Stochastic Environmental Research and Risk Assessment (2024) 38:2331–2358
https://doi.org/10.1007/s00477-024-02682-7

ORIGINAL PAPER

Risk assessment for extreme air pollution events using vine copula

Mohd Sabri Ismail[1] · **Nurulkamal Masseran**[1]

Accepted: 2 February 2024 / Published online: 4 March 2024

Abstract
This study proposes an alternative risk assessment approach for evaluating extreme air pollution events through vine copula modeling. Three characteristics of unhealthy air pollution events (i.e., severity, intensity, and duration) in Klang, Malaysia, are examined. The vine copula is fitted using sequential maximum likelihood estimation and joint maximum likelihood estimation, with a subsequent comparison based on criteria such as log-likelihood, Akaike's information criterion, and Bayesian's information criterion. Model fitting and comparison studies demonstrate that the most well-fitted vine copula model, achieved through joint maximum likelihood estimation, comprises the Joe, Rotated Tawn type 2 (180 degrees), and Rotated BB8 (90 degrees) copulas. The positive Kendall's τ correlation coefficient (0.26) for the obtained vine copula indicates that higher values of one characteristic are likely to be associated with higher values of the other characteristics, and vice versa. Furthermore, with the upper tail dependence coefficient (0.31) surpassing the lower tail dependence coefficient (0.18), indicating stronger dependence in the upper tail of their distribution, this underscores the significance of conducting risk assessments for extreme air pollution events characterized by extreme levels of severity, intensity, and duration. A vine copula-based simulation study is conducted to delve deeper into the risk assessment, revealing that extreme air pollution events are not linked to the highest values of joint and conditional probabilities. These findings suggest that extreme values in those distinct characteristics do not consistently occur simultaneously. The return period measures also indicate that extreme air pollution events have long waiting periods. Despite the current status of extreme air pollution events in Klang being controllable, achieving effective control necessitates ongoing efforts, encompassing regulatory actions, industrial controls, robust public transportation programs, and a dedicated transition to cleaner energy sources. This task is crucial for ensuring continuous clean air quality, sustaining our environment, and avoiding negative impacts on the economy and public well-being.

Keywords Air pollution · Risk assessment · Joint distribution · Multivariate statistical modeling · Vine copula · Copula · Simulation

1 Introduction

An air pollution event pertains to a specific duration when air quality significantly deteriorates, resulting in elevated concentrations of pollutants in the atmosphere. Great concern has been expressed worldwide regarding air pollution event, and various efforts have been made to monitor and control it. One of the main contributors to air pollution event is the burning of fossil fuels (e.g., coal, petrol, and diesel), which produces numerous airborne toxic emissions, particularly pollutants like particulate matters (e.g., PM_{10} and $PM_{2.5}$) (Perera et al. 2019; Shindell and Smith 2019). When inhaled, these hazardous particulate matters penetrate the lungs, enter the bloodstream, and damage vital organs (Kotcher et al. 2019; Zhang et al. 2023). On an extreme scale, air pollution event can lead to a public health crisis, increasing morbidity and mortality rates. Additionally, extreme air pollution event can cause various other problems, including financial deterioration, psychological complications, and social instability (Gautam and Bolia 2020; Lu 2020).

Typically, the Air Pollution Index (API) is employed to monitor unhealthy air pollution events. The API threshold for considering air quality as unhealthy is determined by the concentration levels of specific pollutants. In the context

✉ Mohd Sabri Ismail
sabriismail@ukm.edu.my

Nurulkamal Masseran
kamalmsn@ukm.edu.my

[1] Department of Mathematical Sciences, Faculty of Science and Technology, Universiti Kebangsaan Malaysia, UKM, 43600 Bangi, Selangor, Malaysia

of health impacts, an air pollution event is unhealthy if the API exceeds the threshold of 100. The risk associated with an unhealthy air pollution event increases with its duration, intensity, and severity levels. Here, the duration, intensity, and severity of an unhealthy air pollution event refer to the total period of the unhealthy air pollution event, the highest value of the API within the unhealthy air pollution event, and the total API values within the unhealthy air pollution event, respectively (Shafaei et al. 2017). Furthermore, extreme air pollution events can be defined as those associated with exceptionally extreme duration, intensity, and severity levels. Avoiding such extreme air pollution events is crucial for maintaining stability in social lives, public well-being, the economy, and the environment.

Extreme air pollution events manifest at extreme levels of duration, intensity, and severity, linking the modeling of these occurrences closely with the upper tail distribution (Cirillo and Taleb 2020; Katz 2010). To achieve this, a flexible multivariate statistical model, rich in tail behaviors, becomes essential. Previous studies have demonstrated that duration, intensity, and severity data exhibit skewed and asymmetrical distributions, underscoring the importance of considering skewed and asymmetric models (Ismail and Masseran 2023; Masseran 2021a). Consequently, symmetric distributions such as multivariate normal or multivariate Student's t distributions are not applicable in this context. The implementation of a multivariate copula is highly favorable, as it offers greater flexibility in the dependence structure and various tail properties, thereby presenting an improved risk assessment model (Bhatti and Do 2019; Li et al. 2022). A multivariate copula individually models the marginals and defines a copula function to establish a flexible dependence structure from these marginals (Jaworski et al. 2010; Joe 2014; Patton 2012). Numerous parametric multivariate copula models with parameters controlling the strength of the dependence structure and tail properties are available. Among the most popular are the Clayton, Joe, and Frank copulas (Amini et al. 2022; Kim et al. 2007; Nguyen et al. 2016).

Nevertheless, the aforementioned parametric multivariate copula models exhibit two main shortcomings: 1) they rely on only one or two parameters to describe the intricate dependence structure, and 2) the resulting models, residing in a higher-dimensional space, are not easily visualizable (Lü et al. 2020). To address these limitations, vine copula modeling was introduced. Vine copula, a more powerful multivariate copula, is constructed by combining several bivariate copulas, also known as pair copulas (Czado and Nagler 2022b). This approach capitalizes on the advantages of numerous existing parametric pair copulas, using their combination to model the dependence structure of dependent data across arbitrary dimensions. All parameters of the involved pair copulas are incorporated into the vine copula development, enhancing the model's realism in representing complex dependence structures. Furthermore, the pair copulas of a vine copula can be visualized, making this model highly tractable (Joe and Kurowicka 2011). Algorithms for maximum likelihood estimation (MLE) and simulation of a vine copula were also developed to identify the appropriate vine copula and further simulate the multivariate data modeled by the obtained vine copula, respectively (Czado 2019). Thus, vine copula is highly explainable, practical, and reliable in terms of risk assessment performance (Joe et al. 2010; Pourkhanali et al. 2016).

In the realm of finance, vine copulas have proven effective in modeling tail risk for portfolio optimization (Nguyen and Liu 2023; Semenov and Smagulov 2019; Zhi et al. 2021). Turning to environmental studies, vine copulas have been applied to assess water qualities in various watersheds (Arya Farid and Zhang 2017). The interconnected river systems' water pollution risk has also been scrutinized using vine copulas (Yu et al. 2020). Additionally, vine copulas have played a role in modeling the dependence structure among variables related to pipe conditions, predicting potential pipe leakages induced by extreme weather conditions (Atique and Attoh-Okine 2016). Flood events, too, have been effectively modeled through vine copulas, contributing to the assessment of potential flood hazards (Tosunoglu et al. 2020). The application of vine copulas extends to the examination of compound floods, unraveling the underlying conditions that intensify flood risks (Daneshkhah et al. 2016; Liu et al. 2018). Furthermore, vine copulas have been instrumental in modeling climate-dependent data, with insights drawn from these models applied to forecast crop yields in Australian wheat and predict the occurrence of extreme climate events (Nguyen-Huy et al. 2018). Beyond these domains, vine copulas have found application in determining agricultural water allocation under uncertainties (Shan et al. 2021) and in fault detection and diagnosis within safety processes (Amin et al. 2021).

Focusing on previous studies related to air pollution risk analysis, researchers often explore various indexes associated with this environmental concern, including the API and air pollutant indices (Ma et al. 2023; Zhang et al. 2022). An illustrative example involves the analysis of particulate matter (PM) and nitrogen oxides (NOx) indices, suggesting a potential link between long-term exposure to air pollution and human depression risk (Wu et al. 2023). A comprehensive meta-analysis, using graphical representations and numerical techniques, has recommended investigating air quality indexes and exploring their correlation with the burden of disease across sub-Saharan Africa (Madonsela 2023). Furthermore, $PM_{2.5}$ levels were examined using the TTAinterfaceTrendAnalysis approach to estimate annual and monthly trends in $PM_{2.5}$ across five Indian megacities (Ravindra et al. 2023). In the realm of forecasting, an

ensemble convolutional reinforcement learning gate network has been suggested for predicting the $PM_{2.5}$ index to enhance travel safety (Yu et al. 2023). Additionally, a random forest model, based on satellite Aerosol Optical Depth (AOD) retrievals, has been explored to predict PM_{10} concentrations (Tuna Tuygun and Elbir 2023). Temporal variations of $PM_{2.5}$ and the influence of meteorological parameters on $PM_{2.5}$ concentrations were investigated for six major cities in Central Asia, revealing that coal combustion is the primary source of $PM_{2.5}$ pollution in most cities (Tursumbayeva et al. 2023).

Moving beyond indexes, particularly in Klang, Malaysia, air pollution has been scrutinized by examining the characteristics of unhealthy air pollution events, such as severity, duration, and intensity, which are integral for performing risk analyses on extreme air pollution events. Notably, data on the severity, duration, and intensity of unhealthy air pollution events are computed from the API, following the definition provided in the second paragraph above. The corresponding formulas for these measures are provided in Sect. 3. In the previous study, duration data have been well-fitted using the lognormal distribution, distinguishing it from the exponential, gamma, and Weibull distributions (Masseran et al. 2021). Power law behaviors of duration data indicate that extreme air pollution events typically extend beyond 33 h (Masseran 2021b). Moreover, power law behaviors of severity data suggest that authorities should intervene when extreme severity levels surpass the 1221 thresholds (Masseran and Safari 2022). Considering these studies, extreme air pollution events in Klang are defined to occur at durations exceeding 33 h, severity levels surpassing 1221, and intensities exceeding 300, with reference to Table 1 in Sect. 3 below, considering the potential impact on public health. Employing the intensity–duration–frequency approach has revealed an association between intensity and duration, moving in the same direction (Masseran and Safari 2020a, b). Simultaneously, the generalized extreme value model has been applied to study severity and duration data, revealing a strong positive correlation in Klang. This underscores that prolonged unhealthy air pollution events lead to more severe unhealthy air pollution effects (Masseran and Safari 2022).

Recently, to enhance the understanding of air pollution risk behaviors, researchers have applied multivariate copulas to model the dependence structure between PM10 and other air pollutants (Masseran and Hussain 2020). Furthermore, the application of bivariate copulas has been pivotal in determining the dependence structure of severity and duration (Masseran 2021a). The intricacies of the bivariate relationships among severity, intensity, and duration—encompassing pairs such as duration and intensity, severity and intensity, and duration and severity—have been thoroughly explored using bivariate copula models (Ismail and Masseran 2023). These investigations have unveiled the skewed and asymmetric distribution properties of this data. However, to the best of our knowledge, the trivariate relationship involving duration, severity, and intensity has not yet been modeled using a multivariate copula, specifically the vine copula. Taking a significant stride forward, this study aims to investigate the trivariate relationship using the vine copula and then perform a risk assessment of extreme air pollution events.

To achieve the aforementioned goal, a vine copula model is developed and employed herein to determine the most well-fitted distribution for the intensity, duration, and severity data. The model utilizes dependency measures, such as Kendall's τ correlation, upper tail dependence, and lower tail dependence coefficients, along with other information obtained through vine copula simulation, including probability measures (i.e., joint and conditional probabilities) and return period measures (i.e., joint OR, joint AND, and conditional return periods). These measures are utilized to evaluate the risks associated with extreme air pollution events. By leveraging the information derived from the dependency, probability, and return period measures of a vine copula, policymakers or regulators can establish a systematic framework for assessing the risks of extreme air pollution events. This, in turn, will empower authorities to take any necessary further action to prevent or reduce the effects of extreme air pollution events.

The remainder of this paper is structured as follows: Sect. 2 introduces the vine copula as the main method in

Table 1 API values and their corresponding health effects

API value	Health status	Health effect
0–50	Good	Low pollution with no ill effect on health
51–100	Moderate	Moderate pollution that poses no ill effect on health
101–200	Unhealthy	Worsens the health conditions of high-risk individuals with heart and lung complications
201–300	Very unhealthy	Worsens the health conditions and reduces tolerances to physical exercise in individuals with heart and lung complications; affects public health
> 300	Hazardous	Hazardous to high-risk individuals and the public health in general

this study; Sect. 3 introduces our sample data to analyze the risks associated with extreme air pollution events; Sect. 4 describes the proposed methodology applied herein; Sect. 5 presents and discusses the obtained results; and Sect. 6 provides the conclusion.

2 Vine copula

In this study, the vine copula is employed to assess the risks associated with extreme air pollution events. The vine copula proves to be more potent than standard multivariate distributions and multivariate copulas. It stands out as a more flexible and tractable multivariate copula constructed from pair (conditional) copulas as its fundamental building blocks (Czado 2019; Czado and Nagler 2022b; Joe and Kurowicka 2011).

Before diving into the discussion of the vine copula, we will first introduce a multivariate copula, particularly emphasizing a bivariate copula (or pair copula). Stemming from Sklar's theorem, a d-dimensional joint distribution F is represented as follows by a d-dimensional distribution copula C:

$$F\left(x_1, x_2, \cdots, x_d\right) = C\left(u_1, u_2, \cdots, u_d\right). \tag{1}$$

In Eq. (1), $\boldsymbol{x} = \left(x_1, x_2, \cdots, x_d\right)$ represents the original observation, and $\boldsymbol{u} = \left(u_1, u_2, \cdots, u_d\right)$ denotes a copula observation with uniformly distributed marginals. The copula observation $\boldsymbol{u}$ is derived from a transformation of the original observation $\boldsymbol{x}$, such that $u_i = F_i\left(x_i\right) \in [0, 1]$, where F_i is the marginal distribution of the component x_{i} for $i = 1, \cdots, d$. Therefore, the d-dimensional distribution copula C depends on the d-dimensional hypercube $[0, 1]^d$ (Jaworski et al. 2010; Joe 2014; Sklar 1996).

If the marginal distribution of u_i is continuous for all $i = 1, \cdots, d$, then the distribution copula function C is continuous and unique, and its corresponding density copula function c can be obtained as follows using the partial derivative:

$$c\left(u_1, u_2, \cdots, u_d\right) = \frac{\partial^d}{\partial_{u_1} \partial_{u_2} \cdots \partial_{u_d}} C\left(u_1, u_2, \cdots, u_d\right). \tag{2}$$

In addition to the above, a d-dimensional joint density f is also related to the density copula function c formulated as

$$f\left(x_1, x_2, \cdots, x_d\right) = c\left(u_1, u_2, \cdots, u_d\right) f_1\left(x_1\right) \cdots f_d\left(x_d\right). \tag{3}$$

In Eq. (3), $f_i\left(x_i\right)$ is the marginal density of the original observation's component x_{i} for $i = 1, \cdots, d$.

Therefore, a pair copula is a 2-dimensional distribution copula$C$, denoted by$C\left(u_1, u_2\right)$, with its corresponding density copula function$c$, denoted by $c\left(u_1, u_2\right)$ (Jaworski et al. 2010; Nelsen 2006). As mentioned earlier, pair (conditional) copulas serve as crucial building blocks in vine copula modeling. Previous literature introduces several parametric pair copulas with distinct features (Genest and Rivest 1993; Yan 2023). Two notable examples are Clayton and Frank copulas. The Clayton distribution copula C (Clayton 1978; Genest and Favre 2007; McNeil et al. 2015) is characterized by a parameter $\theta \in (0, \infty)$ that controls its dependence structure. The Clayton copula exhibits positive dependence, reflected in the range of its corresponding Kendall's τ correlation coefficient, which falls within$[0, 1]$. Further extension of the τ dependence range can be achieved through counterclockwise rotations of the density copula function c at 90°, 180°, and 270° angles (Czado 2019). Additionally, the Frank distribution copula (Frank 1979; McNeil et al. 2015; Nelsen 2006) is defined by a control parameter $\theta \in (-\infty, \infty)$ and$\theta \neq 0$. The Kendall's τ correlation coefficient range of the Frank copula is $[-1, 1]$, covering the entire dependency spectrum, eliminating the need for rotation approaches in this case (Czado 2019).

The Kendall's τ correlation coefficient for a pair copula is formally computed using a double integral as follows:

$$\tau = 4 \int_0^1 \int_0^1 C\left(u_1, u_2\right) dC\left(u_1, u_2\right) - 1. \tag{4}$$

In Eq. (4), τ depends solely on the underlying pair distribution copula C and proves valuable in assessing the central dependency of the observed copula data (Czado and Nagler 2022b). Two additional dependency measures, namely the upper (λ^{upper}) and lower (λ^{lower}) tail dependence coefficients, can be employed to ascertain the probabilities of the upper and lower tail parts, respectively. In the context of air pollution, λ^{upper} signifies the probability of extreme air pollution events (Czado and Nagler 2022b). The formulas for λ^{upper} and λ^{lower} are given by:

$$\lambda^{\text{upper}} = \lim_{t \to 1^-} P\left(X_2 > F_2^{-1}(t) | X_1 > F_1^{-1}(t)\right) = \lim_{t \to 1^-} \frac{1 - 2t + C(t, t)}{1 - t} \tag{5}$$

$$\lambda^{\text{lower}} = \lim_{t \to 0^+} P\left(X_2 \leq F_2^{-1}(t) | X_1 \leq F_1^{-1}(t)\right) = \lim_{t \to 0^+} \frac{C(t, t)}{t}. \tag{6}$$

As a consequence of pair copula decompositions and constructions, a d-dimensional density copula c undergoes transformation into a d-dimensional vine copula. This vine copula is a product of the $d(d-1)/2$ building blocks, incorporating pair and conditional pair copulas (Czado 2019; Czado and Nagler 2022b; Joe and Kurowicka 2011). Emphasizing the significance of vine copulas,

this transformation lays the foundation for advanced statistical modeling in various domains. For the sake of simplicity, we will focus our discussion on a three-dimensional (3D) vine copula. This choice is particularly relevant for examining three key characteristics of air pollution: severity, intensity, and duration. The 3D (simplified) density vine copula, with dimension $d = 3$, is represented as:

$$c(u_1,u_2,u_3;\boldsymbol{\theta}) = c_{13;2}\left(C_{1|2}(u_1|u_2), C_{3|2}(u_3|u_2);\theta_{13;2}\right) \times c_{23}(u_2,u_3;\theta_{23}) \times c_{12}(u_1,u_2;\theta_{12}). \tag{7}$$

Here, the control parameter $\boldsymbol{\theta} = (\theta_{13;2}, \theta_{23}, \theta_{12})$; $c_{13;2}$ is a conditional density pair copula that depends on the conditional distribution copulas $C_{1|2}(u_1|u_2)$ and $C_{3|2}(u_3|u_2)$; and c_{23} and c_{12} are the density pair copulas. The pseudo copula data $C_{1|2}(u_1|u_2)$ and $C_{3|2}(u_3|u_2)$ in Eq. (7) are computed using the h function as follows:

$$h_{1|2}(u_1|u_2) = C_{1|2}(u_1|u_2) = \frac{\partial}{\partial u_2} C_{12}(u_1,u_2), \text{ and} \tag{8}$$

$$h_{3|2}(u_3|u_2) = C_{3|2}(u_3|u_2) = \frac{\partial}{\partial u_2} C_{32}(u_3,u_2), \text{ respectively.} \tag{9}$$

These pseudo copula data $C_{1|2}(u_1|u_2)$ and $C_{3|2}(u_3|u_2)$ are crucial for selecting the appropriate conditional pair copula $c_{13;2}$ for the vine copula in Eq. (7) (Joe and Kurowicka 2011). The emphasis on vine copulas underscores their pivotal role in capturing and modeling complex dependencies within the data.

The control parameter $\boldsymbol{\theta} = (\theta_{13;2}, \theta_{23}, \theta_{12})$ for the vine copula corresponds to a set of density (conditional) pair copulas $(c_{13;2}, c_{23}, c_{12})$, and it can be optimized using a sequential approach that involves the inversion of Kendall's τ correlation coefficient (Seq-Itau) or sequential maximum likelihood estimation (Seq-MLE), and joint maximum likelihood estimation (Joint-MLE). Seq-Itau and Seq-MLE employ similar strategies to sequentially optimize the control parameters. In these strategies, the parameters for the density pair copulas c_{23} and c_{12} are first optimized using the inversion of Kendall's τ correlation coefficient (Itau) or the MLE. Subsequently, the remaining density conditional pair copula $c_{13;2}$ is optimized using either the Itau or the MLE method (Czado and Nagler 2022b).

In the Itau approach, the parameter θ is estimated as $\theta = C^{-1}(\hat{\tau})$, where C^{-1} represents the inverse function of the distribution (conditional) pair copula C, and $\hat{\tau}$ is the empirical coefficient of Kendall's τ correlation (Czado 2019). Conversely, the MLE approach is determined as follows:

$$MLE = \max_{\theta\in\Theta}\{\updownarrow(\theta;\boldsymbol{u})\}, \text{ where} \tag{10}$$

$$\updownarrow(\theta;\boldsymbol{u}) = \prod_{j=1}^{N} c_{ab}(u_{a,j}, u_{b,j};\theta), \tag{11}$$

and θ is the parameter from the possible set Θ (Czado 2019).

Differently, the Joint-MLE optimizes the parameter $\boldsymbol{\theta} = (\theta_{13;2}, \theta_{23}, \theta_{12})$ by maximizing the joint likelihood as follows:

$$\text{Joint} - \text{MLE} = \max_{\boldsymbol{\theta}\in\boldsymbol{\Theta}}\{\updownarrow(\boldsymbol{\theta};\boldsymbol{u})\}, \text{ where} \tag{12}$$

$$(\boldsymbol{\theta};\boldsymbol{u}) = \prod_{j=1}^{N} c_{13;2}\left(C_{1|2}(u_{1,j}, u_{2,j}), C_{3|2}(u_{3,j}, u_{2,j});\theta_{13;2}\right) \times c_{23}(u_{2,j}, u_{3,j};\theta_{23}) \times c_{12}(u_{1,j}, u_{2,j};\theta_{12}), \tag{13}$$

and $\boldsymbol{\theta}$ is the parameter set of the possible set $\boldsymbol{\Theta}$ (Czado and Nagler 2022b).

However, under pair copula decompositions and constructions, the vine copula produced is not unique, given the existence of two other distinct vine copulas for the case of the dimension $d = 3$ (Czado and Nagler 2022b). The other two 3D vine copulas with different decompositions are presented as:

$$c(u_1,u_2,u_3;\boldsymbol{\theta}) = c_{12;3}\left(C_{1|3}(u_1|u_3), C_{2|3}(u_2|u_3);\theta_{12;3}\right) \times c_{13}(u_1,u_3;\theta_{13}) \times c_{23}(u_2,u_3;\theta_{23}), \text{ and} \tag{14}$$

$$c(u_1,u_2,u_3;\boldsymbol{\theta}) = c_{23;1}\left(C_{2|1}(u_2|u_1), C_{3|1}(u_3|u_1);\theta_{23;1}\right) \times c_{12}(u_1,u_2;\theta_{12}) \times c_{13}(u_1,u_3;\theta_{13}). \tag{15}$$

The vine copulas described in Eqs. (7), (14), and (15) can also be visualized using a tree sequence (Czado 2019). In graph theory, a tree is an undirected graph in which any two vertices are connected by exactly one path. The set of trees $V = (T_1, T_2, \cdots, T_{d-1})$ generally represents a d-dimensional vine copula, known as a regular vine tree sequence, if it satisfies the following conditions:

1. Each tree $T_i = (N_i, E_i)$ is connected. In other words, for all nodes $a, b \in T_i$, $i = 1, 2, \cdots, d-1$, there exists a path $n_1, n_2, \cdots, n_k$, where $\{n_1, n_2, \cdots, n_k\} \subset N_i$ with $a = n_1$ and $b = n_k$.
2. The node and edge sets for the first tree T_1 are $N_1 = \{1, 2, \cdots, d\}$ and E_1, respectively.
3. For $i \geq 2$, the node and edge sets for the tree T_i are $N_i = E_{i-1}$ and E_i, respectively.

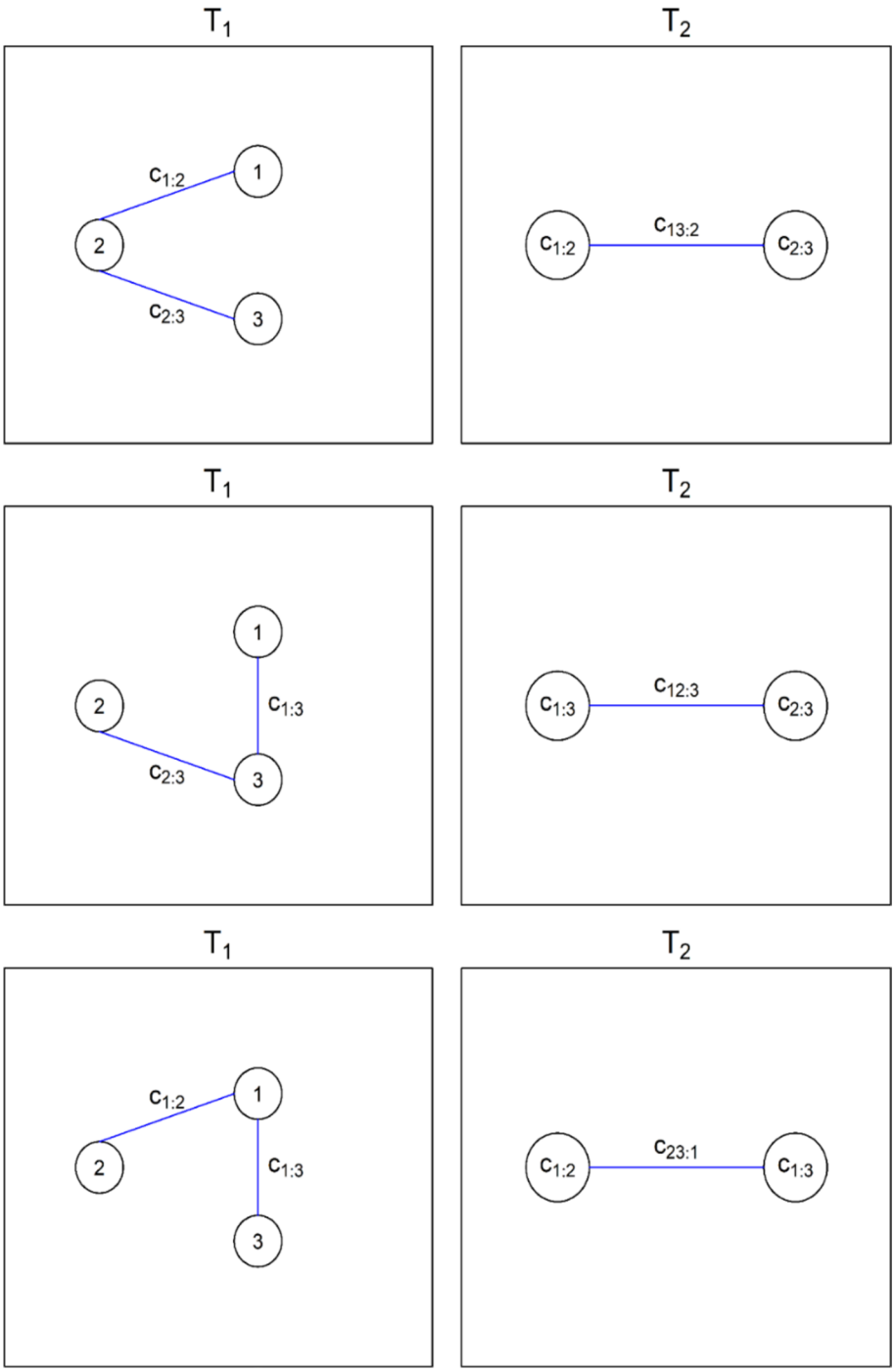

Fig. 1 The regular vine tree sequences for Eqs. (11), (18), and (19) are shown from top to bottom

4. For $i \geq 2$, the edge $\{a, b\} \in E_i$ must share a common node, such that $|a \cap b| = 1$. This condition is also known as the proximity condition (Czado 2019).

A vine copula with a regular vine tree sequence is also referred to as a regular vine copula. For example, the regular vine tree sequences for Eqs. (7), (14), and (15) are provided in the rows of Fig. 1.

Based on the specific structure provided by its tree sequence, a regular vine copula is classified as either a canonical vine copula, a drawable vine copula, or both. A regular vine copula earns the designation of a canonical vine copula if, for each tree T_i, there exists one root node $c \in N_i$, such that $\left|\{e \in E_i | c \in e\}\right| = d - i$, where d is the dimensional of the vine copula. In a canonical vine tree sequence, there exists one root node with a maximal

degree in each tree. Moreover, a regular vine copula is categorized as a drawable vine copula if, for each node $c \in N_i$, the node satisfies $\left|\left\{e \in E_i | c \in e\right\}\right| \leq 2$. Thus, every node in a drawable vine tree sequence has a degree of one or two.

The first row in Fig. 1 depicts a regular vine copula, classified as a canonical vine copula. This categorization is based on its tree sequence, which follows a root node order of 213; where Tree T_1 has a root node of 2, and Tree T_2 has a root node of either 1 or 3. Additionally, this same regular vine copula is identified as a drawable vine copula due to its tree sequence having a node order of 123. In this context, the nodes in Tree T_1 can be arranged horizontally in the order 1, 2, and 3.

3 Sample data

Similar to other countries, Malaysia is also grappling with air pollution (Manga and Awang 2018; Usmani et al. 2020). The Department of Environment (DOE) regularly monitors air pollution in Malaysia using a composite index called the API. The API is continuously measured at 19 different sites, covering rural areas like Jerantut and Kapit, urban areas such as Cheras and Shah Alam, suburban areas including Muar and Tanjung Maling, and industrial areas like Klang, Kuala Lumpur, and Kuching. The API comprises five major air pollutants: carbon monoxide (CO), ozone (O_3), nitrogen dioxide (NO_2), sulfur dioxide (SO_2), and fine particles measuring less than 10 μm (PM_{10}) (Afroz et al. 2003). CO, O_3, NO_2, and SO_2 are measured in the parts per million (ppm), while PM_{10} is measured in micrograms per cubic meter (μg/m^3).

The calculation of the Air Pollution Index (API) is rooted in the Pollution Standard Index (PSI), a globally accepted standard endorsed by the United States Environmental Protection Agency (USEPA). The method involves computing the average for each pollutant within distinct time frames, as different exposure periods are deemed acceptable for human health, resulting in varied concentration breakpoints. The exposure periods for different pollutants are as follows: CO (8 h), O_3 (8 h), NO_2 (1 h), SO_2 (1 h), PM_{10} (24 h). To standardize the average concentration of each pollutant over the specified period, a specific mathematical formula is applied, generating a non-unitary value known as a sub-index. The formulas are provided below:

$$std(\mathrm{CO}) = \begin{cases} \mathrm{CO} \times 11.11111, & \text{if } \mathrm{CO} < 9 \text{ ppm}, \\ 100 + \{[\mathrm{CO} - 9] \times 16.66667\}, & \text{if } 9 \leq \mathrm{CO} < 15 \text{ ppm}, \\ 200 + \{[\mathrm{CO} - 15] \times 6.66667\}, & \text{if } 15 \leq \mathrm{CO} < 30 \text{ ppm}, \\ 300 + \{[\mathrm{CO} - 30] \times 10\} & \text{if } \mathrm{CO} \geq 30 \text{ ppm}. \end{cases} \tag{16}$$

$$std(\mathrm{O_3}) = \begin{cases} \mathrm{O_3} \times 1000, & \text{if } \mathrm{O_3} < 0.2 \text{ ppm}, \\ 200 + \left\{\left[\mathrm{O_3} - 0.2\right] \times 500\right\}, & \text{if } 0.2 \leq \mathrm{O_3} < 0.4 \text{ ppm}, \\ 300 + \left\{\left[\mathrm{O_3} - 0.4\right] \times 1000\right\}, & \text{if } \mathrm{O_3} \geq 0.4 \text{ ppm}. \end{cases} \tag{17}$$

$$std(\mathrm{NO_2}) = \begin{cases} \mathrm{NO_2} \times 588.23529, & \text{if } \mathrm{NO_2} < 0.17 \text{ ppm}, \\ 100 + \left\{\left[\mathrm{NO_2} - 0.17\right] \times 232.56\right\}, & \text{if } 0.17 \leq \mathrm{NO_2} < 0.6 \text{ ppm}, \\ 200 + \left\{\left[\mathrm{NO_2} - 0.6\right] \times 166.667\right\}, & \text{if } 0.6 \leq \mathrm{NO_2} < 1.2 \text{ ppm} \\ 300 + \left\{\left[\mathrm{NO_2} - 1.2\right] \times 250\right\}, & \text{if } \mathrm{NO_2} \geq 1.2 \text{ ppm}. \end{cases} \tag{18}$$

$$std(\mathrm{SO_2}) = \begin{cases} \mathrm{SO_2} \times 2500, & \text{if } \mathrm{SO_2} < 0.04 \text{ ppm}, \\ 100 + \left\{\left[\mathrm{SO_2} - 0.04\right] \times 384.61\right\}, & \text{if } 0.04 \leq \mathrm{SO_2} < 0.3 \text{ ppm}, \\ 200 + \left\{\left[\mathrm{SO_2} - 0.3\right] \times 333.333\right\}, & \text{if } 0.3 \leq \mathrm{SO_2} < 0.6 \text{ ppm}, \\ 300 + \left\{\left[\mathrm{SO_2} - 0.6\right] \times 500\right\}, & \text{if } \mathrm{SO_2} \geq 0.6 \text{ ppm}. \end{cases} \tag{19}$$

$$std(\mathrm{PM_{10}}) = \begin{cases} \mathrm{PM_{10}}, & \text{if } \mathrm{PM_{10}} < 50 \text{ g/m}^3, \\ 50 + \left\{\left[\mathrm{PM_{10}} - 50\right] \times 0.5\right\}, & \text{if } 50 \leq \mathrm{PM_{10}} < 350 \ \mu\text{g/m}^3, \\ 200 + \left\{\left[\mathrm{PM_{10}} - 350\right] \times 1.4286\right\}, & \text{if } 350 \leq \mathrm{PM_{10}} < 420 \ \mu\text{g/m}^3, \\ 300 + \left\{\left[\mathrm{PM_{10}} - 420\right] \times 1.25\right\}, & \text{if } 420 \leq \mathrm{PM_{10}} < 500 \ \mu\text{g/m}^3, \\ 400 + \left[\mathrm{PM_{10}} - 500\right], & \text{if } \mathrm{PM_{10}} \geq 500 \ \mu\text{g/m}^3. \end{cases} \tag{20}$$

The API reading is then determined by selecting the highest relative sub-index. Notably, in Malaysia, API readings are often influenced by the concentration of particulate matter, which tends to be the dominant pollutant, especially during episodes of haze.

In this study, API data for Klang were acquired from the DOE and selected for analysis using the vine copula approach. Globally, Klang ranks as the 13th and 16th busiest city for transshipment and container ports, respectively (AL-Dhurafi et al. 2018b). With trade activities amounting to millions of ringgits, Klang stands as the primary industrial center of Malaysia. However, despite its economic prominence, Klang faces more severe air pollution episodes than other major cities, a consequence of rapid industrialization and dense urbanization (Masseran et al. 2016; Masseran and Safari 2020a). The selection of Klang as the study area is deliberate, driven by its propensity to experience more frequent and intense air pollution events compared to other cities. This strategic choice ensures the availability of a more substantial number of unhealthy air pollution events, providing enhanced sample data for understanding Klang's air pollution situation and facilitating more accurate risk assessments for extreme air pollution events. Figure 2 illustrates the hourly API data spanning from January 1, 1997, to August 31, 2020.

Utilizing the obtained API data, unhealthy air pollution events were initially identified using a threshold at the 100 API level, indicating that the air quality is deemed hazardous to health conditions, see Table 1 (AL-Dhurafi et al. 2018a). The escalation in the API value, as outlined in Table 1, corresponds proportionally to the impact on public health status.

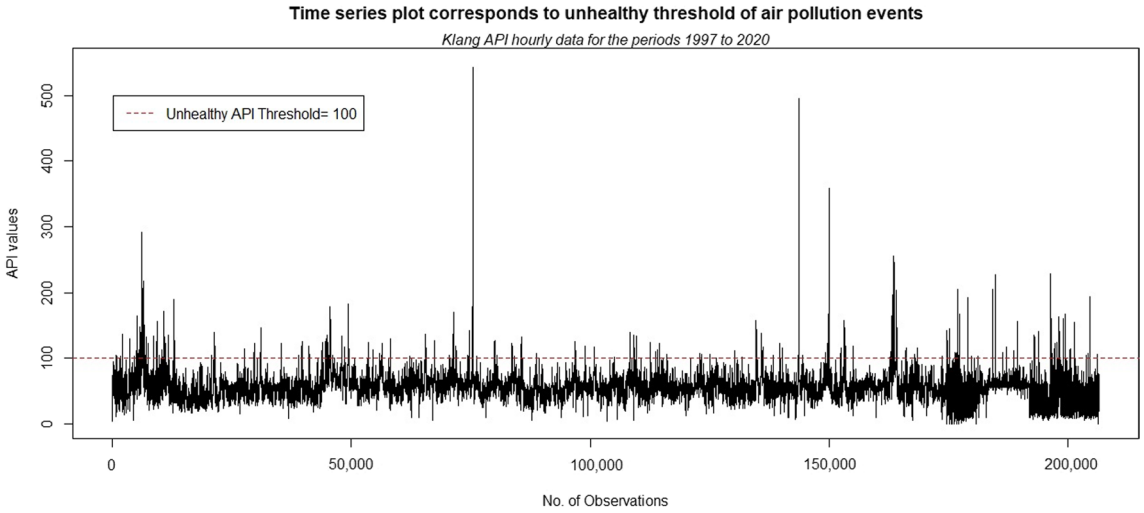

Fig. 2 API in Klang and the threshold indicating unhealthy air pollution events

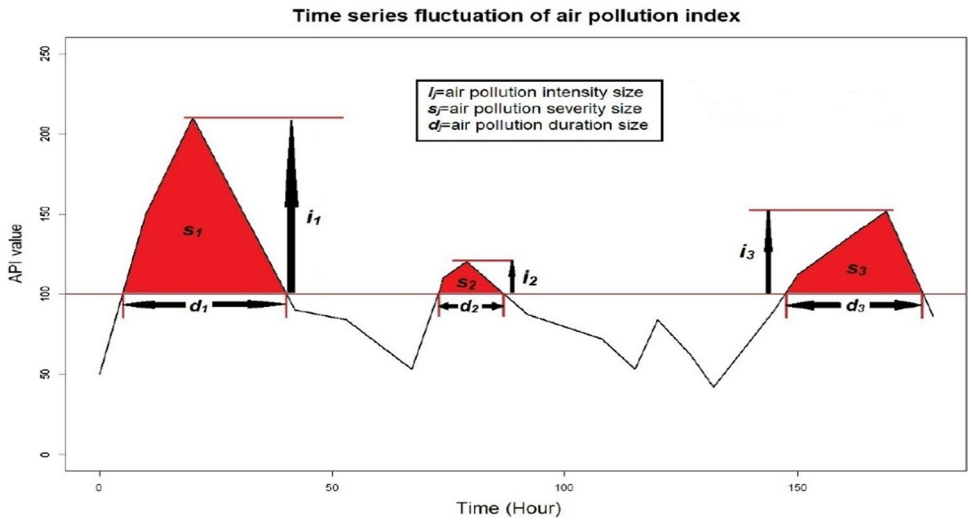

Fig. 3 Determining the duration, severity, and intensity of the first three unhealthy air pollution events (highlighted in red)

This allows for a comprehensive assessment and control of risks associated with extreme air pollution events, particularly noteworthy at an API exceeding 300, signifying a potential public health crisis. Malaysia has declared emergencies during severe haze incidents, where the current API has surpassed 500 (Aghamohammadi and Isahak 2018; Othman et al. 2014).

Consequently, unhealthy air pollution events were identified as nonoverlapping periods during which API values exceeded 100. Subsequently, severity, intensity, and duration were computed for each unhealthy air pollution event. Let $API = \{x_t | t \in T\}$ represent the obtained API data, where x_t is the API value at time t, T is the index period, and $UE_j = \{x_t | t \in T_j\} \subset API$ represents the j th recorded unhealthy air pollution event. For $j = 1, 2, \cdots, N$, the period for the j th unhealthy air pollution event $T_j = \{t | x_t > 100\} \subset T$. The corresponding severity, intensity, and duration for the j th period T_j were determined as follows:

$$sev_j = \sum_{t \in T_j} x_t (\text{sum of all API values within the period} T_j), \tag{21}$$

$$int_j = \max_{t \in T_j} \{x_t\} (\text{maximum API value within the period} T_j), \text{ and} \tag{22}$$

$$dur_j = |T_j| (\text{cardinality of the period} T_j). \tag{23}$$

In this study, the value of N is 301, aligning with the 301 unhealthy air pollution events recorded in Klang from January 1, 1997, to August 31, 2020. Figure 3 demonstrates the process of determining the severity, intensity, and duration of the first three unhealthy air pollution events.

4 Methodology

This study explored three characteristics—intensity, duration, and severity—associated with unhealthy air pollution events, obtained from the API. Starting with these characteristics' data, Fig. 4 illustrates the research flow methodology used in this study to perform risk assessments for extreme air pollution events in Klang, Malaysia using vine copula.

For simplicity, the intensity, duration, and severity data are denoted as $X_1 = \{x_{1,j} | j \in I\}$, $X_2 = \{x_{2,j} | j \in I\}$, and $X_3 = \{x_{3,j} | j \in I\}$ respectively, where $I = \{1, 2, \cdots, N\}$ is an indexing set. To facilitate vine copula modeling, the original data were transformed into copula data through the probability integral transformation (PIT). By applying the PIT, for $k = 1, 2, 3$ and $j \in I$, the j th copula observation for the k th data is represented as $u_{k,j} = \widehat{F}_k(x_{k,j})$, where $\widehat{F}_k$ is the empirical distribution function defined as

$$\widehat{F}_k(x_{k,j}) = \frac{1}{\mathrm{N}+1} \sum_{\mathrm{j}=1}^{N} 1_{\{x \le x_{k,j}\}}, \text{ for all} x \in X_k, \tag{24}$$

and $x_{k,j}$ denotes the j th original observation for the k -th data. Consequently, copula data $U_1 = \{u_{1,j} | j \in I\}$, $U_2 = \{u_{2,j} | j \in I\}$, and $U_3 = \{u_{3,j} | j \in I\}$ were obtained for intensity, duration, and severity, respectively.

In this study, preliminary analysis was conducted to initially examine the original data and copula data. For the original data X_1, X_2, and X_3, descriptive statistics were applied to provide early statistical information about the data. Additionally, the original data was also analyzed

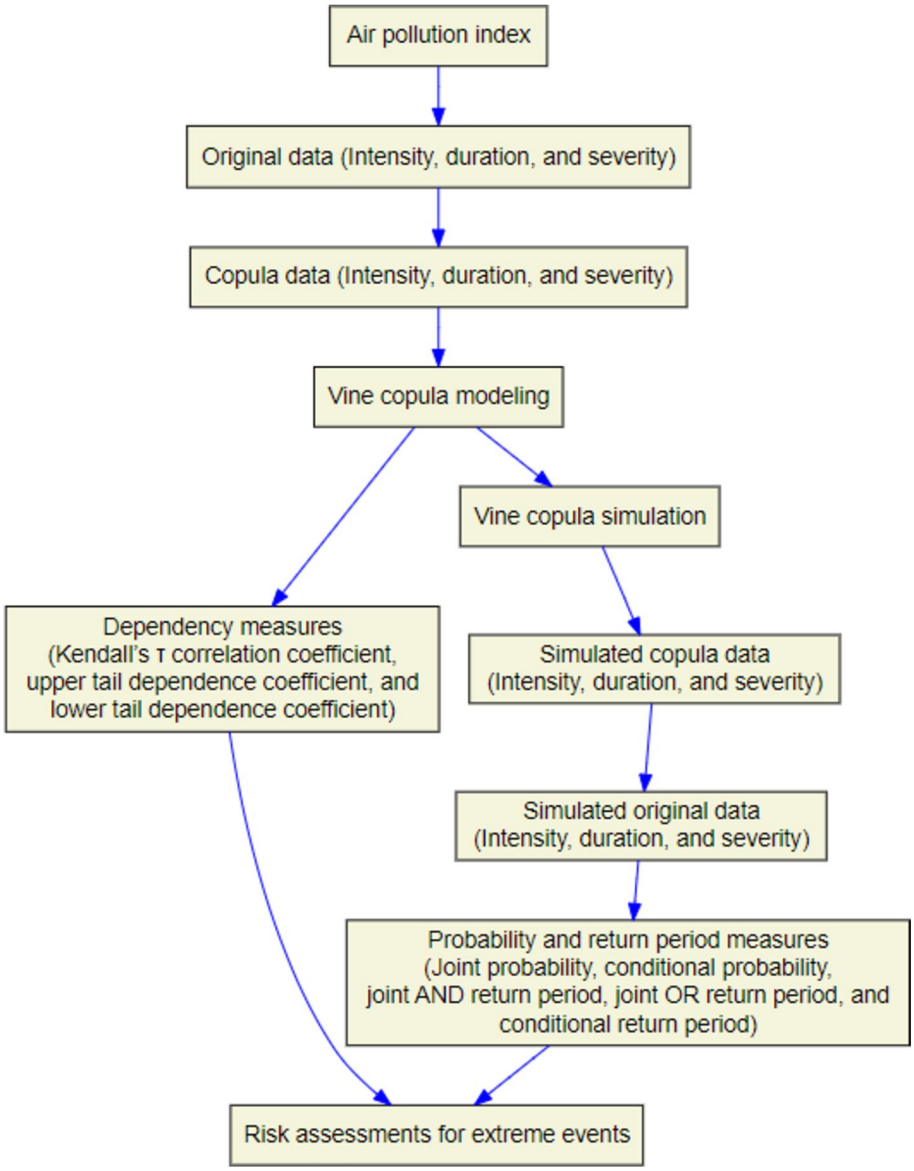

Fig. 4 Research methodology flow

using histograms and probability fits from nonparametric density functions such as the empirical probability density function (epdf) (Mendenhall et al. 2012) and kernel density estimator (kde) with different kernels (Gaussian, cosine, optocosine, and triangular) (Sen 2011; Silverman 2018). To assess whether these density functions exhibited a similar continuous one-dimensional probability distribution, further analysis, including QQ plots and the Kolmogorov–Smirnov test, was carried out. Using the obtained copula data U_1, U_2, and U_3, preliminary analyses were then performed using marginal histograms, pair plots, empirical Kendall's τ correlation coefficients, and normalized contour plots. These analyses were conducted to understand the marginal and pairwise dependency behaviors provided by the obtained copula data. In worth to note here that the empirical distribution function used in this study to obtain copula data, being a nonparametric model, did not require any parameter estimation process prior to its fitting. The reasons for applying the empirical distribution function have been mentioned in the fifth paragraph of the next section.

The appropriate regular vine tree sequence for the copula data obtained in vine copula modeling was determined using Dissman's structure selection algorithm. Dissman's algorithm employed Kendall's τ correlation coefficient to measure the strength of dependence and constructed each tree in a regular vine tree sequence by fitting the strongest dependencies first. In this algorithm, the maximal spanning tree algorithm was used with weights denoted by empirical Kendall's τ correlation coefficients between any copula data pair to build the tree T_1. Consequently, the pair copula data with the highest weight were sequentially chosen to form the structure of tree T_1.

It is crucial to highlight that, in the context of vine copula modeling, a two-step parameter estimation process was essential. Initially, the estimation process involved determining parameters for pair copula models, which function as components within the vine copula model. The first parameter estimation process is discussed in the following paragraph. Subsequently, another parameter estimation process was undertaken to optimize the parameters of the vine copula model itself, utilizing the Joint-MLE Eq. (12), as elaborated upon in the 8th paragraph. Returning to the specifics of vine copula modeling, the most well-fitted pair copula models for all the pair copula data were determined once tree T_1 was formed. All the parametric pair copula models listed in Table 2 were considered to determine the most well-fitted pair copula model.

In the initial parameter estimation process, we fitted all models from Table 3 to each pair of copula data in tree T_1. The parameters of these models were then optimized using the MLE method Eq. (10). Subsequently, all fitted models were compared based on their Akaike's Information Criterion (AIC) scores to determine the most well-fitted pair copula model. The model with the lowest AIC score was considered superior to the others and classified as the most well-fitted pair copula model. The AIC of a pair copula model was measured as follows:

Table 2 List of considered parametric pair copula models

Number	Copula model	Number of parameter (s)
1	Independence	0
2	Gaussian	1
3	Student *t*	2
4	Clayton	1
5	Gumbel	1
6	Frank	1
7	Joe	1
8	BB1	2
9	BB6	2
10	BB7	2
11	BB8	2
12	Survival Clayton	1
13	Survival Joe	1
14	Survival BB6	2
15	Survival BB8	2
16	Rotated Clayton 90 degrees	1
17	Rotated Gumbel 90 degrees	1
18	Rotated Joe 90 degrees	1
19	Rotated BB1 90 degrees	2
20	Rotated BB6 90 degrees	2
21	Rotated BB7 90 degrees	2
22	Rotated BB8 90 degrees	2
23	Rotated Clayton 270 degrees	1
24	Rotated Gumbel 270 degrees	1
25	Rotated Joe 270 degrees	1
26	Rotated BB1 270 degrees	2
27	Rotated BB6 270 degrees	2
28	Rotated BB7 270 degrees	2
29	Rotated BB8 270 degrees	2
30	Tawn type 1	2
31	Rotated Tawn type 1 180 degrees	2
32	Rotated Tawn type 1 90 degrees	2
33	Rotated Tawn type 1 270 degrees	2
34	Tawn type 2	2
35	Rotated Tawn type 2 180 degrees	2
36	Rotated Tawn type 2 90 degrees	2
37	Rotated Tawn type 2 270 degrees	2

Table 3 Descriptive statistics of severity, intensity, and duration data

Variable	Mean	Median	Min. Value	Max. Value	SD	Skewness	Kurtosis
Intensity	125.11	112	100	543	44.77	5.61	44.97
Severity	2241.76	231.27	100	36,677	4948.3	3.92	20.92
Duration	16.74	2	1	224	31.91	3.24	15.73

$$AIC = -2\sum_{\mathrm{j}=1}^{N} \ln\left[c\left(u_{i,j}, u_{i,j};\theta\right)\right] + 2k, \tag{25}$$

where k is the number of parameters of the considered pair copula.

All possible edges allowed by the proximity condition were considered in constructing the tree T_2. Each possible edge had an associated pair of pseudo copula data, wherein two pseudo copula data were obtained using the h function Eqs. (8) and (9). For each pair of pseudo copula data, the corresponding Kendall's τ correlation coefficient was determined and used as a weight for selecting the maximal spanning tree for tree T_2. Subsequently, all of the most well-fitted conditional pair copulas in tree T_2 were selected using the AIC Eq. (25), wherein their parameters were formerly optimized by the MLE Eq. (10).

Vine copula modeling was completed by implementing the RVineStructureSelect function provided in the R-package VineCopula. After establishing the regular vine tree sequence and identifying the best-fitted (conditional) pair copula models, the second parameter estimation process involved optimizing the obtained vine copula model using the Joint-MLE Eq. (12) to determine the most appropriate parameters for the model. Refs. (Czado 2019; Czado and Nagler 2022a) reported that the Joint-MLE provides better parameter estimates than the other two alternative approaches of Seq-Itau and Seq-MLE (Sect. 2). The Joint-MLE computation was performed using R-package VineCopula through the RVineMLE function.

The log-likelihood, AIC, and Bayesian's information criterion (BIC) of the optimized vine copula were computed to observe its fitting performance. For the copula observations $\boldsymbol{u_j} = \left(u_{1,j}, u_{2,j}, u_{3,j}\right)$, the log-likelihood of the 3D vine copula with two trees (T_1 and T_2) and corresponding edge sets (E_1 and E_2) was determined as follows:

$$\begin{aligned} loglik = \sum_{j=1}^{N}\sum_{l=1}^{2}\sum_{E_1}^{E_2} ln\big[c_{r(e),s(e);D(e)}\left(C_{r(e)|D(e)}\left(u_{r(e)}|u_{D(e)}\right),\right. \\ \left. C_{s(e)|D(e)}\left(u_{s(e)}|u_{D(e)}\right);\theta_{r(e),s(e);D(e)}\right)\big], \end{aligned} \tag{26}$$

where $c_{r(e),s(e);D(e)}$ denotes the bivariate copula density associated with an edge e and parameter $\theta_{r(e),s(e);D(e)}$. The two pseudo copula data $C_{r(e)|D(e)}$ and $C_{s(e)|D(e)}$ in Eq. (26) were determined using the h function defined in Eqs. (8) and (9).

The formula for the AIC of the 3D vine copula with the log-likelihood [$loglik$, Eq. (26)] is given as follows:

$$AIC = -2(loglik) + 2k. \tag{27}$$

The BIC of the 3D vine copula with the log-likelihood [$loglik$, Eq. (26)] is presented as

$$BIC = -2(loglik) + log(N)(k), \tag{28}$$

where N is the number of copula observations. In Eqs. (27) and (28), k is the number of parameters of the 3D vine copula. The log-likelihood, AIC, and BIC were estimated herein using R-package VineCopula through the RVineLogLik, RVineAIC, and RVineBIC functions, respectively (Schepsmeier et al. 2015).

Based on the obtained vine copula model, the dependency between intensity, duration, and severity was determined using Kendall's t correlation coefficient. Kendall's t correlation of the obtained vine copula was determined as follows:

$$\tau = \frac{1}{\binom{d}{2}} \sum_{r \neq s} \tau_{r,s}, \tag{29}$$

where $\tau_{r,s}$ is Kendall's τ correlation coefficient of a pair copula or a conditional pair copula Eq. (4). Another important information related to dependency that can be determined by a vine copula are the coefficients of the upper tail and lower tail dependencies obtained as follows:

$$\lambda^{\text{upper}} = \frac{1}{\binom{d}{2}} \sum_{r \neq s} \lambda_{r,s}^{upper}, \text{and} \tag{30}$$

$$\lambda^{\text{lower}} = \frac{1}{\binom{d}{2}} \sum_{r \neq s} \lambda_{r,s}^{lower}, \tag{31}$$

where $\lambda_{r,s}^{upper}$ and $\lambda_{r,s}^{lower}$ are the upper and lower tail dependence coefficients, respectively, of a pair copula or a conditional pair copula Eqs. (5) and (6). In practice, based on the obtained vine copula model, the coefficients of Eqs. (29)–(31) can be applied to measure the central dependency, the probability of the joint occurrence of extremely large values, and probability of the joint occurrence of extremely small values, respectively.

The obtained vine copula was also applied to perform a simulation, utilizing the Rosenblatt transform and its inverse (Czado and Nagler 2022b). In the simulation, the Rosenblatt transform initially mapped a copula vector $\boldsymbol{U} = (u_1, u_2, u_3)$ with a vine copula C into another corresponding vector $\boldsymbol{V} = (v_1, v_2, v_3) = R(\boldsymbol{U})$ that containing independent uniform variables. The transformation function R is given as follows:

$$V_i = C_{i|i-1,\cdots,1}(U_i|U_{i-1}, \cdots, U_i), \tag{32}$$

for $i \in \{1, 2, 3\}$, where $C_{i|i-1,\cdots,1}$ is the conditional distribution of U_i given $\boldsymbol{U}_D = U_{i-1}, \cdots, U_i$. Secondly, the corresponding inverse operation $R^{-1}(\boldsymbol{V}) = \boldsymbol{U}$ converts independent uniform variables $\boldsymbol{V}$ into a vector $\boldsymbol{U}$ with the vine copula C. This inverse operation is provided as

$$U_i = C^{-1}_{i|i-1,\cdots,1}(U_i|U_{i-1}, \cdots, U_i), \tag{33}$$

for $i \in \{1, 2, 3\}$. During computation, the conditional distributions and inverses in the transformations were effectively computed using recursion over the conditional distributions associated with the pair copulas in the model (Czado 2019). The same technique was employed to conditionally stimulate from a vine copula model, provided that the required conditional distributions can be expressed by pair copula terms without needing integration (Czado and Nagler 2022b). The simulation process in this study was conducted using the RVineSim function of R-package VineCopula, resulting in the generation of 20 datasets (Schepsmeier et al. 2015). Each dataset contained the three characteristics of unhealthy air pollution events, namely intensity, duration, and severity, with each characteristic having a sample length of 1000 stimulated copula observations.

All 60,000 stimulated copula observations were transformed back into the corresponding 60,000 observations in the original scale through discrete inverse sampling. Let us assume that X (a set of N observations for each intensity, duration, and severity) was a discrete random variable, such that the probability $P(X = x_j) = p_j$. Given a copula observation $u \in U$, such that $U \sim uniform(0, 1)$, through discrete inverse sampling, we could search for index q as follows:

$$\sum_{j=1}^{q-1} p_j \le u < \sum_{j=1}^{q} p_j. \tag{34}$$

Therefore, the corresponding observation in the original scale (i.e., the scale of the random variable X) for the copula observation u was $x_q \in X$, which was observed at the q position, where X was sorted in an ascending order based on the probability p_j values. Consequently, the corresponding 20 new datasets were obtained. Each dataset contained 1000 observations for intensity, duration, and severity in their original scale.

The five risk measures related to extreme air pollution events, namely, joint probability, conditional probability of severity based on the given intensity and duration, joint OR return period of unhealthy air pollution event characteristics, joint AND return period of unhealthy air pollution event characteristics, and conditional return period of severity based on the given intensity and duration, were determined for each dataset. In practice, the first two probability measures could benefit the authorities by providing them with useful information related to the behaviors of the air pollution severity based on certain levels of air pollution intensity and duration, particularly during the extreme air pollution event. The remaining return period measures were implemented to estimate the return periods of unhealthy air pollution events. These return period measures served as a basis for planning and developing monitoring systems to manage the risk associated with the co-occurrence of the average recurrence interval time and the unhealthy air pollution event,, particularly during the extreme air pollution event.

The joint probability of severity, intensity, and duration, conditional probability of severity based on the given intensity and duration, joint OR return period of unhealthy air pollution event characteristics, joint AND return period of unhealthy air pollution event characteristics, and conditional return period of severity based on the given intensity and duration for each dataset were computed as follows:

$$P_{jointAND} = P(S \ge sev, I \ge int, D \ge dur), \tag{35}$$

$$P_{S|I,D} = P(S \le sev|I \ge int, D \ge dur), \tag{36}$$

$$RP_{jointOR} = \frac{E(L)}{P(S \ge sevorI \ge intorD \ge dur)}, \tag{37}$$

$$RP_{jointAND} = \frac{E(L)}{P(S \ge sev, I \ge int, D \ge dur)}, \text{and} \tag{38}$$

$$RP_{S|I,D} = \frac{E(L)}{P(I \ge int, D \ge dur)} \times \frac{1}{P(S \ge sev, I \ge int, D \ge dur)}, \tag{39}$$

where $E(L)$ is the expected unhealthy air pollution event interarrival time that can be estimated from the N observed unhealthy air pollution events. The probability involved in Eqs. (35)–(39) was empirically computed as the chance of an air pollution event to occur based on the simulated dataset.

The values of parameters *sev*, *int*, and *dur* in Eqs. (35)–(39), which accordingly corresponded to the levels of severity, intensity, and duration, were varied at some values. The level for *sev* was specifically varied in six different values as $sev \in \{100, 1000, 10,000, 20,000, 30,000, 35,000\}$. Meanwhile, the level for *int* was altered to six different values, such that $int \in \{100, 180, 260, 340, 420, 500\}$. The *dur* level was also changed by six different values, which resulted in $dur \in \{1, 45, 89, 133, 177, 221\}$. All possible

level combinations associated with the three characteristics (sev, int, dur) were applied to compute the risk measures. A total of 216 combinations for the triple (sev, int, dur) were obtained.

Therefore, all risk measures in Eqs. (35)–(39) were computed for each dataset and combination of (sev, int, dur). For each combination of (sev, int, dur) and each risk measure, the statistical metrics such as average, median, first quartile, third quartile, and interquartile range (IQR) were then determined based on the corresponding risk measure values obtained from 20 datasets. Concerning practicality in these statistical metrics, the computed values were valuable in providing new insights and inferences regarding the behaviors of extreme air pollution events in the most affected areas, especially in Klang, Malaysia.

5 Results

In this study, the investigation focused on modeling the characteristics of unhealthy air pollution events—specifically, intensity, duration, and severity—to assess the risks associated with extreme air pollution events, employing a vine copula model. Before the vine copula modeling, a preliminary analysis utilizing descriptive statistics for severity, intensity, and duration was conducted to offer early insights into the characteristics. Table 3 displays the descriptive statistics for severity, intensity, and duration data.

In Table 3, both the mean and median values for all characteristics leaned towards the minimum value rather than the maximum value. This observation implies that the majority of the data concentrated on the left side of the range, rather than being evenly distributed across the intermediate value. The standard deviation indicated a notable deviation from the mean, signifying that the data were spread over a broader range compared to the mean. Moreover, the data exhibited skewness, with a majority of values accumulating at extremely low levels, as evidenced by the skewness measure. The kurtosis measure further underscored an asymmetrical distribution. Figure 5 visually illustrates frequency plots for the intensity, duration, and severity data, emphasizing their skewed and asymmetrical characteristics.

In this study, nonparametric models, specifically the empirical probability density function (epdf) and the kernel density estimator (kde) utilizing various kernel functions (Gaussian, cosine, optocosine, and triangular), were applied to the characteristics data. The analysis involved fitting these models, comparing them through histogram plots and probability density fits, and employing the QQ plot and the two-sample Kolmogorov–Smirnov test. Results indicated that the kde with a Gaussian kernel exhibited a distribution similar to the epdf. However, the kde with other kernel functions (cosine, optocosine, and triangular) demonstrated

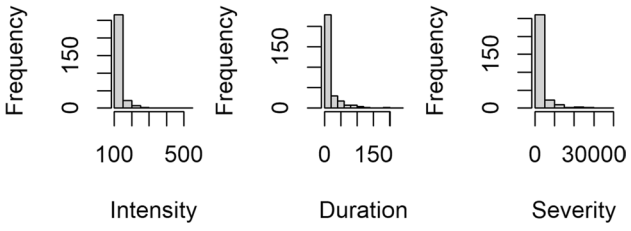

Fig. 5 Frequency plots for the intensity, duration, and severity data

distinct distributions from the epdf. A thorough analysis of these nonparametric models is provided in our supporting document. Despite the considerations regarding the kde, the empirical distribution function was chosen in this study to generate copula data. This decision was influenced by three main reasons:

- Reliability of parametric distributions: Parametric distributions such as Exponential, Gamma, Lognormal, and Weibull distributions were reported as unreliable for representing similar data (Masseran 2021a). This led to the suggestion of the empirical distribution function as a suitable alternative for multivariate copula modeling.
- Computational ease and inverse function possession: The empirical distribution function is easy to compute and possesses an inverse function, specifically the discrete inverse sampling method.
- Support from literature: The current review paper and introduction book in vine copula modelling (Czado 2019; Czado and Nagler 2022a) also recommended the empirical distribution function for this purpose.

As a result, the original data for each characteristic underwent transformation into the corresponding copula data using the PIT approach with an empirical distribution function Eq. (24). The diagonal of Fig. 6 illustrates the marginal distributions of the copula data. When focusing on the marginal distributions, the plots located on the diagonal reveal that the copula data are primarily distributed in an approximately uniform manner. These copula data served as the input for the vine copula model. Consequently, this model has a tri-variate dependence structure represented as a function that does not depend on the marginal effects.

For the preliminary analysis of the copula data, each pair of copula data underwent analysis based on pair distributions, normalized contour plots, and Kendall's τ correlation coefficients. The distributions and Kendall's τ correlation coefficients for each pair were presented above the diagonal in Fig. 6, while the normalized contour plots were positioned below the diagonal in the same figure. Upon focusing on the pair distributions and Kendall's τ correlation coefficients, all pair data exhibited a positive dependence, signifying

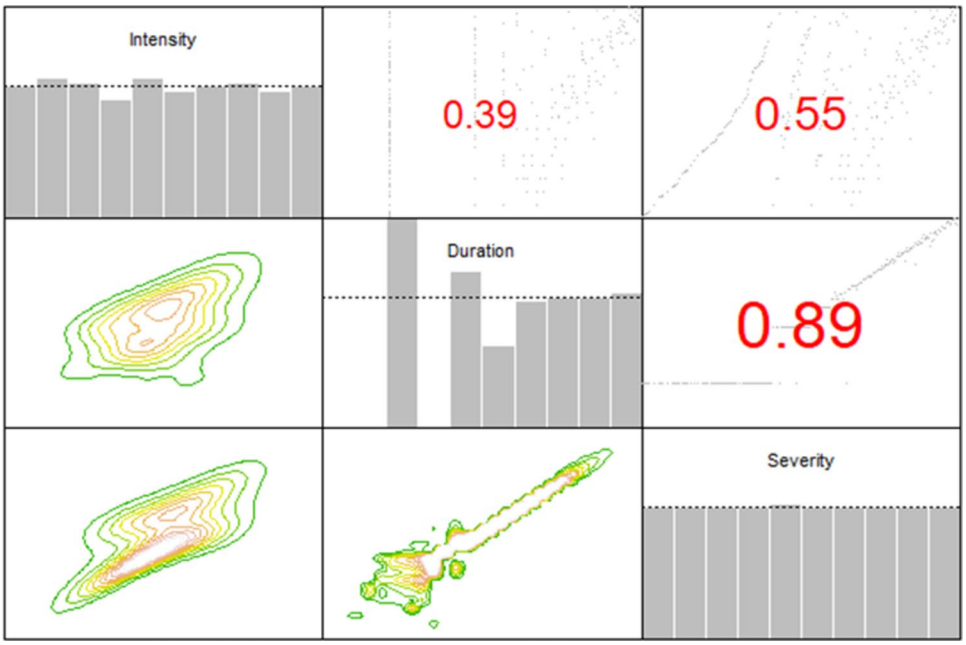

Fig. 6 Marginal distributions (along the diagonal), pair distributions with Kendall's τ correlation coefficient (above the diagonal), and normalized contour plots (below the diagonal) of the copula data of intensity, duration, and severity

that these pairs were positively correlated and tended to move in a similar rank. The relationship between severity and duration showed the strongest positive dependence of 0.89, followed by the association of intensity and severity (0.55). The lowest correlation was observed in the relationship between intensity and duration (0.39). The normalized contour plots indicated that all pairs displayed nonelliptical symmetry, suggesting the consideration of asymmetric pair copula models.

The vine copula for the copula data was subsequently developed using Dissman's structure selection algorithm, specifically the maximal spanning tree algorithm with Kendall's τ correlation coefficients as weights for each tree structure. Given that the data used in this study were three-dimensional, the vine copula comprised two trees. The first tree had three nodes and two edges, while the second tree had two nodes with one connecting edge (Fig. 1). Following Dissman's algorithm, the highest Kendall's tau correlation coefficient, representing the strongest dependency, was initially chosen to develop the tree structure. In Fig. 6, the correlation coefficients provided by Kendall's τ in ascending order were 0.89 (duration and severity), 0.55 (intensity and severity), and 0.39 (intensity and duration). Consequently, the first tree structure involved three nodes represented by the three obtained copula data, and its two edges were represented by the pair relationships of (duration and severity) and (intensity and severity) due to their higher correlations. The first column of Fig. 7 illustrates the structure of Tree 1.

Following the construction of Tree 1, an appropriate pair copula model to fit each pair relationship within this tree was determined using the MLE and the AIC [Eqs. (10) and (25), respectively]. The MLE was employed to optimize parameter selection for all considered parametric pair copula models (Table 2), while the AIC was utilized to select the best model to represent each pair. As depicted in Fig. 7, the best-fitted models for the pair relationships of (duration and severity) and (intensity and severity) were identified as the Joe (J) and Rotated Tawn type 2, 180 degrees (Tawn2_180) copulas, respectively. Table 4 provides details on these two pair copula models. For the pair relationship of (duration and severity), Table 4 indicates that the optimal parameter for the Joe copula is 11.81. The pair exhibited a strong dependency of 0.85, with the model scoring 441.84 in the MLE and -881.68 in the AIC. The Rotated Tawn type 2 180 degrees for the pair relationship of (intensity and severity) had two optimized parameters of 4.70 and 0.58, a moderate dependency of 0.49, an MLE score of 159.57, and an AIC score of -315.14.

The second tree featured only one possible conditional pair, eliminating the need for Dissman's algorithm. In contrast to the previous tree, the second tree structure comprised two nodes, representing the two pair relationships in the first tree. The edge connecting these nodes represented the conditional pair relationship between them. The second column of Fig. 7 illustrates Tree 2. The *h* function, MLE, and AIC were employed to determine the most well-fitted conditional pair copula model for describing the conditional pair relationship in Tree 2. The *h* function generated the pseudo copula data before modeling was carried out by any considered parametric pair copula model listed in Table 2. The MLE and the AIC were implemented with the same objective mentioned earlier. Table 5 indicates that the best-fitted model for the conditional relationship is the Rotated BB8 90 degrees copula. This model had

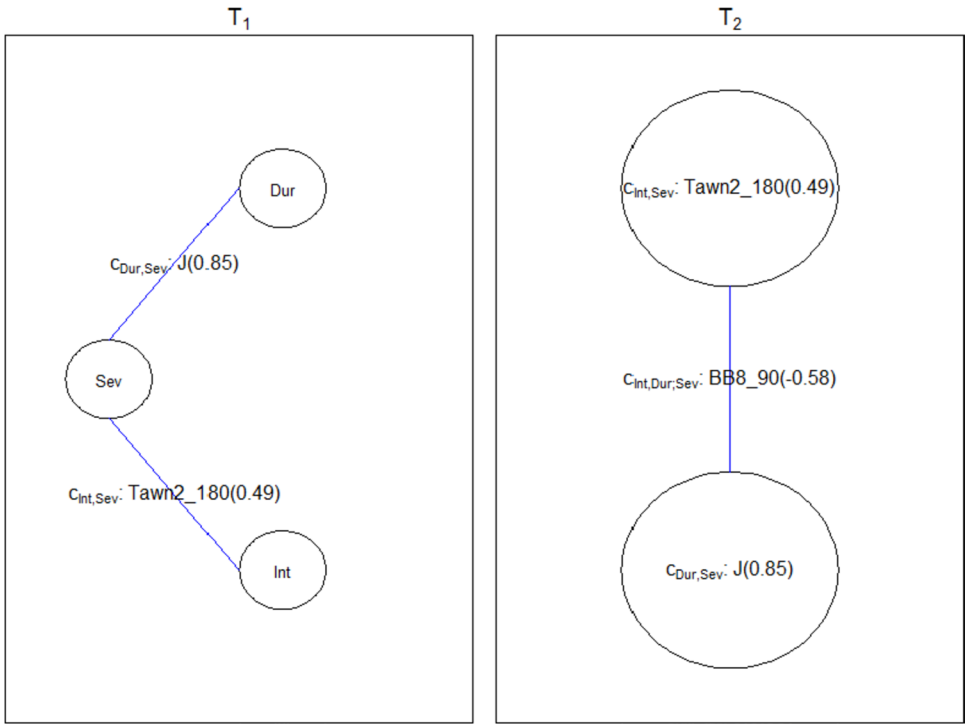

Fig. 7 Regular vine tree sequence of the obtained vine copula model

Table 4 Best-fitted pair copulas in Tree 1

Tree	Pair	The best copula model	Par	Par. 2	τ	MLE	AIC
1	(Sev, Dur)	Joe (J)	11.81	–	0.85	441.84	− 881.68
1	(Int, Sev)	Rotated Tawn type 2 180 degrees (Tawn2_180)	4.70	0.58	0.49	159.57	− 315.14

Table 5 Best-fitted conditional pair copula in Tree 2

Tree	Conditional pair	The copula best model	Par	Par. 2	τ	MLE	AIC
2	(Dur, Int; Sev)	Rotated BB8 90 degrees (BB8_90)	− 5.68	− 0.80	− 0.58	126.80	− 249.60

Table 6 Comparisons of the two optimization methods employed

Optimization method	Log-likelihood	AIC	BIC	Vuong test	p-value of Vuong test
Joint-MLE	730.14	− 1450.28	− 1431.74	− 0.74	0.46
Seq-MLE	728.21	− 1446.42	− 1427.89		

two optimized parameters of − 5.68 and − 0.80, a moderate dependency of − 0.58, an MLE score of 126.80, and an AIC score of − 249.60.

The vine copula model, as depicted in Fig. 7 and detailed in Tables 5 and 6, underwent optimization using the Seq-MLE optimization method. Alternatively, parameter optimization for this vine copula model was carried out using the Joint-MLE approach. Subsequently, the two optimization approaches were compared in terms of determining the best vine copula model based on log-likelihood, AIC, and BIC measures. The Vuong test was also conducted to assess fitting similarities between the two approaches. The results of these two comparisons are presented in Table 6. Based on the criteria (log-likelihood, AIC, and BIC measures), the Joint-MLE (730.14, − 1450.28, and − 1431.74, respectively) provided a superior vine copula model compared to the Seq-MLE (728.21, − 1446.42, and − 1427.89, respectively). The statistics of the Vuong test for these two approaches

Table 7 Components of the obtained vine copula model optimized by the Joint-MLE

Tree	(Conditional) Pair	The copula best model	Par	Par. 2	τ	Ltd	Utd
1	(Sev, Dur)	Joe (J)	12.30	–	0.85	0.00	0.94
1	(Int, Sev)	Rotated Tawn type 2 180 degrees (Tawn2_180)	5.38	0.56	0.49	0.55	0.00
2	(Dur, Int; Sev)	Rotated BB8 90 degrees (BB8_90)	− 6.00	− 0.77	− 0.58	0.00	0.00

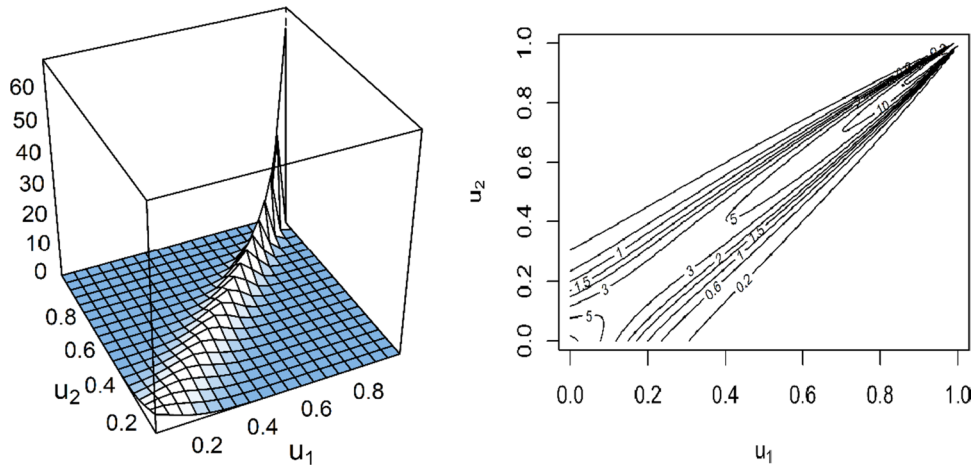

Fig. 8 Density plot (left) and contour plot (right) for the Joe copula model applied to the (severity, duration) pair

were − 0.74 with a corresponding p-value of 0.46. At the 5% significance level test, the latter comparison showed insufficient evidence to reject the null hypothesis (i.e., the two approaches have the same fitting), leading to the conclusion that the approaches have same fittings.

Considering the log-likelihood, AIC, and BIC measures, the vine copula model optimized by the Joint-MLE method was selected for further interpretation of the trivariate relationship involving the severity, intensity, and duration of unhealthy air pollution events. This model was also employed for conducting a risk assessment of extreme air pollution events in Klang. Table 7 provides a summary of the components, optimized parameters obtained through Joint-MLE, and dependency coefficients of the resulting vine copula model. Although a minor discrepancy was identified in the optimal parameters obtained via Joint-MLE (Table 7) compared to those from Seq-MLE (Tables 5 and 6), their Kendall's τ correlation coefficients were identical. This suggests that the regular vine tree sequence of the vine copula model optimized by the Joint-MLE can be accurately represented by Fig. 7.

In the first row of Table 7, the Kendall's τ correlation coefficient for the Joe copula model of the (severity, duration) pair was 0.85, indicating a strong positive dependence. This suggests that the pair exhibited a high level of correlation and predominantly moved in the same order. Table 7 also provides the lower (Ltd) and upper (Utd) tail correlation coefficients for the Joe copula model, which are 0.00 and 0.94, respectively. The lower tail signifies a very rare probability of the joint occurrence of severity and duration related to unhealthy air pollution at extremely low values. Conversely, the upper tail indicates a high probability of the severity of unhealthy air pollution being at a very high level when this pollution occurs over an extended period. These characteristics are clearly illustrated in Fig. 8, where the density in the upper tail domain is significantly higher than in other domains. This observation underscores the importance for authorities to carefully consider these properties for effective control and monitoring of unhealthy air pollution events in Klang.

As illustrated in the second row of Table 7, the Kendall's τ correlation coefficient for the Rotated Tawn type 2 180 degrees copula model of the (intensity, severity) pair was 0.49, indicating a moderate dependency strength, suggesting that the pair is likely to move in the same direction. The Ltd and Utd coefficients for this copula model were 0.55 and 0.00, respectively. The Ltd value suggests that the simultaneous occurrence of intensity and severity related to unhealthy air pollution, which occurred at exceptionally low values, has an intermediate probability of 0.55. Additionally, the Utd value implies that the joint manifestation of extreme severity and intensity in Klang is notably infrequent, as indicated by a probability of 0.00. Figure 9 illustrates the density plot (left) and contour plot (right) of the Rotated Tawn type 2 180 degrees copula model for the intensity and severity relationship.

As observed in the last row of Table 7, for the conditional pair of (duration, intensity; severity), the obtained Rotated

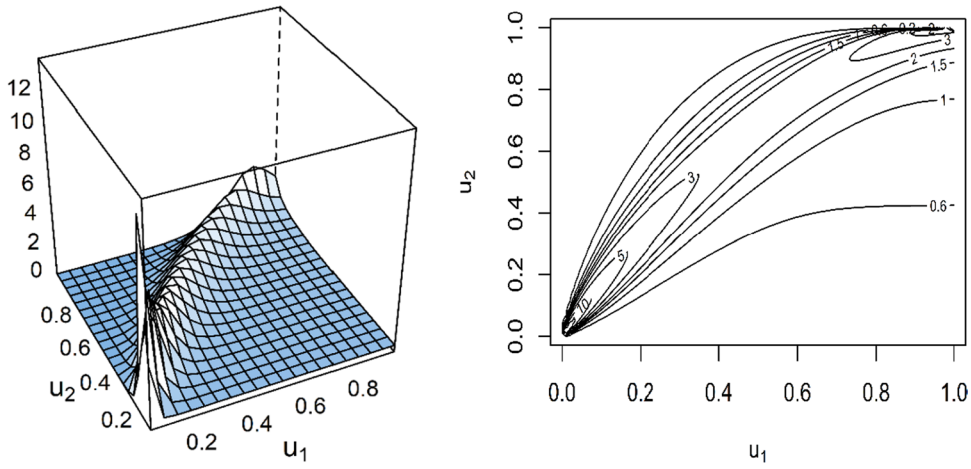

Fig. 9 Density plot (left) and contour plot (right) for the Rotated Tawn type 2 180 degrees copula model applied to the (intensity, severity) pair

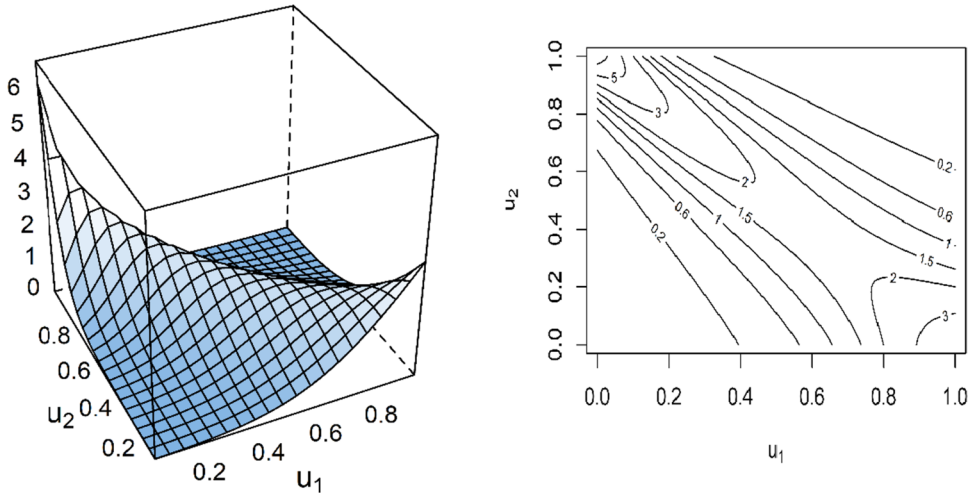

Fig. 10 Density plot (left) and contour plot (right) for the Rotated BB8 90 degrees copula model applied to the conditional pair (duration, intensity; severity)

BB8 90 degrees copula model had a negative Kendall's τ correlation coefficient of −0.58. This implies that this conditional pair was negatively correlated, and their values mostly moved in opposite ranks. The obtained model also provided zero coefficients for both lower and upper tails. In either the lower or upper tail parts, given the certain condition of air pollution severity, the probability of the joint occurrence of duration and intensity in these tail domain areas was extremely low. Figure 10 illustrates these properties.

The obtained vine copula model, combining all the previously mentioned components, is expressed as follows:

$$\begin{aligned} c\left(u_{Int},u_{Dur},u_{Sev};\boldsymbol{\theta}\right) = {} & c_{Int,Dur;Sev}\left(C_{Int|Sev}\left(u_{Int}|u_{Sev}\right),\right. \\ & \left. C_{Dur|Sev}\left(u_{Dur}|u_{Sev}\right);\theta_{Int,Dur;Sev}\right) \\ & \times c_{Int,Sev}\left(u_{Int},u_{Sev};\theta_{Int,Sev}\right) \\ & \times c_{Dur,Sev}\left(u_{Dur},u_{Sev};\theta_{Dur,Sev}\right), \end{aligned} \tag{40}$$

where the components of $c_{Int,Dur;Sev}$, $c_{Int,Sev}$, and $c_{Dur,Sev}$ are represented by Joe, Rotated Tawn type 2 180 degrees, and Rotated BB8 90 degrees, respectively (Table 7). Dependency measurements including Kendall's τ correlation and lower and upper tail dependency coefficients, for the tri-variate relationships of intensity, duration, and severity were computed in Eqs. (29)–(31), respectively. Consequently, Kendall's τ correlation and the lower and upper tail dependency coefficients for the tri-variate relationship were found to be 0.26, 0.18, and 0.31, respectively (Table 8). In Table 8, the moderate positive correlation value (0.26) for the relationship among intensity, duration, and severity indicated that these variables still tended to move in the same rank, implying that unhealthy air pollution severity was related to intensity and duration. Furthermore, the upper tail dependency (0.31) was higher than the lower tail dependency (0.18),

Table 8 Dependency coefficients of the obtained regular vine copula

Method	Kendall's τ	Ltd	Utd
Vine copula model	0.26	0.18	0.31

suggesting that the probability of the joint occurrence of intensity, duration, and severity related to unhealthy air pollution at extremely high values was higher than that at extremely low values. Therefore, a further risk assessment of extreme air pollution events is deemed significantly necessary.

A simulation study based on the Rosenblatt transform (Eq. 32) and its inverse (Eq. 33) was conducted to generate 20 datasets using the obtained vine copula model. Each dataset consisted of three samples of intensity, duration, and severity, with each sample comprising 1000 simulated copula observations. Subsequently, all 60,000 stimulated copula observations were transformed back into the corresponding 60,000 observations in the original scale through discrete inverse sampling (Eq. 34). As a result, 20 new datasets were obtained, with each dataset containing 1000 observations for intensity, duration, and severity in the original scale.

Five risk measures, namely, the joint probability of severity, intensity, and duration, conditional probability of severity based on the given intensity and duration, joint OR return period of unhealthy air pollution event characteristics, joint AND return period of unhealthy air pollution event characteristics, and conditional return period of severity based on the given intensity and duration, were determined for each dataset. All possible level combinations associated with the three characteristics (sev, int, dur) were applied to compute the risk measures. For each combination of (sev, int, dur) and each risk measure, the statistical metrics such as average, median, first quartile, third quartile, and IQR were then determined based on the corresponding risk measure values obtained from 20 datasets. In this study, these statistical metrics (average, median, first quartile, third quartile, and IQR) were utilized to assess the risks related to extreme air pollution events in Klang.

The comprehensive outcomes encompassing joint probability of severity, intensity, and duration, conditional probability of severity based on the given intensity and duration, joint OR return period of unhealthy air pollution event characteristics, joint AND return period of unhealthy air pollution event characteristics, and conditional return period of severity based on the given intensity and duration, are detailed in the supporting document. For risk assessments on extreme air pollution events purposes, the computed average is first employed. After that, the dynamics of the average, median, and IQR during extreme levels of severity, intensity, and duration are observed to understand extreme air pollution events better. The two discussions on the risk assessments on extreme air pollution events are consecutively provided in below paragraphs.

The comprehensive results encompassing joint probabilities of severity, intensity, and duration, conditional probabilities of severity based on the given intensity and duration, joint OR return period of unhealthy air pollution event characteristics, joint AND return period of unhealthy air pollution event characteristics, and conditional return period of severity based on the given intensity and duration, are detailed in the supporting document. Initially, for the purpose of risk assessments on extreme air pollution events, the computed average is utilized. Subsequently, the dynamics of the average, median, and IQR during extreme levels of severity, intensity, and duration are observed to enhance the understanding of extreme air pollution events. The two discussions on the risk assessments of extreme air pollution events are sequentially provided in the paragraphs below.

Table 9 Top five highest averages of the joint occurrence probabilities

Num	Intensity	Duration	Severity	Top 5 highest averages
1	100	1	100	1.00
2	100	1	1000	0.32
3	100	45	100	0.12
4	100	45	1000	0.12
5	100	1	10,000	0.06

The initial analysis involves utilizing the computed average for conducting risk assessments on extreme air pollution events. By using the average, Tables 9, 10, 11, 12, and 13 below present the top five highest averages of the joint probabilities, top five highest averages of the conditional probabilities, top five lowest averages of the joint AND return periods, top five lowest averages of the joint OR return periods, and top five lowest averages of the conditional return periods, respectively.

Table 9 reveals that the highest average (1.00) of the joint occurrence probability of $S \geq sev$, $I \geq int$, and $D \geq dur$ is provided by the values of (int, dur, sev) at (100, 1, 100). In practical terms, a joint probability of 1 implies that every instance in the stimulated data involves levels of S, I, and D equal to or greater than (100, 1, 100). Consequently, the most prevalent ground-level unhealthy air pollution occurs at the lowest threshold of (100, 1, 100). It's noteworthy that, while this represents a significant concern, these pollution levels are still considered less severe than extreme air pollution events causing public health issues ($I \geq 300$) or haze emergencies ($I \geq 500$). This is followed by a 0.32 average provided by the second joint probability of (100, 1, 1000), indicating that the joint occurrence of $I \geq 100$, $D \geq 1$, and $S \geq 1000$ is also likely to occur. The third and fourth averages (0.12) are given by the joint probabilities of (100, 45, 100) and (100, 45, 1000), respectively. Therefore, at $I \geq 100$ and $D \geq 45$, an unhealthy air pollution event with $S \geq 100$ or $S \geq 1000$ has the same possibility of occurrence. Lastly,

Table 10 Top five highest averages of the conditional probabilities

Num	Intensity	Duration	Severity	Top 5 highest averages
1	100	1	35,000	1.00
2	100	1	30,000	0.99
3	100	1	20,000	0.98
4	100	1	10,000	0.94
5	100	1	1000	0.68

Table 11 Top five lowest averages of the joint AND return periods

Num	Intensity	Duration	Severity	Top 5 lowest averages
1	100	1	100	106.00
2	100	1	1000	330.40
3	100	45	100	871.51
4	100	45	1000	871.51
5	100	1	10,000	1765.19

Table 12 Top five lowest averages of the joint OR return periods

Num	Intensity	Duration	Severity	Top 5 lowest averages
1	100	1	100	106.00
2	100	1	1000	106.00
3	100	1	10,000	106.00
4	100	1	20,000	106.00
5	100	1	30,000	106.00

Table 13 Top five lowest averages of the conditional return periods

Num	Intensity	Duration	Severity	Top 5 lowest averages
1	100	1	100	106
2	100	1	1000	330.4014
3	100	1	10,000	1765.188
4	100	1	20,000	4453.46
5	100	45	100	7191.52

the joint occurrence of $I \geq 100$, $D \geq 1$, and $S \geq 10,000$ can also occur in Klang with a low 0.06 probability. In summary, unhealthy air pollution events exceeding the minimum level (100, 1, 100) are likely to occur in the city.

Examining Table 10, the highest average (1.00) of the conditional probability for the occurrence of $S \leq sev$, given that $I \geq int$ and $D \geq dur$, is associated with the (int, dur, sev) values of (100, 1, 35,000). In this context, when the conditional probability is 1, it signifies that all simulated data instances display I and D levels equal to or exceeding 100 and 1, respectively, resulting in a S level less than 35,000. Thus, it is highly likely to observe a severity level below 35,000 when the intensity level is greater than or equal to 100, and the duration level is greater than or equal to 1. The second average (0.99) is derived from the conditional probability of (100, 1, 30,000), indicating a strong chance of the 30,000-severity level occurring when the intensity level is greater than or equal to 100 and the duration level is greater than or equal to 1. This pattern persists with the third (100, 1, 20,000) and fourth (100, 1, 10,000) conditional probabilities, boasting strong occurrence rates at 0.98 and 0.94, respectively. Finally, the fifth conditional probability average (0.68) corresponds to (100, 1, 1000). As the intensity level surpasses 100 and the duration level exceeds 1, the probability of air pollution severity reaching levels of 35,000, 30,000, 20,000, 10,000, or 1000 decreases in that order.

In this study, return period measures based on certain probabilities were also investigated. Table 11 displays the top five lowest averages of the joint AND return periods. The highest average (106) in the table corresponds to the joint AND return periods of the level set (int, dur, sev) at (100, 1, 100). The probability of the joint occurrence of (int, dur, sev) equal to or greater than (100, 1, 100) is 1, leading to the return period for this case (106) being equal to $E(L)$, which represents the expected unhealthy air pollution event interarrival time. This is followed by the second average (330.40) provided by the level set (int, dur, sev) of (100, 1, 1000). The third and fourth averages (871.51) are given by the joint AND return periods of (100, 45, 100) and (100, 45, 1000), respectively. Therefore, at $I \geq 100$ and $D \geq 45$, an air pollution event with $S \geq 100$ or $S \geq 1000$ shares the same return period. Lastly, the return period of the joint AND occurrence of $I \geq 100$, $D \geq 1$, and $S \geq 10,000$ is 1765.19. In summary, unhealthy air pollution events in Klang that exceed the minimum level (100, 1, 100), have the lowest return period. Note that the return period in Table 11 is inversely related to that in Table 9. Hence, if a probability from Table 9 is higher, then the corresponding return period in Table 11 is lower, approaching the $E(L)$ value of 106 in this work.

Table 12 presents the results obtained from the joint OR return periods, which are the top five lowest averages. In this case, all joint OR return periods share a common value, which is $E(L)$, specifically 106. The results reveal that the joint OR probabilities of (100, 1, 100), (100, 1, 1000), (100, 1, 10,000), (100, 1, 20,000), and (100, 1, 30,000) are very likely to occur, as this numerical analysis assigns a probability of 1 to each. Consequently, their joint OR return periods

are 106, indicating that the joint OR occurrences of the mentioned level sets (*int*, *dur*, *sev*) have a higher return period and are most likely to recur in Klang.

Table 13 lists the top five lowest averages of the conditional return periods. The highest conditional return period (106.00) in this table is associated with the level set (*int*, *dur*, *sev*) at (100, 1, 100). In other words, the product of the joint probabilities $P(I \geq 100, D \geq 1)$ and $P(S \geq 100, I \geq 100, D \geq 1)$ equals 1 (happens almost certainly), resulting in the conditional return period being identical to the $E(L)$ value, which is 106 (very highly to occur again). Following this, the second conditional return period average (330.40) is attributed to the level set (*int*, *dur*, *sev*) of (100, 1, 1000). The third conditional return period average (1765.19) is derived from the level set (*int*, *dur*, *sev*) of (100, 1, 10,000). The more extended conditional return period (4453.46) is associated with the level set (*int*, *dur*, *sev*) of (100, 1, 20,000). Lastly, the last average of the conditional return period (7191.52) is provided by the level set (*int*, *dur*, *sev*) of (100, 45, 100). Based on these outcomes, the most likely conditional unhealthy air pollution event to occur again in Klang is provided by the level set (*int*, *dur*, *sev*) of (100, 1, 100). Given $I \geq 100$ and $D \geq 1$, the return period for the upcoming unhealthy air pollution with $S \geq 100$ is at a higher reoccurrence level of 106 days.

To enhance our comprehension of the dynamics of extreme air pollution events, the average, median, and interquartile range (IQR) during extreme levels of severity, intensity, and duration are also examined and analyzed. These extreme levels of severity, intensity, and duration are defined as $sev \in \{20,000, 30,000, 35,000\}$, $int \in \{340, 420, 500\}$, and $dur \in \{133, 177, 221\}$, respectively. The outcomes of these examinations, pertaining to joint probabilities of severity, intensity, and duration, conditional probabilities of severity based on the given intensity and duration, joint OR return period of unhealthy air pollution event characteristics, joint AND return period of unhealthy air pollution event characteristics, and conditional return period of severity based on the given intensity and duration, are presented in Fig. 11 until Fig. 15 below.

Figure 11 illustrates the dynamics of joint occurrence probabilities for extreme air pollution events, presented

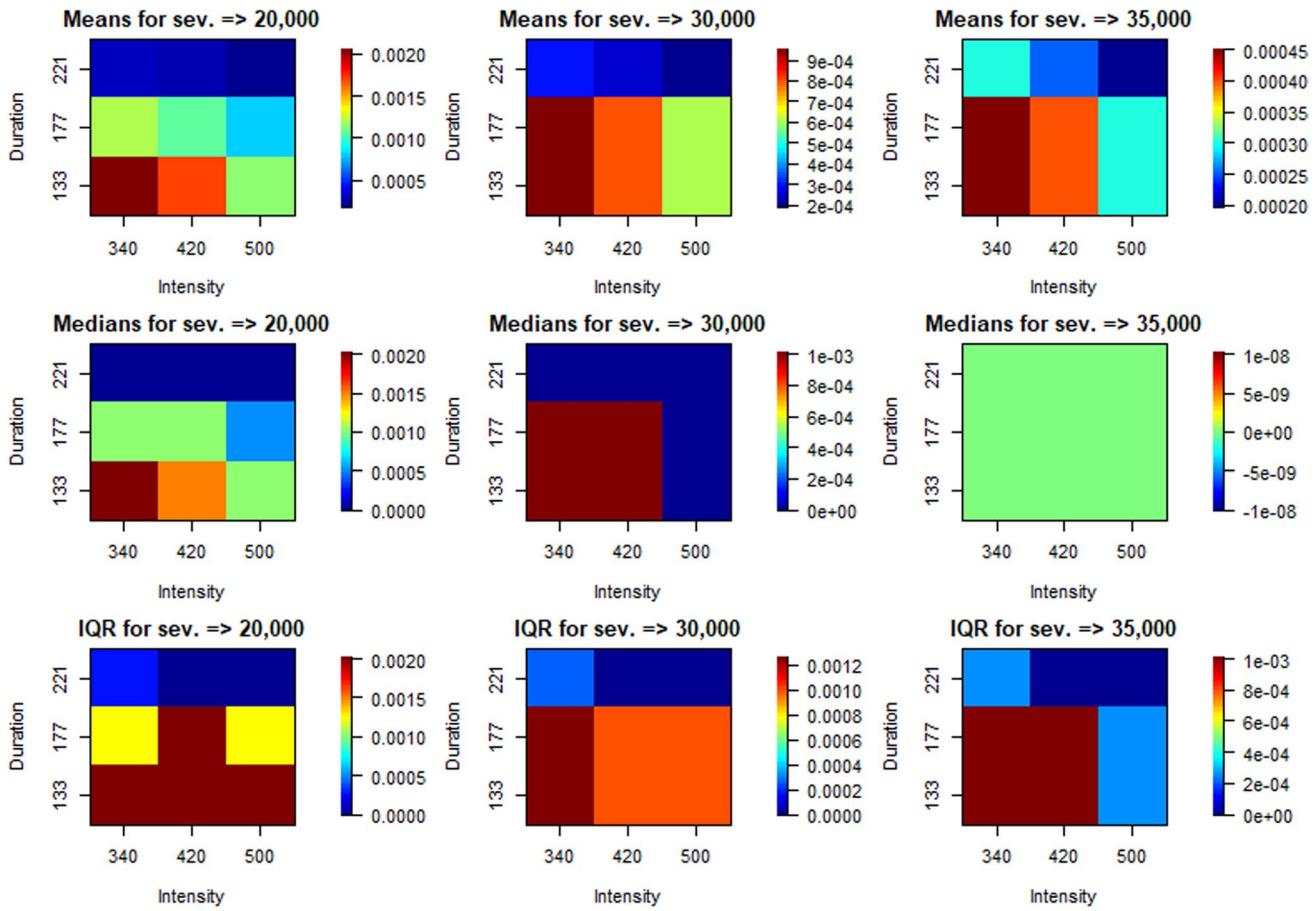

Fig. 11 Dynamics of joint occurrence probabilities for extreme air pollution events depicted through mean (top row), median (middle row), and IQR (bottom row)

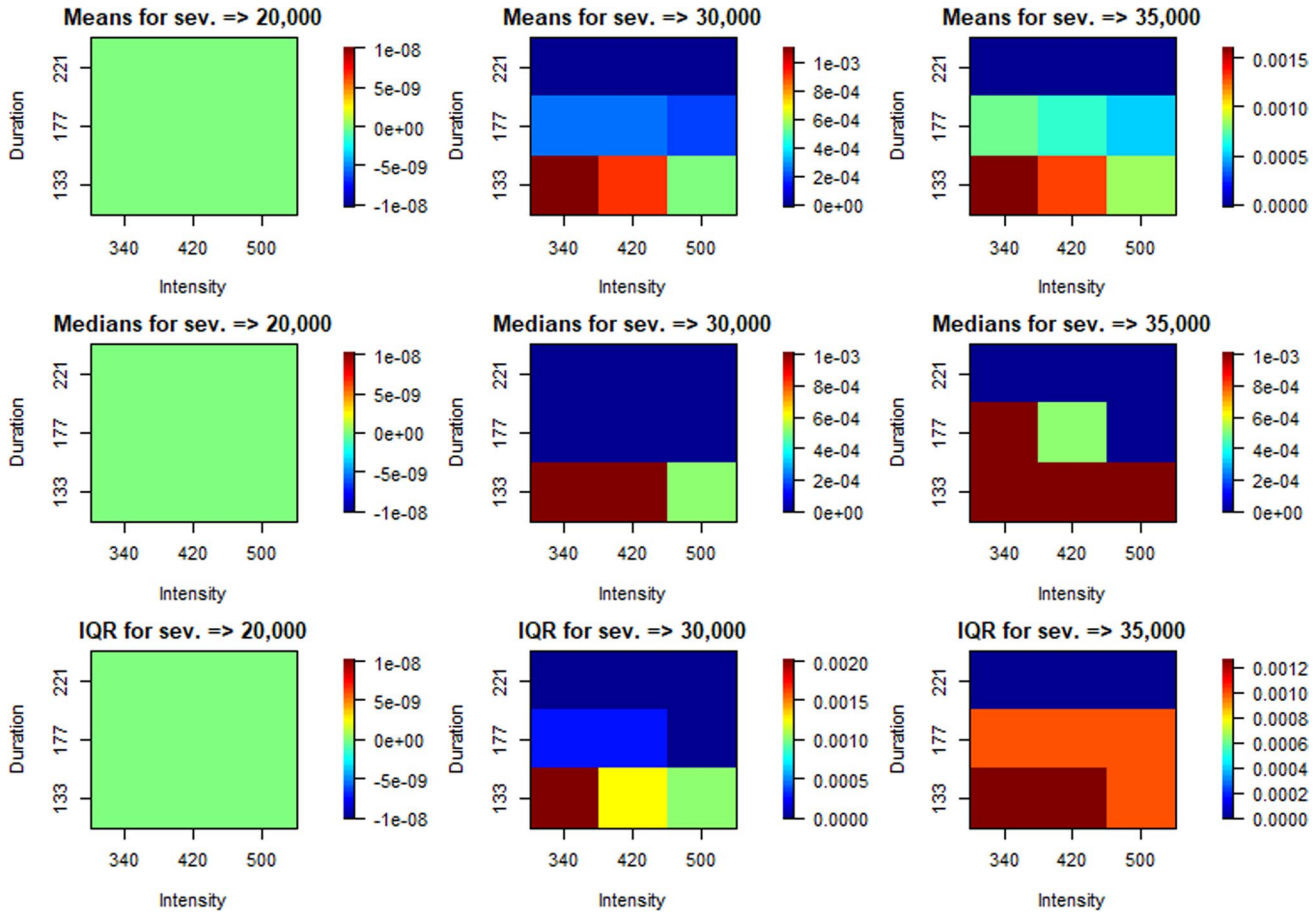

Fig. 12 Dynamics of conditional probabilities for extreme air pollution events depicted through mean (top row), median (middle row), and IQR (bottom row)

through mean (top row), median (middle row), and interquartile range (IQR) (bottom row). In the top row, utilizing the mean, the recorded probabilities range from a minimum of 0.00020 to a maximum of 0.00205. Additionally, the plots in this row demonstrate a decrease in joint occurrence probabilities as the levels of severity, intensity, and duration increase. Moving to the middle row and employing the median, the recorded probabilities vary from a minimum of 0 to a maximum of 0.002. Similar to the mean, the medians indicate a decline in recorded probabilities with increasing levels of severity, intensity, and duration. Results from both central tendency measures, mean and median, consistently indicate that more extreme air pollution events are less likely to occur compared to less extreme air pollution events. In this study, the IQR, representing the statistical dispersion between the first quartile and the third quartile, is employed for the bottom row. Using the IQR, the dispersion in joint occurrence probabilities ranges from a minimum of 0 to a maximum of 0.002. Furthermore, the IQRs reveal a reduction in the dispersion of joint occurrence probabilities as severity, intensity, and duration increase. This suggests that more extreme air pollution events exhibit a lower IQR, signifying lower variability in joint occurrence probabilities compared to less extreme air pollution events.

Figure 12 illustrates the dynamics of conditional probabilities for extreme air pollution events, presented through mean (top row), median (middle row), and interquartile range (IQR) (bottom row). In the top row, utilizing the mean, the recorded probabilities range from a minimum of 0 to a maximum of 0.0016. Discounting the first plot (as it shares identical probabilities), the subsequent two plots in this row reveal a general reduction in conditional probabilities with increasing levels of severity, intensity, and duration. Moving to the middle row and employing the median, the recorded probabilities vary from a minimum of 0 to a maximum of 0.001. Similar to the mean, the medians suggest a decrease in recorded probabilities with rising levels of severity, intensity, and duration, with the exception of the first plot. Overall, both central tendency measures (mean and median) consistently show that extreme air pollution events with higher severity levels are less likely to occur compared to those with lower severity levels. For the bottom row, utilizing

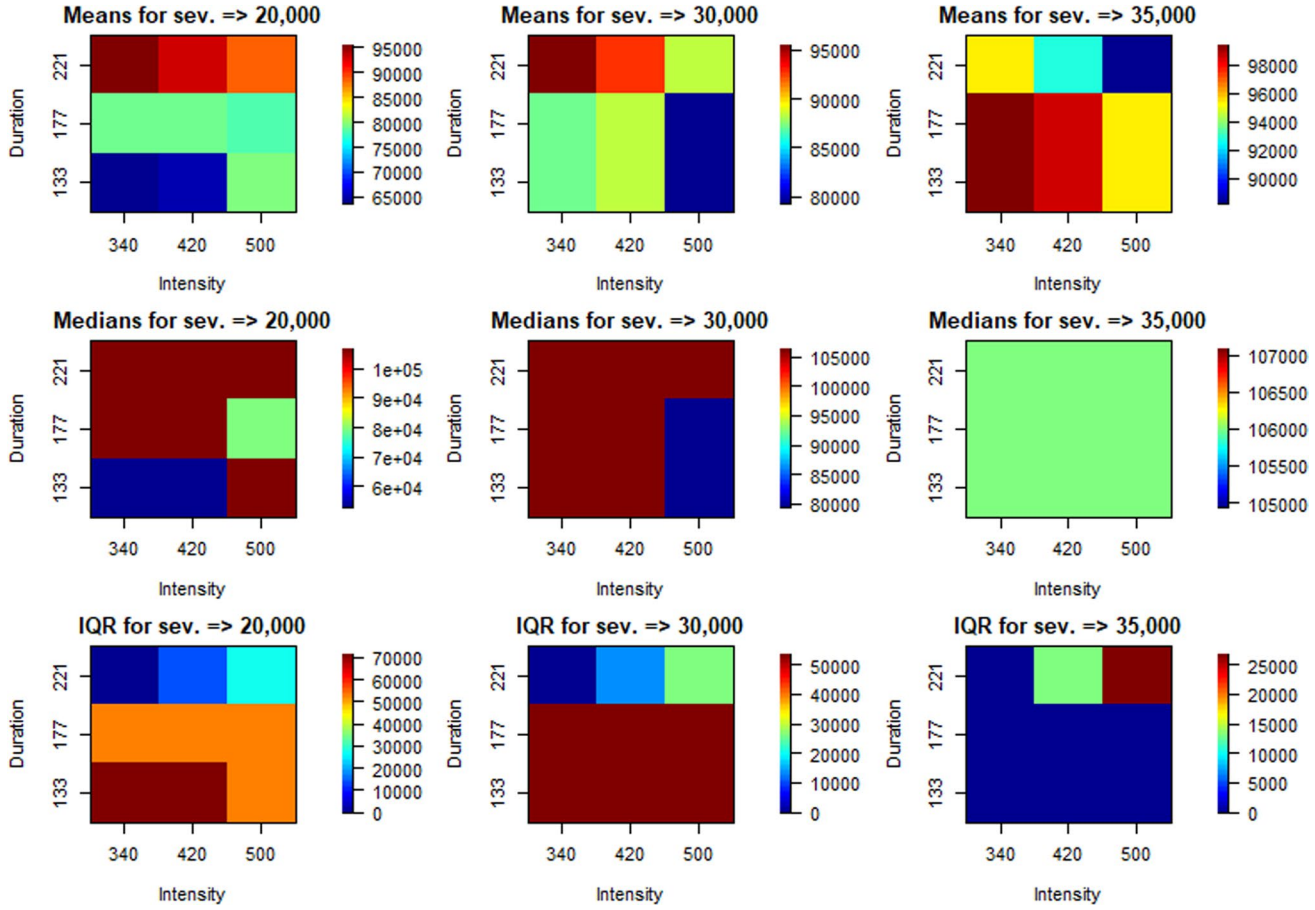

Fig. 13 Dynamics of joint AND return periods for extreme air pollution events depicted through mean (top row), median (middle row), and IQR (bottom row)

the IQR, the dispersion in conditional probabilities ranges from a minimum of 0 to a maximum of 0.002. Furthermore, excluding the first plot, the IQRs demonstrate a reduction in the dispersion of conditional probabilities as severity, intensity, and duration increase. This implies that extreme air pollution events with higher severity levels exhibit a lower IQR, indicating decreased variability in conditional probabilities compared to extreme air pollution events with lower severity levels.

Figure 13 illustrates the dynamics of joint AND return periods for extreme air pollution events, represented through mean (top row), median (middle row), and interquartile range (IQR) (bottom row). In the top row, utilizing the mean, the recorded return periods range from a minimum of 63,992.59 to a maximum of 99,375. Additionally, in general, the plots in the first row suggest an increase in joint AND return periods with elevated levels of severity, intensity, and duration. Moving to the middle row and employing the median, the recorded return periods vary from a minimum of 53,000 to a maximum of 106,000. Similar to the mean, the medians indicate an increase in recorded return periods with rising levels of severity, intensity, and duration, except for the last plot, which displays a similar return period. The results from both central tendency measures (mean and median) consistently indicate that more extreme air pollution events are less likely to recur compared to less extreme air pollution events. For the bottom row, using the IQR, the dispersion in joint AND return periods ranges from a minimum of 0 to a maximum of 70,666.67. The IQRs also reveal that the dispersion of joint AND return periods is mostly varied and does not exhibit a specific pattern as severity, intensity, and duration increase. This suggests that extreme air pollution events can exhibit different variability depending on their severity, intensity, and duration levels.

Figure 14 illustrates the dynamics of joint OR return periods for extreme air pollution events, presented through mean (top row), median (middle row), and interquartile range (IQR) (bottom row). In the top row, employing the mean, the recorded return periods range from a minimum of 3384.907 to a maximum of 20,739.4. Additionally, the plots in the first row indicate an increase in joint OR return periods with higher levels of severity, intensity, and duration. Moving

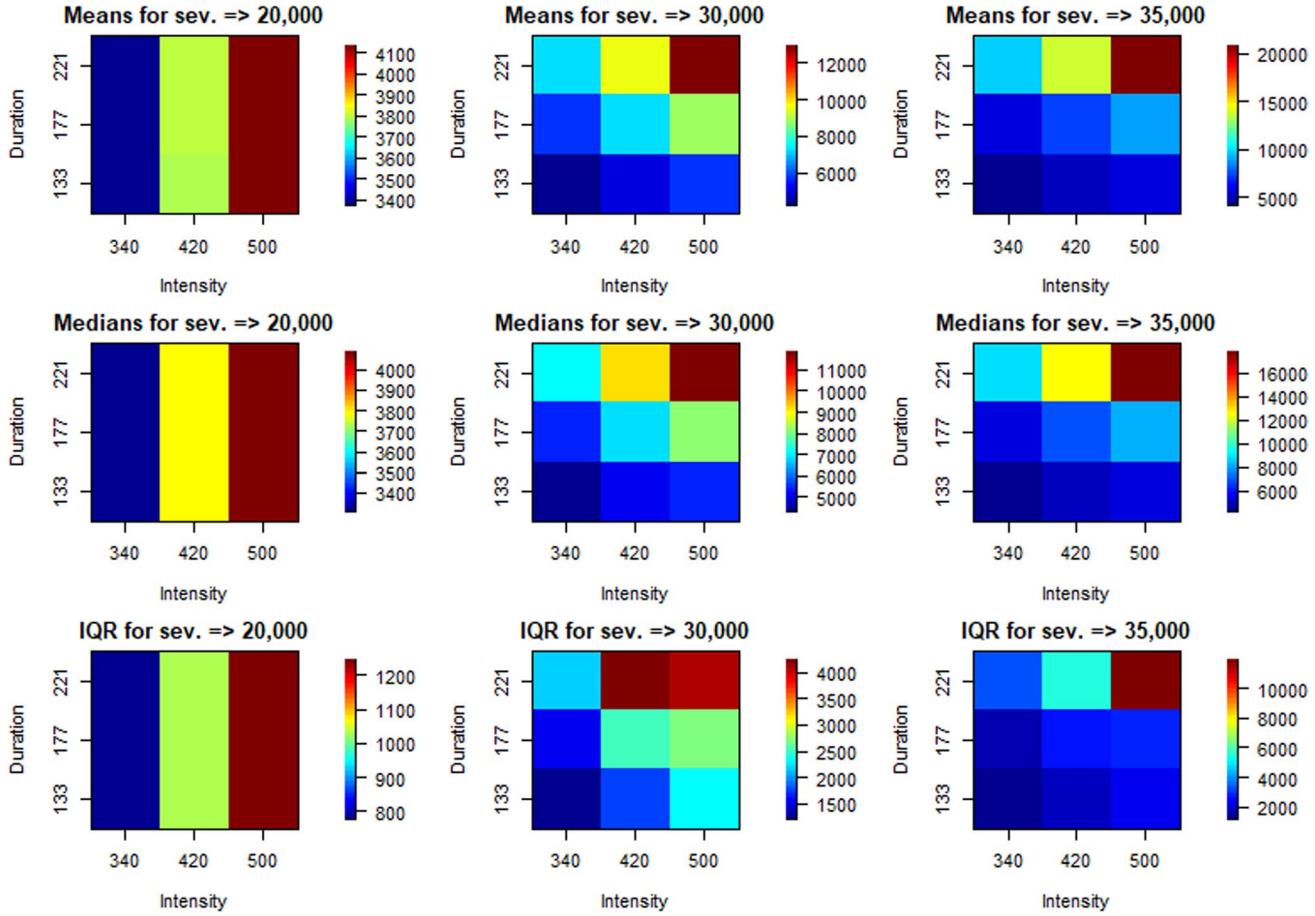

Fig. 14 Dynamics of joint OR return periods for extreme air pollution events depicted through mean (top row), median (middle row), and IQR (bottom row)

to the middle row and utilizing the median, the recorded return periods vary from a minimum of 3312.5 to a maximum of 17,666.67. Similar to the mean, the medians suggest an increase in recorded return periods with escalating levels of severity, intensity, and duration. Results from both central tendency measures (mean and median) consistently demonstrate that more extreme air pollution events are less likely to recur compared to less extreme air pollution events. For the bottom row, using the IQR, the dispersion in joint OR return periods ranges from a minimum of 778.4163 to a maximum of 11,830.36. However, in contrast to previous observations, the IQRs reveal an increase in the dispersion of joint OR return periods as severity, intensity, and duration increase. This implies that more extreme air pollution events exhibit a higher IQR, indicating increased variability in joint OR return periods compared to less extreme air pollution events.

Figure 15 below illustrates the dynamics of conditional return periods for extreme air pollution events, presented through mean (top row), median (middle row), and interquartile range (IQR) (bottom row). In the top row, utilizing the mean, the recorded return periods range from a minimum of 46,059,524 to a maximum of 90,100,000. Additionally, the plots in the first row suggest a general increase in conditional return periods with higher levels of severity, intensity, and duration. Moving to the middle row and employing the median, the recorded return periods vary from a minimum of 26,500,000 to a maximum of 106,000,000. Similar to the mean, the medians indicate an increase in recorded return periods with escalating levels of severity, intensity, and duration. Results from both central tendency measures (mean and median) consistently demonstrate that extreme air pollution events with higher severity levels are less likely to recur compared to extreme air pollution events with lower severity levels. For the bottom row, using the IQR, the dispersion in conditional return periods ranges from a minimum of 0 to a maximum of 94,222,222. However, unlike previous observations, the IQRs reveal a varied dispersion of conditional return periods that does not follow a specific pattern as severity, intensity, and duration increase. This implies that extreme air pollution events can exhibit

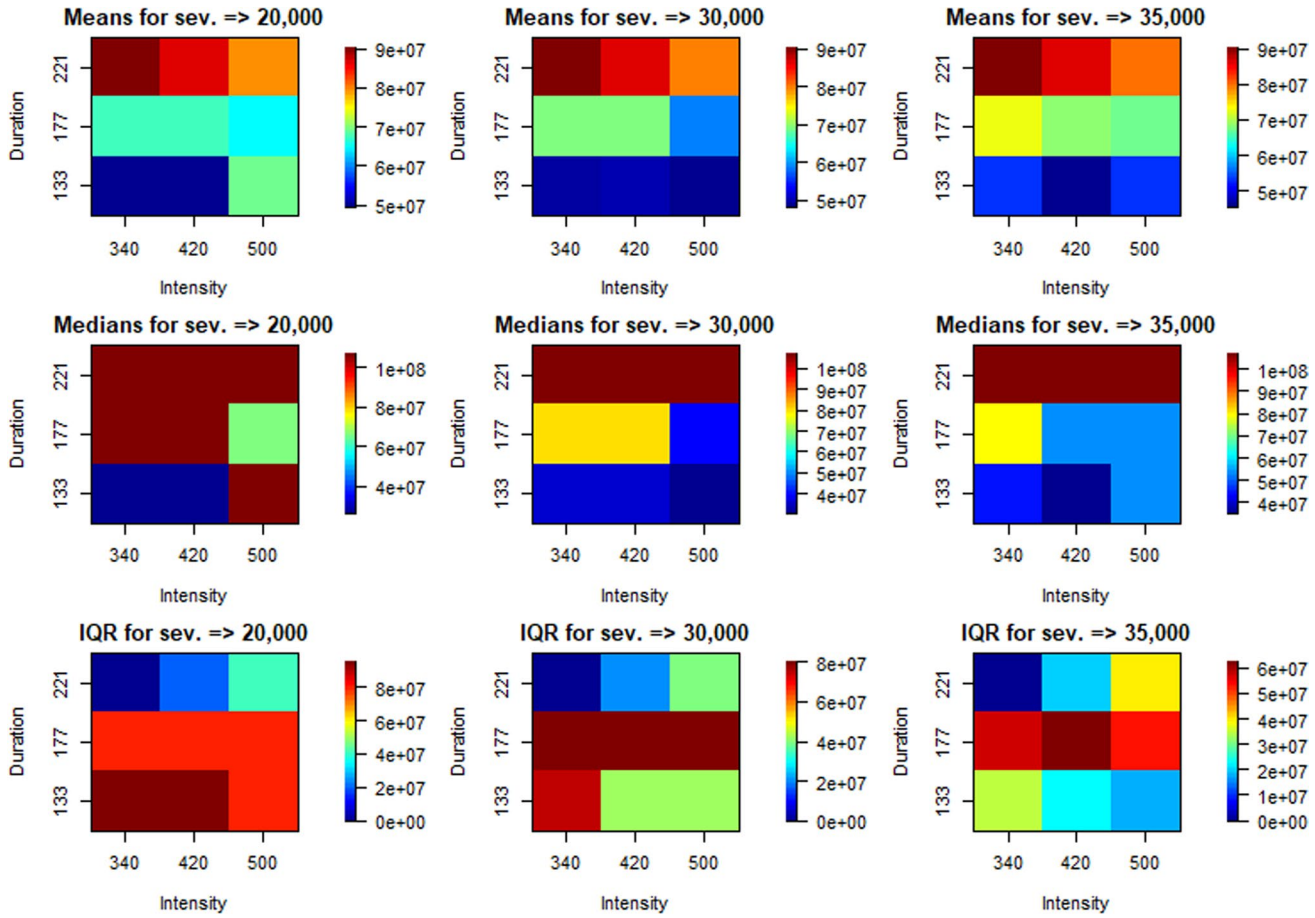

Fig. 15 Dynamics of conditional return periods for extreme air pollution events depicted through mean (top row), median (middle row), and IQR (bottom row)

different variability depending on their severity, intensity, and duration levels.

In summary, the outcomes derived from the simulation study, based on the derived vine copula and risk assessment of extreme air pollution events, reveal that these extreme air pollution events are not consistently associated with the highest values of joint and conditional probabilities. This suggests that extreme values observed across different characteristics do not necessarily coincide. Additionally, the analysis of return period measures indicates that extreme air pollution events typically exhibit very long waiting periods, making their recurrence in Klang, Malaysia, infrequent. This encouraging finding implies that the management of extreme air pollution events in Klang, Malaysia, remains within manageable limits. Consequently, it is crucial for authorities to persist in their efforts to mitigate the risks associated with extreme air pollution events. The complexity of air pollution behaviors, as highlighted by previous studies (Alyousifi et al. 2018; Amato et al. 2020; West et al. 2021; Yu et al. 2011). Moreover, the essential means to effectively combat air pollution involves the implementation of practical strategies advocated in previous literature. This includes regulatory actions and industrial controls, the establishment of robust public transportation programs, and a dedicated effort to transition to cleaner energy sources (Jonidi Jafari et al. 2021; Kumar and Gupta 2016; Mukhopadhyay and Pandit 2014; Ou et al. 2020; Yu et al. 2019). This undertaking to control and manage extreme air pollution events plays a critical role in maintaining a continuous supply of clean air, preserving our environment, and mitigating potential adverse effects on the economy and public well-being.

6 Conclusion

This study conducted a risk assessment for extreme air pollution events in Klang, Malaysia, utilizing vine copula modeling. Three characteristics of unhealthy air pollution events—intensity, duration, and severity—were examined. Firstly, these characteristics were analyzed through preliminary analysis. Following that, an empirical distribution function was employed to generate the copula data for the

vine copula modeling. Subsequently, a vine copula model was developed and applied to capture the dependency structure of the tri-variate relationship of intensity, duration, and severity.

The vine copula, a flexible and tractable multivariate copula constructed using pair (conditional) copulas as its structural components, proves more potent than standard multivariate distributions and copulas, particularly in modeling tail behaviors related to extreme air pollution events. In this study, thirty-seven different types of parametric pair copula models were considered as components for the vine copula. The optimization of the vine copula model fitting employs a two-step parameter estimation process. In the initial step, parameters for parametric pair copula models are determined, and subsequently, the second step focuses on optimizing the parameters of the vine copula model itself. In this second step, two fitting approaches, namely sequential maximum likelihood estimation and joint maximum likelihood estimation, are utilized. These two approaches are later compared based on criteria such as log-likelihood, Akaike's information criterion, and Bayesian's information criterion to determine the most well-fitted vine copula model.

To evaluate the occurrence risk of extreme air pollution events in Klang, dependency measures were computed based on the most well-fitted vine copula model, including Kendall's τ correlation coefficient, lower tail dependency, and upper tail dependency coefficients. Additionally, a comprehensive risk assessment of extreme air pollution events in Klang was conducted through a simulation study. Consequently, the obtained vine copula was simulated, and five statistical measures for assessing extreme air pollution events were proposed. These risk assessment measures encompass joint probability, conditional probability, joint AND return period, joint OR return period, and conditional return period. Furthermore, statistical metrics including average, median, first quartile, third quartile, and interquartile range (IQR) were calculated based on the risk assessment measures. This process contributes to offering new insights and inferences concerning the behaviors of extreme air pollution events, particularly in the most affected areas such as Klang, Malaysia.

The preliminary analysis results highlight the interconnected nature of the studied characteristics. Notably, air pollution severity demonstrates a significantly strong dependency on both intensity and duration. Following the model fitting evaluation, it is evident that the vine copula with components represented by the Joe, Rotated Tawn type 2 180 degrees, and Rotated BB8 90 degrees copulas is the optimal model for fitting the tri-variate relationship of intensity, duration, and severity. The positive Kendall's τ correlation coefficient (0.26) for the obtained vine copula indicates that higher values of one characteristic are likely to be associated with higher values of the other characteristics, and vice versa. Furthermore, the upper tail dependence coefficient (0.31) is greater than the lower tail dependence coefficient (0.18). This implies that the characteristics exhibit stronger dependence in the upper tail of their distribution. This underscores the significance of conducting risk assessments for extreme air pollution events, characterized by the extreme levels of severity, intensity, and duration.

For a more comprehensive risk assessment of extreme air pollution events in Klang, Malaysia, employing the five risk assessment measures, various properties of these events were documented. Notably, instances of extreme air pollution events occurring at elevated levels of intensity, severity, and duration are not necessarily correlated with the highest joint probability. Furthermore, extreme air pollution events characterized by extreme severity levels are also not associated with the highest values of conditional probability. These findings indicate that extreme levels of these distinct characteristics do not always coincide. Moreover, the analysis of return period measures revealed that extreme air pollution events in Klang did not display a waiting period close to the minimum point of the return period, which was identified as 106 days. In essence, extreme air pollution events in Klang were infrequent and required a considerable amount of time to recur.

In conclusion, while the air pollution level in Klang, Malaysia, is currently within manageable limits, it remains crucial for authorities to maintain vigilance in their ongoing risk assessment efforts. Effectively mitigating the risks associated with extreme air pollution events demands a nuanced understanding of the complex behaviors inherent in air pollution dynamics, emphasizing the need for continuous monitoring and proactive measures. Furthermore, the implementation of practical strategies recommended in the existing literature, such as regulatory actions and industrial controls, the establishment of robust public transportation programs, and a dedicated effort to transition to cleaner energy sources, is paramount. This concerted effort to control and manage extreme air pollution events assumes a critical role not only in ensuring a sustained supply of clean air but also in safeguarding our environment. Additionally, it serves as a proactive measure to mitigate potential adverse effects on the economy and public well-being. The commitment to these measures is essential for the long-term health and sustainability of the region.

To enhance future risk assessment methodologies, researchers might consider incorporating the vine copula to encompass data from other variables closely linked to extreme air pollution events. Additionally, given the profound impact of air pollution on health and businesses—resulting in reduced workforce productivity, work absences, premature deaths, and diminished crop yields, leading to severe economic consequences—future investigations could

delve into the relationship between air pollution and economic factors, utilizing the versatility of the vine copula. Moreover, exploring the performance of nonparametric vine copulas alongside their parametric counterparts in modeling air pollution data could provide valuable insights. Specific nonparametric copula densities, such as Beta kernel copula density, Bernstein estimator, and others, may be considered for comparison with parameter copula densities during vine copula construction. Furthermore, the approach employed in this study holds applicability beyond air pollution assessment. It can be extended to evaluate risks associated with various natural disasters, including droughts, earthquakes, floods, hurricanes, and landslides. These events also possess the capacity to result in fatalities, property damage, and substantial social and environmental disruptions.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s00477-024-02682-7.

Acknowledgements The authors acknowledge the Malaysia Department of Environment for providing data on the air pollution index in the study area and the University Kebangsaan Malaysia for the Dana Impak Perdana 2.0 (Grant Number DIP-2022-002).

Author contributions All authors contributed to the study conception and design. Software, formal analysis, and investigation were performed by Mohd Sabri Ismail. Material preparation, data collection, supervision, and funding acquisition were performed by Nurulkamal Masseran. The first draft of the manuscript was written by Mohd Sabri Ismail and Nurulkamal Masseran commented on previous versions of the manuscript. All authors read and approved the final manuscript.

Funding This research was funded by the University Kebangsaan Malaysia through the Dana Impak Perdana 2.0 (Grant Number DIP-2022-002).

Declarations

Conflict of interest The authors have no relevant financial or non-financial interests to disclose.

References

Afroz R, Hassan MN, Ibrahim NA (2003) Review of air pollution and health impacts in Malaysia. Environ Res 92:71–77

Aghamohammadi N, Isahak M (2018) Climate change and air pollution in Malaysia. Climate Change and Air Pollution: the impact on human health in developed and developing countries, pp 241–254

Al-Dhurafi NA, Masseran N, Zamzuri ZH (2018a) Compositional time series analysis for air pollution index data. Stoch Environ Res Risk Assess 32:2903–2911

Al-Dhurafi NA, Masseran N, Zamzuri ZH, Razali AM (2018b) Modeling unhealthy air pollution index using a peaks-over-threshold method. Environ Eng Sci 35:101–110

Alyousifi Y, Masseran N, Ibrahim K (2018) Modeling the stochastic dependence of air pollution index data. Stoch Environ Res Risk Assess 32:1603–1611

Amato F, Laib M, Guignard F, Kanevski M (2020) Analysis of air pollution time series using complexity-invariant distance and information measures. Physica A 547:124391

Amin MT, Khan F, Ahmed S, Imtiaz S (2021) Risk-based fault detection and diagnosis for nonlinear and non-Gaussian process systems using R-vine copula. Process Saf Environ Prot 150:123–136

Amini S, Bidaki RZ, Mirabbasi R, Shafaei M (2022) Flood risk analysis based on nested copula structure in Armand Basin. Iran Acta Geophysica 70:1385–1399

Arya Farid K, Zhang L (2017) Copula-based markov process for forecasting and analyzing risk of water quality time series. J Hydrol Eng 22:04017005

Atique F, Attoh-Okine N (2016) Using copula method for pipe data analysis. Constr Build Mater 106:140–148

Bhatti MI, Do HQ (2019) Recent development in copula and its applications to the energy, forestry and environmental sciences. Int J Hydrogen Energy 44:19453–19473

Cirillo P, Taleb NN (2020) Tail risk of contagious diseases. Nat Phys 16:606–613

Clayton DG (1978) A model for association in bivariate life tables and its application in epidemiological studies of familial tendency in chronic disease incidence. Biometrika 65:141–151

Czado C, Nagler T (2022a) Vine copula based modeling. Ann Rev Stat Appl 9:453–477

Czado C (2019) Analyzing dependent data with vine copulas. Lecture Notes in Statistics, vol 222. Springer

Daneshkhah A, Remesan R, Chatrabgoun O, Holman IP (2016) Probabilistic modeling of flood characterizations with parametric and minimum information pair-copula model. J Hydrol 540:469–487

Frank MJ (1979) On the simultaneous associativity of F (x, y) and x+ y− F (x, y). Aequationes Math 19:194–226

Gautam D, Bolia BN (2020) Air pollution: impact and interventions. Air Qual Atmos Health 13:209–223

Genest C, Favre A-C (2007) Everything you always wanted to know about copula modeling but were afraid to ask. J Hydrol Eng 12:347–368

Genest C, Rivest L-P (1993) Statistical inference procedures for bivariate Archimedean copulas. J Am Stat Assoc 88:1034–1043

Ismail MS, Masseran N (2023) Modeling the characteristics of unhealthy air pollution events using bivariate copulas. Symmetry 15:907

Jaworski P, Durante F, Hardle WK, Rychlik T (2010) Copula theory and its applications. Springer

Joe H (2014) Dependence modeling with copulas. CRC Press

Joe H, Kurowicka D (2011) Dependence modeling: vine copula handbook. World Scientific

Joe H, Li H, Nikoloulopoulos AK (2010) Tail dependence functions and vine copulas. J Multivar Anal 101:252–270

Jonidi Jafari A, Charkhloo E, Pasalari H (2021) Urban air pollution control policies and strategies: a systematic review. J Environ Health Sci Eng 19:1911–1940

Katz RW (2010) Statistics of extremes in climate change. Clim Change 100:71–76

Kim G, Silvapulle MJ, Silvapulle P (2007) Comparison of semiparametric and parametric methods for estimating copulas. Comput Stat Data Anal 51:2836–2850

Kotcher J, Maibach E, Choi W-T (2019) Fossil fuels are harming our brains: identifying key messages about the health effects of air pollution from fossil fuels. BMC Public Health 19:1079

Kumar R, Gupta P (2016) Air pollution control policies and regulations. In: Kulshrestha U, Saxena P (eds) Plant responses to air pollution. Springer Singapore, Singapore, pp 133–149

Li Z, Beirlant J, Yang L (2022) A new class of copula regression models for modelling multivariate heavy-tailed data. Insurance Math Econom 104:243–261

Liu Z, Cheng L, Hao Z, Li J, Thorstensen A, Gao H (2018) A framework for exploring joint effects of conditional factors on compound floods. Water Resour Res 54:2681–2696

Lü T-J, Tang X-S, Li D-Q, Qi X-H (2020) Modeling multivariate distribution of multiple soil parameters using vine copula model. Comput Geotech 118:103340

Lu JG (2020) Air pollution: a systematic review of its psychological, economic, and social effects. Curr Opin Psychol 32:52–65

Ma Y, Cheng B, Li H, Feng F, Zhang Y, Wang W, Qin P (2023) Air pollution and its associated health risks before and after COVID-19 in Shaanxi Province China. Environ Pollut 320:121090

Madonsela BS (2023) A meta-analysis of particulate matter and nitrogen dioxide air quality monitoring associated with the burden of disease in sub-Saharan Africa. J Air Waste Manag Assoc 73:737–749

Manga E, Awang N (2018) Bayesian autoregressive spatiotemporal model of PM10 concentrations across Peninsular Malaysia. Stoch Environ Res Risk Assess 32:3409–3419

Masseran N (2021a) Modeling the characteristics of unhealthy air pollution events: a copula approach. Int J Environ Res Public Health 18:8751

Masseran N (2021b) Power-law behaviors of the duration size of unhealthy air pollution events. Stoch Environ Res Risk Assess 35:1499–1508

Masseran N, Hussain SI (2020) Copula modelling on the dynamic dependence structure of multiple air pollutant variables. Mathematics 8:1910

Masseran N, Safari MAM (2020a) Intensity–duration–frequency approach for risk assessment of air pollution events. J Environ Manage 264:110429

Masseran N, Safari MAM (2020b) Risk assessment of extreme air pollution based on partial duration series: IDF approach. Stoch Environ Res Risk Assess 34:545–559

Masseran N, Safari MAM (2022) Statistical modeling on the severity of unhealthy air pollution events in Malaysia. Mathematics 10:3004

Masseran N, Razali AM, Ibrahim K, Latif MT (2016) Modeling air quality in main cities of Peninsular Malaysia by using a generalized Pareto model. Environ Monit Assess 188:1–12

Masseran N, Safari M, Hussain S (2021) Modeling the distribution of duration time for unhealthy air pollution events. J Phys Confer Ser IOP Publishing, p 012088

McNeil AJ, Frey R, Embrechts P (2015) Quantitative risk management: concepts, techniques and tools-revised edition. Princeton University Press

Mendenhall W, Beaver RJ, Beaver BM (2012) Introduction to probability and statistics. Cengage Learning

Mukhopadhyay A, Pandit V (2014) Control of industrial air pollution through sustainable development. Environ Dev Sustain 16:35–48

Nelsen RB (2006) An introduction to copulas. Springer

Nguyen PM, Liu W-H (2023) Portfolio management using time-varying vine copula: an application on the G7 equity market indices. Eur J Finance 29:1303–1329

Nguyen C, Bhatti MI, Komorníková M, Komorník J (2016) Gold price and stock markets nexus under mixed-copulas. Econ Model 58:283–292

Nguyen-Huy T, Deo RC, Mushtaq S, An-Vo D-A, Khan S (2018) Modeling the joint influence of multiple synoptic-scale, climate mode indices on Australian wheat yield using a vine copula-based approach. Eur J Agron 98:65–81

Othman J, Sahani M, Mahmud M, Sheikh Ahmad MK (2014) Transboundary smoke haze pollution in Malaysia: Inpatient health impacts and economic valuation. Environ Pollut 189:194–201

Ou Y, West JJ, Smith SJ, Nolte CG, Loughlin DH (2020) Air pollution control strategies directly limiting national health damages in the US. Nat Commun 11:957

Patton AJ (2012) A review of copula models for economic time series. J Multivar Anal 110:4–18

Perera F, Ashrafi A, Kinney P, Mills D (2019) Towards a fuller assessment of benefits to children's health of reducing air pollution and mitigating climate change due to fossil fuel combustion. Environ Res 172:55–72

Pourkhanali A, Kim J-M, Tafakori L, Fard FA (2016) Measuring systemic risk using vine-copula. Econ Model 53:63–74

Ravindra K, Vakacherla S, Singh T, Upadhya AR, Rattan P, Mor S (2023) Long-term trend of PM2.5 over five Indian megacities using a new statistical approach. Stoch Environ Res Risk Assess

Schepsmeier U, Stoeber J, Brechmann EC, Graeler B, Nagler T, Erhardt T, Almeida C, Min A, Czado C, Hofmann M (2015) Package 'vinecopula'. R package version 2

Semenov M, Smagulov D (2019) Copula models comparison for portfolio risk assessment. In: Kaz M, Ilina T, Medvedev GA (eds) Global economics and management: transition to economy 4.0. Springer International Publishing, Cham, pp 91–102

Sen PK (2011) Introduction to Nonparametric Estimation by Alexandre B. Tsybakov. Wiley Online Library

Shafaei M, Fakheri-Fard A, Dinpashoh Y, Mirabbasi R, De Michele C (2017) Modeling flood event characteristics using D-vine structures. Theoret Appl Climatol 130:713–724

Shan B, Guo S, Wang Y, Li H, Guo P (2021) Vine copula and cloud model-based programming approach for agricultural water allocation under uncertainty. Stoch Environ Res Risk Assess 35:1895–1915

Shindell D, Smith CJ (2019) Climate and air-quality benefits of a realistic phase-out of fossil fuels. Nature 573:408–411

Silverman BW (2018) Density estimation for statistics and data analysis. Routledge

Sklar A (1996) Random variables, distribution functions, and copulas: a personal look backward and forward. Lecture notes-monograph series, pp 1–14

Tosunoglu F, Gürbüz F, İspirli MN (2020) Multivariate modeling of flood characteristics using Vine copulas. Environ Earth Sci 79:459

Tuna Tuygun G, Elbir T (2023) Estimation of particulate matter concentrations in Türkiye using a random forest model based on satellite AOD retrievals. Stoch Environ Res Risk Assess 37:3469–3491

Tursumbayeva M, Muratuly A, Baimatova N, Karaca F, Kerimray A (2023) Cities of Central Asia: new hotspots of air pollution in the world. Atmos Environ 309:119901

Usmani RSA, Saeed A, Abdullahi AM, Pillai TR, Jhanjhi NZ, Hashem IAT (2020) Air pollution and its health impacts in Malaysia: a review. Air Qual Atmos Health 13:1093–1118

West SE, Bowyer CJ, Apondo W, Büker P, Cinderby S, Gray CM, Hahn M, Lambe F, Loh M, Medcalf A, Muhoza C, Muindi K, Njoora TK, Twigg MM, Waelde C, Walnycki A, Wainwright M, Wendler J, Wilson M, Price HD (2021) Using a co-created transdisciplinary approach to explore the complexity of air pollution in informal settlements. Human Soc Sci Commun 8:285

Wu J, Grande G, Triolo F, Pyko A, Sjöberg L, Ljungman P, Eneroth K, Bellander T, Rizzuto D (2023) Air pollution, social engagement, and depression in older adults: results from a Swedish population-based cohort study. Environ Pollut 336:122394

Yan J (2023) Multivariate modeling with copulas and engineering applications. In: Pham H (ed) Springer handbook of engineering statistics. Springer, London, pp 931–945

Yu B, Huang C, Liu Z, Wang H, Wang L (2011) A chaotic analysis on air pollution index change over past 10 years in Lanzhou, northwest China. Stoch Environ Res Risk Assess 25:643–653

Yu M, Zhu Y, Lin C-J, Wang S, Xing J, Jang C, Huang J, Huang J, Jin J, Yu L (2019) Effects of air pollution control measures on air quality improvement in Guangzhou, China. J Environ Manage 244:127–137

Yu R, Yang R, Zhang C, Špoljar M, Kuczyńska-Kippen N, Sang G (2020) A vine copula-based modeling for identification of multivariate water pollution risk in an interconnected river system network. Water 12:2741

Yu C, Yan G, Ruan K, Liu X, Yu C, Mi X (2023) An ensemble convolutional reinforcement learning gate network for metro station PM2.5 forecasting. Stoch Environ Res Risk Assess

Zhang Y, Han A, Deng S, Wang X, Zhang H, Hajat S, Ji JS, Liang W, Huang C (2023) The impact of fossil fuel combustion on children's health and the associated losses of human capital. Global Transitions 5:117–124

Zhang Z, Zhang G, Li L (2022) The spatial impact of atmospheric environmental policy on public health based on the mediation effect of air pollution in China. Environ Sci Pollut Res

Zhi B, Wang X, Xu F (2021) Portfolio optimization for inventory financing: copula-based approaches. Comput Oper Res 136:105481