Hindawi
Journal of Probability and Statistics
Volume 2023, Article ID 1780404, 18 pages
https://doi.org/10.1155/2023/1780404

Research Article

Flexible Lévy-Based Models for Time Series of Count Data with Zero-Inflation, Overdispersion, and Heavy Tails

Confort Kollie,[1] **Philip Ngare**,[2] **and Bonface Malenje**[3]

[1]*Pan African University Institute for Basic Sciences, Technology and Innovation (PAUSTI), Nairobi, Kenya*
[2]*School of Mathematics, University of Nairobi, Nairobi, Kenya*
[3]*Jomo Kenyatta University of Agriculture and Technology, Department of Statistics and Actuarial Sciences, Juja, Kenya*

Correspondence should be addressed to Confort Kollie; kollietinaconfort@gmail.com

Received 3 October 2023; Revised 6 November 2023; Accepted 9 November 2023; Published 30 November 2023

Academic Editor: Zacharias Psaradakis

The explosion of time series count data with diverse characteristics and features in recent years has led to a proliferation of new analysis models and methods. Significant efforts have been devoted to achieving flexibility capable of handling complex dependence structures, capturing multiple distributional characteristics simultaneously, and addressing nonstationary patterns such as trends, seasonality, or change points. However, it remains a challenge when considering them in the context of long-range dependence. The Lévy-based modeling framework offers a promising tool to meet the requirements of modern data analysis. It enables the modeling of both short-range and long-range serial correlation structures by selecting the kernel set accordingly and accommodates various marginal distributions within the class of infinitely divisible laws. We propose an extension of the basic stationary framework to capture additional marginal properties, such as heavy-tailedness, in both short-term and long-term dependencies, as well as overdispersion and zero inflation in simultaneous modeling. Statistical inference is based on composite pairwise likelihood. The model's flexibility is illustrated through applications to rainfall data in Guinea from 2008 to 2023, and the number of NSF funding awarded to academic institutions. The proposed model demonstrates remarkable flexibility and versatility, capable of simultaneously capturing overdispersion, zero inflation, and heavy-tailedness in count time series data.

1. Introduction

Time series of count data arises in different disciplines, where observed counts are recorded over time, such as economics, epidemiology, finance, and insurance. Several aspects of count time data have been the subject of extensive research as evidenced by the rich literature on this area. A major issue entails modeling dependence arising from the observations' discrete nature, which renders the autoregressive structure for continuous data incoherent. The efforts directed towards handling time series of count data are aimed at ensuring the validity of inference and consequently data-driven decisions. The challenge of handling serial correlation in count data continues to attract the attention of many researchers and scholars, who are inspired by the difficulties that arise when dealing with these data in various situations, including the trend of daily COVID-19 deaths in Ghana [1], stock market trends [2], road accident counts [3], and crime analysis [4]. Count data exhibit features such as nonnegativity, integer-valued, and frequently overdispersed which indicates that the variance is greater than the corresponding mean, zero-inflation which is a high occurrence of zero values in the dataset, and heavy-tailedness which refers to higher probability relative frequency of having extremes values or outliers in the dataset. The presence of zero-inflation, extreme values, or outliers in count data can have an effect on mean and variance estimations, as well as the validity of statistical inferences. Consequently, complexities arise in this setting due to the requirement to provide a modeling strategy capable of capturing the dependence patterns and simultaneous modeling as well as the marginal features of the observations. There are various modeling strategies that have been proposed to deal with the issues arising when handling time series for count data.There

are two paradigms predominant in the literature for handling serial dependence in count data, the first is the discrete autoregressive moving-average models introduced by Jacobs and Lewis [5], and a given ARMA (p, q) model is defined as follows:

$$X_t = c + \sum_{i=1}^{p} \phi_i X_{t-i} + \varepsilon_t + \sum_{i=1}^{q} \theta_i \varepsilon_{t-i}, \tag{1}$$

where, in the given time series model, X_t represents the value at time t, with c as the intercept and ϕ_i and θ_i as coefficients to estimate autoregressive and moving-average lags. The error term ε_t is a statistically independent random variable, uncorrelated both with itself over time and with other random variables in the model. The second is based on a thinning operator and was introduced by McKenzie [6] and Al-Osh and Alzaid [7]. The corresponding model used for modeling the dynamics of an integer-time sequence Y is defined as follows:

$$Y_t = \vartheta \ominus Y_{t-1} + \varepsilon_t, \quad t \in \mathbb{Z}, \tag{2}$$

where $0 \le \vartheta < 1$, Y_{t-1} represents the value of the sequence at the time step preceding t, $\ominus$ is a thinning operator, and $\{\varepsilon_t\}_{t=n} \in \mathbb{Z}$ is a sequence of random variables. The advantage of the first approach is that in such a stationary process, their marginal distribution can be of any kind shown by McKenzie [8]. However, count data's drawback includes long runs of constant values, making sample paths unrealistic in many applications. The second provides a diverse set of models. Sample pathways from thinning models frequently appear more realistic than those from discrete autoregressive moving-average processes. Thinned models, on the other hand, cannot generate an arbitrary marginal distribution for integer-valued data [9].

Efforts to apply the ARMA framework to continuous data have emerged despite challenges, yielding promising results, particularly with Gaussian data. This adaptation reflects innovative strategies to accommodate continuous data's distinct nature while retaining ARMA's core principles. Successful application in Gaussian contexts demonstrates the model's potential for capturing temporal dependencies, it has limitations in the count data field, and Gaussian ARMA-type processes are insufficient for capturing features of integer-valued time series, such as overdispersion and zero inflation [10]. Since this model does not generate integer predictions, it is prone to approximation errors when applied to count data. This has led to the creation of specific data counting approaches, some of which draw concepts from the autoregressive modeling of continuous data. Some popular approaches such as the integer autoregressive modeling framework (INAR) strive to keep the data's distinct nature. Several researchers have applied this technique in both univariate and multivariate contexts. However, a significant challenge emerges when attempting to capture higher-order dependencies, especially in the extension to multivariate cases. This introduces complexities in implementation within this framework. To tackle this problem, some authors adopt Markov modeling for example [11]. However, Markov models have limited memory and do not explicitly capture long-term dependencies or past events beyond the current state. In situations of systems with intricate temporal relationships or require considerable historical data for appropriate modeling and prediction, this can be an obstacle. Hidden Markov models (HMMs) reduce this issue in part by including hidden states that collect more information [12], but they fail to capture long-term dependencies in some cases. In addition, they work on the assumption that transitions between states are independent events which is also not realistic. Another solution would be to use copula-based modeling, which accounts for dependence in multivariate count data. Although copulas allow for different kinds of dependence structures, finding parametric distributions for high-dimensional random vectors remains difficult because any type of high-dimensional multivariate distribution is limited in covariance structure [13].

Time series of count data recorded in various applications exhibits diverse characteristics and features such as overdispersion, zero inflation, heavy-tailedness, volatility, nonstationarity, and complex dependence structures. In response to this, numerous models have been introduced to effectively handle count time series data by accounting for zero inflation and overdispersion. However, the aspect of heavy-tailedness has received less attention, as noted by Qian et al. [10]. However, ignoring the extreme values or outliers that characterize the feature of the heavy tail may result in the loss of useful information since it is not feasible to ignore the tail probability or assume that it decreases very slowly. Modeling heavy-tailed data present a challenge because it necessitates identifying distributions that can capture both the major portion of the data and the extreme values or outliers [14]. Zhu and Joe [15] introduced a family of distributions known as the generalized Poisson-inverse Gaussian distribution. This distribution is constructed to efficiently capture heavy-tailed count data and provides a flexible strategy for such scenarios. Building upon this work, Qian et al. [10] introduced a novel approach called the GPIG-INAR model of the first order, which involves an INAR process incorporating innovations from the generalized Poisson-inverse Gaussian distribution. For GPIG-INAR to successfully model time series data with heavy-tailed count distributions, this methodology was developed. However, it was considered only in the short-range dependence INAR (1), whereby the motivation of this work considers both short- and long-range dependence.

Additionally, simultaneous modeling of two or more of these aspects when present in the data presents challenges in model specification and estimation. This is further aggravated by the need to specify a modeling strategy that respects the integer nature of the data when accounting for the serial correlation over time. Existing frameworks, such as Markov modeling, INAR strategy, and GLM framework, though successful in their own right, encounter challenges in accommodating numerous features as well as capturing certain dependence patterns such as a long-range serial correlation. The Lévy-based modeling approach was first introduced in the area of turbulence modeling by Barndorff-Nielsen and

Schmiegel [16] where they found that the Lévy-based framework allows very flexible autocorrelation structures and can produce any kind of marginal distribution within the class of integer-valued infinitely divisible distributions. In the context of time series analysis, the Lévy-based framework has been adopted in modeling time series of count data in recent years, see Barndorff-Nielsen et al. [9]; Veraart [17]; Bennedsen et al. [18]; Leonte and Veraart [19]. This approach entails modeling serially correlated count and integer-valued data in continuous time and offers several advantages including flexibility of the autocorrelation structure, simplicity, and accommodating short or long memory processes. This framework due to its simplicity and dynamical control can be enhanced to accommodate various features to develop flexible models within the count time series landscape. Zero inflation and overdispersion are common aspects in various application areas, and failure to account for them, if present in the data, may result in misleading inference.

To the best of our knowledge, there are existing gaps in the literature. First, how can heavy-tailed count data be modeled, considering both short-range and long-range dependence under stationary conditions in the data? In other words, how can we account for all memory ranges in count data, given that it exhibits stationarity and heavy-tailedness? Another question is this how can these features be handled in a simultaneous modeling framework?

In this work, we consider marginal distributions that can account for more features in the data such as zero inflation and overdispersion within the stationary setting and heavy-tailedness in both short- and long-range dependence. To achieve this aim, we develop stationary Poisson inverse-Gaussian Lévy-based models for time series of count data with heavy-tailed characteristics; and stationary semi-Poisson Lévy-based models for zero inflation and overdispersion time series of count data for simultaneous modeling.

The article is structured as follows: Section 2 provides brief preliminaries and components of the Lévy-based modeling framework. In Section 3, model specification consists of choosing the distributions and the kernel set. We estimate the parameters of our models using moments-based methods and composite pairwise likelihood in Section 4. In Section 5, a simulation study is presented. Real data applications are considered in Section 6. We give a conclusion in Section 7.

2. Lévy Bases

This section briefly introduces the Lévy-based framework. This framework can accommodate any kind of marginal distribution within the class of integer-valued infinitely divisible distributions. Further details are provided in Barndorff-Nielsen and Schmiegel [16]. A Lévy basis is a random measure that is infinitely divisible and independently scattered, meaning it can be decomposed into an infinite number of smaller independent random measures. This characteristic is useful for modeling a variety of phenomena, such as disease spread, traffic movement, and customer arrivals at stores. Additional information about independently scattered random measures can be found in Rajput and Rosinski [20] and Kwapien and Woyczynski [21].

Let $(\Omega, \mathcal{F}, P)$ be probability space, and let $(S, \mathcal{S}, |\cdot|)$ denote a Lebesgue-Borel space and $|D|$ denotes the Lebesgue measure of. We assume that S is a subset of $\mathbb{R}^d$, i.e., $S \subset \mathbb{R}^d$ with $(d \in \mathbb{N})$. The set $\mathcal{B}_{\text{Leb}}(S)$ represents the collection of Borel measurable sets with finite Lebesgue measure contained in S. We can think of S as a collection of events that have a time and location in space. The measure l is finite if $l(\mathbb{R}) < \infty$. Lévy measure on $\mathbb{R}$ is Borel measure such that $l(0) = 0$ and $\int_{\mathbb{R}} \min(1, y^2) l(\mathrm{d}y) < \infty$. Finally, $\mathcal{B}_b(S)$ defines the bounded Borel sets of S such that:

$$\mathcal{B}_b(S) = \{D \in S: |D| < \infty\}. \tag{3}$$

The cumulant transform of a random variable X is given by $C(\theta, X) = \log(\mathbb{E}(e^{i\theta X}))$ [22], denoted $X \overset{d}{=} Y$, if X and Y are identically distributed.

2.1. Definition. Lévy basis Λ on (S, S) is a collection of random variables $\Lambda = \{\Lambda(D): D \in \mathcal{B}_{\text{Leb}}(S)\}$ such that:

(i) The law of Λ is infinitely divisible for all $D \in \mathcal{B}_{\text{Leb}}(S)$. Thus, for any natural number $n \in \mathbb{N}$, the measure can be expressed as the sum of n independent and identically distributed random measures $\Lambda_{n,k}$, where $k = 1, \ldots, n$. Otherwise $\Lambda \overset{d}{=} \Lambda_{n,1} + \cdots + \Lambda_{n,n}$.

(ii) The random variables $\Lambda(D_1), \ldots, \Lambda(D_n)$ are independent whenever $D_1, D_2, \ldots, D_n \in \mathcal{B}_{\text{Leb}}(S)$ are disjoint (Independent scattering property).

(iii) For every disjoint sequence $D_1, D_2, \ldots, D_n \in \mathcal{B}_{\text{Leb}}(S)$ with bounded union $\cup_{i=1}^{\infty} D_i \in \mathcal{B}_{\text{Leb}}(S)$ then we have,

$$\Lambda\left(\cup_{i=1}^{\infty} D_i\right) \overset{a.s}{=} \sum_{i=1}^{\infty} \Lambda(D_i). \tag{4}$$

(Additivity property).

The Lévy basis controls the marginal distribution of the resulting stochastic process and is specified by an infinitely divisible distribution. This is the only restriction to its marginal distribution. This offers a wide range of stochastic processes that can be supported on integer and real number states that have light or heavy tail properties. In addition, a Lévy basis Λ on S is homogeneous if it is stationary. Their statistical properties remain unchanged across different points in time.

2.2. Definition. A Lévy basis Λ on (S, S) is said to be stationary if for any $j \in S$ and $D \in \mathcal{B}_b(S)$ such that $j + D = \{j + x \mid x \in D\}$ then

$$\Lambda(D) \overset{d}{=} \Lambda(j + D). \tag{5}$$

A Lévy basis is considered homogeneous if it exhibits stationarity, and its characteristic function follows the following form:

$$C(\theta, \Lambda(D)) = \left(i\theta\varrho - \frac{1}{2}\theta^2 u + \int_R \left(e^{i\theta v} - 1 - i\theta y \mathbf{1}_{[-1,1]}(v) \right) l(\mathrm{d}v) \right) \mathrm{Leb}(D), \tag{6}$$

where $\varepsilon \in \mathbb{R}$, $u \in \mathbb{R}_{\geq 0}$, and l is a Lévy measure on $\mathbb{R}$. The condition for a Lévy basis to be homogeneous is that its characteristic function must be of the form given above called Lévy–Khinchine. This condition ensures that the distribution of $\Lambda(D)$ is the same for all sets D that have the same size and shape, regardless of their location in S.

3. Models Specification

Lévy-based framework has two key components which consist of the choice of the marginal distribution which has to be infinitely divisible, and the kernel set where we consider shapes able to induce a flexible autocorrelation structure, which is expected to be a flexible and valid autocorrection structure, and finally possible to induce both long-range and short-range dependence. For various choices of kernel sets, see Barndorff-Nielsen et al. [9] and Veraart [17]. The criterion for choosing the kernel set in this framework, firstly, is that it must have a finite Lebesgue measure, and secondly, as we concentrate on stationary processes in this context, we make the assumption that the shape of the kernel set remains constant over time.

The Lévy basis determines the marginal law of the process with the chosen distribution depending on the problem at hand. It can handle various marginal distributions as long as they are infinitely divisible. The semi-Poisson distribution is effective in addressing overdispersion and zero-inflation scenarios, while in cases of heavy-tailedness modeling, we consider the Poisson-inverse Gaussian distribution.

3.1. Stationary Poisson-Inverse Gaussian Process. Using Lévy based framework, we define the following observation-driven model Y_t with a Poisson-inverse Gaussian process:

$$Y_t = \mathrm{PIG}(D_t) = \int_{D_t} \mathrm{PIG}(s)\mathrm{d}s, \quad s \in \mathbb{R}, \tag{7}$$

where PIG is a homogeneous Poisson-inverse Gaussian Lévy basis, $\mathrm{PIG}(s)$ is the value of a stochastic process following a Poisson-inverse Gaussian distribution at a specific location s within the real numbers with $D_t \subset R^d$, and $d \in \mathbb{N}$ is a kernel set.

More specifically, the Poisson-inverse Gaussian process is given by

$$Y_t = \mathrm{PIG}(D_t) \sim \text{Poisson} - \text{inverse Gaussian}(\mu|D|, \sigma), \tag{8}$$

with $\mu \in \mathbb{R}_+$.

Moreover, a Lévy basis PIG on (S, S) is said to be stationary if for any $j \in S$ and $D \in \mathscr{B}_b(S)$ such that $j + D = \{j + x \mid x \in D\}$ then

$$\mathrm{PIG}(D) \overset{d}{=} \mathrm{PIG}(j + D). \tag{9}$$

More specifically, the Poisson-inverse Gaussian basis satisfies $\mathrm{PIG}(D) \sim \text{Poisson} - \text{inverse Gaussian}(\mu|D|, \sigma)$, where probability mass function (pmf) of Poisson-inverse Gaussian ($\mathrm{PIG}(\mu, \sigma)$) distribution is derived from the mixed Poisson distribution. The proposed PIG(D) model offers enhanced flexibility in capturing complex autocorrelation structures through the incorporation of a kernel set D, potentially providing a better fit for time series data compared to the standard PIG distribution.

3.2. Definition. A discrete random variable X follows a Poisson-inverse Gaussian (PIG) distribution parameterized by two positive real numbers, μ and σ. The stochastic representation of X given $Y = y$ is Poisson with a mean (μ, y), where Y is a random variable with an inverse Gaussian distribution with $\mathbb{E}[Y] = 1$. We denote $X \sim \mathrm{PIG}(\mu|D|, \sigma)$, and the moment-generating function of X is given by

$$\Phi_X(t) \equiv \mathbb{E}\left[e^{\mathrm{tX}}\right] = \exp\left(\sigma\left(1 - \sqrt{1 - 2\sigma^{-1}\mu\left(e^t - 1\right)|D|} \right) \right), \tag{10}$$

with $t < \log(1 + \sigma/2\mu|D|)$.

The probability mass function is given by

$$p(k) \equiv P(\mathrm{PIG} = k) = \frac{\sqrt{2}}{\sqrt{\pi}} [\sigma(\sigma + 2\mu|D|)]^{-(k-1/2)/2} \left(\frac{e^{\sigma}(\mu|D|)^k}{k!} \right) K_{k-1/2}\left(\sqrt{\sigma(\sigma + 2\mu|D|)} \right), \tag{11}$$

for $k = 0, 1, \ldots$ where $K_\lambda(t) = 1/2 \int_0^\infty u^{\lambda-1} \exp(-1/2u^2)$ $u + 1/u\mathrm{d}u$ the altered Bessel function of the third kind is a mathematical function that can be calculated using software such as Maple and Mathematica.

The mean is defined as follows:

$$\mathbb{E}(\mathrm{PIG}(D)) = \mu|D|. \tag{12}$$

The variance is defined as follows:

$$\mathbb{V}(\mathrm{PIG}(D)) = \mu|D| + \frac{(\mu|D|)^2}{\sigma}. \tag{13}$$

The heavy tail (HT) is defined as follows:

$$\mathrm{HT} = \frac{P(\mathrm{PIG}(D) = k+1)}{P(\mathrm{PIG}(D) = k)}, \quad k \longmapsto \infty,$$

$$p(h) = \mathrm{Cov}(Y_t, Y_{t+h}) = |D \cap D_h| \left(\mu|D| + \frac{(\mu|D|)^2}{\sigma} \right), \quad (14)$$

$$r(h) = \mathrm{Corr}(Y_t, Y_{t+h}) = \frac{|D_t \cap D_{t+h}|}{D_t}.$$

In this scenario, the condition $0 \leq |D_t \cap D_{t+h}| \leq 1$ indicates that the model cannot exhibit negative correlations. Additionally, considering the expression, for $Y_t = \mathrm{PIG}(D_t)$, as the distance $d_{\mathrm{th}} = \|t - h\|$ grows infinitely large, the overlapping region $|D_t \cap D_h|$ tends towards 0. This characteristic guarantees that the process follows an α-mixing pattern.

3.3. Stationary Semi-Poisson Process. Using the Lévy-based framework, we define the following observation-driven model Y_t with a semi-Poisson distribution:

$$Y_t = \mathrm{SP}(D_t) = \int_{D_t} \mathrm{SP}(s)\mathrm{d}s, \quad s \in \mathbb{R}, \quad (15)$$

where SP is a homogeneous semi-Poisson Lévy basis, and $D_t \subset R^d$, $d \in \mathbb{N}$ is a kernel set defined by $k(.)$ an exponential and sup-IG.

More specifically, the semi-Poisson process is given by

$$Y_t = \mathrm{SP}(D_t) \sim \mathrm{Semi-Poisson}(\lambda|D|), \quad \lambda \in \mathbb{R}_+. \quad (16)$$

Furthermore, a Lévy basis SP on (S, S) is said to be stationary if for any $a \in S$ and $D \in \mathscr{B}_b(S)$ such that $a + D = \{a + x \mid x \in D\}$, then

$$\mathrm{SP}(D) \overset{d}{=} \mathrm{SP}(a + D). \quad (17)$$

More specifically, the semi-Poisson basis satisfies $\mathrm{SP}(D) \sim \mathrm{Semi-Poisson}(\lambda|D|)$ where probability mass function (pmf) of semi-Poisson $(\mathrm{SP}(\lambda))$ distribution is defined by

$$P(\mathrm{SP}(D) = k) = C_\lambda \frac{(k+1)(\lambda|D|)^{k+2}}{(k+2)k!}, \quad k \in \mathbb{R}, \quad (18)$$

where $C_\lambda = 1/((\lambda|D|)^2 - \lambda|D| + 1)e^{\lambda|D|} - 1$, and $\lambda > 0$.

The cumulative function is given by

$$C(y, \mathrm{SP}(D)) = E_\lambda\left(\frac{Y+1}{Y+2}\right) - (\lambda|D|)^y E\left[\left(\frac{t!}{(Y+t)}\right)\left(\frac{Y+t+1}{Y+t+2}\right)\right]. \quad (19)$$

The mean is defined as follows:

$$\mathbb{E}(\mathrm{SP}(D)) = e^{\lambda|D|} \frac{\left\{-2 + 2\lambda|D| - (\lambda|D|)^2 + (\lambda|D|)^3\right\} + 2}{e^{\lambda|D|}\left\{1 - \lambda|D| + (\lambda|D|)^2\right\} - 1},$$
$$\mathbb{E}\left(\mathrm{SP}(D)^2\right) = \frac{\mathrm{e}^{\lambda|D|}\lambda|D|^4 + 2\mathrm{e}^{\lambda}(\lambda|D|)^2 - 4\mathrm{e}^{\lambda|D|}\lambda|D| + 4\mathrm{e}^{\lambda|D|} - 4}{\mathrm{e}^{\lambda|D|}\left((\lambda|D|)^2 - \lambda|D| + 1\right) - 1}. \quad (20)$$

Hence, the variance can be obtained as follows:

$$\mathbb{V}(\mathrm{SP}(D)) = \frac{e^{\lambda|D|}(\lambda|D|)^2\left[-2 - 4\lambda|D| - (\lambda|D|)^2 + e^{\lambda|D|}\left\{2 + 2\lambda|D| - 2(\lambda|D|)^2 + (\lambda|D|)^3\right\}\right]}{\left[e^{\lambda|D|}\left\{1 - \lambda|D| + (\lambda|D|)^2\right\} - 1\right]^2}. \quad (21)$$

The index of dispersion is defined as follows:

$$\mathbb{I} = \frac{\mathbb{V}(\mathrm{SP}(D))}{\mathbb{E}(\mathrm{SP}(D))} = \frac{e^{\lambda|D|}(\lambda|D|)^2\left(-2 - 4\lambda|D| - (\lambda|D|)^2 + e^{\lambda|D|}\left(2 + 2\lambda|D| - 2(\lambda|D|)^2 + (\lambda|D|)^3\right)\right)}{\left(e^{\lambda|D|}\left(1 - \lambda|D| + (\lambda|D|)^2\right) - 1\right)\left(e^{\lambda|D|}\left(-2 + 2\lambda|D| - (\lambda|D|)^2 + (\lambda|D|)^3\right) + 2\right)}. \quad (22)$$

We introduce the zero inflation index as follows:

$$\mathrm{ZI} = 1 + \frac{\log P(\mathrm{SP}(D) = 0)}{\mathbb{E}(\mathrm{SP}(D))}. \quad (23)$$

The zero inflation index (ZI) is a measure of the excess of zeros in a dataset. A negative ZI indicates that there are more zeros than expected, a zero ZI indicates no excess zeros, and a positive ZI indicates fewer zeros than expected.

Let $h > 0$. For each component, the autocovariance and autocorrelation functions are given by

$$p(h) = \mathrm{Cov}(Y_t, Y_{t+h}) = |D \cap D_h| \frac{e^{\lambda|D|}(\lambda|D|)^2\left[-2 - 4\lambda|D| - (\lambda|D|)^2 + e^{\lambda|D|}\left\{2 + 2\lambda|D| - 2(\lambda|D|)^2 + (\lambda|D|)^3\right\}\right]}{\left[e^{\lambda|D|}\left\{1 - \lambda|D| + (\lambda|D|)^2\right\} - 1\right]^2}. \tag{24}$$

Hence,

$$r(h) = \mathrm{Corr}(Y_t, Y_{t+h}) = \frac{|D_t \cap D_{t+h}|}{D_t}. \tag{25}$$

In this work, for the kernel sets, we consider a parametric specification of the form:

$$D = \left\{(s, v):\ s \in \mathbb{R}, 0 \le v < \frac{1}{\phi} k\left(\frac{s-\mu}{\phi}\right)\right\}, \tag{26}$$

where μ and ϕ are location and scale parameters, respectively.

For a short-range dependence, we consider the exponential shape in the form:

$$D_t = \left\{(s, v):\ s \le t, 0 \le v < e^{-\lambda(t-s)}\right\}, \quad \lambda > 0,\ t \le 0, \tag{27}$$

and the autocorrelation function is given by $r(h) = \exp(-\lambda h)$ with $h \ge 0$; consequently, for $t \ge 0$, $|D| = 1/\lambda$, $|D_t \cap D| = 1/\lambda e^{-\lambda t}$ and $|D_t/D| = 1/\lambda(1 - e^{-\lambda t})$.

For a long-range process dependence, we have

$$D_t = \left\{(s, v):\ s \le t, 0 \le v < \left(1 - \frac{2t}{\gamma^2}\right)^{1/2} \exp\left(\delta_\gamma\left(1 - \sqrt{1 - \frac{2t}{\gamma^2}}\right)\right)\right\}, \tag{28}$$

for $t \le 0$. The autocorrelation function is as follows:

$$r(h) = \mathrm{Cor}(Y_t, Y_{t+h}) = \exp\left(\delta_\gamma\left(1 - \sqrt{1 - \frac{2h}{\gamma^2}}\right)\right), \tag{29}$$

with

$$|D| = \frac{\gamma}{\delta}, |D_t \cap D| = \frac{\gamma}{\delta} e^{\delta\gamma(1-\eta_t)}, |D_t \backslash D| = \frac{\gamma}{\delta}\left(1 - e^{\delta\gamma(1-\eta_t)}\right), \tag{30}$$

where $\eta_t = \sqrt{2t/\gamma^2 + 1}$.

The choice of our kernel set is due to both analytical tractability and modeling flexibility. The exponential kernel is the simplest, while the super-GIG kernel is flexible, consistent with data properties, and computationally tractable. For the process's realization, we have the set that is moving along the time axis via the location parameter μ governing the movement and temporal dependence of the process with the shape parameter controlling the strength and pattern of this dependence via the scale parameter ϕ.

4. Parameters Estimation

This section looks into the statistical properties of the moments-based methods and composite likelihood based on pairs of observations for the estimation of parameters. We give a thorough overview of moments-based methods. We also review pairwise likelihood methods and highlight their advantages over the standard likelihood method. Indeed, the maximum likelihood becomes impractical when the number of observations is very large. This is mainly due to computational challenges that arise with the increased size of the dataset. Pairwise likelihood can be useful in situations with large datasets or complex models where computing the whole likelihood function is difficult or when data are sparse or partial.

4.1. Composite Pairwise Likelihood. The introduction of composite likelihood methods is important because there are a number of situations, such as time series models, where the computation of the full likelihood is very difficult and too time-consuming [23]. The term composite likelihood denotes a general class of pseudolikelihoods [24] based on likelihood-type objects. Consider an m–dimensional vector random variable Y, with probability density function $f(z; \theta)$ for unknown p–dimensional parameter vector $\theta \in \Theta$.

Let $A_1, A_2, \ldots, A_k$ be a collection of marginal or conditional events, with associated composite likelihoods $L_k(\theta; y)$ proportional to $f(y \in A_k; \theta)$ that is $L_k(\theta; y) = L(\theta; A_k(y))$ with $k = 1, 2, \ldots$.

A composite likelihood is defined as follows:

$$\mathrm{CL}(\theta; y) = \prod_{k=1}^{K} L_k(\theta; y)^{w_k}, \tag{31}$$

where w_k are suitable nonnegative weights that do not depend on θ. Here, we discuss an alternative strategy based on a simple pseudolikelihood known as "pairwise likelihood." Its advantage is that it reduces the computational burden so that it is possible to fit highly structured statistical models. The pairwise likelihood is a statistical technique that breaks down the joint likelihood function into a product of pairwise conditional or marginal likelihoods. This simplification allows for more manageable parameter estimation and inference, making it particularly useful in situations with complex or

high-dimensional data where traditional likelihood methods become computationally infeasible. For more details, we point to Lindsay [24]; Varin et al. [25]; and Davis and Yau [26]; since the bivariate distributions are available in closed form, this technique has also been used in Bennedsen et al. [27] to make conclusions about related procedures. For the pairs, $(Y_i, Y_j; i < j)$ can be decomposed as follows:

$$\begin{aligned} Y_j &= \widetilde{Y}_{j\backslash i} + \widetilde{Y}_0, \\ Y_i &= \widetilde{Y}_{i\backslash j} + \widetilde{Y}_0, \end{aligned} \tag{32}$$

where $\widetilde{Y}_{j\backslash i} \sim \text{Semi} - \text{Pois}(\lambda|D_j\backslash D_i|)$, $\widetilde{Y}_{i\backslash j} \sim \text{Semi} - \text{Pois}(\lambda|D_i\backslash D_j|)$, and $\widetilde{Y}_0 \sim \text{Semi} - \text{Pois}(\lambda|D_i \cap D_j|)$ are random variables, and $\widetilde{Y}_j$ and $\widetilde{Y}_i$ have $\widetilde{Y}_0$ in common. Since the semi-Poisson Lévy basis is independently scattered, the sets are disjoint and independent. The pairwise composite likelihood function of order k ($k = \max$ time lag)

$$\text{pcL}(\theta; y) = \sum_{i=1}^{n-k} \sum_{j=i+1}^{i+k} \Pr\left(Y_i = y_i, Y_j = y_j; \theta\right), \tag{33}$$

where y is the observed data, $\theta = (\lambda)$ is the parameter to be estimated, and Y_i and Y_j represent random variables at time i and j, respectively.

$$\begin{aligned} \Pr\left(Y_j = y_j, Y_i = y_i\right) &= \sum_{w=0}^{\min\left(y_i, y_j\right)} \Pr\left(\widetilde{Y}_{j\backslash i} = y_j - w\right) \\ &\quad \cdot \Pr\left(\widetilde{Y}_{i\backslash j} = y_i - w\right) \Pr\left(\widetilde{Y}_0 = w\right) \\ &= \sum_{w=0}^{\min\left(y_i, y_j\right)} C_\lambda \frac{\left(y_j - w + 1\right)\left(\lambda\left|D_j\backslash D_i\right|\right)^{y_j - w + 2}}{\left(y_j - w + 2\right)\left(y_j - w\right)!} \\ &\quad \times \sum_{w=0}^{\min\left(y_i, y_j\right)} C_\lambda \frac{\left(y_i - w + 1\right)\left(\lambda\left|D_i\backslash D_j\right|\right)^{y_i - w + 2}}{\left(y_i - w + 2\right)\left(y_i - w\right)!} \\ &\quad \times \sum_{w=0}^{\min\left(y_i, y_j\right)} C_\lambda \frac{\left(w + 1\right)\left(\lambda\left|D_j \cap D_i\right|\right)^{w+2}}{\left(w + 2\right)\left(w\right)!}, \end{aligned} \tag{34}$$

with

$$\begin{aligned} C_\lambda = &\frac{1}{\left(\left(\lambda\left|D_j\backslash D_i\right|\right)^2 - \lambda\left|D_j\backslash D_i\right| + 1\right)e^{\lambda\left|D_j\backslash D_i\right|} - 1} \\ &\times \frac{1}{\left(\left(\lambda\left|D_i\backslash D_j\right|\right)^2 - \lambda\left|D_i\backslash D_j\right| + 1\right)e^{\lambda\left|D_i\backslash D_j\right|} - 1} \\ &\times \frac{1}{\left(\left(\lambda\left|D_j \cap D_i\right|\right)^2 - \lambda\left|D_j \cap D_i\right| + 1\right)e^{\lambda\left|D_j \cap D_i\right|} - 1} \lambda > 0. \end{aligned} \tag{35}$$

Davis and Yau [26] provided evidence that it is not necessary to use all possible lags.

To find the maximum likelihood estimator (MLE) for the parameter λ based on the pairwise composite likelihood function, we need to maximize the log-likelihood function at the form:

$$\log p\text{cL}(\lambda; y) = \log\left(\sum_{i=1}^{n-k} \sum_{j=i+1}^{i+k} \sum_{w=0}^{\min\left(y_i, y_j\right)} \frac{\left(y_j - w + 1\right)\left(\lambda\left|D_j\backslash D_i\right|\right)^{y_j - w + 2}}{\left(y_j - w + 2\right)\left(y_j - w\right)!} \cdot \frac{\left(y_i - w + 1\right)\left(\lambda\left|D_i\backslash D_j\right|\right)^{y_i - w + 2}}{\left(y_i - w + 2\right)\left(y_i - w\right)!} \cdot \frac{\left(w + 2\right)\left(w + 1\right)\left(\lambda\left|D_j \cap D_i\right|\right)^{w+2}}{\left(w + 2\right)!}\right), \tag{36}$$

where C_λ is the constant defined as follows:

$$C_\lambda = \frac{1}{\left(\left(\lambda\left|D_j\backslash D_i\right|\right)^2 - \lambda\left|D_j\backslash D_i\right| + 1\right)e^{\lambda\left|D_j\backslash D_i\right|} - 1} \times \frac{1}{\left(\left(\lambda\left|D_i\backslash D_j\right|\right)^2 - \lambda\left|D_i\backslash D_j\right| + 1\right)e^{\lambda\left|D_i\backslash D_j\right|} - 1} \times \frac{1}{\left(\left(\lambda\left|D_j \cap D_i\right|\right)^2 - \lambda\left|D_j \cap D_i\right| + 1\right)e^{\lambda\left|D_j \cap D_i\right|} - 1}. \tag{37}$$

4.2. Method of Moments Estimation. While the pairwise composite likelihood has proven effective within the univariate context and has demonstrated efficient computational performance in terms of both time and estimation, its application encounters challenges when transitioning to the presence of multiple components. The approach faces computational challenges of increased magnitude, especially when closed-form solutions are unavailable due to the nontrivial interaction of two distributions like in Poisson-inverse Gaussian. Another approach to bridge this gap is the method of moments estimation (MME), introduced by Karl Pearson in 1894. MME provides a flexible framework for parameter estimation and gives the estimate by comparing the functions of the sample and their theoretical moments. MME is a specific case of GMM and is often used to estimate the parameters of a distribution by equating the sample moments to the theoretical moments of the distribution. Given that it is often impractical to gather data from an entire population, we rely on a sample taken from that population to estimate its moments. The notion of a moment is fundamental for describing features of a population.

Suppose an observation $X_n = X_i\colon i = 1, \ldots, n$ is a sample from a population with mean μ for which we aim to estimate an unknown parameter vector $\theta \in \Theta \subset \mathbb{R}^d$. This estimation involves using a vector $T_n = T_n(X_n)$. These statistics have an expected value $\gamma(\theta) = E[T_n]$ in the theoretical context, where $\gamma(\theta)$ represents their theoretical counterparts under the specific model. The concept of a "moment condition" is introduced, which involves the expectation of a function $f_n(\theta; X_n)$ and is defined as $E[f_n(\theta; X_n)] = 0$. In this case, $f_n(\theta; X_n)$ is a continuous $1 \times k$ vector function of θ, and $E[f_n(\theta; X_n)]$ is finite and exists for all values of i and θ. In practical terms, this moment condition is approximated using its sample equivalent: $1/n\sum_{i=1}^n (f_n\theta; X_i) = 0$. This equation allows us to obtain the estimator $\widehat{\theta}$. For the dimension of $j = k$, we arrive at what is known as the method of moments (MM) estimator. The MM estimator θ_n is obtained by minimizing the expression:

$$\widehat{\theta}_n = \arg\min_\theta \left[\frac{1}{n}\sum_{i=1}^n f_n(\theta; X_n)^T Z_n \sum_{i=1}^n f_n(\theta; X_n)\right], \tag{38}$$

where Z_n represents a symmetric and positive definite weight matrix of size $k \times k$. It can depend on the data and should converge in probability to a positive definite matrix Z.

In our case for the moment conditions, with unknown parameters $\theta = (\mu, \sigma)$, let m_1 and m_2 denote the first- and second-order moments respectively, then

$$\mathbb{E}(X) = \mu|D|,$$

$$\mathbb{V}(X) = \mu|D| + \frac{(\mu|D|)^2}{\sigma},$$

1st moment:

$$m_1 = \frac{1}{n}\sum_{i=1}^n X_i,$$

$$\mathbb{E}(\text{PIG}) = m_1 \Longrightarrow \widehat{\mu}|D| = m_1 \Longrightarrow \widehat{\mu} = \frac{1}{|D|} \times \frac{\sum_{i=1}^n X_i}{n},$$

$$\widehat{\mu} = \frac{\sum_{i=1}^n X_i}{n|D|},$$

2nd moment:

$$m_2 = \frac{1}{n} \times \sum_{i=1}^n (X_i - m_1)^2,$$

$$m_2 = \text{Var}(\text{PIG}) = \widehat{\mu}|D| + \frac{(\widehat{\mu}|D|)^2}{\widehat{\sigma}}$$

$$= \frac{m_1}{|D|} \times |D| + \left(\frac{m_1}{|D|}\right)^2 \times \frac{1}{\widehat{\sigma}}$$

$$= m_1 + \frac{m_1^2}{|D|^2\widehat{\sigma}} \Longrightarrow \widehat{\sigma} = \frac{m_1^2}{|D|^2(m_2 - m_1)},$$

$$\widehat{\sigma} = \frac{\left(1/n\sum_{i=1}^n X_i\right)^2}{|D|^2\left(1/n\sum_{i=1}^n (X_i - 1/n\sum_{i=1}^n X_i)^2 - 1/n\sum_{i=1}^n X_i\right)}. \tag{39}$$

By substituting the first-moment estimator m_1 into the equation of $\widehat{\sigma}$, we have

$$\widehat{\sigma} = \frac{m_1^2}{|D|^2\left(1/n\sum_{i=1}^n (X_i - m_1)^2 - m_1\right)}. \tag{40}$$

This equation now expresses the sample variance estimator $\widehat{\sigma}$ in terms of the first-moment estimator m_1 and the differences between each data point and the first moment. In this case, the criterion function seems to relate to the first-moment estimator m_1 and the second-moment estimator $\widehat{\sigma}$ through a formula that involves the differences between each

data point and the first moment. The purpose of the criterion function would likely be related to parameter estimation or evaluating the goodness-of-fit of a distribution to the given data.

5. Simulation Studies

5.1. Slice Partition. This section presents the simulation approach based on slice partition, let us start by decomposing the sets $A_\psi, \ldots, A_{k\psi}$ into distinct slices denoted as S collected in S. This allows for simulating the values of the Poisson-inverse Gaussian Lévy basis Λ over each slice, and the process leads to the formulation of the computation for $X_{l\psi}$ as the sum of $\Lambda(S)$ across slices contained within $A_{l\psi}$:

$$Xl_\psi = \sum_{S \subset A_{l\psi}} \Lambda(S). \tag{41}$$

Exploiting the independent-scattered nature of the Lévy basis, we can independently sample $L(S)_{S \in \mathbb{S}}$. This allows us to utilize the additivity property of the Lévy basis [19]. As a result, we can reconstruct the value of the trawl $X_{l\psi}$ by summing the values derived from the Lévy basis simulations over slices contained by $A_{l\psi}$.

For instance, if there exists a $T < 0$ such that $\zeta(T) = 0$, let $I = \lceil -T/\psi \rceil$, using the ceiling function $\lceil \cdot \rceil$, and this consequently defines the slice partition as follows:

$$\begin{aligned} S_{t1} &= \{t \le \psi\} \cap \left(\frac{A_t}{A_{t+1}}\right), \\ S_{tj} &= \frac{\left(\{(j-1)\cdot\psi < t \le j\cdot\psi\} \cap A_{t+j-1}\right)}{A_{t+j}}, \quad \text{for } j \ge 2. \end{aligned} \tag{42}$$

Consequently, I consecutive trawl sets share nonempty intersections, with each A_ψ containing exactly I slices, culminating in a total of kI slices. Defining $s_{tj} = \mathrm{Leb}(S_{tj})$, the translational invariance of the Lebesgue measure establishes $s_{tj} = s_{tj'}$ for $j, j' \ge 2$. This simplifies the process of determining the slice areas S_{ij} by computing s_{tj} for $t \in \{1, \ldots, k\}$ and $j \in \{1, 2\}$. The calculation is defined as follows:

$$\begin{aligned} s_{t1} &= \int_{-i\cdot\psi}^{(-t+1)\cdot\psi} \zeta(t)\mathrm{d}t, \\ s_{t2} &= s_{t1} - s_{t+1,1}, \end{aligned} \tag{43}$$

in which we set $s_{I+1,1} = 0$.

The equations given above completely explain the values of the areas s_{tj}.

5.2. Inverse Transform Method. For the semi-Poisson Lévy basis, we generate random variables using the inverse transform method.

Theorem 1. *Let X be a random variable with cumulative distribution function (cdf) $F(x), x \in \mathbb{R}$ (continuous or not). Then,*

$$F(X) \sim U(0, 1). \tag{44}$$

Proof. Let $Y = F(X)$ and suppose that Y has cumulative distribution function (cdf) $K(y)$. Then,

$$\begin{aligned} K(y) &= P(Y \le y) = P(F(X) \le y) \\ &= P\left(X \le F^{-1}(y)\right) \\ &= F\left(F^{-1}(y)\right) = y. \end{aligned} \tag{45}$$

From the above, the inverse c.d.f. can be defined as follows:

$$F^{-1}(y) = \min\{x : F(x) \ge y\}, \quad y \in [0, 1]. \tag{46}$$

□

Proposition 2. *Let $F(x), x \in \mathbb{R}$ denote any given cumulative distribution function (cdf) and let $F^{-1}(y)$ with $y \in [0, 1]$ be the inverse function defined in 5.4. Let $U \sim U(0, 1)$. Define $X = F^{-1}(U)$ means X is distributed as F, that is, $F(x) = P(X \le x)$.*

Proof. Let us show that $P(F^{-1}(U) \le x) = F(x)$ with $x \in \mathbb{R}$. First, we assume F to be continuous. Saying so let us show that $\{F^{-1}(U) \le x\} = \{U \le F(x)\}$, by taking probabilities (and letting $b = F(x)$ in $P(U \le b) = b$) results in what follows:

$$P\left(F^{-1}(U)\right) \le x = P(U \le F(x)) = F(x). \tag{47}$$

Finally, the equation $F(F^{-1}(y)) = y$ implies (by monotonicity of F) that if $F^{-1}(U) \le x$, then $U = F(F^{-1}(U)) \le F(x)$, or $U \le F(x)$. Likewise, we can observe $F^{-1}(F(x)) = x$, and as a result, when $U \le F(x)$, then $F^{-1}(U) \le x$. This establishes the equality of the two events as was sought to prove. In the general context, it is straightforward to illustrate that

$$\{U < F(x)\} \subseteq \left\{F^{-1}(U) \le x\right\} \subseteq \{U \le F(x)\}. \tag{48}$$

This leads to the same outcome when considering probabilities (since $P(U = F(x)) = 0$ due to the continuous nature of the random variable U).

In this case, we simulate the semi-Poisson process following the scheme outlined.

We first need to identify the cumulative distribution function (cdf) for the probability distribution we want to generate random numbers from, that is, $F(x) = P(X \le x)$. Next, we compute the inverse of the cdf as $F^{-1}(u) = x$ such that $F(x) = u$, which is also referred to as the quantile function or percent-point function. This inverse cdf takes a probability value as input and outputs the corresponding value from the desired distribution. Once we have the inverse cdf, we generate a random number using a uniform distribution that ranges between 0 and 1 such as $x = F^{-1}(u)$.

The cumulative distribution function of a semi-Poisson random variable is given as follows:

$$F(y,\lambda) = E_\lambda\left(\frac{Y+1}{Y+2}\right) - (\lambda|D|)^y E\left[\left(\frac{t!}{(Y+t)}\right)\left(\frac{Y+t+1}{Y+t+2}\right)\right],$$

Pose $U = F(y,\lambda)$,

$$A = E_\lambda\left(\frac{Y+1}{Y+2}\right),$$

$$B = E\left[\left(\frac{t!}{Y+t}\right)\left(\frac{Y+t+1}{Y+t+2}\right)\right]. \tag{49}$$

Then, from a given equation:

$$U = A - \lambda^y B, \tag{50}$$

we can manipulate it as follows: $A - U = \lambda^y B$, which leads to

$$\frac{A-U}{B} = \lambda^y = e^{y\ln\lambda},$$

$$\ln\left(\frac{A-U}{B}\right) = y\ln\lambda, \tag{51}$$

$$y = \frac{1}{\ln\lambda}\ln\left(\frac{A-U}{B}\right).$$

This expression represents the inverse of our cdf and U, A, B defined above with λ the parameter.

Now, we simulate our models by considering both the exponential kernel for the short-range dependence and sup-IG for the long-range dependence. Then, we estimate the parameter vectors $\theta = (\lambda, \mu, \sigma)$, $\theta = (\delta, \gamma, \mu, \sigma)$, $\theta = (\lambda, L = \lambda)$, and $\theta = (\delta, \gamma, L = \lambda)$, respectively, where $L = \lambda$ is the parameter of the semi-Poisson. We compute the mean, bias, and mean-squared error (MSE) given by $\widehat{\theta} = 1/N\ \sum_{i=1}^{N}\widehat{\theta}_i$, Bias $(\widehat{\theta}) = 1/N\sum_{i=1}^{N}(\widehat{\theta}_i - \theta)$, and MSE $(\widehat{\theta}) = 1/N\sum_{i=1}^{N}\ (\widehat{\theta}_i - \theta)^2$where $\widehat{\theta}$ represents the parameter vector values that have been estimated from the data of the i^{th} simulated series. $0 \leqslant \vartheta < 1$.

The study investigated the stationarity of data generated from a Poisson-inverse Gaussian Lévy-based model using an exponential kernel, as depicted in Figures 1 and 2. This investigation was extended to include a sup-IG kernel, illustrated in Figure 3. Stationarity was confirmed through the observation of exponential decay in the autocorrelation function. The study also included graphical summaries of time series data for parameter estimates, presented in Tables 1 and 2. These results indicate that larger sample sizes enhance the accuracy of parameter estimation. Additionally, Figures 4 and 5 demonstrate the stationarity of data from both the semi-Poisson Lévy-based model with an exponential kernel and the sup-IG kernels, respectively. Parameter estimation was also conducted using the pairwise likelihood estimation method, which yielded more accurate estimates with increasing sample sizes, as shown in Tables 3 and 4. Finally, the trawl process of the super-IG model, depicted in Figure 6, exhibits sustained long-term dependence under the estimated parameters. □

6. Real-Data Applications

This section discusses the real-data application of the proposed models: Poisson-inverse Gaussian Lévy-based and semi-Poisson Lévy-based. Data for the second model were obtained from the Meteorological Services Department of Guinea Conakry. For the first model, we analyzed data consisting of the numbers of NSF funding awarded to academic institutions, which is discussed in Qian et al. [10] and available at https://www.nsf.gov.

6.1. Dataset 1. These authors used the number of NSF award data, which is presented in Figure 7, for modeling the GPIG-INAR model. Figure 8 shows that the data are stationary. The proposed Poisson-inverse Gaussian Lévy-based model exhibits flexibility in capturing both short-term (Table 5; Figure 9) and long-term (Table 5; Figure 10) dependencies within the same dataset. Table 5 shows that the PIG-Lévy-based model is particularly more suitable for the given data as it has a lower value of the AIC information criterion. The mean predictions from the two models are comparable.

We also fitted the PIG-Lévy-based model to the data (Tables 6 and 7) using an exponential and sup-IG kernel, respectively, to capture both short-range and long-range dependencies nature. Mean estimates from the model were slightly higher, while variance estimates were slightly lower (Table 5). The AIC was found to be slightly high. Despite the fact that a comparison with existing models could not be made because they are not constructed within the long-range framework, the model still provided more precise predictions for the numbers of NSF award data (Figure 10). The extension to long-range dependence is an advantage of our framework.

6.2. Dataset 2. We estimated weekly rainfall frequencies using daily rainfall data from the N'zerekore region in Guinea Conakry between 2008 and 2023 (Figure 11) and the autocorrelation and partial autocorrelation functions of the data (Figure 12).

We considered a rainfall amount of at most 2.54 mm to represent a dry day. Thus, the data are zero-inflated and overdispersed. The parameter estimate showed that $L = 1.42$ (Table 8), indicating that the semi-Poisson random variable from this model has the highest probability of yielding many zeros. The semi-Poisson Lévy-based model captured this tendency very well by predicting quantiles with a high degree of closeness (Table 9; Figures 13 and 14). Additionally, the predicted values were shown to follow the same probability distribution as the raw data, as depicted in Figures 15 and 16.

6.3. Goodness of Fit of the Model. To verify the accuracy of the model, we used the mean absolute error (equation (52)) and the coefficient of determination (equation (53))

$$\text{MAE} = \frac{1}{N}\sum_{i=1}^{N}\left|y_i - \widehat{y}_i\right|, \tag{52}$$

$$R^2 = 1 - \frac{\sum_{i=1}^{n}\left(y_i - \widehat{y}_i\right)^2}{\sum_{i=1}^{n}\left(y_i - \overline{y}_i\right)^2}. \tag{53}$$

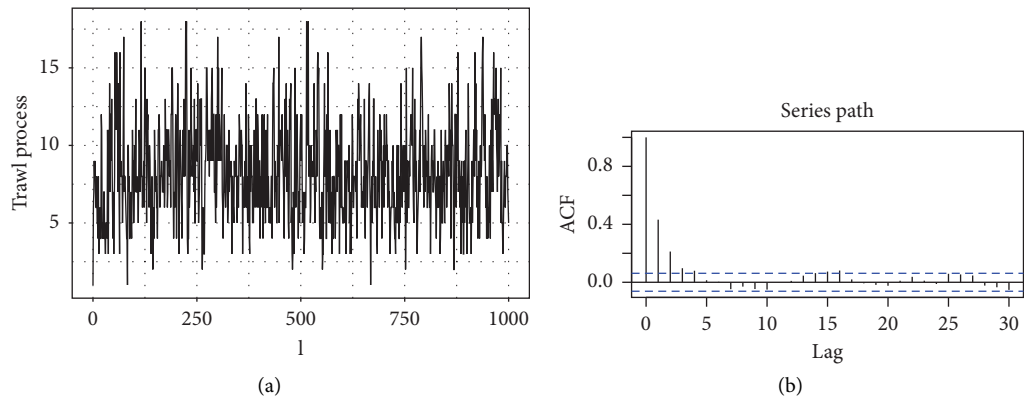

Figure 1: Time series and autocorrelation function plots of simulated data from Poisson-inverse Gaussian Lévy-based with an exponential kernel for the short memory process ($N = 1000, \lambda = 0.6, \mu = 5, \sigma = 0.5$). (a) Time series. (b) ACF.

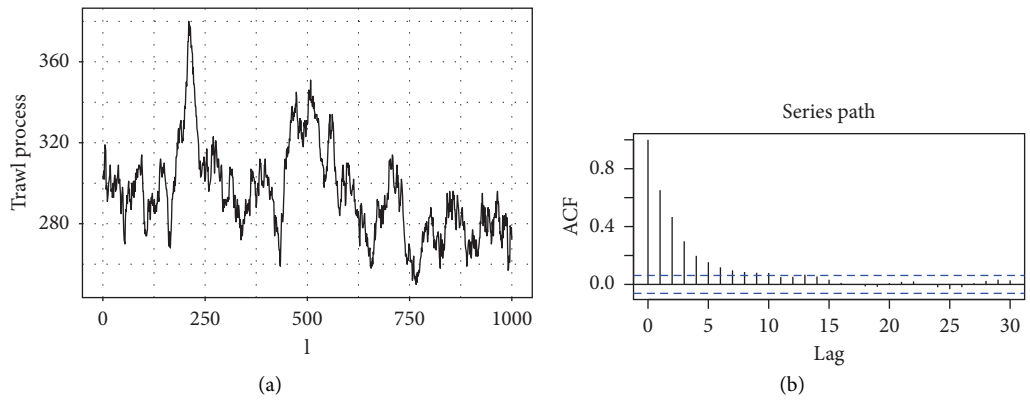

Figure 2: Time series and autocorrelation function plots of simulated data from Poisson-inverse Gaussian Lévy-based with an exponential kernel ($N = 1000, \lambda = 0.3, \mu = 9, \sigma = 0.7$). (a) Time series. (b) ACF.

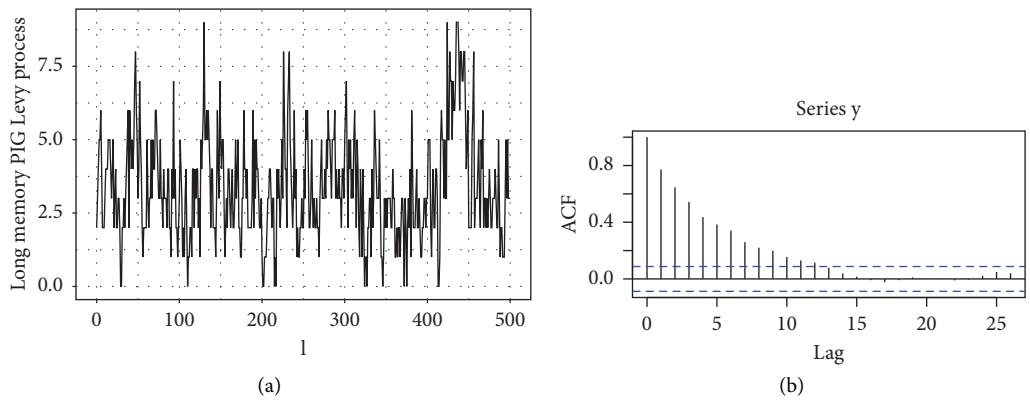

Figure 3: Time series and autocorrelation function plots of simulated data from Poisson-inverse Gaussian Lévy-based with a long memory process ($N = 500, \delta = 0.62, \gamma = 0.95, \mu = 2.26, \sigma = 0.25$). (a) Time series. (b) ACF.

For the semi-Poisson Lévy-based model, the MAE is 0.128, implying that model values are much closer to data values on average. The value of R^2 is 0.823, indicating that the model behaves much closer to the data than the center line of the data. The confidence interval for R^2 was calculated using the following formula [28]:

$$\mathrm{CI} = R^2 \pm Z_{\alpha/2}\widehat{\sigma}_{R^2},$$
$$\sigma_{R^2} = \sqrt{\frac{4R^2\left(1-R^2\right)^2(N-p-1)^2}{\left(N^2-1\right)(N+3)}}. \tag{54}$$

Table 1: Summary statistics for the MM estimator for different parameter values $\theta = (\lambda, \mu, \sigma)$, using an exponential kernel (representing short-range dependence) at different sample lengths of the series N for a Poisson-inverse Gaussian Lévy-based process.

N	True value	λ 1.9	μ 5.5	σ 0.03	λ 2	μ 9	σ 0.04	λ 9	μ 0.8	σ 6.7
	Mean	1.8666	5.4666	0.0288	1.9666	8.9666	0.0388	8.9666	0.7666	6.6988
30	Bias	0.0022	−0.0004	−0.0000	0.0055	−0.0011	−0.0153	−0.0011	−0.1744	0.2066
	MSE	0.0001	0.0000	0.0000	0.0009	0.0000	0.0070	0.0000	0.9129	1.2808
	Mean	1.8900	5.4900	0.0299	1.9900	8.9900	0.0399	8.9900	0.7900	6.6999
100	Bias	0.0009	−0.0001	−0.0000	0.0019	−0.0001	−0.0046	−0.0001	−0.0521	0.0619
	MSE	0.0000	0.0000	0.0000	0.0003	0.0000	0.0021	0.0000	0.2714	0.3843
	Mean	1.8975	5.4975	0.0300	1.9975	8.9975	0.0399	8.9975	0.7975	6.6999
400	Bias	0.0002	−0.0000	0.0000	0.0004	−0.0000	−0.0011	−0.0000	−0.0130	0.0154
	MSE	0.0000	0.0000	0.0000	0.0000	0.0000	0.0005	0.0000	0.0676	0.0960
N	True value	7	0.4	0.05	10	0.7	3	3	6	9
	Mean	6.9900	0.3900	0.0499	9.9900	0.6900	2.9999	2.9900	5.9900	8.9999
100	Bias	0.0519	−0.0561	−0.0045	0.0819	−0.0531	0.0249	0.0119	−0.0001	0.0849
	MSE	0.2693	0.3147	0.0020	0.6707	0.2819	0.0624	0.0141	0.0000	0.7224
	Mean	6.9950	0.3950	0.0499	9.9950	0.6950	2.9999	2.9950	5.9950	8.9999
200	Bias	0.0259	−0.0280	−0.0022	0.0409	−0.0265	0.0124	0.0059	−0.0000	0.0424
	MSE	0.1349	0.1570	0.0010	0.3357	0.1407	0.0312	0.0071	0.0000	0.3612
	Mean	6.9975	0.3975	0.0499	9.9975	0.6975	2.9999	2.9975	5.9975	8.9999
400	Bias	0.0129	−0.0140	−0.0011	0.0204	−0.0132	0.0062	0.0029	−0.0000	0.0212
	MSE	0.0675	0.0784	0.0005	0.1679	0.0702	0.0156	0.0035	0.0000	0.1806

Table 2: Summary statistics for the MM estimator for different parameter values $\theta = (\delta, \gamma, \mu, \sigma)$ with different N for a Poisson-inverse Gaussian Lévy-based process with long-range dependence.

N	True value	δ 5	γ 2	μ 4	σ 0.7	δ 7	γ 0.3	μ 1.9	σ 0.05
	Mean	4.9962	1.9962	3.9962	0.6999	7.0003	0.3004	1.9004	0.0500
100	Bias	0.0099	0.0196	−0.0200	0.0019	0.0300	0.0027	−0.0409	−0.0044
	MSE	0.0099	0.0386	0.0401	0.0003	0.0900	0.0007	0.1680	0.0020
N	True value	5	2	4	0.7	7	0.3	1.9	0.05
	Mean	5.0002	2.0002	4.0002	0.7000	7.0004	0.3004	1.9004	0.0500
500	Bias	0.0020	0.0039	−0.0039	0.0004	0.0060	0.0005	−0.0081	−0.0008
	MSE	0.0020	0.0077	0.0079	0.0000	0.0180	0.0001	0.0336	0.0004
N	True value	5	2	4	0.7	7	0.3	1.9	0.05
	Mean	4.9997	1.9997	3.9997	0.6999	7.0003	0.3003	1.9003	0.0500
1000	Bias	0.0009	0.0019	−0.0020	0.0001	0.0030	0.0002	−0.0040	−0.0004
	MSE	0.0009	0.0038	0.0040	0.0000	0.0090	0.0000	0.0168	0.0002

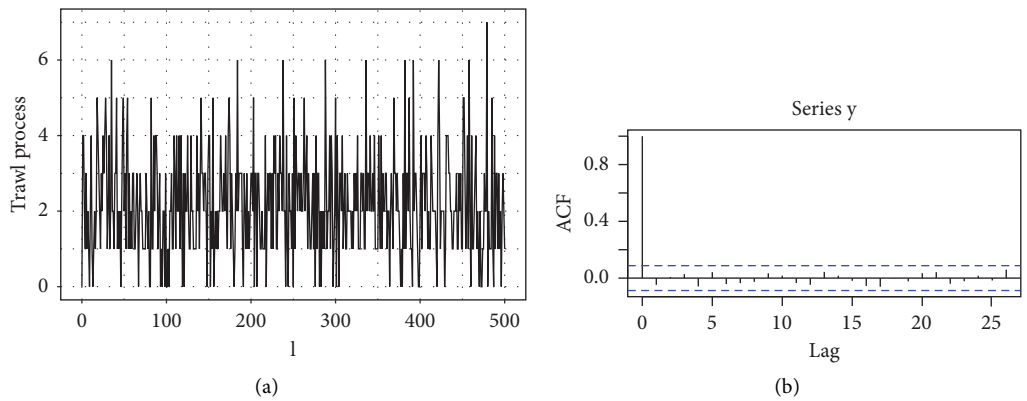

Figure 4: Semi-Poisson Lévy-based time series and autocorrelation function plots of simulated data with an exponential shape for the short memory process ($N = 500, \lambda = 1.5, L = 2$). (a) Time series. (b) ACF.

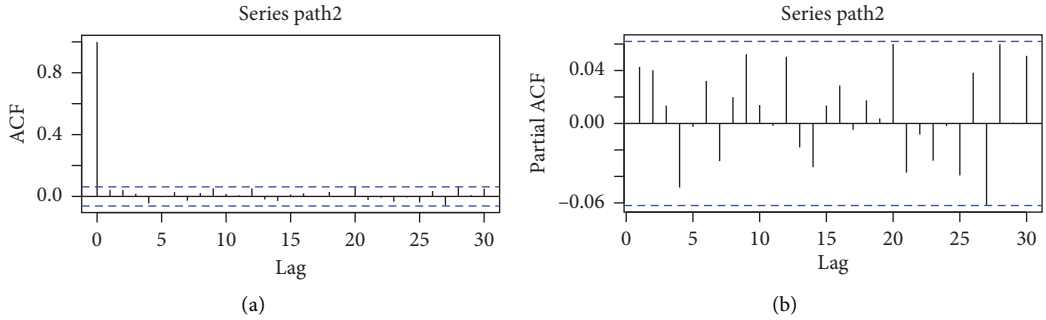

Figure 5: Autocorrelation and partial autocorrelation function plots of simulated data from semi-Poisson Lévy-based with a long memory process where $(N = 1000, L = 2, \delta = 0.6, \gamma = 1)$. (a) ACF. (b) PACF.

Table 3: Summary statistics for the PL estimator for different parameter values $\theta = (\lambda, L)$, using an exponential kernel (representing short-range dependence) at various sample lengths of the series N for a semi-Poisson Lévy-based process.

N	True value	λ 5	L 2	λ 4	L 0.7	λ 9	L 0.3	λ 15	L 0.05
200	Mean	4.9292	1.9996	3.9292	0.6996	8.9292	0.2996	14.9292	0.0496
	Bias	0.0151	0.0097	0.0101	0.0032	0.0356	−0.0010	0.0656	−0.0097
	MSE	0.0458	0.0190	0.0205	0.0021	0.2541	0.0002	0.8618	0.0190
N	True value	5	2	4	0.7	9	0.3	15	0.05
500	Mean	4.9552	1.9251	3.9552	0.6999	8.9552	0.2999	14.9552	0.0499
	Bias	0.0061	0.0038	0.0041	0.0012	0.0143	−0.0004	0.0261	−0.0019
	MSE	0.0186	0.0076	0.0084	0.0008	0.1023	0.0000	0.3408	0.0018
N	True value	5	2	4	0.7	9	0.3	15	0.05
1000	Mean	4.9683	1.9999	3.9683	0.6999	8.9983	0.2999	14.9683	0.0499
	Bias	0.0030	0.0019	0.0021	0.0001	0.0071	−0.0002	0.0131	−0.0004
	MSE	0.0094	0.0038	0.0047	0.0000	0.0513	0.0000	0.1734	0.0002

Table 4: Summary statistics for the PL estimator for different parameter values $\theta = (\delta, \gamma, L)$ and different sample lengths of the series N with long-range dependence semi-Poisson Lévy-based process.

N	True value	L 2	δ 0.8	γ 0.2	L 3	δ 1.6	γ 0.6	L 1.3	δ 3	γ 0.3
100	Mean	1.8161	0.8414	0.1329	3.1862	1.2904	0.6178	1.3012	2.8738	0.2210
	Bias	0.1838	−0.0414	0.0671	−0.1862	0.3096	−0.0178	−0.0012	0.1261	0.3
	MSE	0.0654	0.0209	0.0574	0.0346	0.0958	0.0003	0.0280	0.0359	0.09
250	Mean	1.9398	0.8203	0.2081	3.0856	1.2957	0.5884	1.2911	2.973	0.3410
	Bias	0.0601	−0.0203	−0.0081	−0.0855	0.3043	0.0115	−0.0000	0.0269	0.0003
	MSE	0.0077	0.0140	0.0044	0.0283	0.0104	0.0077	0.0000	0.0209	0.0000
500	Mean	2.0032	0.8111	0.2029	3.0065	1.5594	0.6222	1.3001	3.0001	0.3041
	Bias	0.0223	0.0102	0.0011	−0.0064	0.0004	0.0081	0.0000	0.0000	0.0000
	MSE	0.0049	0.0079	0.0021	0.0038	0.0058	0.0031	0.0000	0.0000	0.0000
N	True value	1.5	1	0.6	1	0.4	0.2	3	0.6	0.9
30	Mean	1.6158	0.8636	0.4518	0.9091	0.6454	0.1524	3.3537	0.5939	0.7143
	Bias	−0.1158	0.1363	0.1482	0.0908	−0.2454	0.0475	−0.3536	0.0060	0.1856
	MSE	0.0383	0.0478	0.0541	0.0354	0.0751	0.0448	0.1274	0.0072	0.0561
500	Mean	1.5808	0.8860	0.6172	0.9671	0.3774	0.1899	3.0793	0.6211	0.8059
	Bias	−0.0807	0.1139	−0.0172	0.0328	0.0225	−0.0022	−0.0792	−0.0211	0.0740
	MSE	0.0065	0.0380	0.0241	0.0010	0.0005	0.0000	0.0064	0.0124	0.0160
1000	Mean	1.5025	0.9933	0.6002	0.9953	0.4029	0.2001	3.0431	0.6065	0.9026
	Bias	−0.0025	0.0766	0.0053	−0.0000	0.0000	0.0000	−0.0431	−0.0056	0.0167
	MSE	0.0000	0.0058	0.0000	0.0000	0.0000	0.0000	0.0027	0.0021	0.0042

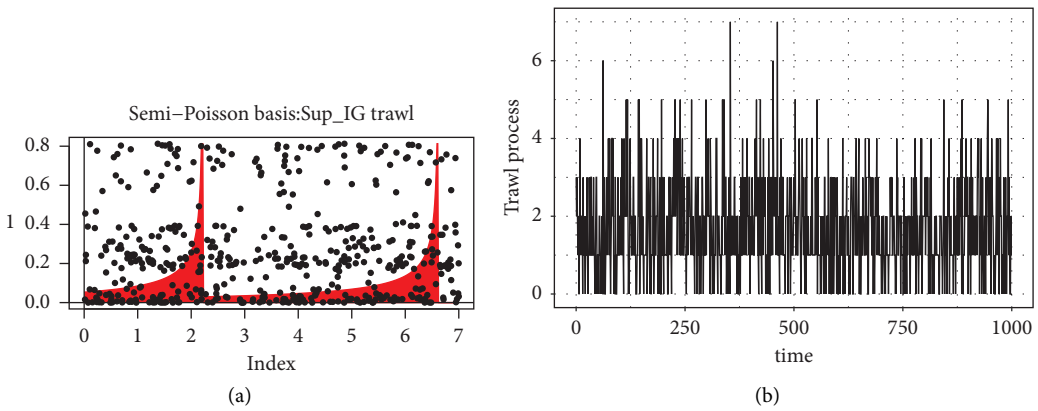

Figure 6: Semi-Poisson Lévy basis and semi-Poisson Lévy-based time series plots of simulated data with a long memory process where $(N = 1000, L = 2, \delta = 0.6, \gamma = 1)$. (a) Illustration of the semi-Poisson processes. (b) Time series plot.

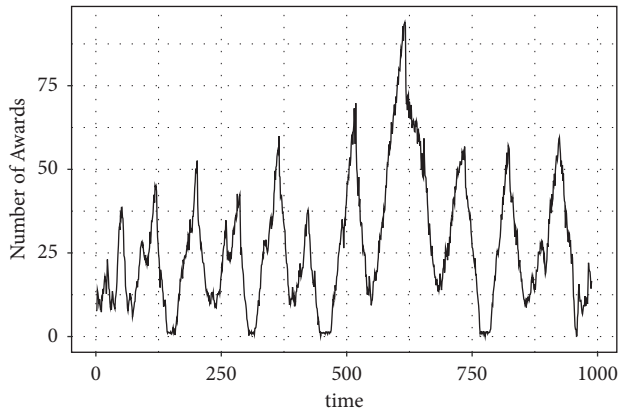

Figure 7: The number of NSF funding awarded to academic institutions.

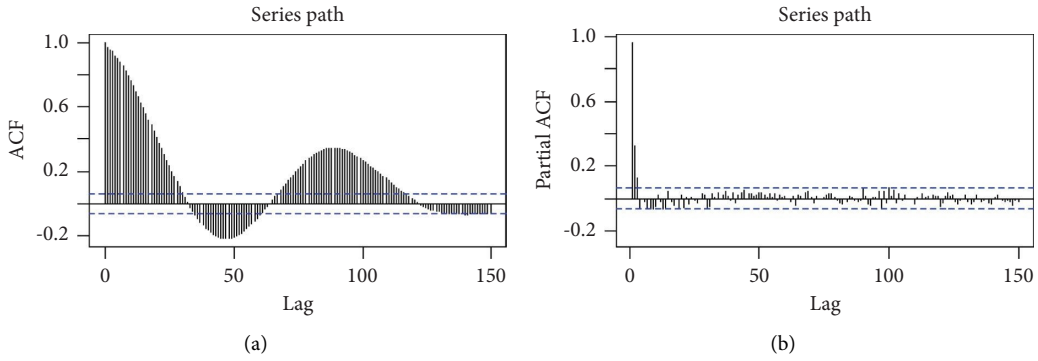

Figure 8: Autocorrelation and partial autocorrelation functions of the data.

Table 5: Comparison of Lévy-based models with GPIG-INAR (1).

Model	Mean	Variance	AIC	RMSE
Empirical	14.5208	251.1662	—	—
GPIG-INAR (1)	14.5383	292.5875	1033.1376	1.6462
PIG-Lévy-based with an exponential kernel	14.5634	149.8246	616.9301	1.2065
PIG-Lévy-based with sup-IG kernel	15.1953	168.0466	1038.2727	1.7370

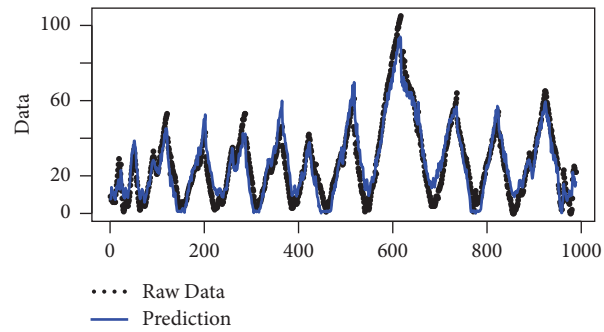

Figure 9: Prediction of the numbers of NSF fundings under the PIG-Lévy-based model with an exponential kernel.

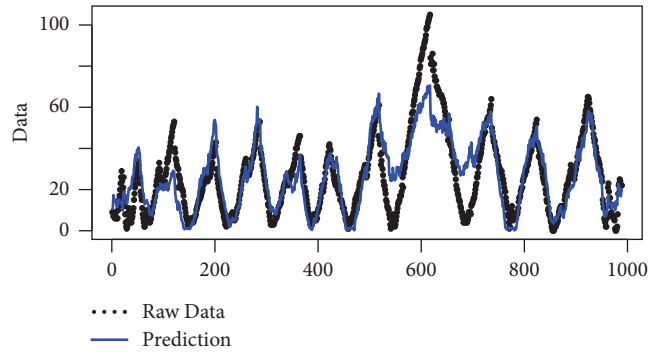

Figure 10: Prediction of the numbers of NSF fundings under the PIG-Lévy-based model with a sup-IG kernel set.

Table 6: Parameter estimates for the PIG-Lévy-based model with an exponential kernel after fitting to NSF data.

Parameters	λ	μ	σ
Estimate values	0.6219	1.0351	2.5396
Standard error	0.0037	0.0147	0.0840

Table 7: Parameter estimates for the PIG-Lévy-based model with a sup-IG kernel.

Parameters	γ	μ	σ	δ
Estimate values	0.6503	0.7863	0.8848	0.7225
Standard error	0.0502	0.0806	0.0734	0.0550

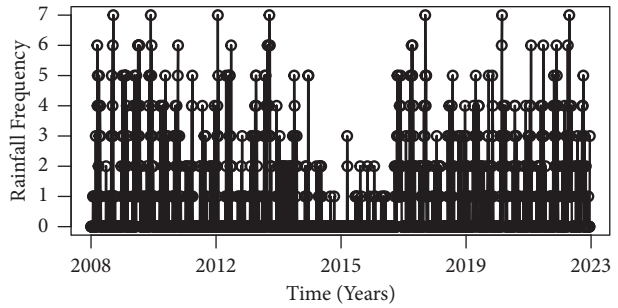

Figure 11: Weekly rainfall frequencies for N'zerekore region.

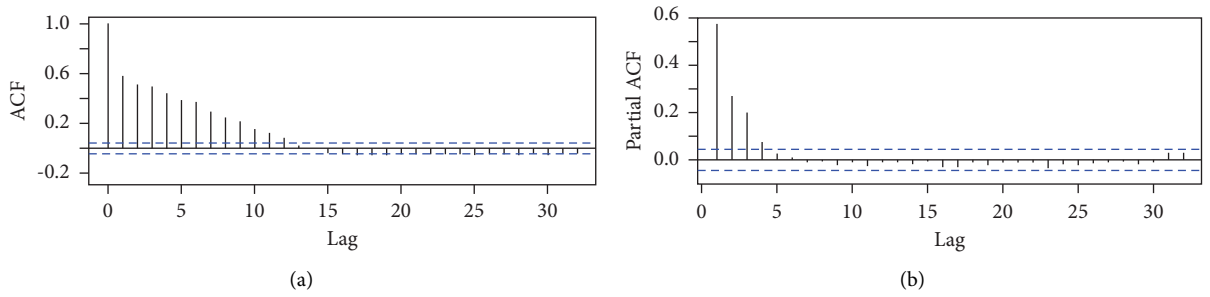

Figure 12: Autocorrelation and partial autocorrelation functions of the data.

Table 8: Parameter estimates for the semi-Poisson Lévy-based model.

Parameters	L	λ
Estimate values	1.4254	0.0231
Standard error	0.00016	0.00019

Table 9: Data statistics.

	Min	1st Qu.	Median	3rd Qu.	Max	Var	Mean
Model	0	0	0.318	1.431	6.711	1.869	0.925
Actual data	0	0	0	1	7	1.958	0.896

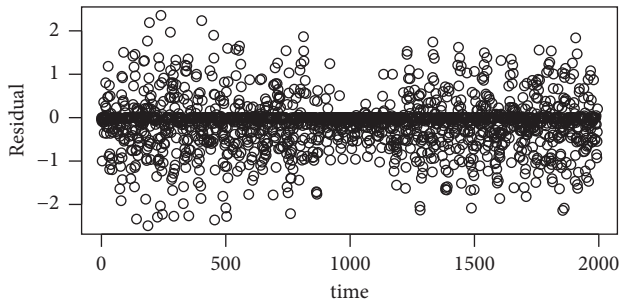

Figure 13: Residuals of the model.

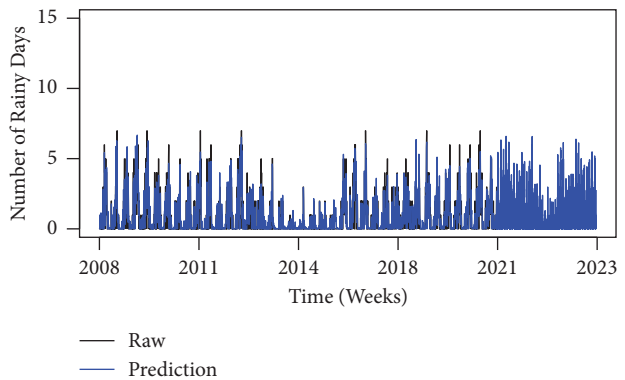

Figure 14: Data and model prediction.

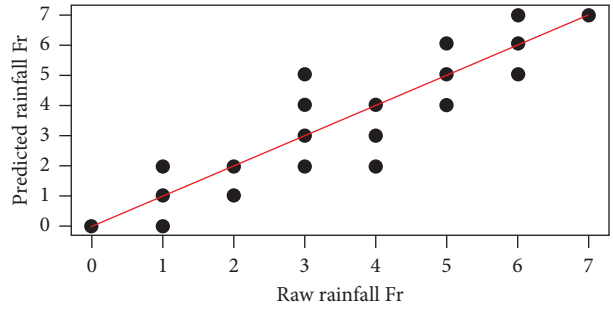

Figure 15: Accuracy of the model.

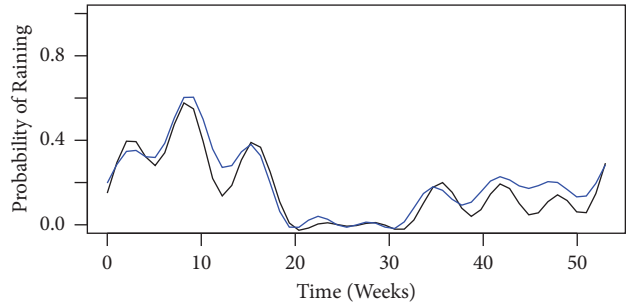

Figure 16: Probability of frequent rainfall: blue represents the model, and black represents the data.

The 90%, 95%, and 99% confidence intervals were found as $R^2 \pm 0.038$, $R^2 \pm 0.045$, and $R^2 \pm 0.046$, meaning that R^2 is much closer to the value of 82.3%.

7. Summary and Conclusions

In conclusion, our exploration of the Lévy-based modeling framework for time series analysis of count data has revealed its remarkable flexibility and versatility. The developing time series model offers a powerful tool for simultaneously capturing diverse characteristics and features in count time series, including complex dependence structures, and critical aspects such as heavy-tailedness, overdispersion, and zero inflation. Our study has also emphasized the importance of achieving realism in modeling by incorporating Lévy basis which is infinitely divisible marginal distributions and the kernel set for dependence modeling. Moreover, we have underscored the significance of stationary and homogeneous Lévy bases to ensure statistical consistency across time and space. Our simulations and real-data applications have demonstrated this approach's practical relevance and potential advantages and flexibility. There are potential directions for future research. One compelling direction is the extension of Lévy-based models to multivariate settings, where higher-order dependencies can be effectively addressed. This can allow for comprehensive modeling of both short-term and long-term serial correlation structures. Finally, theoretical advances in comprehending these models' features and limitations will contribute to a better understanding of their applicability and resilience. In conclusion, our findings highlight the promise of the Lévy-based paradigm and motivate further research into its application to count data analysis.

Data Availability

The data supporting the current study are available from the corresponding author upon request.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

The authors express their appreciation to the African Union, as well as to anonymous reviewers and editors, for their support in funding the research and for providing valuable feedback that contributed to the refinement of the paper in its current state. This work was funded by the Pan African University Institute for Basic Sciences, Technology, and Innovation.

Supplementary Materials

Supplementary materials include derivations and R-codes for reference. (*Supplementary Materials*)

References

[1] K. Tawiah, W. A. Iddrisu, and K. Asampana Asosega, "Zero-inflated time series modelling of covid-19 deaths in Ghana," *Journal of Environmental and Public health*, vol. 2021, Article ID 5543977, 9 pages, 2021.

[2] S. M. Idrees, M. A. Alam, and P. Agarwal, "A prediction approach for stock market volatility based on time series data," *IEEE Access*, vol. 7, pp. 17287–17298, 2019.

[3] M. A. Quddus, "Time series count data models: an empirical application to traffic accidents," *Accident Analysis and Prevention*, vol. 40, no. 5, pp. 1732–1741, 2008.

[4] G. Saridakis, "Violent crime in the United States of America: a time-series analysis between 1960–2000," *European Journal of Law and Economics*, vol. 18, no. 2, pp. 203–221, 2004.
[5] P. A. Jacobs and P. A. Lewis, "Discrete time series generated by mixtures. iii. autoregressive processes (dar (p))," Technical report, Naval Postgraduate School, Monterey CA, USA, 1978.
[6] E. McKenzie, "Some simple models for discrete variate time series 1," *JAWRA Journal of the American Water Resources Association*, vol. 21, no. 4, pp. 645–650, 1985.
[7] M. A. Al-Osh and A. A. Alzaid, "First-order integer-valued autoregressive (inar (1)) process," *Journal of Time Series Analysis*, vol. 8, no. 3, pp. 261–275, 1987.
[8] P. J. McKenzie, "A model of information practices in accounts of everyday-life information seeking," *Journal of Documentation*, vol. 59, no. 1, pp. 19–40, 2003.
[9] O. E. Barndorff-Nielsen, A. Lunde, N. Shephard, and A. E. Veraart, "Integer-valued trawl processes: a class of stationary infinitely divisible processes," *Scandinavian Journal of Statistics*, vol. 41, no. 3, pp. 693–724, 2014.
[10] L. Qian, Q. Li, and F. Zhu, "Modelling heavy-tailedness in count time series," *Applied Mathematical Modelling*, vol. 82, pp. 766–784, 2020.
[11] C. Velasco-Gallego and I. Lazakis, "Radis: a real-time anomaly detection intelligent system for fault diagnosis of marine machinery," *Expert Systems with Applications*, vol. 204, Article ID 117634, 2022.
[12] J. Liu, J. Huang, Y. Zhou et al., "From distributed machine learning to federated learning: a survey," *Knowledge and Information Systems*, vol. 64, no. 4, pp. 885–917, 2022.
[13] C. Schölzel and P. Friederichs, "Multivariate non-normally distributed random variables in climate research–introduction to the copula approach," *Nonlinear Processes in Geophysics*, vol. 15, no. 5, pp. 761–772, 2008.
[14] L. Qian, B. Zhou, and H. T. Yang, "Cardiomyocyte proliferation and reprogramming for cardiac regeneration," *Journal of Molecular and Cellular Cardiology*, vol. 179, pp. 1–24, 2023.
[15] R. Zhu and H. Joe, "Modelling heavy-tailed count data using a generalised Poisson-inverse Gaussian family," *Statistics and Probability Letters*, vol. 79, no. 15, pp. 1695–1703, 2009.
[16] O. E. Barndorff-Nielsen and J. Schmiegel, "Lévy-based spatial-temporal modelling, with applications to turbulence," *Russian Mathematical Surveys*, vol. 59, no. 1, pp. 65–90, 2004.
[17] A. E. Veraart, "Modeling, simulation and inference for multivariate time series of counts using trawl processes," *Journal of Multivariate Analysis*, vol. 169, pp. 110–129, 2019.
[18] M. Bennedsen, A. Lunde, N. Shephard, and A. E. Veraart, "Inference and forecasting for continuous-time integer-valued trawl processes," 2021, https://arxiv.org/abs/2107.03674.
[19] D. Leonte and A. E. Veraart, "Simulation methods and error analysis for trawl processes and ambit fields," 2022, https://arxiv.org/abs/2208.08784.
[20] B. S. Rajput and J. Rosinski, "Spectral representations of infinitely divisible processes," *Probability Theory and Related Fields*, vol. 82, no. 3, pp. 451–487, 1989.
[21] S. Kwapien and W. A. Woyczynski, *Random Series and Stochastic Integrals: Single and Multiple*, Birkhäuser, Boston, MA, USA, 1992.
[22] S. Ken-Iti, *Lévy Processes and Infinitely Divisible Distributions*, Cambridge University Press, Cambridge, UK, 1999.
[23] C. Varin and P. Vidoni, "Pairwise likelihood inference for general state space models," *Econometric Reviews*, vol. 28, no. 1-3, pp. 170–185, 2008.
[24] B. G. Lindsay, "Composite likelihood methods," *Comtemporary Mathematics*, vol. 80, no. 1, pp. 221–239, 1988.
[25] C. Varin, N. Reid, and D. Firth, "An overview of composite likelihood methods," *Statistica Sinica*, vol. 21, pp. 5–42, 2011.
[26] R. A. Davis and C. Y. Yau, "Comments on pairwise likelihood in time series models," *Statistica Sinica*, vol. 21, pp. 255–277, 2011.
[27] M. Bennedsen, A. Lunde, N. Shephard, and A. E. Veraart, *Estimation of Integer-Valued Trawl Processes*, 2017.
[28] P. Chidzalo, P. O. Ngare, and J. K. Mung'atu, "Trivariate stochastic weather model for predicting maize yield," *Journal of Applied Mathematics*, vol. 2022, Article ID 3633658, 32 pages, 2022.