data and actual data are shown in Table 2 for comparisons.

The main objective of simulation is to demonstrate that the Poisson-Gamma can be used to predict and forecast rainfall occurrence and intensity simultaneously. Based on the results above (Figure 8), the model has shown that it works well in predicting the rainfall intensity and hence can be used in agriculture, actuarial science, hydrology, and so on.

However the model performed poorly in predicting probability of rainfall occurrence as it underestimated the probability of rainfall occurrence. It is suggested here that probably the use of truncated Fourier series can improve this estimation as compared to the sinusoidal.

But it performed better in predicting probability of no rainfall on days where there was little or no rainfall as indicated in Figure 8.

It can also be observed that the model produces synthetic precipitation that agrees with the four characteristics of a stochastic precipitation model as suggested by [4] as follows. The probability of rainfall occurrence obeys a seasonal pattern (Figure 8); in addition we can also tell that a probability of a rain in a day is higher if the previous day was wet which is the basis of precipitation models that involve the

Table 2: Data statistics.

	Min	1st Qu.	Median	Mean	3rd Qu.	Max
Predicted data	0.00	0.00	0.00	3.314	0.00	116.5
Actual data [10 yrs]	0.00	0.00	0.00	3.183	0.300	123.7
Actual data [2015]	0.00	0.00	0.00	3.328	0.00	84.5

Markov process. From Figure 7 we can also observe variation of rainfall intensity based on time of the season.

In addition the model allows modeling of exact zeros in the data and is able to predict a probability of no rainfall event simultaneously.

6. Conclusion

A daily stochastic rainfall model was developed based on a compound Poisson process where rainfall events follow a Poisson distribution and the intensity is independent of events following a Gamma distribution. Unlike several researches that have been carried out into precipitation modeling whereby two models are developed for occurrence and intensity, the model proposed here is able to model both processes simultaneously. The proposed model is also able to model the exact zeros, the event of no rainfall, which is not the case with the other models. This precipitation model is an important tool to study the impact of weather on a variety of systems including ecosystem, risk assessment, drought predictions, and weather derivatives as we can be able to simulate synthetic rainfall data. The model provides mechanisms for understanding the fine scale structure like number and mean of rainfall events, mean daily rainfall, and probability of rainfall occurrence. This is applicable in agriculture activities, disaster preparedness, and water cycle systems.

The model developed can easily be used for forecasting future events and, in terms of weather derivatives, the weather index can be derived from simulating a sample path by summing up daily precipitation in the relevant accumulation period. Rather than developing a weather index which is not flexible enough to forecast future events, we can use this model in pricing weather derivatives.

Rainfall data is generally zero inflated in that the amount of rainfall received on a day can be zero with a positive probability but continuously distributed otherwise. This makes it difficult to transform the data to normality by power transforms or to model it directly using continuous distribution. The Poisson-Gamma distribution has a complicated probability density function whose parameters are difficult to estimate. Hence expressing it in terms of a Tweedie distribution makes estimating the parameters easy. In addition, Tweedie distributions belong to the exponential family of distributions upon which generalized linear models are based; hence there is an already existing framework in place for fitting and diagnostic testing of the model.

The model developed allows the information in both zero and positive observations to contribute to the estimation of all parts of the model unlike the other model [3, 4, 9] which conditions rainfall intensity based on probability of occurrence. In addition the introduction of the dispersion parameter in the model helps in reducing underestimation of overdispersion of the data which is also common in the aforementioned models.

Conflicts of Interest

The authors declare that there are no conflicts of interest regarding the publication of this paper.

Acknowledgments

The authors extend their gratitude to Pan African University Institute for Basic Sciences, Technology and Innovation for the financial support.

References

[1] A. Hussain, "Stochastic modeling of rainfall processes: a markov chain-mixed exponential model for rainfalls in different climatic conditions".

[2] M. Cao, A. Li, and J. Z. Wei, "Precipitation modeling and contract valuation: A frontier in weather derivatives," *The Journal of Alternative Investments*, vol. 7, no. 2, pp. 93–99, 2004.

[3] G. Leobacher and P. Ngare, "On modelling and pricing rainfall derivatives with seasonality," *Applied Mathematical Finance*, vol. 18, no. 1, pp. 71–91, 2011.

[4] M. Odening, O. Musshoff, and W. Xu, "Analysis of rainfall derivatives using daily precipitation models: Opportunities and pitfalls," *Agricultural Finance Review*, vol. 67, no. 1, pp. 135–156, 2007.

[5] D. S. Wilks, "Multisite generalization of a daily stochastic precipitation generation model," *Journal of Hydrology*, vol. 210, no. 1–4, pp. 178–191, 1998.

[6] C. Onof, R. E. Chandler, A. Kakou, P. Northrop, H. S. Wheater, and V. Isham, "Rainfall modelling using poisson-cluster processes: a review of developments," *Stochastic Environmental Research and Risk Assessment*, vol. 14, no. 6, pp. 384–411, 2000.

[7] R. Carmona and P. Diko, "Pricing precipitation based derivatives," *International Journal of Theoretical and Applied Finance*, vol. 8, no. 7, pp. 959–988, 2005.

[8] F. E. Benth and J. S. Benth, *Modeling and pricing in financial markets for weather derivatives*, vol. 17, World Scientific, 2012.

[9] B. López Cabrera, M. Odening, and M. Ritter, "Pricing rainfall futures at the CME," Technical report, Humboldt University, Collaborative Research Center.

[10] T. I. Harrold, A. Sharma, and S. J. Sheather, "A nonparametric model for stochastic generation of daily rainfall occurrence," *Water Resources Research*, vol. 39, no. 10, 2003.

[11] D. Lord, "Modeling motor vehicle crashes using Poisson-gamma models: examining the effects of low sample mean values and small sample size on the estimation of the fixed

dispersion parameter," *Accident Analysis & Prevention*, vol. 38, no. 4, pp. 751–766, 2006.

[12] J.-P. Wang, "Estimating species richness by a Poisson-compound gamma model," *Biometrika*, vol. 97, no. 3, pp. 727–740, 2010.

[13] C. S. Withers and S. Nadarajah, "On the compound Poisson-gamma distribution," *Kybernetika*, vol. 47, no. 1, pp. 15–37, 2011.

[14] E. W. Frees, R. Derrig, and G. Meyers, *Regression modeling with actuarial and financial applications*, vol. 1, Cambridge University Press, Cambridge, UK, 2014.

[15] P. K. Dunn and G. K. Smyth, "Evaluation of Tweedie exponential dispersion model densities by Fourier inversion," *Statistics and Computing*, vol. 18, no. 1, pp. 73–86, 2008.

[16] A. Agresti, *Foundations of linear and generalized linear models*, John Wiley & Sons, 2015.

[17] A. J. Dobson and A. G. Barnett, *An Introduction to Generalized Linear Models*, Texts in Statistical Science Series, CRC Press, Boca Raton, Fla, USA, 3rd edition, 2008.