Article

Modeling the Characteristics of Unhealthy Air Pollution Events: A Copula Approach

Nurulkamal Masseran

Department of Mathematical Sciences, Faculty of Science and Technology, Universiti Kebangsaan Malaysia, UKM, Bangi 43600, Selangor, Malaysia; kamalmsn@ukm.edu.my; Tel.: +60-3-8921-3424

Abstract: This study proposes the concept of duration (D) and severity (S) measures, which were derived from unhealthy air pollution events. In parallel with that, the application of a copula model is proposed to evaluate unhealthy air pollution events with respect to their duration and severity characteristics. The bivariate criteria represented by duration and severity indicate their structural dependency, long-tail, and non-identically marginal distributions. A copula approach can provide a good statistical tool to deal with these issues and enable the extraction of valuable information from air pollution data. Based on the copula model, several statistical measurements are proposed for describing the characteristics of unhealthy air pollution events, including the Kendall's τ correlation of the copula, the conditional probability of air pollution severity based on a given duration, the joint OR/AND return period, and the conditional D|S and conditional S|D return periods. A case study based on air pollution data indices was conducted in Klang, Malaysia. The results indicate that a copula approach is beneficial for deriving valuable information for planning and mitigating the risks of unhealthy air pollution events.

Keywords: air pollution characteristics; pollution risk assessment; statistical modeling

1. Introduction

Air pollution is an important issue that needs to be addressed worldwide, particularly in urban areas. This issue relates to the condition of the air environment corresponding to unhealthy air pollution events [1,2]. Unhealthy air pollution events can be characterized based on their duration and severity. The duration of an air pollution event is determined based on the period in which the air pollution index (API) values are harmful to human health for consecutive periods [3], while severity measures the magnitude of the air pollution event based on the cumulative effect of an unhealthy API during a particular air pollution event. These measures are related because the severity of an air pollution event always depends on its duration. Separate analyses of the duration and severity of an air pollution event cannot reveal their significant association. To model their relationship, joint modeling is required, in which the marginal distributions of both variables are combined [4,5]. Several different bivariate distributions have been proposed by previous researchers in order to model the joint characteristics of bivariate variables. For example, bivariate normal distribution has been applied in many research areas, particularly for the cases where each considered variable can be easily described by a Gaussian/normal distribution [6–9]. In a similar vein, bivariate lognormal distribution has been employed to describe a phenomenon that can be represented by a lognormal-distributed marginal [10,11]. The Gumbel logistic model [12,13] and the Gumbel mixed model [14,15] were proposed to describe a phenomenon that can be represented by random variables corresponding to a Gumbel-distributed marginal. A bivariate exponential model [16,17] and a Nagao–Kadoya bivariate exponential model [7] have been proposed to model a phenomenon where their marginal random variables can be represented by an exponential model [18]. Apart from that, there are various forms of bivariate gamma models, such as Royen's Bivariate Gamma model [19], the Izawa Bigamma model [20], the Moran model [21], Schmeiser and Lal's

Int. J. Environ. Res. Public Health **2021**, 18, 8751

2 of 18

Bivariate Gamma model [22], and the Loaciga and Leipnik Bivariate Gamma model [23], have been proposed to model the bivariate random variable with a combination of marginal Gamma distribution [24–26]. However, if the marginal distribution among the variables is non-identically distributed, obtaining the analytical or closed form of the joint probability model is difficult [27], which makes it challenging to use the available bivariate models to describe the statistical properties of the data. Fortunately, this issue can be resolved using the copula approach [28,29].

Copula functions have the advantage of providing a platform for representing a multi-variate distribution for any type or form of marginal variable distribution involved [30,31]. Several researchers have employed the copula model to analyze the air pollution data. Sak et al. [32], for instance, used a copula model to quantify the pollution risk of $PM_{2.5}$ concentration in several cities in China. Chan and So [33] used a copula approach to model extreme spatial air pollution events in Guangdong province, China. Falk et al. [34] used a copula approach to evaluate the joint exceedance probabilities for air pollutants in Milan, Italy. Kim et al. [35] characterized the dependence between several cities in China based on a measure of directional dependence estimated from the $PM_{2.5}$ pollutant variable. Masseran and Hussain [36] modelled and visualized the dependency fluctuation among the air pollution variables in Malaysia using a dynamic Copula approach. He at al. [37] employed a copula model to investigate the time-varying correlations between meteorological factors and atmospheric pollutants in the cities of Beijing and Guangzhou, China. However, most of the available literature focuses on the copula analysis of real values or the magnitude of air pollution events. Thus, this study attempts look at a different perspective by attempting to investigate another important aspect of air pollution events, which is characterized by their duration and severity size. In summary, all of the previous studies found that the copula approach could provide good modeling flexibility for the multivariate distribution of random variables. However, when the dimensionality of the dataset increases, a single copula model parameter is unreliable to describe a simultaneous relationship among the variables [28,29]. To overcome this problem, a vine-copula approach needs to be adopted [38,39]. Fortunately, for the case of bivariate data analysis, the problem of dimensional complexity does not occur [40,41].

A bivariate copula of air pollution severity and duration can effectively reveal the significant relationship between these characteristics. In addition, the copula model can be used to measure the conditional probability and return period of air pollution episodes, which are essential for predicting air pollution events and mitigating their impacts. This information can be beneficial to environmental management authorities in planning for and measuring the risks associated with unhealthy air pollution events. The rest of this paper is organized as follows: Section 2 describes the study area and data. Section 3 provides a description of copula modeling on air pollution characteristics. A variety of copula models are presented in Section 4. In Section 5, an application of flexible parameter estimation and model selection are proposed based on pseudo maximum likelihood approach. In Section 6, the performance of each fitted copula model on the characteristics of unhealthy air pollution data are evaluated. Apart from that, Section 6 also presents several valuable statistical measures for assessing air pollution risk, including (i) the Kendall's τ correlation of the copula to measure the dependency between the duration and severity of an air pollution event, (ii) the conditional probability of a certain air pollution severity given the air pollution duration, (iii) the joint OR/AND return period, and (iv) the conditional D|S and S|D return periods. Finally, in Section 7, the conclusions about the overall findings are provided.

2. Study Area and Data

This study investigates the API data in the area of Klang, Peninsular Malaysia, which is located at latitude $101°26'44.023$ E and longitude $3°2'41.701$ N. Klang is one of the largest cities in Malaysia, with a land area of approximately 573 km^2. This city has a dense population and is very actively involved in a range of important economic and industrial

activities, particularly with the import and export trade. Klang has also been recognized as the 13th busiest trans-shipment port and the 16th busiest container port in the world [42]. Although many advantages are associated with its rapid development, economic activity, and dense population, Klang is prone to poor air quality, which makes it an ideal case for the investigation and continuous evaluation of the Malaysia's air quality [43]. Figure 1 shows maps of Klang and Peninsular Malaysia [44].
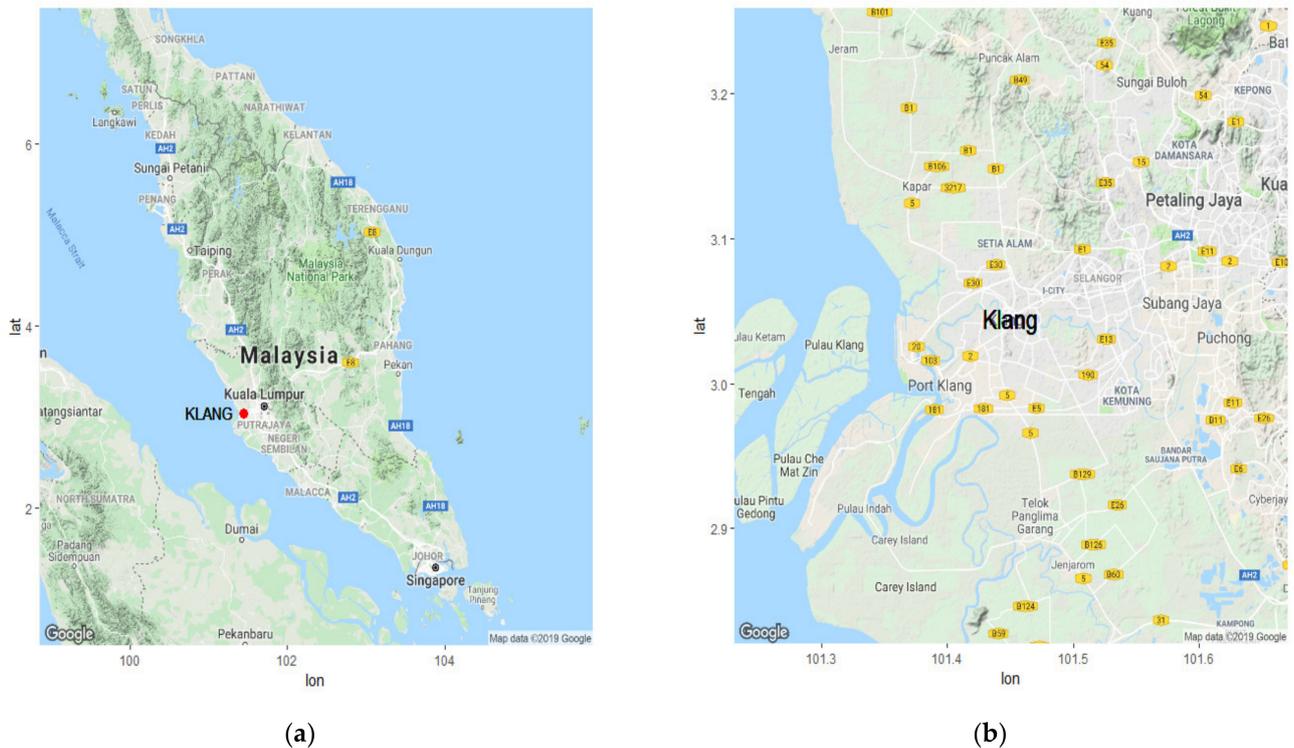

(**a**)　　　　　　　　　　　　　　　　　　　　　(**b**)

Figure 1. Maps of (**a**) Peninsular Malaysia (Klang location identified by red dot) and (**b**) the city of Klang.

In Malaysia, the responsibility for collecting, supervising, and reporting API data is held by the Department of Environment (DOE), Malaysia. As a case study, hourly API data in Klang for the period of 1 January 1997 to 31 August 2020 was used in the analysis. A total of five major pollutant variables are recorded hourly, including carbon monoxide (CO), ozone (O_3), nitrogen dioxide (NO_2), sulfur dioxide (SO_2), and particulate matter less than 10 microns in size (PM_{10}). The observed data for O_3, CO, NO_2, and SO_2, have been measured in terms of the parts per million (ppm) unit mass of a contaminant, while the observed PM_{10} data have been measured in terms of micrograms per cubic meter ($\mu g/m^3$). Thus, to measure the API indices, these pollutant variables need to be standardized to derive individual indices. Based on each individual indices, their values can then be integrated based on the highest sub-indices to determine the API indices at particular times [45]. A detailed calculation on the standardization and determination of the sub-indices values for each pollutant variable can be referred to in Masseran and Safari [46]. Figure 2 illustrate the schematic process for determining the API value based on five pollutant variables [47].
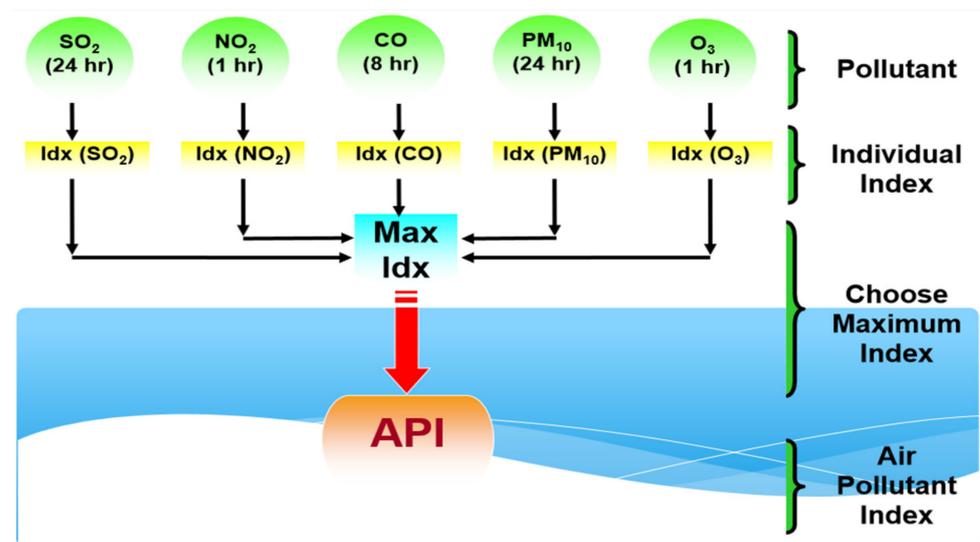
Figure 2. The process of determining the API value.

Based on the API index values, the DOE classifies API values higher than 100 as unhealthy air pollution events [48], and consecutive API values greater than 100 in a given period indicate the duration of an air pollution event. Assuming that the random variable D_i represents the duration of any pollution event, then $i = 1, 2, 3, \ldots, n$. Mathematically, the duration of an air pollution event can be determined as follows:

$$D_i = \sum_{j=1}^{N} I_i\big(API_j\big), \ for \ i = 1, 2, 3, \ldots, n, \tag{1}$$

where N is the total number of observations, and $I_i\big(API_j\big)$ is an indicator function, which is determined as follows:

$$I_i\big(API_j\big) = \begin{cases} 1, \ if \ API_j > 100, \\ 0, \ if \ API_j \leq 100. \end{cases} \tag{2}$$

For any pollution event of a certain duration, the severity of that event can be determined based on the cumulative API values obtained during that period using the following equation:

$$S_i = \sum_{j=1}^{D_i} API_j, \tag{3}$$

where S_i is a random variable representing the severity of a pollution event, $i = 1, 2, 3, \ldots, n$. Figure 3 shows a graphical representation of the process used to determine the severity and duration of an air pollution event.
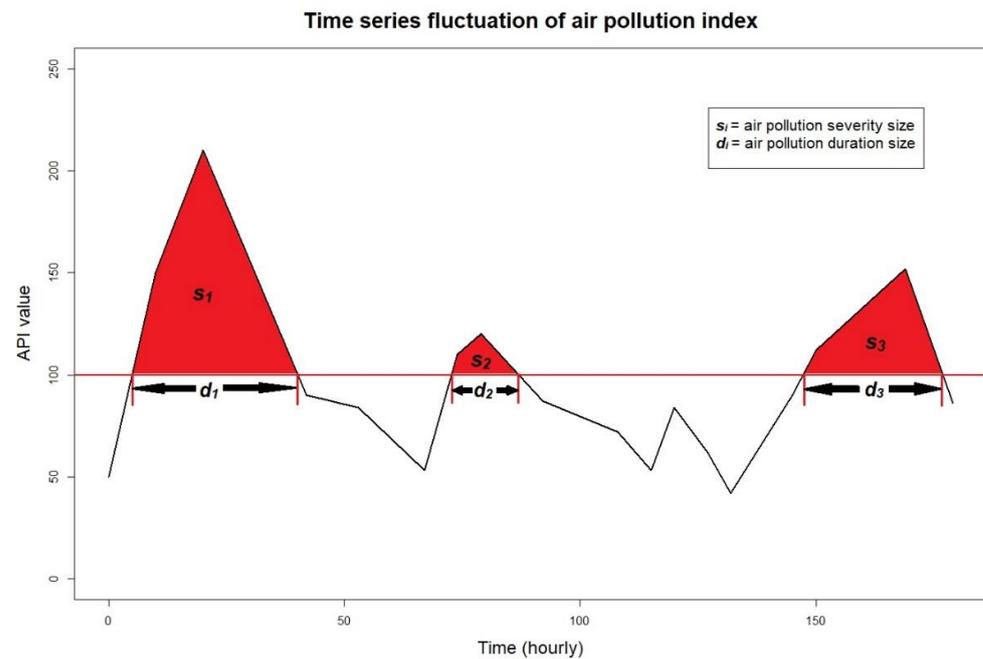
Figure 3. Time series plot corresponding to unhealthy air pollution event threshold.

3. Copula Description of Air Pollution Characteristics

Sklar's theorem [49] provides an important foundation for the theory and application of the copula model. Copulas play a significant role as mapping functions that enable the combination of uniformly distributed marginal models. Specifically, copulas represent a joint distribution and describe the dependence structure of any arbitrarily distributed dependent variables [50]. Based on Sklar's theorem, a copula function can be constructed using the following equation:

$$H(x_1, x_2, \ldots, x_2) = C[F_1(x_1), F_2(x_2), \ldots, F_n(x_n)] = C(u_1, u_2, \ldots, u_n) \tag{4}$$

where $H(x_1, x_2, \ldots, x_2)$ denotes an *n*-dimensional distribution function, $F_i(x_i) = u_i$ is a specified univariate marginal distribution function for random variable X_i with $U \sim uniform(0, 1)$, and C denotes the copula function [29]. In the context of the bivariate model used in air pollution studies, assuming that the random variables $X_1 = D$ and $X_2 = S$ represent the duration and severity of an air pollution event, respectively, the copula model described in Equation (4) can be simplified as follows:

$$H(d, s) = C[F_D(d), F_S(s)] = C(u_1, u_2). \tag{5}$$

Equation (5) defines a two-dimensional copula as a mapping function of $[0, 1]^2 \rightarrow [0, 1]$, which implies that a bivariate distribution function can be defined in $[0, 1]^2$ with the standard univariate margins u_1 and u_2 [31]. Thus, the bivariate density function of a copula can be obtained as follows:

$$c(u_1, u_2) = \frac{\partial^2 C(u_1, u_2)}{\partial u_1 \partial u_2}. \tag{6}$$

Based on the copula model, the dependency between the pollutant duration and severity can be determined using Kendall's τ correlation. This correlation provides a scale-invariant measure for determining this dependency based on the concept of concordance [51]. As mentioned by Klein et al. [52], Kendall's τ has good resistance to outliers

and provides a wider class of dependencies. Generally, the Kendall's τ correlation can be related to the copula parameter, θ, using the following equation:

$$\tau = 4 \iint\limits_{[0,1]} C(u,v;\theta)c(u_1,u_2;\theta)du_1 du_2 - 1. \tag{7}$$

Another important piece of information that can be determined is the conditional probability of the air pollution severity given a certain air pollution duration. This information can benefit authorities who are tasked with evaluating the patterns and trends of air pollution behavior, particularly during critical air pollution events. Mathematically, this conditional probability can be computed using the following formula [40]:

$$P(S \leq s | D \geq d\prime) = \frac{F_S(s) - C(F_D(d\prime), F_S(s))}{1 - F_D(d\prime)} = \frac{u_2 - C(u_1', u_2)}{1 - u_1'}. \tag{8}$$

The copula model can also be used to estimate the return periods of air pollution characteristics. An understanding of the information provided by the return period can serve as a basis for planning and developing monitoring systems to reduce the risk associated with extreme air pollution events. The return period provides insight into the average recurrence interval between unhealthy air pollution events. This information could be beneficial for evaluating the patterns and trends of air pollution behaviors, particularly during critical air pollution events. According to Shiau [40], the copula model is useful for providing information about the joint and conditional return periods of a bivariate event. The joint return period for $D \geq d$ or $S \geq s$ can be determined as follows:

$$T'_{DS} = \frac{E(L)}{P(D \geq d \cup S \geq s)} = \frac{E(L)}{1 - F_{DS}(d,s)} = \frac{E(L)}{1 - C(F_D(d), F_S(s))}, \tag{9}$$

where $E(L)$ is the expected inter-arrival time of an air pollution event. Likewise, the joint return period for $D \geq d$ and $S \geq s$ can be determined as follows:

$$\begin{aligned} T_{DS} &= \frac{E(L)}{P(D \geq d \cap S \geq s)} = \frac{E(L)}{1 - F_D(d) - F_S(s) + F_{DS}(d,s)} \\ &= \frac{E(L)}{1 - F_D(d) - F_S(s) + C(F_D(d), F_S(s))}. \end{aligned} \tag{10}$$

On the other hand, the conditional return period for D, given $S \geq s$, can be determined as follows:

$$\begin{aligned} T_{D|S \geq s} &= \frac{E(L)}{1 - F_S(s)} \times \frac{1}{1 - F_D(d) - F_S(s) + F_{DS}(d,s)} \\ &= \frac{1}{[1 - F_S(s)][1 - F_D(d) - F_S(s) + C(F_D(d), F_S(s))]}, \end{aligned} \tag{11}$$

Additionally, the conditional return period for S given $D \geq d$ can be determined as follows:

$$\begin{aligned} T_{S|D \geq d} &= \frac{E(L)}{1 - F_D(d)} \times \frac{1}{1 - F_D(d) - F_S(s) + F_{DS}(d,s)} \\ &= \frac{1}{[1 - F_D(d)][1 - F_D(d) - F_S(s) + C(F_D(d), F_S(s))]}. \end{aligned} \tag{12}$$

4. Copula Models

Various copula models are available in the literature for modeling the relationship between the duration–severity variables. However, to obtain an accurate result, comparing several copula models is important to find the one that provides the best representation of the dataset.

4.1. Clayton Copula

The Clayton copula model is useful for capturing the positive dependence of bivariate variables in which the strength of the dependency is dictated by the parameter $0 \leq \theta < \infty$. With a particular rotation, this copula is appropriate for modeling either a positive or

negative dependence structure that exists in a dataset [53]. The bivariate joint cumulative distribution function (CDF) and probability density function (PDF) of a Clayton copula are given as:

$$C(u_1, u_2) = \left(u_1^{-\theta} + u_2^{-\theta} - 1 \right)^{-\frac{1}{\theta}}, \tag{13}$$

$$c(u_1, u_2) = (\theta + 1) \left(u_1^{-\theta} + u_2^{-\theta} - 1 \right)^{-\left(\frac{1+2\theta}{\theta} \right)} (u_1 u_2)^{-\theta - 1}. \tag{14}$$

The Kendall's τ correlation determined by the Clayton copula is given as:

$$\tau = \frac{\theta}{\theta + 2} \tag{15}$$

4.2. Ali–Mikhail–Haq (AMH) Copula

The AMH copula is suitable for datasets in which there is a weaker dependence among the variables [53]. However, this copula is always used for comparison with other copulas in terms of model fitting. The bivariate joint CDF and PDF of the AMH copula are given as:

$$C(u_1, u_2) = \frac{u_1 u_2}{1 - \theta(1 - u_1)(1 - u_2)}, \tag{16}$$

$$c(u_1, u_2) = \frac{[1 - \theta(1 - u_1)(1 - u_2)] + 2\theta u_1 u_2}{[1 - \theta(1 - u_1)(1 - u_2)]^3}, \tag{17}$$

With the parameter space $[-1,1)$, the Kendall's τ correlation determined by the AMH copula is given as:

$$\tau = \left[\frac{3\theta - 2}{3\theta} \right] - \left[\frac{2(1 - \theta)^2}{3\theta^2} \right] [\ln(1 - \theta)]. \tag{18}$$

4.3. Frank Copula

The Frank copula provides a versatile dependency measure because it can accommodate the entire range of dependencies, $\tau\theta \in [1, -1]$ [41]. Thus, with the correct rotation, the Frank copula is an appropriate model for describing either a positive or negative dependence structure in the dataset [54]. The bivariate joint CDF and PDF of the Frank copula are given as:

$$C(u_1, u_2) = -\frac{1}{\theta} \ln \left[1 + \frac{\left(e^{-\theta u_1} - 1 \right) \left(e^{-\theta u_2} - 1 \right)}{e^{-\theta} - 1} \right], \tag{19}$$

$$c(u_1, u_2) = \frac{-\theta e^{-\theta(u_1 + u_2)} \left(e^{-\theta} - 1 \right)}{\left[e^{-\theta(u_1 + u_2)} - e^{-\theta u_1} - e^{-\theta u_2} + e^{-\theta} \right]^2}, \tag{20}$$

The Kendall's τ correlation determined by the Frank copula is given as:

$$\tau = 1 + \frac{4[D_1(\theta) - 1]}{\theta}, \tag{21}$$

where $D_1(\theta) = \frac{\int_0^\theta \frac{t^k}{e^{(t-1)}} dt}{\theta}$ is a Debye function [50].

Int. J. Environ. Res. Public Health **2021**, *18*, 8751

8 of 18

4.4. Plackett Copula

The Plackett copula is also appropriate for modeling either the positive or negative dependence structures in a dataset [53]. The bivariate joint CDF and PDF of the Plackett copula are given as:

$$C(u_1, u_2) = \frac{\left\{ 1 + (\theta - 1)(u_1 + u_2) - \left[(1 + (\theta - 1)(u_1 + u_2))^2 - 4\theta(\theta - 1)(u_1 u_2) \right]^{\frac{1}{2}} \right\}}{2(\theta - 1)}, \tag{22}$$

$$c(u_1, u_2) = \left[(1 + (\theta - 1)(u_1 + u_2))^2 - 4\theta(\theta - 1)(u_1 u_2) \right]^{-\frac{3}{2}} \theta [1 + (\theta - 1)(u_1 + u_2 - 2u_1 u_2)] \tag{23}$$

with the parameter space $(0, \infty)$. However, the Plackett copula has no analytical form for Kendall's τ correlation. Thus, the measure of Kendall's τ correlation must be obtained by the numerical integration of Equation (7) [50].

4.5. Gumbel–Hougaard (GH) Copula

The GH copula is suitable to be used to model the positive dependence structure among variables [41,53]. The GH copula can be extended to model the negative dependence by using the concept of rotations [38]. The bivariate joint CDF and PDF of the GH copula are given as:

$$C(u_1, u_2) = e^{-[(-\ln u_1)^\theta + (-\ln u_1)^\theta]^{\frac{1}{\theta}}}, \tag{24}$$

$$c(u_1, u_2) = \frac{C(u_1, u_2)[(-\ln u_1) + (-\ln u_2)]^{\theta - 1} [\vartheta]^{\frac{2}{\theta} - 2} \left\{ [(\theta - 1)][\vartheta]^{-\frac{1}{\theta}} + 1 \right\}}{u_1 u_2}, \tag{25}$$

with the parameter space $[1, \infty)$, and $\vartheta = (-\ln u_1)^\theta + (-\ln u_2)^\theta$. The Kendall's τ correlation determined by the GH copula is given as:

$$\tau = 1 - \frac{1}{\theta}. \tag{26}$$

4.6. Joe Copula

The Joe copula is also suitable for describing the positive dependence among variables [41,53]. As reported by McNeil et al. [55], the Joe copula is a good choice for cases in which the dataset exhibits a higher positive correlation among the random variables involved. Apart from that, the Joe copula can be extended to model negative dependence by using the concept of rotations [38]. The bivariate joint CDF and PDF of the Joe copula are given as:

$$C(u_1, u_2) = 1 - \left[(1 - u_1)^\theta + (1 - u_2)^\theta - (1 - u_1)^\theta (1 - u_2)^\theta \right]^{\frac{1}{\theta}}, \tag{27}$$

$$c(u_1, u_2) = \frac{\partial \left\{ 1 - \left[(1 - u_1)^\theta + (1 - u_2)^\theta - (1 - u_1)^\theta (1 - u_2)^\theta \right]^{\frac{1}{\theta}} \right\}}{\partial u_1 \partial u_2}, \tag{28}$$

with the parameter space $[1, \infty)$. The Kendall's τ correlation for the Joe copula is obtained as follows:

$$\tau = 1 + \left(\frac{-2 + 2\gamma + 2\ln(2) + \Psi\left(\frac{1}{\theta}\right) + \Psi\left(\frac{1}{\theta}\left(\frac{2 + \theta}{\theta}\right)\right) + \theta}{\theta - 2} \right), \tag{29}$$

where $\gamma = \lim_{n \to \infty} \left(\sum_{i=1}^{n} \frac{1}{i} - \ln(n) \right) \approx 0.57721$ is a Euler constant, and $\Psi(x) = \frac{d}{dx} \ln(\Gamma(x))$ is a digamma function [38].

5. Parameter Estimation and Model Selection

Several methods are available in the literature for estimating copula parameters. In this study, the pseudo maximum likelihood estimation (pseudo-MLE) method is used to estimate the parameter of each copula model. This method has the advantage of not requiring accurate information of the parametric form in the marginal model for each variable. Thus, this method provides a more flexible approach than other estimation methods [31].

5.1. Pseudo Maximum Likelihood Estimation (Pseudo-MLE)

In general, the pseudo-MLE is a semi-parametric method that uses a nonparametric empirical distribution (NED) to represent the marginal distribution of the variables involved in the copula model [56]. The NED for both the duration and severity variables is determined as follows:

$$\hat{F}_i(u) = \frac{\sum_{j=1}^{n} 1(U_{ij} \leq u)}{n+1}, \ i = 1, 2. \tag{30}$$

Next, using the NED as a basis for the marginal models, the copula parameter is estimated using the maximum likelihood approach by maximizing the pseudo-log-likelihood function as follows:

$$\log L(\theta) = \sum_{j=1}^{n} \ln \left[c\left(\hat{F}_1(u_{1j}), \hat{F}_2(u_{2j}); \theta \right) \right]. \tag{31}$$

5.2. Cross-Validation Copula Information Criterion (cvCIC)

The cvCIC criterion is used to select the copula model that best fits a dataset. Hofert et al. [29] briefly reported that the cvCIC measure for a fitted copula model can be determined as follows:

$$xv_n = \frac{\sum_{i=1}^{n} \log \left[c_{\theta_{n,-i}}(\mathbf{F}_{n,-i}(\mathbf{U}_i)) \right]}{n}, \tag{32}$$

where $\theta_{n,-i}$ represents the maximum pseudo-likelihood estimate, and $\mathbf{F}_{n,-i}(\mathbf{U}_i) = [F_{n,1,-i}(u_1), F_{n,2,-i}(u_2)]$ can be computed using the following equation:

$$F_{n,j,-i}(u) \begin{cases} \dfrac{\sum_{k=1}^{n} 1(U_{kj} \leq u)}{n}, & if \ u \geq \min_{k \in [1,2,\dots,n] \setminus [i]} U_{kj}, \\ \dfrac{1}{n}, & otherwise. \end{cases} \tag{33}$$

The copula model that maximizes the cvCIC criterion provides the best-fitted model to the dataset [57]. A detailed discussion of the cvCIC criterion can be found in [58,59].

6. Results and Discussion

Before conducting a detailed analysis of the structure of the dependency of the variables for duration and severity, providing a statistical summary of the dataset is useful. Figure 4 show that the time series plot corresponds to an unhealthy air pollution events threshold in Klang for the period of January 1, 1997 to August 31, 2020. Based on the observed API data, the proportion of unhealthy air pollution events is found to be 2.44%. As described by Masseran and Safari [48], although this proportion is quite small, however, these events provide the most valuable information, particularly on pollution risk management and mitigation.

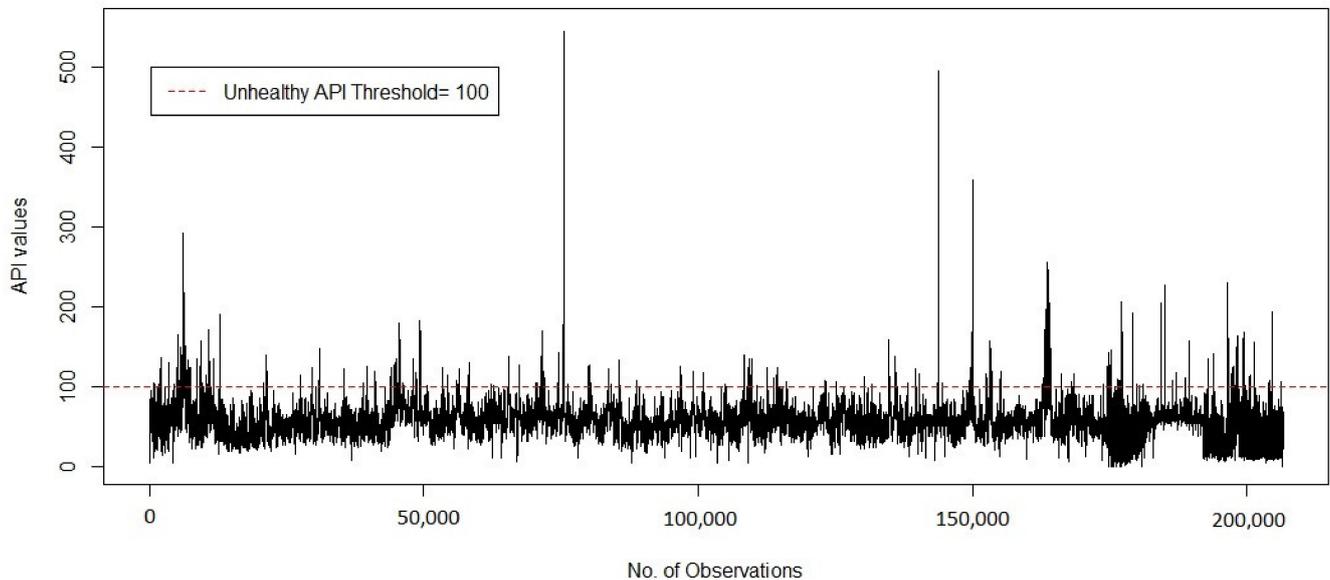*Int. J. Environ. Res. Public Health* **2021**, *18*, 8751

10 of 18

Figure 4. Time series plot corresponds to unhealthy threshold of air pollution events.

Table 1 show the descriptive statistics for the severity and duration data. Based on Table 1, the center measures for both variables provided by their means and medians show a large discrepancy, and the measure of spread shows a very large range between the maximum and minimum values. In fact, their standard deviations are also very large, which indicates significant variation in both variables. These findings are more pronounced for severity, but both have significant skewness with a long right-tail distribution. These results are based on measures of skewness and kurtosis. A long right-tail distribution also indicates the presence of extreme and rare events for both variables.

Table 1. Descriptive statistics for the severity and duration data.

Variable	Mean	Median	Min. Value	Max. Value	Std. Deviation	Skewness	Kurtosis
Duration (hours)	21.39	3.00	1.00	224.00	64.81	2.78	9.23
Severity	2876.50	367.00	102.00	36,677.00	15,193.68	3.38	13.12

The empirical distribution plots shown in Figure 5 illustrate these findings. The scatter plot in Figure 5 also indicates that the duration and severity have a strong positive dependence, with a Pearson correlation coefficient of 0.95. Note that the asterisks *** indicate the significant correlation at 1% significance level. However, the Pearson correlation measure is not a reliable method for dealing with data that exhibit skewness, long-tail, and non-identically marginal distributions [28]. Kendall's τ correlation, which is derived from a fitted copula model that takes into account the dependence structure among the variables, could provide a more reliable measurement [29]. In addition, some large data points are observed to deviate from the diagonal line, which could be because of the skewness of the data set. To check further, a K-plot is used to reveal the dependent relationships between the points close to the curve, which imply high dependence. If the points are close to the diagonal, this indicates relatively weaker dependence [60]. Figure 6 shows that all of the data points for air pollution severity–duration are very close to the curve, thus confirming the positive relationship between the two.
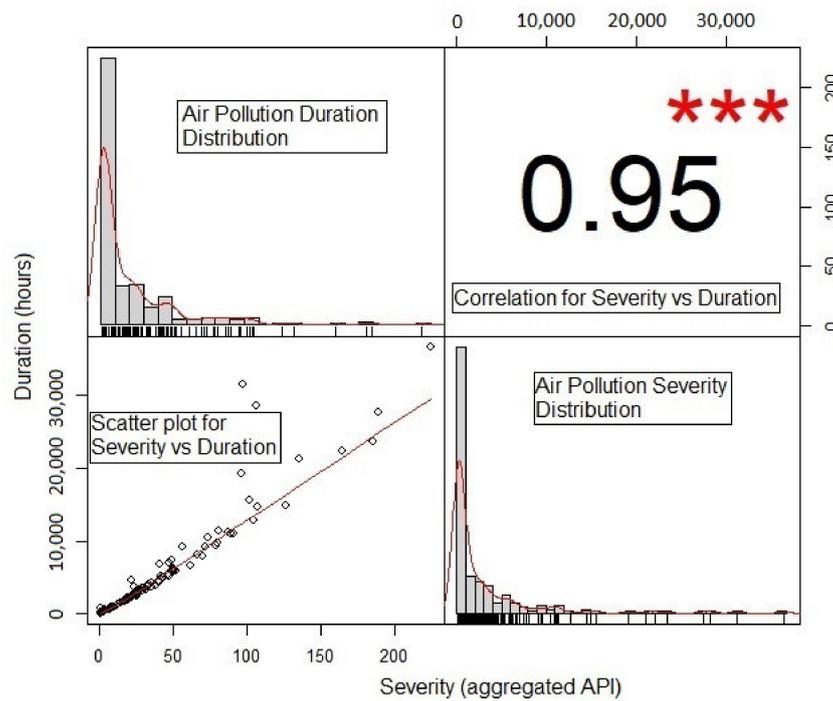
Figure 5. Empirical plots of the relationship of air pollution severity and duration.
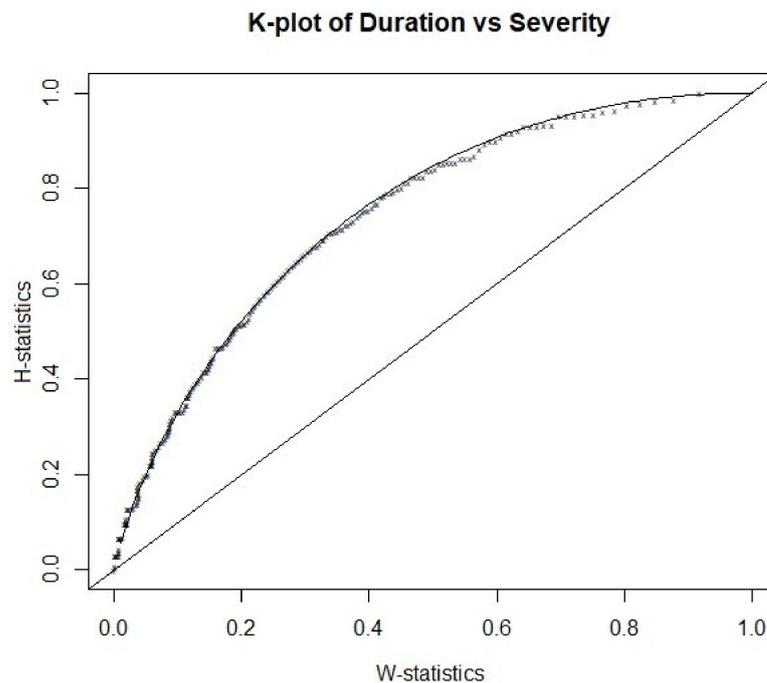

Figure 6. K-plot of the relationship between the duration and severity of air pollution events.

Based on the empirical statistics and plots, first, the marginal model is determined for the duration–severity data, which serves as a building block for constructing a copula model. As indicated by Table 1 and Figure 5, the distributions of both the duration and severity data are highly skewed to the right. In the literature, the exponential, gamma, lognormal, and Weibull distributions of the models are used to represent marginal models of the duration and severity data [40,61]. Figure 7 shows the fitted marginal distribution used to represent the marginal model. The lognormal and gamma distributions seem to

provide an approximate representation of the duration and severity datasets. However, to ensure the accuracy of these fitted models, their goodness of fit needs to be evaluated.
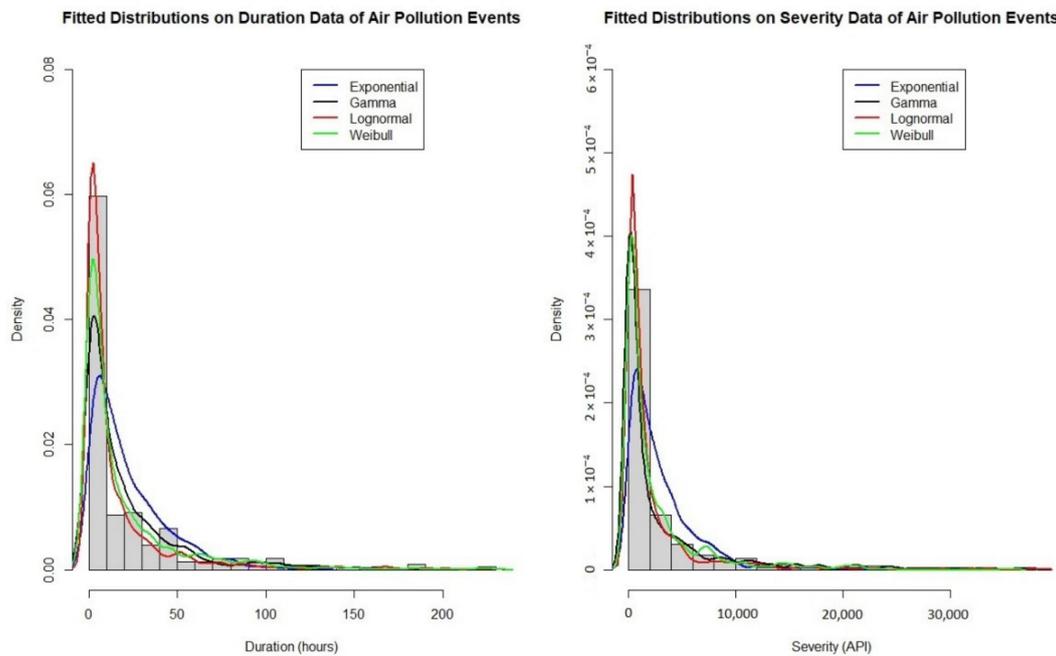

Figure 7. Fitted statistical distributions of duration and severity data.

Table 2 presents the results of the goodness-of-fit evaluation on each of the fitted statistical distributions obtained using the Kolmogorov–Smirnov (KS) test. We found that the *p*-values corresponding to the K–S statistics for all of the fitted distributions of both variables are not significant at $\alpha = 0.05$, which means that the null hypotheses (i.e., the data follow a specified distribution) is rejected. Thus, although the plots in Figure 7 show a good approximation, these fitted distributions cannot reliably represent a marginal model. An accurate marginal model is critical for ensuring precision in copula modeling [62]. To overcome this weakness, an empirical distribution determined using Equation (30) will be used to represent a marginal model.

Table 2. Goodness-of-fit evaluations of the fitted statistical distributions.

Variable	Fitted Distribution	KS-Statistic	*p*-Value
Duration	Exponential	0.3843	0.0000
	Gamma	0.1834	0.0009
	Lognormal	0.1877	0.0006
	Weibull	0.2227	0.0002
Severity	Exponential	0.3973	0.0000
	Gamma	0.2969	0.0000
	Lognormal	0.1834	0.0009
	Weibull	0.2139	0.0005

Based on the empirical distribution, a pseudo-MLE method is used to estimate the parameters of each fitted copula model for the dataset. Table 3 shows the parameter estimates, cvCIC, and Kendall's τ correlation obtained from each fitted copula model. Based on the estimated parameters for each fitted copula model, the Joe copula model was found to obtain the maximum cvCIC criterion value. Thus, we can conclude that the Joe copula is better than the other copula models in describing the severity–duration relationship. This result is supported by the graphical representation given in Figure 8

of each fitted copula model, in which it is clearly shown that the relationship between the duration and severity of air pollution events is of highly positive dependence with a stronger structural dependency in the right-upper tails of their distributions. Except for the AMH copula model, all of the fitted copula models provide a good representation of the dataset. However, as determined by the cvCIC criterion shown in Table 3, the model that best represents the data is the Joe copula, which has the maximum cvCIC criterion value. The Kendall's τ correlations determined for each copula model do not differ very much, with the exception of that of the AMH copula. The Kendall's τ correlation (0.9195) determined by the Joe copula model is slightly lower that that obtained for the Pearson correlation (0.9500) and is more reliable because it considers the properties of data skewness and long-tail behavior in its computation. Next, based on the fitted copula model, the dependence structure of the severity–duration relationship can be analyzed. This information could benefit the governing authorities who are responsible for managing the risk of extreme air pollution, particularly during critical air pollution events. In addition, the copula approach can work as a simulated model for air pollution risk assessment. For any given air pollution duration size, the analyst can employ the copula model to determine the estimated severity size of air pollution event. Figure 9 shows the PDF and CDF of the fitted Joe copula as density and contour plots.

Table 3. Parameter estimates and cvCIC and Kendall' τ measures obtained by each fitted copula model.

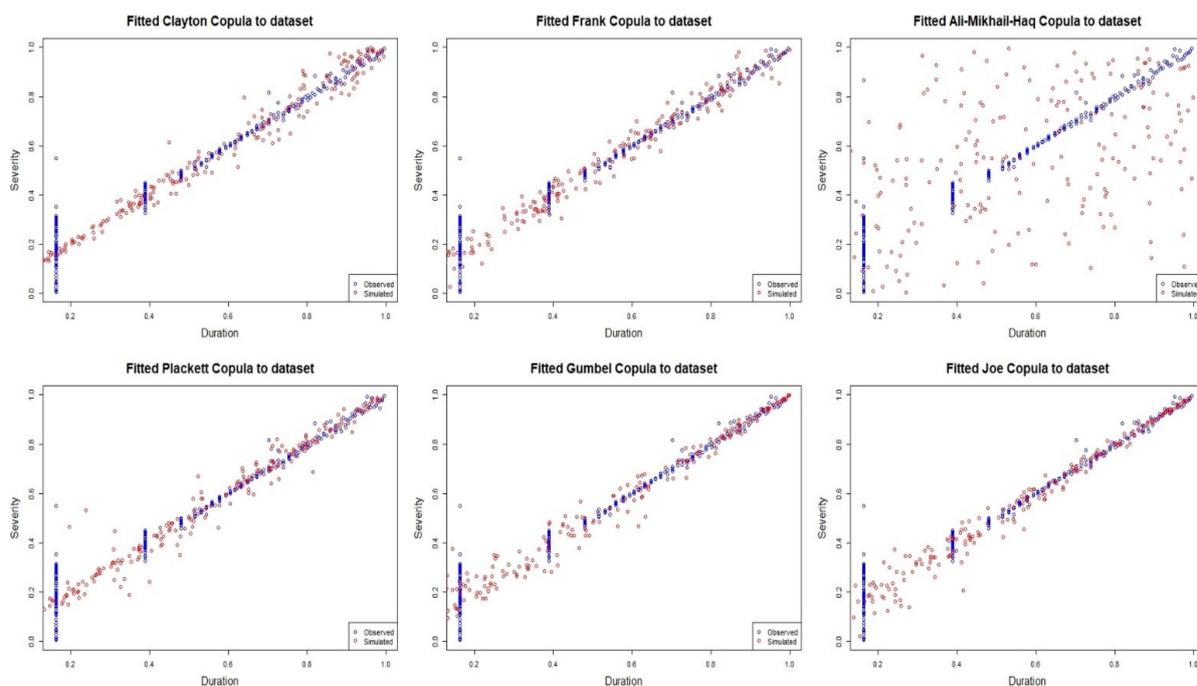| Copula Model	Parameter Estimate (θ)	cVCIC	Kendall's τ
Clayton	22.85	49.108	0.9195
Ali–Mikhail–Haq	1	35.689	0.3333
Frank	48	142.708	0.9195
Placket	846.5	221.530	0.9195
Gumbel	12.42	199.361	0.9194
Joe	25.38	255.258	0.9195

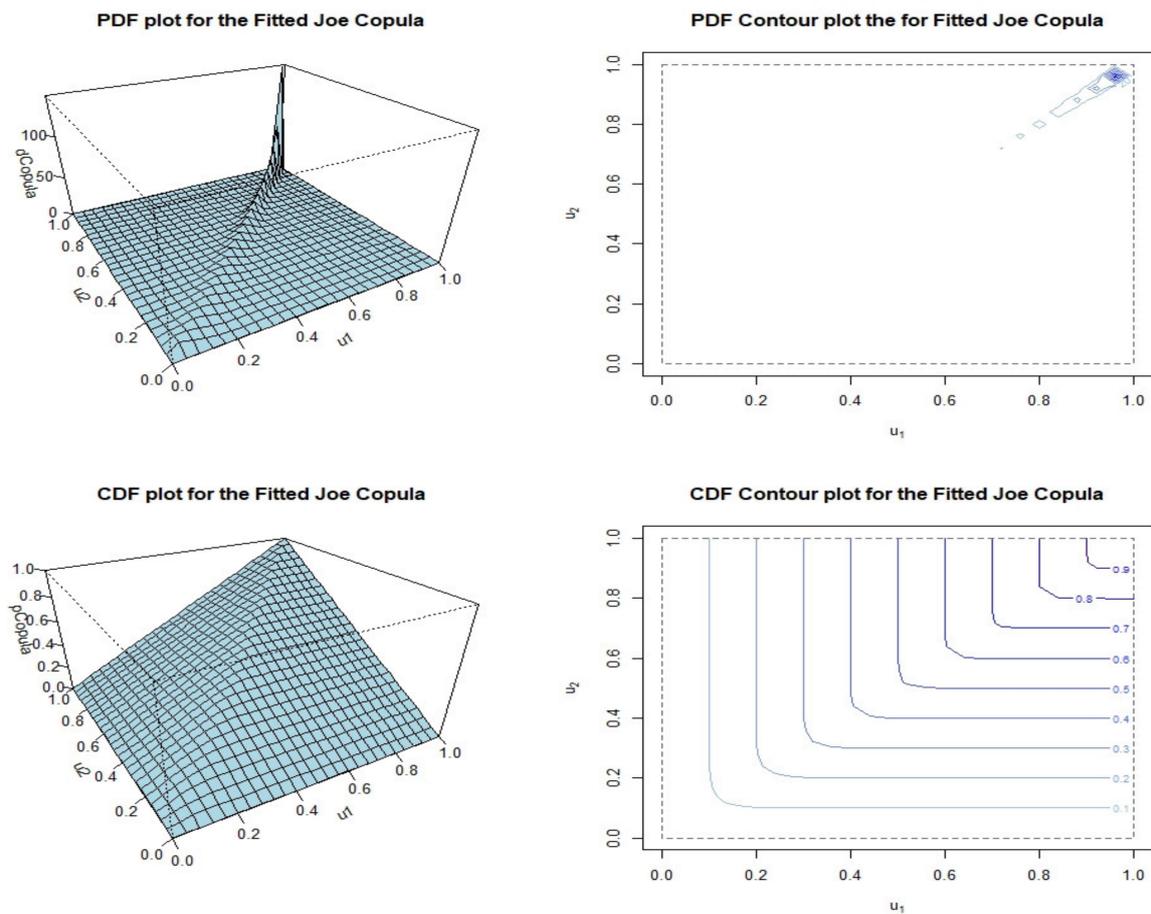**Figure 8.** Fitted copula models for the severity–duration dataset.

Figure 9. PDF and CDF plots for the fitted Joe copula model corresponding to their contour representations.

The PDF and CDF density and contour plots depict the probable behaviors of the duration–severity relationship and confirm the dependence structure shown in the scatter plot of Figure 8. However, the information provided by the PDF and CDF are described with respect to their joint probability distribution of the duration–severity relationship. Therefore, PDF and CDF information are referred to as the likelihood or probability of a certain duration period and severity level occurring simultaneously. The PDF and CDF contour plots efficiently illustrate the positive dependence and stronger structural dependency in the right-upper tails of the duration–severity data. From the fitted Joe copula, the probabilities of air pollution events under certain circumstances with a specific severity or duration can determined. Figure 10 shows the curves of various conditional probabilities of a certain severity level given the duration of an air pollution event. For example, the authorities might want to determine the risk probability for an air pollution severity of less than 8000 and 15,000 for a given duration exceeding 64 h. By normalizing these values using the probability integral transform to obtain a uniform (0,1), the computed probabilities obtained are $\Pr(X_2 \leq 0.90 | X_1 \geq 0.89) = 0.908$ and $\Pr(X_2 \leq 0.9608 | X_1 \geq 0.89) = 1$, respectively.

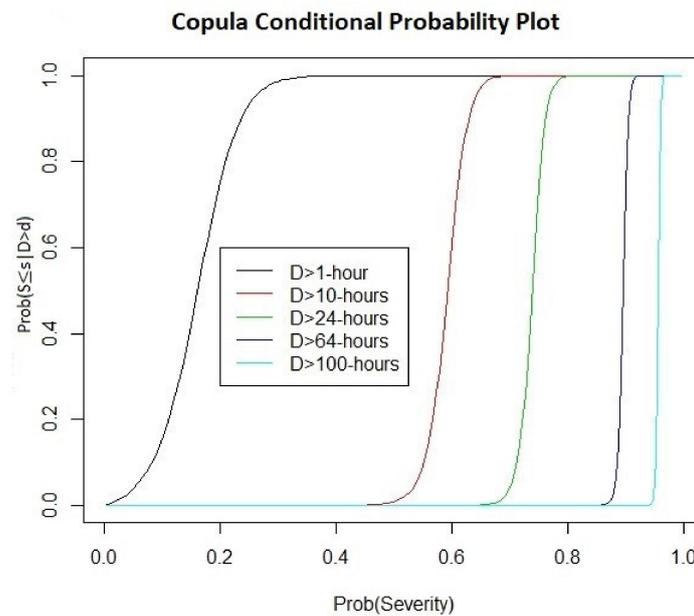*Int. J. Environ. Res. Public Health* **2021**, *18*, 8751

15 of 18

Figure 10. Conditional probability of severity level given the duration.

Based on the fitted Joe copula model, the return periods for recurrent air pollution events of different durations and severities can be computed, as shown in Table 4. The rows correspond to the first and second columns in Table 4, indicating given specific levels of duration and severity sizes that represent various scenarios related to air pollution events. On the other hand, all of the values represented in the third, fourth, fifth, and sixth columns indicate the estimated joint and conditional return periods for the air pollution duration–severity pairs in Klang. If the authorities want to evaluate the expected recurrence of air pollution events of a particular duration and severity, four types of information can be determined, as described in Equations (9)–(12). For example, for an air pollution event with a duration of 50 h, with a possibility that this event will achieve a severity level of 20,000, its joint OR return period ($D \geq 50h \cup S \geq 20,000$) is approximately 34.4 days, and its AND return period ($D \geq 50h \cap S \geq 20,000$) is approximately 135.4 days. The conditional D|S return period for D = 50 h given $S \geq 20,000$ is approximately 4153.6 days, and the conditional S|D return period for S = 20,000 given $D \geq 50$ h is approximately 1056 days. The above information could be very beneficial to the authorities for air pollution even planning and for mitigating the risks associated with unhealthy air pollution events.

Table 4. Joint and conditional return periods for the air pollution duration–severity pairs in Klang using the Joe copula.

Duration Size	API Severity	Joint OR Return Period, T'_{DS} (Days)	Joint AND Return Period, T_{DS} (Days)	Conditional D|S Return Period, $T_{D\|S \geq s}$ (Days)	Conditional S|D Return Period, $T_{S\|D \geq d}$ (Days)
	100	4.4	34.4	34.6	268.5
	1000	10.8	34.4	84.3	268.5
50-h	10,000	34.4	56.4	720.9	439.9
	20,000	34.4	135.4	4153.6	1056.0
	30,000	34.4	406.3	4328.5	3168.0
	100	4.4	56.4	56.7	721.1
	1000	10.8	56.4	138.1	721.1
80-h	10,000	54.8	58.2	743.2	743.2
	20,000	56.4	135.4	4153.6	1730.7
	30,000	56.4	406.3	7382.4	5192.0

Int. J. Environ. Res. Public Health **2021**, *18*, 8751

16 of 18

Duration Size	API Severity	Joint OR Return Period, T'_{DS} (Days)	Joint AND Return Period, T_{DS} (Days)	Conditional D|S Return Period, $T_{D\|S\geq s}$ (Days)	Conditional S|D Return Period, $T_{S\|D\geq d}$ (Days)
100-h	100	4.4	101.6	102.0	2336.4
	1000	10.8	101.6	248.6	2336.4
	10,000	56.4	101.6	1297.9	2336.4
	20,000	101.5	135.5	4153.9	3115.4
	30,000	101.6	406.3	7382.4	9345.6
120-h	100	4.4	169.3	170.0	6490.0
	1000	10.8	169.3	414.2	6490.0
	10,000	56.4	169.3	2163.1	6490.0
	20,000	135.4	169.3	5193.5	6491.8
	30,000	169.3	406.3	7382.4	15,576.0
150-h	100	4.4	225.7	226.7	11,537.9
	1000	10.8	225.7	552.3	11,537.9
	10,000	56.4	225.7	2884.1	11,537.9
	20,000	135.4	225.7	6922.7	11,537.9
	30,000	225.7	406.3	7382.4	20,768.1

7. Conclusions

This study proposes the concept of duration and severity measures for evaluating unhealthy air pollution events. Since these bivariate criteria indicate the properties of structural dependency, long-tail, and non- identically marginal distributions, a copula model is proposed to deal with these issues. In particular, this study presents the application of a copula model in evaluating the behaviors of unhealthy air pollution events with respect to their duration and severity characteristics. The duration of an air pollution event was defined as a period in which the air pollution index values qualify as unhealthy for a period of consecutive days. On the other hand, the severity was defined as a magnitude of an air pollution event based on the cumulative effect of an unhealthy API during air pollution events. These two characteristics are related because the severity of an air pollution event always depends on its duration. Thus, considering their dependency, copula models were used to jointly model a combination of the marginal distributions of both variables.

A case study was conducted using data from Klang, Malaysia. To ensure the accuracy of the marginal model, an empirical distribution approach was proposed in the copula model. A total of six types of copula models, namely the Clayton, Ali–Mikhail–Haq, Frank, Plackett, Gumbel–Hougaard, and Joe copulas, were considered for evaluating the relationship between the duration and severity of unhealthy air pollution events in Klang. Based on its efficient fit, the Joe copula was found to best fit the data. Based on the Joe copula model, several valuable statistical measures for assessing air pollution risk were proposed, including (i) the Kendall's τ correlation of the copula to measure the dependency between the duration and severity of an air pollution event, (ii) the conditional probability of a certain air pollution severity given the air pollution duration, (iii) the joint OR/AND return period, and (iv) the conditional D\|S and S\|D return periods. All of the these measures could be very beneficial for planning and mitigating the risks associated with unhealthy air pollution events. On the other hand, this approach is also applicable in other areas such as for natural hazards, extreme wind speeds, climate change, extreme temperature analysis, etc. Particularly, for the purpose of investigating the severity impact of some particular risky events corresponding to their duration size. Further research is recommended to evaluate how the trivariate relationship corresponds to the duration, severity, and intensity of air pollution event. A vine-copula approach could be used to analyze various combination structures represented by a duration–severity–intensity relationship.

Int. J. Environ. Res. Public Health **2021**, *18*, 8751

17 of 18

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Conflicts of Interest: The author declares no conflict of interest.

References

1. Murena, F. Measuring air quality over large urban areas: Development and application of an air pollution index at the urban area of Naples. *Atmos. Environ.* **2004**, *38*, 6195–6202. [CrossRef]
2. Masseran, N. Modeling fluctuation of PM10 data with existence of volatility effect. *Environ. Eng. Sci.* **2017**, *34*, 816–827. [CrossRef]
3. Masseran, N. Power-law behaviors of the duration size of unhealthy air pollution events. *Stoch. Environ. Res. Risk Assess.* **2021**, *35*, 1499–1508. [CrossRef]
4. Genest, C.; Quessy, J.-F.; Rémillard, B. Goodness-of-fit procedures for copula models based on the probability integral transformation. *Scand. J. Stat.* **2006**, *32*, 337–366. [CrossRef]
5. Huard, D.; Évin, G.; Favre, A.-C. Bayesian copulas selection. *Comput. Stat. Data Anal.* **2006**, *51*, 809–822. [CrossRef]
6. Marcotte, D.; Henry, E. Automatic joint set clustering using a mixture of bivariate normal distributions. *Int. J. Rock Mech. Min. Sci.* **2002**, *39*, 323–334. [CrossRef]
7. Lai, C.D.; Balakrishnan, N. *Continuous Bivariate Distributions*; Springer: Berlin/Heidelberg, Germany, 2009. [CrossRef]
8. Tran, K.P.; Castagliola, P.; Celano, G. Monitoring the ratio of population means of a bivariate normal distribution using CUSUM type control charts. *Stat. Pap.* **2016**, *59*, 387–413. [CrossRef]
9. Choi, J.-E.; Shin, D.W. Three regime bivariate normal distribution: A new estimation method for co-value-at-risk, CoVaR. *Eur. J. Financ.* **2019**, *25*, 1817–1833. [CrossRef]
10. Yue, S. The bivariate lognormal distribution to model a multivariate flood episode. *Hydrol. Process.* **2000**, *14*, 2575–2588. [CrossRef]
11. Pundir, S.; Amala, R. Detecting diagnostic accuracy of two biomarkers through a bivariate log-normal ROC curve. *J. Appl. Stat.* **2015**, *42*, 2671–2685. [CrossRef]
12. Gumbel, E.J. Bivariate logistic distributions. *J. Am. Stat. Assoc.* **1961**, *56*, 335–349. [CrossRef]
13. Yue, S. The Gumbel logistic model for representing a multivariate storm event. *Adv. Water Resour.* **2000**, *24*, 179–185. [CrossRef]
14. Yue, S.; Ouarda, T.; Bobée, B.; Legendre, P.; Bruneau, P. The gumbel mixed model for flood frequency analysis. *J. Hydrol.* **1999**, *226*, 88–100. [CrossRef]
15. Yue, S. A bivariate extreme value distribution applied to flood frequency analysis. *Hydrol. Res.* **2001**, *32*, 49–64. [CrossRef]
16. Marshall, A.W.; Ingram, O. A multivariate exponential distribution. *J. Am. Stat. Assoc.* **1967**, *62*, 30–44. [CrossRef]
17. Bacchi, B.; Becciu, G.; Kottegoda, N. Bivariate exponential model applied to intensities and durations of extreme rainfall. *J. Hydrol.* **1994**, *155*, 225–236. [CrossRef]
18. Won, J.; Choi, J.; Lee, O.; Park, M.J.; Kim, S. Two ways to quantify Korean drought frequency: Partial duration series and bivariate exponential distribution, and application to climate change. *Atmosphere* **2020**, *11*, 476. [CrossRef]
19. Royen, T. Expansions for the multivariate chi-square distribution. *J. Multivar. Anal.* **1991**, *38*, 213–232. [CrossRef]
20. Izawa, T. Two or multi-dimensional gamma-type distribution and its application to rainfall data. *Pap. Meteorol. Geophys.* **1965**, *15*, 167–200. [CrossRef]
21. Moran, P.A.P. Statistical inference with bivariate gamma distribution. *Biometrika* **1969**, *54*, 385–394. [CrossRef]
22. Schmeiser, B.W.; Lal, R. Bivariate gamma random vectors. *Oper. Res.* **1982**, *30*, 355–374. [CrossRef]
23. Loaciga, H.A.; Leipnik, R.B. Correlated gamma variables in the analysis of microbial densities in water. *Adv. Water Resour.* **2005**, *28*, 329–335. [CrossRef]
24. Yue, S.; Ouarda, T.; Bobée, B. A review of bivariate gamma distributions for hydrological application. *J. Hydrol.* **2001**, *246*, 1–18. [CrossRef]
25. Nadarajah, S. A bivariate gamma model for drought. *Water Resour. Res.* **2007**, *43*, W08501. [CrossRef]
26. Antão, E.M.; Soares, C.G. Approximation of the joint probability density of wave steepness and height with a bivariate gam-ma distribution. *Ocean. Eng.* **2016**, *126*, 402–410. [CrossRef]
27. Lambert, P.; Vandenhende, F. A copula-based model for multivariate non-normal longitudinal data: Analysis of a dose titra-tion safety study on a new antidepressant. *Stat. Med.* **2002**, *21*, 3197–3217. [CrossRef] [PubMed]
28. Joe, H. *Dependence Modeling with Copulas*; Chapman and Hall/CRC: New York, NY, USA, 2014. [CrossRef]
29. Hofert, M.; Kojadinovic, I.; Mächler, M.; Yan, J. *Elements of Copula Modeling with R*; Springer: Berlin/Heidelberg, Germany, 2018. [CrossRef]

30. Kojadinovic, I.; Yan, J. Modeling multivariate distributions with continuous margins using the copula R package. *J. Stat. Softw.* **2010**, *34*, 17192. [CrossRef]

31. Zhang, L.; Singh, V.P. *Copulas and Their Applications in Water Resources Engineering*; Cambridge University Press: Cambridge, UK, 2019. [CrossRef]

32. Sak, H.; Yang, G.; Li, B.; Li, W. A copula-based model for air pollution portfolio risk and its efficient simulation. *Stoch. Environ. Res. Risk Assess.* **2017**, *31*, 2607–2616. [CrossRef]

33. Chan, R.K.S.; So, M.K.P. Multivariate modelling of spatial extremes based on copulas. *J. Stat. Comput. Simul.* **2018**, *88*, 2404–2424. [CrossRef]

34. Falk, M.; Padoan, S.A.; Wisheckel, F. Generalized pareto copulas: A key to multivariate extremes. *J. Multivar. Anal.* **2019**, *174*, 104538. [CrossRef]

35. Kim, J.-M.; Lee, N.; Xiao, X. Directional dependence between major cities in China based on copula regression on air pollution measurements. *PLoS ONE* **2019**, *14*, e0213148. [CrossRef] [PubMed]

36. Masseran, N.; Hussain, S. Copula modelling on the dynamic dependence structure of multiple air pollutant variables. *Mathematics* **2020**, *8*, 1910. [CrossRef]

37. He, S.; Li, Z.; Wang, W.; Yu, M.; Liu, L.; Alam, M.N.; Gao, Q.; Wang, T. Dynamic relationship between meteorological condi-tions and air pollutants based on a mixed Copula model. *Int. J. Climatol.* **2021**, *41*, 2611–2624. [CrossRef]

38. Czado, C. *Analyzing Dependent Data with Vine Copulas*; Springer: Berlin/Heidelberg, Germany, 2019. [CrossRef]

39. Aas, K. Pair-copula constructions for financial applications: A review. *Econometrics* **2016**, *4*, 43. [CrossRef]

40. Shiau, J.T. Fitting drought duration and severity with two-dimensional copulas. *Water Resour. Manag.* **2006**, *20*, 795–815. [CrossRef]

41. Latif, S.; Mustafa, F. Bivariate flood distribution analysis under parametric copula framework: A case study for Kelantan River basin in Malaysia. *Acta. Geophys.* **2020**, *68*, 821–859. [CrossRef]

42. Al-Dhurafi, N.A.; Masseran, N.; Zamzuri, Z.H.; Razali, A.M. Modeling unhealthy air pollution index using a peaks-over-threshold method. *Environ. Eng. Sci.* **2018**, *35*, 101–110. [CrossRef]

43. Masseran, N.; Safari, M.A.M. Intensity–duration–frequency approach for risk assessment of air pollution events. *J. Environ. Manage.* **2020**, *264*, 110429. [CrossRef] [PubMed]

44. Google Maps. 2019. Available online: https://www.google.com/maps/place/Klang,+Selangor/@3.2467558,101.2650693,9.1z/data=!4m5!3m4!1s0x31cc534c4ffe81cf:0xeb61f5772fd54514!8m2!3d3.044917!4d101.4455621 (accessed on 25 March 2019).

45. AL-Dhurafi, N.A.; Masseran, N.; Zamzuri, Z.H. Hierarchical-generalized Pareto model for estimation of unhealthy air pol-lution index. *Environ. Model. Assess.* **2020**, *25*, 555–564. [CrossRef]

46. Masseran, N.; Safari, M. Mixed POT-BM approach for modeling unhealthy air pollution events. *Int. J. Environ. Res. Public Heal.* **2021**, *18*, 6754. [CrossRef] [PubMed]

47. A guide to air pollutant index in Malaysia (API). Available online: https://aqicn.org/images/aqi-scales/malaysia-api-guide.pdf (accessed on 4 June 2020).

48. Masseran, N.; Safari, M.A.M. Risk assessment of extreme air pollution based on partial duration series: IDF approach. *Stoch. Environ. Res. Risk Assess.* **2020**, *34*, 545–559. [CrossRef]

49. Sklar, A. Fonctions de repartition à n dimensionls et leurs marges. *Publ. Inst. Statist. Univ. Paris* **1959**, *8*, 229–231.

50. Chowdhary, H.; Escobar, L.A.; Singh, V.P. Identification of suitable copulas for bivariate frequency analysis of flood peak and flood volume data. *Hydrol. Res.* **2011**, *42*, 193–216. [CrossRef]

51. Nelsen, R.B. *An Introduction to Copulas*; Springer: Berlin/Heidelberg, Germany, 2006.

52. Klein, B.; Schumann, A.H.; Pahlow, M. Copulas—New risk assessment methodology for Dam safety. *Flood Risk Assess. Manag.* **2010**, 149–185. [CrossRef]

53. Yusof, F.; Hui-Mean, F.; Suhaila, J.; Yusof, Z.; Jamaludin, S.S.S. Characterisation of drought properties with bivariate copula analysis. *Water Resour. Manag.* **2013**, *27*, 4183–4207. [CrossRef]

54. Tosunoglu, F.; Can, I. Application of copulas for regional bivariate frequency analysis of meteorological droughts in Turkey. *Nat. Hazards* **2016**, *82*, 1457–1477. [CrossRef]

55. McNeil, A.J.; Frey, R.; Embrechts, P. *Quantitative risk Management: Concepts, Techniques and Tools—Revised Edition*; Princeton University Press: Princeton, NJ, USA, 2015.

56. Genest, C.; Ghoudi, K.; Rivest, L. A semi-parametric estimation procedure of dependence parameters in multivariate families of distributions. *Biometrika* **1995**, *82*, 543–552. [CrossRef]

57. Hofert, M.; Kojadinovic, I.; Maechler, M.; Yan, J.; Nešlehová, J.G. Copula: Multivariate Dependence with Copulas. R Package Version 0.999-19.1. Available online: https://cran.r-project.org/web/packages/copula/index.html (accessed on 13 January 2021).

58. Grønneberg, S.; Hjort, N.L. The copula information criteria. *Scand. J. Stat.* **2014**, *41*, 436–459. [CrossRef]

59. Jordanger, L.A.; Tjøstheim, D. Model selection of copulas: AIC versus a cross validation copula information criterion. *Stat. Probab. Lett.* **2014**, *92*, 249–255. [CrossRef]

60. Genest, C.; Boies, J.-C. Detecting dependence with Kendall plots. *Am. Stat.* **2003**, *57*, 275–284. [CrossRef]

61. Xu, Y.; Huang, G.; Fan, Y. Multivariate flood risk analysis for Wei River. *Stoch. Environ. Res. Risk Assess.* **2015**, *31*, 225–242. [CrossRef]

62. Kim, G.; Silvapulle, M.J.; Silvapulle, P. Comparison of semiparametric and parametric methods for estimating copulas. *Comput. Stat. Data Anal.* **2006**, *51*, 2836–2850. [CrossRef]