

## Article

# Copula Modelling on the Dynamic Dependence Structure of Multiple Air Pollutant Variables

Nurulkamal Masseran \* and Saiful Izzuan Hussain 

Department of Mathematical Sciences, Faculty of Science and Technology, Universiti Kebangsaan Malaysia, Bangi, Selangor 43600, Malaysia; sih@ukm.edu.my

\* Correspondence: kamalmsn@ukm.edu.my; Tel.: +60-3-8921-3424

Received: 29 September 2020; Accepted: 28 October 2020; Published: 31 October 2020



**Abstract:** A correlation analysis of pollutant variables provides comprehensive information on dependency behaviour and is thus useful in relating the risk and consequences of pollution events. However, common correlation measurements fail to capture the various properties of air pollution data, such as their non-normal distribution, heavy tails, and dynamic changes over time. Hence, they cannot generate highly accurate information. To overcome this issue, this study proposes a combination of the Generalized Autoregressive Conditional Heteroskedasticity model, Generalized Pareto distribution, and stochastic copulas as a tool to investigate the dependence structure between the  $PM_{10}$  variable and other pollutant variables, including  $CO$ ,  $NO_2$ ,  $O_3$ , and  $SO_2$ . Results indicate that the dynamic dependence structure between  $PM_{10}$  and other pollutant variables can be described with a ranking of  $PM_{10}-CO > PM_{10}-SO_2 > PM_{10}-NO_2 > PM_{10}-O_3$  for the overall time paths ( $\delta$ ) and the upper tail ( $\tau^U$ ) or lower tail ( $\tau^L$ ) dependency measures. This study reveals an evident correlation among pollutant variables that changes over time; such correlation reflects dynamic dependency.

**Keywords:** copula model; dynamic dependence; multiple correlation measurement; pollution risk assessment

## 1. Introduction

The air pollution problem has long been the centre of discussions all over the world. This issue is particularly alarming for the urban areas of developed and developing countries [1–3]. In Malaysia, air pollution is generally determined from five major types of pollutants, namely, carbon monoxide ( $CO$ ), suspended particulate matter ( $PM_{10}$ ), sulphur dioxide ( $SO_2$ ), ozone ( $O_3$ ), and nitrogen dioxide ( $NO_2$ ) [4,5]. These five pollutant variables are observed simultaneously and monitored constantly to provide relevant information about air quality [6]. As these pollutants are simultaneously being observed and monitored, the investigation into the relationship and correlation among these five pollutant variables is expected to contribute to a comprehensive understanding of the behaviour of air pollution events at any particular area. For example, Marković et al. [7] have investigated the behaviour of  $CO$ ,  $O_3$ ,  $NO_2$ ,  $SO_2$ , and  $PM_{10}$  in the Belgrade urban area during the autumnal period of 2005. They reported the existence of a positive correlation between the  $PM_{10}$ ,  $SO_2$ ,  $NO_2$ , and  $CO$ . On the contrary, the  $O_3$  variable was found to indicate a negative correlation with  $SO_2$ ,  $NO_2$ ,  $PM_{10}$  and  $CO$ . With the basis of the correlation measured, they contended that  $SO_2$ ,  $NO_2$ ,  $CO$ , and  $PM_{10}$  could originate from similar sources. Rich et al. [8] found a high association between myocardial infarction disease and particulate matter concentration in the presence of other pollutants, such as  $O_3$ ,  $CO$ ,  $SO_2$ , and  $NO_2$ . The analysis by Xie et al. [9] for 31 Chinese cities showed the following: the correlation between particulate matter and  $NO_2$  and  $SO_2$  is either high or moderate, the correlation between particulate matter and  $CO$  is diverse, and the correlation between particulate matter and  $O_3$  is either weak or uncorrelated. These results varied spatio-temporally across all the cities.

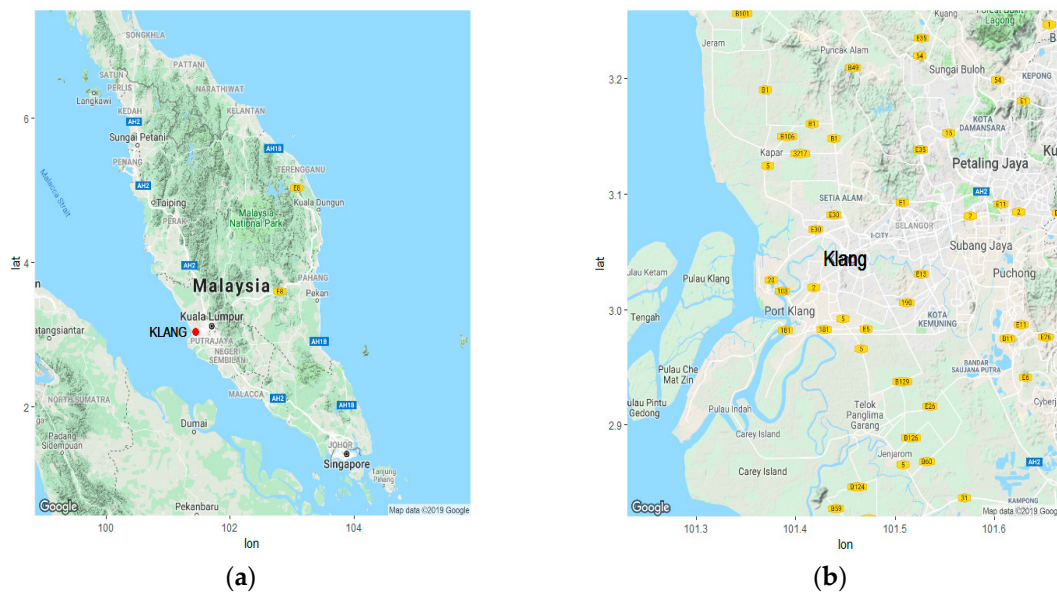
Generally, the dependency between two or more pollutant variables is measured by the Pearson, Spearman, or Kendall correlation methods. These common methods are widely popular because of their computational simplicity and easily interpretable results. However, despite their simplicity, the Pearson correlation method can only measure the strength of the linear dependence between the underlying random variables. In fact, the Pearson correlation method fails to provide an accurate measurement if any of the variables involved do not follow a normal distribution [10,11]. Unfortunately, the distribution of air pollutant data rarely follows a normal distribution. Instead, they often follow fat tailed distributions, such as the extreme value distribution, especially during periods of extreme pollution events [12–14]. However, the empirical measure of the Spearman or Kendall correlation method provides rank correlation measures which do not consider the properties of marginal distributions and time-varying properties among the random variables.

Apart from the common methods, several studies have considered a long-memory approach to investigate the temporal relationship and dependency behaviour of air pollutant data [15]. For example, Lee [16] showed the existence of scale-invariant behaviour in the air pollution concentration time series based on the concept of multifractal characteristics and long-term memory. Weng et al. [17] found that the time series of the ozone in Southern Taiwan indicates a long and persistent memory process involving nonlinearity and fractal time series based on R/S analyses. In a similar vein, Liu et al. [18] have shown the existence of persistence in the time-scaling behaviour for three pollution indices ( $\text{SO}_2$ ,  $\text{PM}_{10}$ ,  $\text{NO}_2$ ) in Shanghai, China, using the method of detrended fluctuation analysis (DFA) and the multifractal approach. However, these methods are bound by some limitations. Particularly, they fail to integrate the behaviour of asymmetric co-movements and contagion effects that exist in the dynamic behaviour among variables. The existence of this behaviour could affect the results of the analysis [19]. Thus, to overcome this problem, the method of dynamic conditional correlation (DCC) proposed by Engle [20] and Engle and Colacito [21] seems to be a good alternative. However, the data indicate the existence of extreme values and long tail properties; hence, an asymmetrical measure of a tail correlation may lead to biased estimates in DCC models [22,23].

As the probabilistic behaviour of extreme events can be effectively described using extreme value models, such as the generalized Pareto distribution (GPD), this study proposes a combination of a GPD model with copulas to establish a model that is able to describe the dependence structure between the  $\text{PM}_{10}$  variable and a set of four major pollutant variables, namely,  $\text{CO}$ ,  $\text{NO}_2$ ,  $\text{SO}_2$ , and  $\text{O}_3$ . This study adopts the copula approach because it is flexible to use with various types of marginal distributions without being constrained by normality limitations. A copula model can easily provide accurate information on the joint distribution of several pollution variables. The Gaussian copula was used in this study to describe the behaviour of the dependence structure among the pollutant variables. To investigate the tail dependence among the pollutant variables, this work employed the symmetrized Joe–Clayton (SJC) copulas. The generalized autoregressive conditional heteroskedasticity (GARCH) model was also combined with the GPD model to improve the distribution modelling. This combination method integrates all the properties of fat tail behaviour, asymmetric co-movements, stochastic volatility, as well as the contagion effects that exist on the dynamic behaviour among variables.

## 2. Study Area and Data

Klang is a large city in Malaysia located at a latitude of  $101^\circ 26' 44.023''$  E latitude and  $3^\circ 2' 41.701''$  N longitude. Klang is characterized by a dense population and an area of approximately  $573 \text{ km}^2$ . It is the centre of the import and export activities in Malaysia. Various important industrial, commercial, and economic activities are carried out in Klang. Moreover, Klang has been recognized as the 16th busiest container port and the 13th busiest trans-shipment port in the world [24]. However, its rapid urbanization has increased its risk of atmospheric pollution. Thus, given the importance of Klang, the behaviour of the air pollutants in the region should be evaluated and analysed. Figure 1 shows the location of Klang in peninsular Malaysia.



**Figure 1.** (a) Map of peninsular Malaysia (the location of Klang is marked by the red dotted point); and (b) map of Klang.

This study used the daily data of five main pollutant variables, namely,  $\text{SO}_2$ ,  $\text{NO}_2$ ,  $\text{CO}$ ,  $\text{PM}_{10}$ , and  $\text{O}_3$  for the period 1 January 2002–31 December 2016. A small percentage of missing values with a random pattern were found in the data. Thus, the single imputation method based on the average of the last known and next known observations was used to estimate the missing data. In addition to its easy implementation, the method provides satisfactory results for missing data with a random mechanism [25].

### 3. Generalized Pareto Distribution (GPD) Model

Extreme events are generally rare events that occur in the upper or lower tails of the distribution of data. Meanwhile, the extreme pollution data refer to an environment with a large air pollution index (API) at particular periods.  $Y_1, Y_2, \dots, Y_n$  represent the independent and identically distributed random variables of the hourly pollutants. The distribution of  $Y_i$  is governed by an unknown density function  $F$ . High value pollutant variables that exceed the unhealthy level ( $\geq 100$ ) indicate a pollution event. Mathematically, this phenomenon could represent a conditional event that is larger than some threshold  $u$ , and its conditional exceedance distribution function  $F^{[u]}$  can be written as follows:

$$\begin{aligned} F^{[u]}(y) &= \Pr(Y \leq y | Y > u) \\ &= \frac{\Pr\{Y \leq y, Y > u\}}{\Pr\{Y > u\}} \\ &= \frac{F(y) - F(u)}{1 - F(u)}; \quad y \geq u \end{aligned} \quad (1)$$

The extreme value approach is also known as the peaks-over-threshold method and is used for events with a threshold of  $u$  [26]. In this study, the observed excesses over the threshold are fitted to the GPD given by the following equation:

$$G_{\xi, \alpha}(y) = \begin{cases} 1 - \left(1 + \frac{\xi y}{\alpha}\right)^{-\frac{1}{\xi}}, & \text{if } \xi \neq 0, \\ 1 - \exp\left(-\frac{y}{\alpha}\right), & \text{if } \xi = 0, \end{cases} \quad (2)$$

where  $y \geq 0$ ,  $1 + \frac{\xi y}{\alpha} > 0$ . The parameters  $\xi$  and  $\alpha$  refer to the shape and scale parameter, respectively [27].

#### 4. Generalized Autoregressive Conditional Heteroskedasticity (GARCH) Model

The GJR–GARCH model was used as a marginal distribution for each pollutant variable. Let  $y_t$  represent the time series data for the pollutant variables, and let  $h_t^2$  denote the conditional variance for the period of  $t$ . Then, the GJR–GARCH model can be written as

$$y_t = \mu + \nu y_{t-1} + \varepsilon_t, \quad (3)$$

$$h_t^2 = c + \gamma h_{t-1}^2 + \eta_1 \varepsilon_{t-1}^2 + \eta_2 s_{t-1} \varepsilon_{t-1}^2, \quad (4)$$

where  $s_{t-1} = 1$  when  $\varepsilon_{t-1}$  is negative and 0 otherwise. Then,  $df$  denotes the degree of freedom, and  $\Omega_{t-1}$  represents the previous time data by  $t-1$ . Then, the standardize residual of the series  $z_t$  can be described as a  $t$ -distribution given as

$$z_t | \Omega_{t-1} = \sqrt{\frac{df}{\sigma_t^2(df-2)}} \varepsilon_t z_t \sim t_{df} \quad (5)$$

The GDP model is used to model the tail behaviour in the distribution of each pollutant variable. In sum, our approach is a combination of the GJR–GARCH and GPD models. The GJR–GARCH model is used to describe the interior part of the marginal distribution for each pollutant variable while the GPD model is used to describe the tail behaviour of each pollutant variable. In this study, we used the 10th percentile as the lower threshold  $u^L$  to indicate a healthy air environment and the 90th percentile as the upper threshold  $u^U$  to indicate an unhealthy air environment. Thus, the combination of the GJR–GARCH and GPD models can be represented as the following cumulative function:

$$F(z) = \begin{cases} \frac{k^L}{n} \left(1 + \xi \frac{u^L - z}{\alpha}\right)^{-\frac{1}{\xi}}, & \text{for } z < u^L, \\ f(z_t), & \text{for } u^L < z < u^U, \\ 1 - \frac{k^U}{n} \left(1 + \xi \frac{z - u^U}{\alpha}\right)^{-\frac{1}{\xi}}, & \text{for } z > u^U, \end{cases} \quad (6)$$

where  $n$  is the sample size of the data,  $k^L$  represents the volume of data below the threshold  $u^L$  and  $k^U$  represents the volume of data above the threshold  $u^U$ ,  $f(z_t)$  represents the distribution function determined from the GJR–GARCH model, and  $k^L(k^U)$  is the number of observations below (exceeding) the threshold  $u^L(u^U)$ .

#### 5. Copula Model

A copula model can be used to provide accurate information about the joint distribution of several pollution variables because it is not tied to the assumption of normality for the datasets involved. Moreover, the ability of the copula model to extract information regarding the dependence structure from the joint probability distribution function makes it a useful method for air pollution modelling. Mathematically, for a couple of random variables  $X_1$  and  $X_2$ , a bivariate copula model can be determined as

$$F(X_1, X_2) = C(F_1(X_1), F_2(X_2)), \quad (7)$$

where  $C$  is the copula function; and  $F_1$  and  $F_2$ , are the marginal distributions of the random variables  $X_1$  and  $X_2$ , respectively [10]. In this study, a Gaussian model was used to describe the overall dependence structure between the pollutant variables. The SJC copula is used to describe the behaviour of the dependence structure for the upper and lower tails of each pollutant variable.

#### 6. Interrelationship Behaviour among Pollutant Variables

To evaluate the interrelationship behaviour among the pollutant variables, we use a time-varying model for the Gaussian and SJC copulas.

### 6.1. Gaussian Copula

The Gaussian copula is a well known copula that has been used in applied research. It is associated with a multivariate normal distribution. For a bivariate distribution, which involves random variables  $u$  and  $v$ , its dependence structure based on the Gaussian copula can be determined as

$$C(u, v) = \int_{-\infty}^{\Phi^{-1}(u)} \int_{-\infty}^{\Phi^{-1}(v)} \frac{1}{2\pi\sqrt{1-\delta^2}} \exp\left(-\frac{x^2 - 2\delta xy + y^2}{2(1-\delta^2)}\right) dx dy, \quad (8)$$

with:

$$C = \Phi_{\delta}[\Phi^{-1}(u), \Phi^{-1}(v)], \quad (9)$$

where  $\delta$  is the linear correlation coefficient and  $\Phi$  is the standard normal cumulative density function. To include the time-varying properties, Patton [28] proposed a modification on the Gaussian dependence parameter by assuming that it evolves over time; the modified parameter is given as

$$\delta_t = \lambda \left[ \omega + \beta \delta_{t-1} + \frac{\alpha}{10} \sum_{j=1}^{10} \Phi^{-1}(u_{t-j}) \Phi^{-1}(v_{t-j}) \right], \quad (10)$$

where  $\lambda = (1 - e^{-x}) / (1 + e^{-x})$  is the modified logistic transformation that ensures  $\delta_t$  in the interval of  $(-1, 1)$ . The parameter of  $\beta \delta_{t-1}$  plays a role to capture the persistence effect while the mean of the model for the 10 observations of the transformed variables  $\Phi^{-1}(u_{t-j})$  and  $\Phi^{-1}(v_{t-j})$  represents the variation effect of the dependence series.

### 6.2. SJC Copula

The SJC copula is determined from the modification of the Joe–Clayton copula. The Joe–Clayton copula is given as

$$C_{JC}(u, v | \tau^U, \tau^L) = 1 - \left[ 1 - \left\{ (1 - (1 - u)^k)^{-\gamma} + (1 - (1 - v)^k)^{-\gamma} - 1 \right\}^{-\frac{1}{\gamma}} \right]^{\frac{1}{k}}, \quad (11)$$

where

$$k = \frac{1}{\log_2(2 - \tau^U)}, \quad (12)$$

$$\gamma = -\frac{1}{\log_2(\tau^L)} \quad (13)$$

The terms  $\tau^U \in (0, 1)$  and  $\tau^L \in (0, 1)$  represent the upper tail dependence and lower tail dependence, respectively. However, the Joe–Clayton copula cannot be used to evaluate the behaviour of the upper and lower tail dependence simultaneously. Thus, the SJC copula is used herein to overcome the weakness of the Joe–Clayton copula. The SJC copula is given as

$$C_{SJC}(u, v | \tau^U, \tau^L) = 0.5 [C_{JC}(u, v | \tau^U, \tau^L) + C_{JC}(1 - u, 1 - v | \tau^U, \tau^L) + u + v - 1], \quad (14)$$

where  $C_{JC}$  represents the Joe–Clayton copula in Equation (11). To include the time-varying properties, Patton [28] proposed the use of the evolution parameters in the SJC copula; the formula is written as follows:

$$\tau^{U/L} = \tilde{\lambda} \left( \omega^{U/L} + \beta^{U/L} \tau_{t-1}^{U/L} + \alpha^{U/L} \frac{\sum_{i=1}^{10} |u_{1,t-i} - u_{2,t-i}|}{10} \right), \quad (15)$$

where  $\tilde{\lambda}$  is the logistic transformation obtained as  $\tilde{\lambda}(x) = (1 + e^{-x})^{-1}$ ; it ensures that the dependence parameter  $\tau^{U/L}$  is in the range of (0, 1). Equation (15) specifies that  $\tau^{U/L}$  follows the Autoregressive-Moving Average (ARMA) type process with the order of (1,10), in which  $\beta^{U/L} \tau_{t-1}^{U/L}$

represents the autoregression,  $\alpha^{U/L}$  is the forcing variable and  $\frac{\sum_{i=1}^{10} |u_{1,t-i} - u_{2,t-i}|}{10}$  represents the persistence effect and variation in dependence.

The parameters for each model are estimated by a two-step estimation procedure. First, the parameters of the GARCH model corresponding to the GPD model for the upper and lower tails are estimated, and the standardized residuals  $\hat{z}_1, \hat{z}_2, \dots, \hat{z}_k$  are determined. Second, a transformation is performed using the distribution functions to create pseudo-uniform variables. On the basis of these pseudo-uniform variables, a copula model is estimated by maximizing the log-likelihood function given by

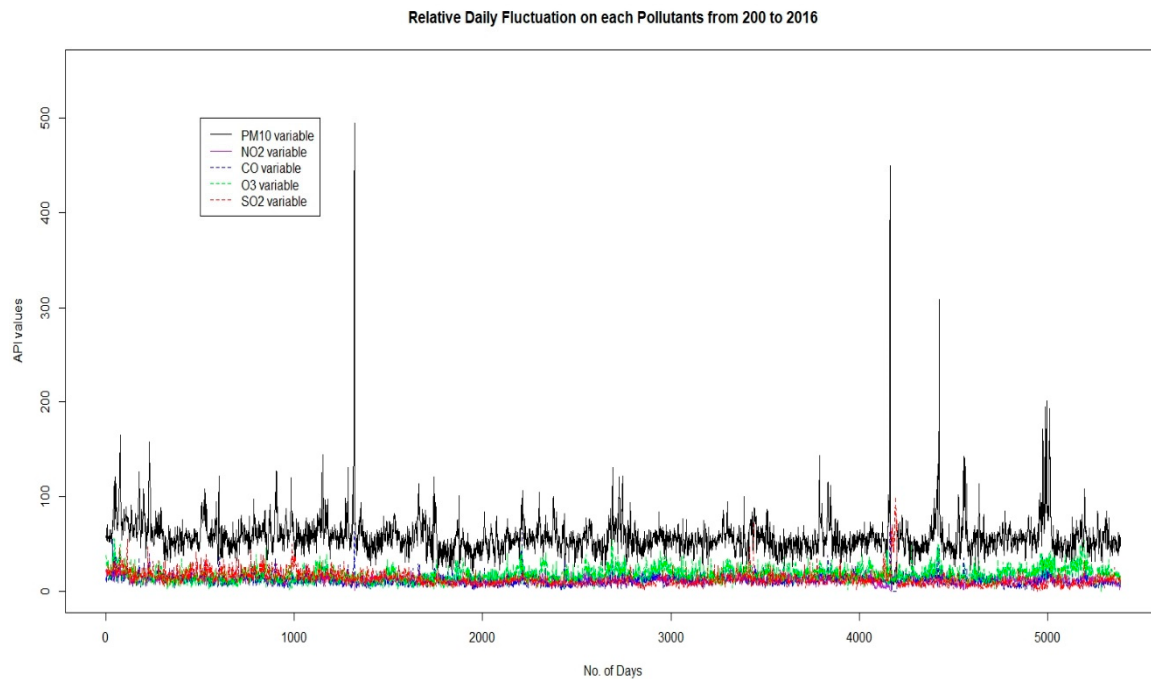
$$L(\xi; \hat{z}_1, \hat{z}_2, \dots, \hat{z}_k) = \sum_{i=m}^T \log \left[ c \left( F_1(\hat{z}_{1,i}), F_2(\hat{z}_{2,i}), \dots, F_k(\hat{z}_{k,i}), \xi \right) \right], \quad (16)$$

where  $\xi$  represents the parameter vector for each copula model,  $m = \max(p_{i,k}, q_{i,k})$  for  $i = 1, 2$ , and  $k = 1, 3, \dots, K$  [29,30].

## 7. Results and Discussion

A preliminary statistical analysis was carried out prior to the detailed discussion of the analysis results. Figure 2 shows that the volatility of PM<sub>10</sub> is higher than those of the other pollutant variables. Table 1 presents the descriptive statistics of all pollutant variables. The mean and median for PM<sub>10</sub> were found to be higher than those of the other pollutant variables. Hence, this pollutant exerts the greatest influence on the status of air quality. In addition, the standard deviation of PM<sub>10</sub> is the highest among all variables, thus implying that PM<sub>10</sub> presents more volatile behaviour than the other pollutant variables. NO<sub>2</sub> shows the lowest standard deviation and thus exhibits the most stable dynamic behaviour. The PM<sub>10</sub> pollutant also shows the highest of kurtosis value, which indicates that it appears most frequently in unhealthy pollution events. All the pollutant variables were found to have a positive skewness. A high positive skewness means that the pollutants exhibit long tailed behaviour to the right of the distribution data. The Jarque–Bera test results confirm that all the pollutant variables are not normally distributed. In particular, high Jarque–Bera test statistics are found for PM<sub>10</sub> and SO<sub>2</sub>. The results in Table 1 indicate that all the pollutant variables exhibit a nonlinear phenomenon and extreme behaviour. Thus, a model that functions under the normality assumption is not appropriate to use in representing these types of data.

Table 2 shows the results of the correlation among all the pollutant variables. The linear correlation values for the PM<sub>10</sub> pollutant in relation to the other pollutants range from −0.077 to 0.684. The PM<sub>10</sub> pollutant-related pairs are strongly correlated with CO, followed by O<sub>3</sub>, NO<sub>2</sub>, and SO<sub>2</sub>. The NO<sub>2</sub> pollutant is found to have a moderate correlation with the CO pollutant. For the other pairs of pollutant variables, they present a positive low correlation. However, the measurement of linear correlation may generate misleading results because it assumes that the pairs of data share linearity properties. Moreover, linear correlation measures generally work well for pairs of data that satisfy a normality assumption. For most air pollutant data, the existence of skewness cannot be neglected, as presented in the results in Table 1. A linear correlation measure also indicates a constant correlation behaviour over time for each pollutant variable. However, one of the most important characteristics of air pollution data is that they always fluctuate with volatility behaviour over time. Particularly during unhealthy air pollution events, some of the pollution variables can increase significantly and thereby correspond to extreme values. These behaviours are clearly depicted in Figure 2. Thus, we believe that a linear correlation measure may not be a reliable tool for describing the relationship among pollutant variables. As mentioned previously, we address these issues by proposing a time-varying copula with a combination of the GPD and GARCH models.



**Figure 2.** Time series plot for the relative daily fluctuation of each pollutant variable.

**Table 1.** Descriptive statistics for datasets 1, 2, 3 and 4.

| Statistic          | PM <sub>10</sub> | CO    | NO <sub>2</sub> | O <sub>3</sub> | SO <sub>2</sub> |
|--------------------|------------------|-------|-----------------|----------------|-----------------|
| Mean               | 57.46            | 11.67 | 12.77           | 17.62          | 13.09           |
| Median             | 55.13            | 10.84 | 12.53           | 16.52          | 11.50           |
| Standard Deviation | 20.14            | 5.29  | 3.90            | 7.07           | 7.37            |
| Kurtosis           | 99.39            | 12.23 | 0.26            | 1.65           | 21.64           |
| Skewness           | 6.35             | 2.13  | 0.29            | 0.94           | 3.19            |
| Min                | 6.04             | 0.02  | 0.37            | 0.42           | 0.00            |
| Max                | 494.88           | 70.33 | 31.86           | 56.32          | 100.29          |
| Jarque–Bera Stat   | 37,598.00        | 91.37 | 1401.90         | 2,249,500      | 114,110         |
| Count              | 5387             | 5387  | 5387            | 5387           | 5387            |

**Table 2.** Correlation among the pollutant variables.

|                  | PM <sub>10</sub> | CO    | NO <sub>2</sub> | O <sub>3</sub> | SO <sub>2</sub> |
|------------------|------------------|-------|-----------------|----------------|-----------------|
| PM <sub>10</sub> | 1                |       |                 |                |                 |
| CO               | 0.684            | 1     |                 |                |                 |
| NO <sub>2</sub>  | 0.200            | 0.445 | 1               |                |                 |
| O <sub>3</sub>   | 0.261            | 0.143 | 0.151           | 1              |                 |
| SO <sub>2</sub>  | 0.088            | 0.199 | 0.247           | −0.077         | 1               |

In the analysis, the parameters for the tails of each pollutant variable are estimated on the basis of the GPD model. Table 3 shows the estimated parameters describing the upper and lower tail behaviour of the data of each pollutant variable. The sign of the shape parameter shows how fast the tail decreases. A negative value indicates that the tail is finite, whereas a positive value indicates that the tail decreases as a polynomial. The higher the absolute value of the shape parameter is, the heavier the tail distribution will be. As shown in Table 3, for the lower tails, all pollutant variables are found to have a negative estimate of parameter  $\xi$ . This result indicates that all pollution variables exhibit finite short tail behaviour, with the PM<sub>10</sub> pollutant exhibiting the lowest value of  $\xi$  (−0.0626). For the upper tails, the CO, NO<sub>2</sub>, and SO<sub>2</sub> pollutants exhibit short tail behaviour. The PM<sub>10</sub> pollutant is found to

have the highest positive value  $\xi = 0.1142$ , followed by the  $\text{SO}_2$  pollutant with  $\xi = 0.0179$ . These values indicate that the  $\text{PM}_{10}$  and  $\text{SO}_2$  pollutants demonstrate heavy upper tail behaviour. These results indicate that all pollutant variables present the same pattern for a low API at a particular time. In the case of extreme pollution events, the  $\text{PM}_{10}$  variable is most likely to have the highest API value among all pollution variables at a particular time. Hence, the  $\text{PM}_{10}$  pollutant is highly volatile, particularly at the upside of its distribution.

**Table 3.** Correlation among pollutant variables.

|                   | $\text{PM}_{10}$ | CO      | $\text{NO}_2$ | $\text{O}_3$ | $\text{SO}_2$ |
|-------------------|------------------|---------|---------------|--------------|---------------|
| <b>Lower tail</b> |                  |         |               |              |               |
| $\xi$             | −0.0626          | −0.1768 | −0.1291       | −0.0766      | −0.0894       |
| Std. Error        | 0.0302           | 0.0313  | 0.0374        | 0.0424       | 0.0444        |
| $\alpha$          | 0.5213           | 0.3819  | 0.4629        | 0.4574       | 0.408         |
| Std. Error        | 0.0274           | 0.0201  | 0.0263        | 0.0276       | 0.0252        |
| <b>Upper tail</b> |                  |         |               |              |               |
| $\xi$             | 0.1142           | −0.0182 | −0.0758       | −0.0768      | 0.0179        |
| Std. Error        | 0.0465           | 0.0436  | 0.0395        | 0.0328       | 0.045         |
| $\alpha$          | 0.5714           | 0.7471  | 0.6111        | 0.6973       | 0.7576        |
| Std. Error        | 0.0361           | 0.0458  | 0.0357        | 0.0377       | 0.0472        |

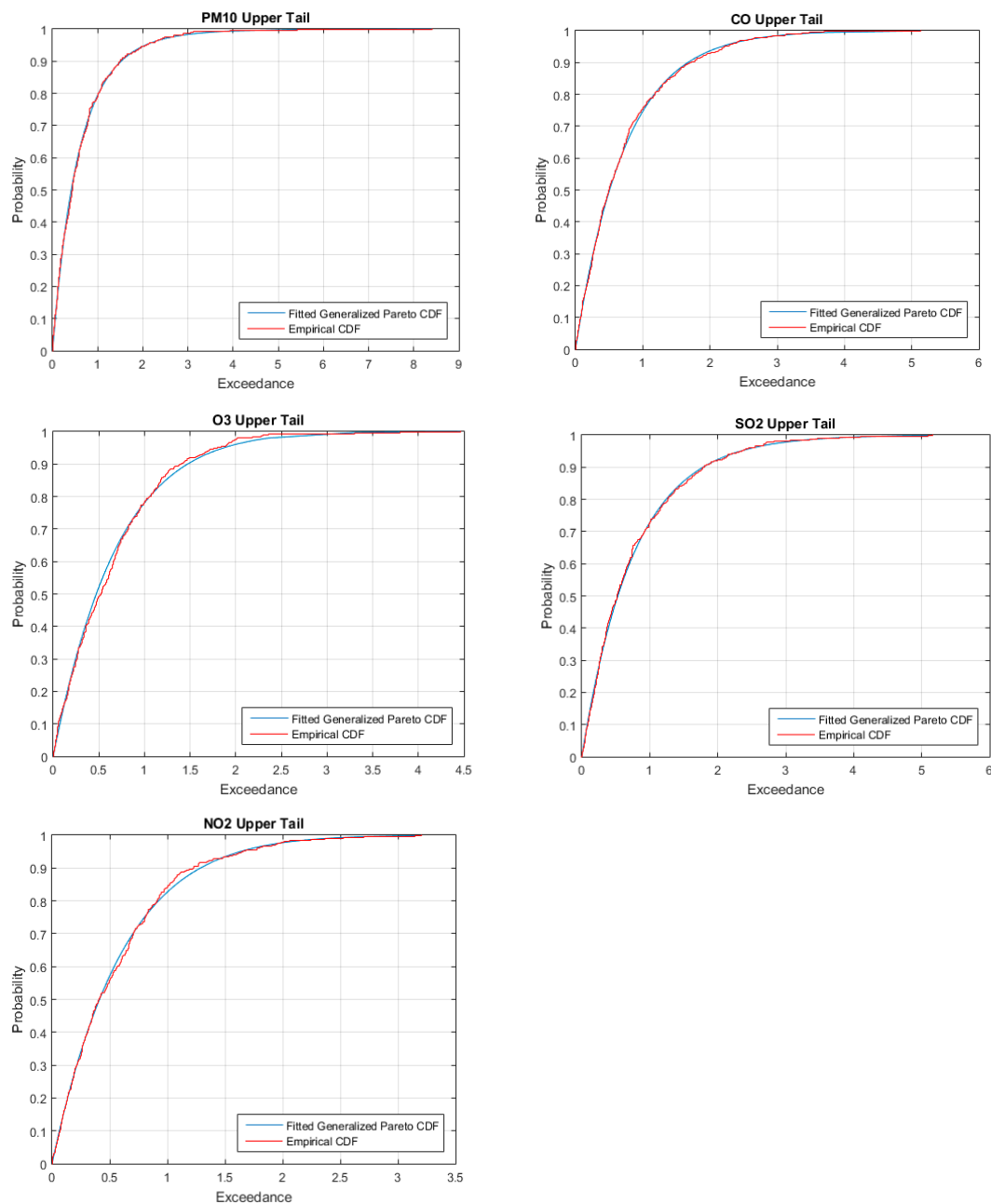
Figure 3 shows a comparison of the empirical and GPD plots for the cumulative distribution function (CDF) of the exceedance of the residuals in the upper tail of each pollutant variable. The fitted GPD closely follows the exceedance of the residuals, although only 10% of the standardized residuals were employed. Thus, the GPD model is a suitable choice for the upper tail data of each pollutant. This result is also valid for the lower tails. On the basis of the GPD model, the overall CDF of the semi-parametric models is obtained (Figure 4). In particular, the lower and upper tails are determined by the fitted GPD model while the interior part is estimated by the GARCH model.

The results shown in Table 3 and Figures 3 and 4 demonstrate that our approach can adequately model the marginal distributions of the time series data of the air pollutant variables in Malaysia. However, the estimated marginal distribution alone is not able to describe the dependence structure that exists among the pollutant variables. Thus, a copula needs to be adopted in the model. As mentioned previously, the Gaussian and SJC copulas are employed in this study. Given the available properties on each copula model, the Gaussian copula is useful in assessing the overall dependence structure of pollutant variables, as described in Equation (10). For the SJC copula model, it is useful in exploring the dependence behaviour of the lower and upper tails, as described in Equations (14) and (15).

Table 4 shows the results of the parameter estimates for the Gaussian and SJC copulas. The most important parameters for the constant copula are determined by  $\delta$  for the Gaussian copula and by the  $\tau^U$  and  $\tau^L$  parameters for the SJC copula. For the time-varying copulas, the important parameters for the Gaussian copula are  $\omega$ ,  $\alpha$ , and  $\beta$ ; and those for the SJC copula are  $\omega^U$ ,  $\omega^L$ ,  $\alpha^U$ ,  $\alpha^L$ ,  $\beta^U$  and  $\beta^L$ . As shown in Table 3, the  $\text{PM}_{10}$  variable is likely to have the highest API value among all pollution variables at a particular time. Thus, investigating the dynamic dependency of  $\text{PM}_{10}$  on other pollutant variables could yield informative results. In this regard, the parameters of  $\omega$ ,  $\omega^U$ , and  $\omega^L$  are useful in describing the magnitude of dependence between  $\text{PM}_{10}$  and other pollutant variables. The adjustment in the dependence measure is captured by the parameters of  $\alpha$ ,  $\alpha^U$  and  $\alpha^L$ . Moreover, the parameters  $\beta$ ,  $\beta^U$  and  $\beta^L$  are useful to represent the degree of the persistence of the dependence.

The best fitting copula model is determined by goodness-of-fit measures, namely, Akaike's information criterion (AIC) and Bayesian information criterion (BIC). The lowest value determined by the AIC or BIC indicates a well fitted model (Table 4). Two important points can be derived from the information provided in Table 4 and Figures 5–8. First, the overall dependence ( $\delta$ ) between  $\text{PM}_{10}$  and the other pollutants can be ranked in decreasing order as  $\text{PM}_{10}\text{--CO} > \text{PM}_{10}\text{--SO}_2 > \text{PM}_{10}\text{--NO}_2 > \text{PM}_{10}\text{--O}_3$ . Second, the rank in decreasing order for the upper tail ( $\tau^U$ ) and lower tail ( $\tau^L$ ) dependence between

PM<sub>10</sub> and the other pollutants is found to be same as that of the overall dependence, that is, PM<sub>10</sub>–CO > PM<sub>10</sub>–SO<sub>2</sub> > PM<sub>10</sub>–NO<sub>2</sub> > PM<sub>10</sub>–O<sub>3</sub>. This ranking implies that the dynamic fluctuations between PM<sub>10</sub> and CO over time have the strongest overall dependency and tail dependency; they are followed by the relationship of PM<sub>10</sub> with SO<sub>2</sub> and by PM<sub>10</sub> with NO<sub>2</sub>. The weakest overall dependency and tail dependency are those between PM<sub>10</sub> and O<sub>3</sub>. Thus, we can conclude that CO has the greatest influence on the behaviour of PM<sub>10</sub> among all the pollutants in a normal air quality, good air quality (lower tail), or bad air quality (upper tail). The results in Table 4 also show that the value of  $\tau^U$  is larger than that of  $\tau^L$  for all possible pairs. Hence, the upper tail dependency between PM<sub>10</sub> and the other pollutants is stronger than the lower tail dependency.



**Figure 3.** Fitted generalized Pareto distribution (GPD) model of the upper tail distributions for each pollutant variable.

To further investigate the dynamic dependence structure among the pollutant variables, we present in Figures 5–8 the time paths for the overall, lower, and upper tail dependencies based on the Gaussian and SJC time-varying copulas. In each figure, the red dashed line denotes the dependence parameter for the constant copula while the solid blue line denotes the dynamic parameter values under the time-varying copula. As mentioned previously, the time variations of the dependency measures between the pollutant variables are described by the parameters of  $\alpha$ ,  $\alpha^U, \alpha^L$ ,  $\beta$ ,  $\beta^U$  and  $\beta^L$ . Then, as shown in Table 4, the results of the time-varying Gaussian copula model reveal that most of the time paths are close to a white noise series as the values of the variation coefficient  $\alpha$  are relatively higher than those of the persistence coefficient  $\beta$ , except for the  $\text{PM}_{10}$ – $\text{NO}_2$  relationship. For the time-varying SJC copula, all the persistence coefficients  $\beta$  are larger than the coefficients  $\alpha$  for all pairs. This result gives some insights into the changes of the dependence structure for the upper and lower tails over a time period.

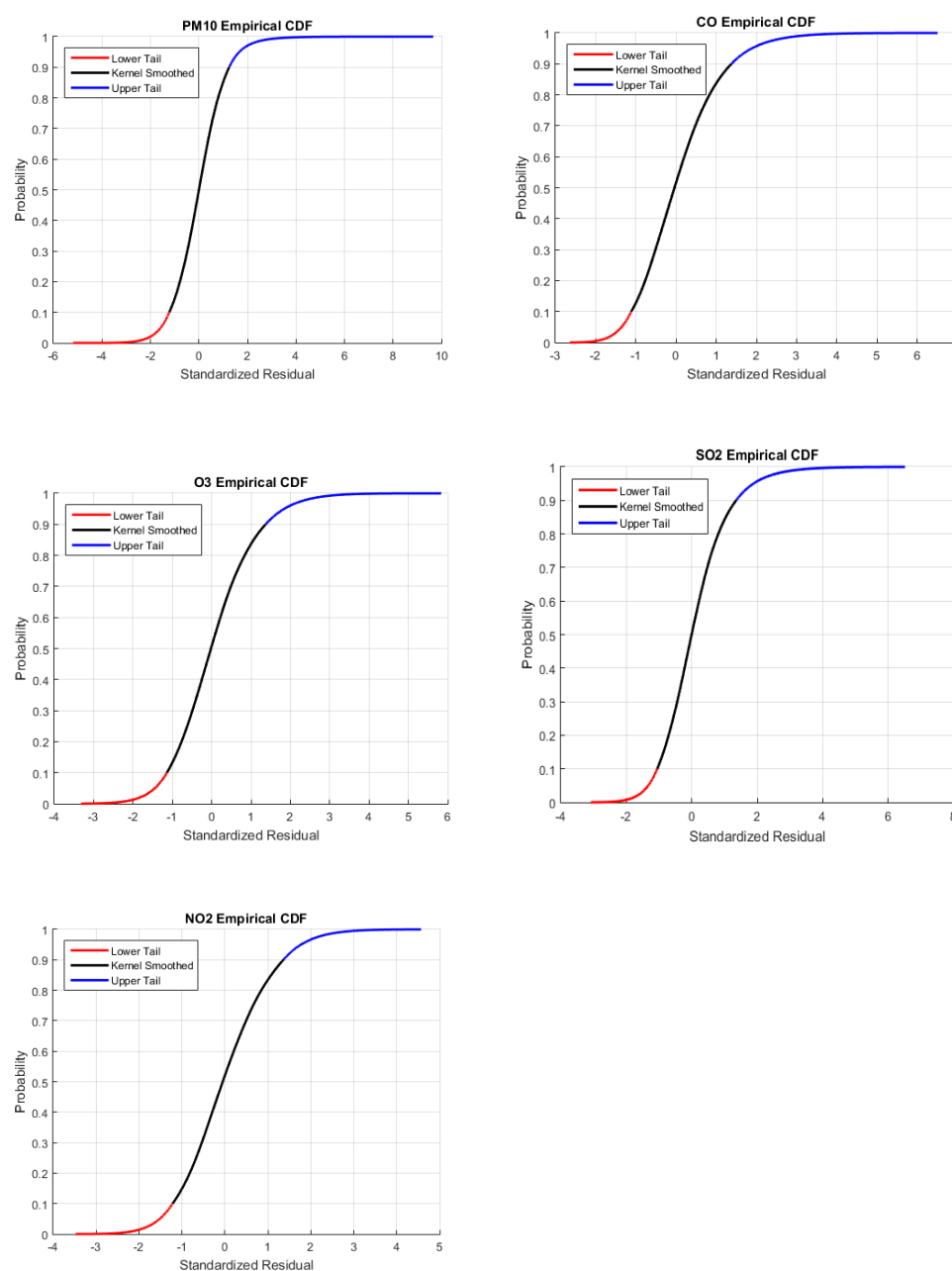
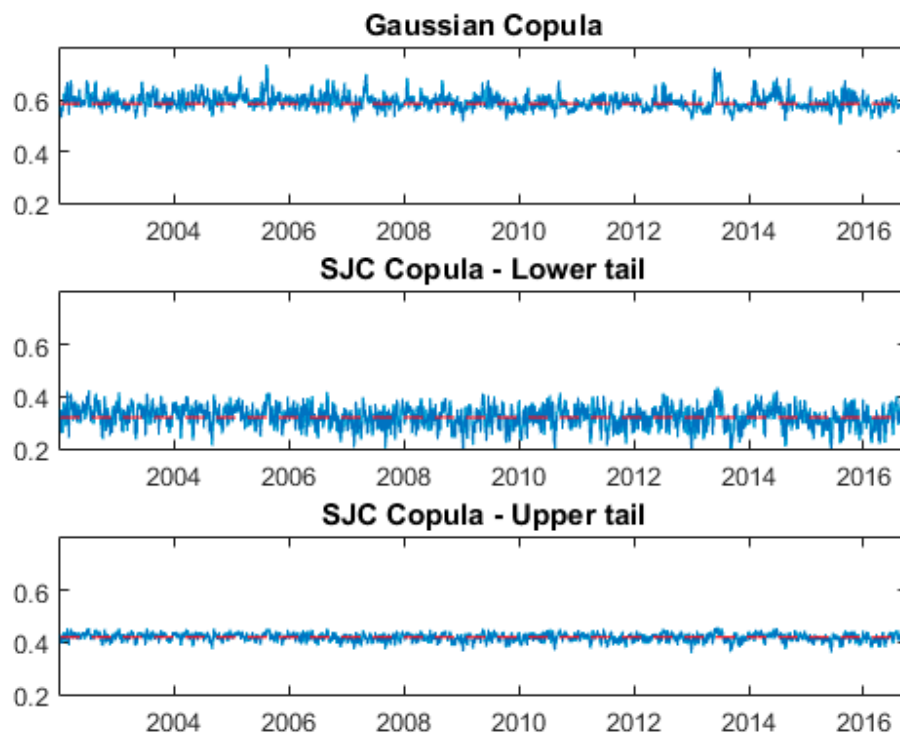


Figure 4. Fitted semi-parametric models for each pollutant variable.

**Table 4.** Parameter estimates for constant and time-varying copulas.

|                                | PM <sub>10</sub> –CO | PM <sub>10</sub> –NO <sub>2</sub> | PM <sub>10</sub> –O <sub>3</sub> | PM <sub>10</sub> –SO <sub>2</sub> |
|--------------------------------|----------------------|-----------------------------------|----------------------------------|-----------------------------------|
| <b>Gaussian Copula</b>         |                      |                                   |                                  |                                   |
| $\delta$                       | 0.586                | 0.307                             | 0.111                            | 0.352                             |
| AIC                            | −2272.546            | −535.598                          | −67.044                          | −716.037                          |
| BIC                            | −2272.545            | −535.597                          | −67.043                          | −716.036                          |
| <b>SJC Copula</b>              |                      |                                   |                                  |                                   |
| $\tau^U$                       | 0.419                | 0.1268                            | 0.0186                           | 0.1838                            |
| $\tau^L$                       | 0.319                | 0.1237                            | 0.0004                           | 0.1550                            |
| AIC                            | −2125.325            | −492.516                          | −59.392                          | −677.540                          |
| BIC                            | −2112.142            | −479.332                          | −46.208                          | −664.357                          |
| <b>Dynamic Gaussian Copula</b> |                      |                                   |                                  |                                   |
| $\omega$                       | 1.162                | 0.025                             | 0.409                            | 1.400                             |
| $\alpha$                       | 0.237                | 0.055                             | 0.164                            | 0.485                             |
| $\beta$                        | 0.125                | 1.949                             | −1.765                           | −2.222                            |
| AIC                            | −2289.200            | −561.787                          | −68.637                          | −743.441                          |
| BIC                            | −2289.196            | −561.783                          | −68.633                          | −743.437                          |
| <b>Dynamic SJC Copula</b>      |                      |                                   |                                  |                                   |
| $\omega^U$                     | 0.735                | −0.790                            | −2.358                           | 0.777                             |
| $\alpha^U$                     | −1.954               | −8.302                            | −0.057                           | −10.000                           |
| $\beta^U$                      | −0.463               | −0.878                            | 0.330                            | −0.430                            |
| $\omega^L$                     | 0.600                | 0.499                             | −0.794                           | 1.067                             |
| $\alpha^L$                     | −4.672               | −8.699                            | 2.741                            | −9.999                            |
| $\beta^L$                      | −0.193               | −0.130                            | 1.039                            | −0.094                            |
| AIC                            | −2125.447            | −512.480                          | −62.110                          | −727.503                          |
| BIC                            | −2085.896            | −472.929                          | −22.559                          | −687.953                          |

**Figure 5.** Dependence path of the time-varying copula for PM<sub>10</sub>–CO.

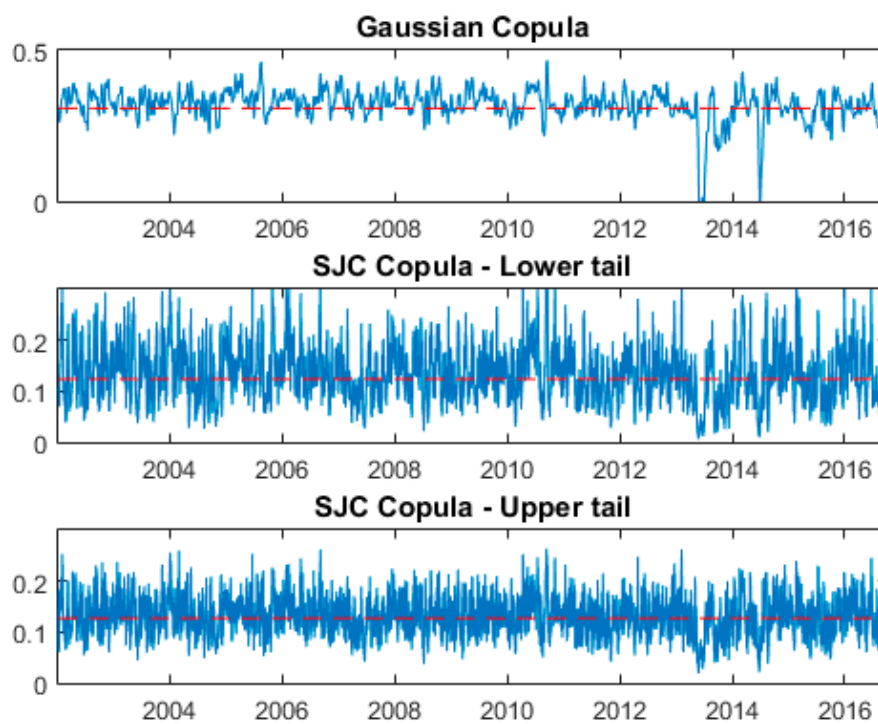


Figure 6. Dependence path of the time-varying copula for  $\text{PM}_{10}\text{--NO}_2$ .

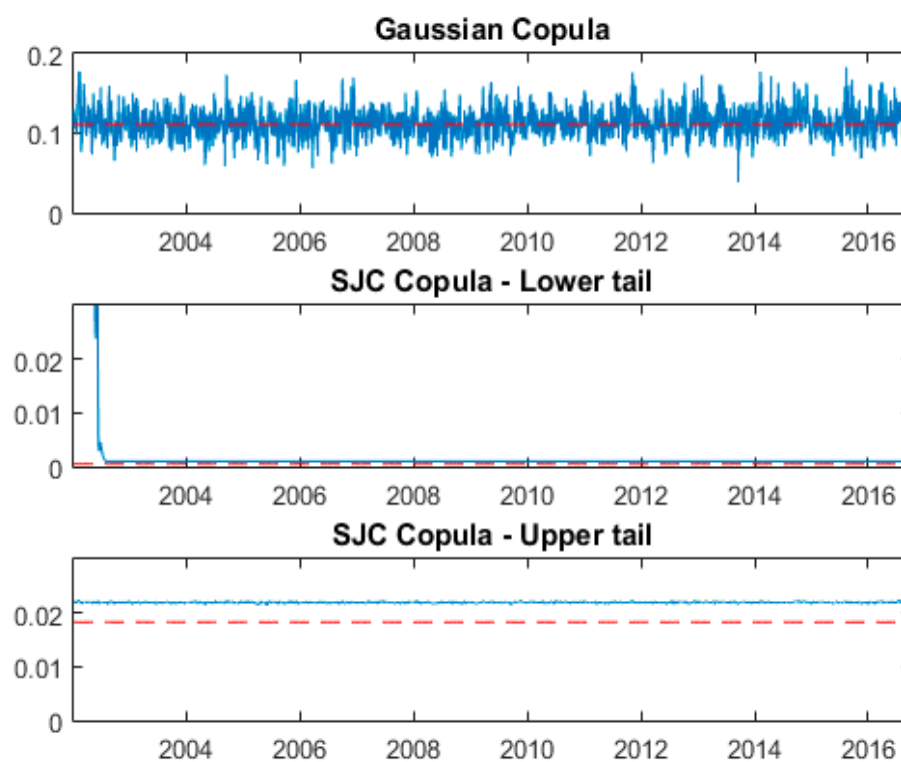


Figure 7. Dependence path of the time-varying copula for  $\text{PM}_{10}\text{--O}_3$ .

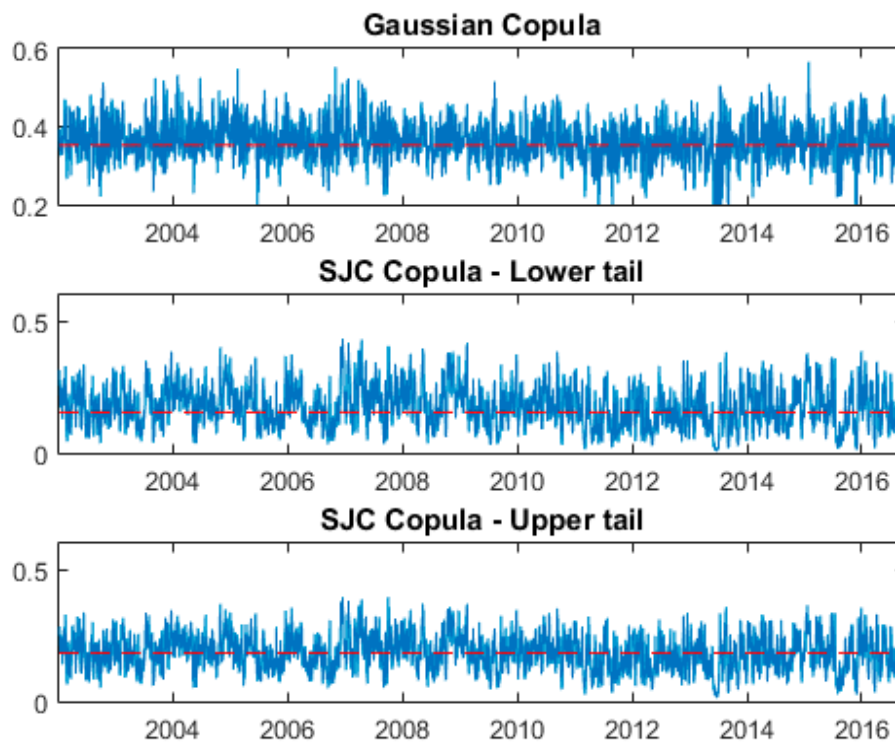


Figure 8. Dependence path of the time-varying copula for PM<sub>10</sub>–SO<sub>2</sub>.

For the PM<sub>10</sub>–CO pair (Figure 5), the parameters realized for the overall time-varying ( $\delta_t$ ) dependence are found to be in the mean of 0.6 and range of 0.5–0.7. For a lower tail dependence, the time-varying parameters ( $\tau^L$ ) fluctuate around the mean of 0.35 and range from 0.2 to 0.4. For an upper tail dependence, the time-varying parameters ( $\tau^U$ ) fluctuate around the mean of 0.42 and range from 0.39 to 0.42. The volatility properties of the time-varying parameters of the upper tail dependence are more stable than those of the overall and lower tail dependencies. They fluctuate around the mean of 0.6 and range from 0.5 to 0.7. Thus, the overall time-varying dependence for PM<sub>10</sub>–CO is quite high. For the lower tail and upper tail, their time-varying dependencies are relatively moderate. The dependence path of the time-varying copula for PM<sub>10</sub>–NO<sub>2</sub> is shown in Figure 6. The overall time-varying properties of the dependence parameters ( $\delta_t$ ) are found to be in the range of 0 to 0.5 with a mean of 0.35. Meanwhile, the time-varying lower dependence ( $\tau^L$ ) fluctuates between 0 and 0.3 with a mean of 0.13. The time-varying upper dependence ( $\tau^U$ ) fluctuates between 0.02 and 0.25 with a mean of 0.12. The fluctuations of the lower and upper dependency behaviour are more volatile than those of the overall time-varying dependence parameters ( $\delta_t$ ). Thus, we can conclude that the time-varying overall, lower tail, and upper tail dependencies for PM<sub>10</sub>–NO<sub>2</sub> are low.

As mentioned previously, Figure 7 shows the dependence path of the time-varying copula for the PM<sub>10</sub>–O<sub>3</sub> pair. The overall dependence parameters ( $\delta_t$ ) are found to be more volatile than those of the time-varying lower dependence ( $\tau^L$ ) and upper dependence ( $\tau^U$ ). The values of the parameter fluctuate around the mean of 0.12 within a range of 0.05–0.19. The parameters for the time-varying lower dependence ( $\tau^L$ ) and upper dependence ( $\tau^U$ ) fluctuate around a very low mean and very low range, except for some early points of the lower tail dependency before the year 2004. However, in general, these values are considerably lower than those for the PM<sub>10</sub>–CO and PM<sub>10</sub>–NO<sub>2</sub> pairs. Thus, we can conclude that the behaviour of the time-varying overall, lower tail, and upper tail dependencies for PM<sub>10</sub>–O<sub>3</sub> are low. The dependency behaviour of the PM<sub>10</sub>–NO<sub>2</sub> pair is shown in Figure 8. The values of the time-varying overall dependence parameter ( $\delta_t$ ), lower dependence ( $\tau^L$ ), and upper dependence ( $\tau^U$ ) indicate similar volatilities in the range of 0–0.5. The mean of the

overall dependence parameter is about 0.36, which is higher than the mean of the lower and upper dependence parameters. These values indicate the low magnitudes of the overall, lower tail, and upper tail dependencies for the PM<sub>10</sub>–NO<sub>2</sub> pair. The results shown in Figures 5–8 also show that the measure based on the constant copula (red dashed line) is not sufficient to describe the dependency behaviour among the pollutant variables over time. Nevertheless, the constant measure of dependency is found to provide a good approximation of the means of the fluctuations of the plots for overall dependence, lower tail dependence, or upper tail dependence (except for the lower and upper tails of PM<sub>10</sub>–O<sub>3</sub> pair).

## 8. Conclusions

This study investigated the behaviour of the dynamic dependence structure between several important pollutant variables, namely, PM<sub>10</sub>, CO, O<sub>3</sub>, NO<sub>2</sub>, and SO<sub>2</sub>. The distribution of each pollutant variable is non-normal, heavy tailed, and dynamically changes over time. Thus, a measurement determined by a common linear method, such as a Pearson correlation, which is subject to the normality assumption, will fail to provide accurate results. The Pearson correlation method is a rigid approach and cannot be used to evaluate the dynamic changes of the dependency behaviour among pollutant variables over time. Thus, this study proposes a method that combines the GARCH, GPD, and copula models to overcome air pollution data's non-normal, heavy tailed, and dynamically changing properties over time. The results in this work indicate that compared with the linear correlation method, the proposed method provides more information about the behaviour of the dependency structure among pollutant variables, particularly in terms of the overall time paths and lower and upper tail dependencies.

**Author Contributions:** Conceptualization, formal analysis, investigation, methodology and writing, funding acquisition, review and editing, N.M.; software, review and editing, S.I.H. All authors have read and agreed to the published version of the manuscript.

**Funding:** This work is supported by the Universiti Kebangsaan Malaysia (grant number FRGS/1/2014/SG04/UKM/03/1 and DIP-2018-038).

**Acknowledgments:** The authors would like to thank the editor for their time spent on reviewing our manuscript. The authors would also like to thank the reviewers for the careful and insightful review of our manuscript.

**Conflicts of Interest:** The authors declared no potential conflicts of interest with respect to the research.

## References

1. Torres, J.M.; Pérez, J.P.; Val, J.S.; McNabola, A.; Comesaña, M.M.; Gallagher, J. A functional data analysis approach for the detection of air pollution episodes and outliers: A case study in Dublin, Ireland. *Mathematics* **2020**, *8*, 225. [\[CrossRef\]](#)
2. Mannucci, P.M.; Franchini, M. Health effects of ambient air pollution in developing countries. *Int. J. Environ. Res. Public Health* **2017**, *14*, 1048. [\[CrossRef\]](#) [\[PubMed\]](#)
3. Masseran, N. Modeling fluctuation of PM<sub>10</sub> data with existence of volatility effect. *Environ. Eng. Sci.* **2017**, *34*, 816–827. [\[CrossRef\]](#)
4. AL-Dhurafi, N.A.; Masseran, N.; Zamzuri, Z.H. Compositional time series analysis for air pollution index data. *Stoch. Environ. Res. Risk Assess.* **2018**, *32*, 2903–2911. [\[CrossRef\]](#)
5. AL-Dhurafi, N.A.; Masseran, N.; Zamzuri, Z.H. Hierarchical-Generalized Pareto model for estimation of unhealthy air pollution index. *Environ. Model. Assess.* **2020**, *25*, 555–564. [\[CrossRef\]](#)
6. Department of Environment. *A Guide to Air Pollutant Index in Malaysia (API)*; Ministry of Science, Technology and the Environment: Kuala Lumpur, Malaysia, 1997.
7. Marković, D.M.; Marković, D.A.; Jovanović, A.; Lazić, L.; Mijić, Z. Determination of O<sub>3</sub>, NO<sub>2</sub>, SO<sub>2</sub>, CO and PM<sub>10</sub> measured in Belgrade urban area. *Environ. Monit. Assess.* **2008**, *145*, 349–359. [\[CrossRef\]](#)
8. Rich, D.Q.; Ozkaynak, H.; Crooks, J.; Baxter, L.; Burke, J.; Ohman-Strickland, P.; Thevenet-Morrison, K.; Kipen, H.M.; Zhang, J.; Kostis, J.B.; et al. The triggering of myocardial infarction by fine particles is enhanced when particles are enriched in secondary species. *Environ. Sci. Technol.* **2013**, *47*, 9414–9423. [\[CrossRef\]](#)

9. Xie, Y.; Zhao, B.; Zhang, L.; Luo, R. Spatiotemporal variations of PM<sub>2.5</sub> and PM<sub>10</sub> concentrations between 31 Chinese cities and their relationships with SO<sub>2</sub>, NO<sub>2</sub>, CO and O<sub>3</sub>. *Particuology* **2015**, *20*, 141–149. [[CrossRef](#)]
10. Hofert, M.; Kojadinovic, I.; Machler, M.; Yan, J. *Elements of Copula Modeling with R*; Springer: Cham, Switzerland, 2018.
11. Cherubini, U.; Mulinacci, S.; Gobbi, F.; Romagnoli, S. *Dynamic Copula Methods in Finance*; Wiley: West Sussex, UK, 2011.
12. AL-Dhurafi, N.A.; Masseran, N.; Zamzuri, Z.H.; Safari, M.A.M. Modeling the air pollution index based on its structure and descriptive status. *Air Qual. Atmos. Health* **2018**, *11*, 171–179. [[CrossRef](#)]
13. Masseran, N.; Safari, M.A.M. Risk assessment of extreme air pollution based on partial duration series: IDF approach. *Stoch. Environ. Res. Risk Assess.* **2020**, *34*, 545–559. [[CrossRef](#)]
14. Masseran, N.; Safari, M.A.M. Intensity–duration–frequency approach for risk assessment of air pollution events. *J. Environ. Manag.* **2020**, *264*, 110429. [[CrossRef](#)] [[PubMed](#)]
15. Chelani, A. Long-memory property in air pollutant concentrations. *Atmos. Res.* **2016**, *171*, 1–4. [[CrossRef](#)]
16. Lee, C.-K. Multifractal characteristics in air pollutant concentration time series. *Water Air Soil Pollut.* **2002**, *135*, 389–409. [[CrossRef](#)]
17. Weng, Y.-C.; Chang, N.-B.; Lee, T.Y. Nonlinear time series analysis of ground-level ozone dynamics in Southern Taiwan. *J. Environ. Manag.* **2008**, *87*, 405–414. [[CrossRef](#)] [[PubMed](#)]
18. Liu, Z.; Wang, L.; Zhu, H. A time–scaling property of air pollution indices: A case study of Shanghai, China. *Atmos. Pollut. Res.* **2015**, *6*, 886–892. [[CrossRef](#)]
19. Poon, S.H.; Rockinger, M.; Tawn, J. Extreme value dependence in financial markets: Diagnostics, models, and financial implications. *Rev. Financ. Stud.* **2004**, *17*, 581–610. [[CrossRef](#)]
20. Engle, R. Dynamic conditional correlation: A simple class of multivariate generalized autoregressive conditional heteroskedasticity models. *J. Bus. Econ. Stat.* **2002**, *20*, 339–350. [[CrossRef](#)]
21. Engle, R.; Colacito, R. Testing and valuing dynamic correlations for asset allocation. *J. Bus. Econ. Stat.* **2006**, *24*, 238–253. [[CrossRef](#)]
22. Silvennoinen, A.; Teräsvirta, T. Modeling multivariate autoregressive conditional heteroskedasticity with the double smooth transition conditional correlation GARCH model. *J. Financ. Econom.* **2009**, *7*, 373–411. [[CrossRef](#)]
23. Tsafack, G. Asymmetric dependence implications for extreme risk management. *J. Deriv.* **2009**, *17*, 7–20. [[CrossRef](#)]
24. Masseran, N.; Safari, M.A.M. Modeling the transition behaviors of PM<sub>10</sub> pollution index. *Environ. Monit. Assess.* **2020**, *192*, 1–15. [[CrossRef](#)] [[PubMed](#)]
25. Masseran, N.; Razali, A.M.; Ibrahim, K.; Zaharim, A.; Sopian, K. Application of the single imputation method to estimate missing wind speed data in Malaysia. *Res. J. Appl. Sci. Eng. Technol.* **2013**, *6*, 1780–1784. [[CrossRef](#)]
26. Masseran, N.; Razali, A.M.; Ibrahim, K.; Latif, M.T. Modeling air quality in main cities of Peninsular Malaysia by using a generalized Pareto model. *Environ. Monit. Assess.* **2016**, *188*, 65. [[CrossRef](#)] [[PubMed](#)]
27. Zhao, X.; Zhang, Z.; Cheng, W.; Zhang, P. A new parameter estimator for the Generalized Pareto Distribution under the peaks over threshold framework. *Mathematics* **2019**, *7*, 406. [[CrossRef](#)]
28. Patton, A.J. Modelling asymmetric exchange rate dependence. *Int. Econ. Rev.* **2006**, *47*, 527–556. [[CrossRef](#)]
29. McNeil, A.J.; Frey, R. Estimation of tail-related risk measures for heteroscedastic financial time series: An extreme value approach. *J. Empir. Financ.* **2000**, *7*, 271–300. [[CrossRef](#)]
30. Pfaff, B. *Financial Risk Modelling and Portfolio Optimization with R*, 2nd ed.; Wiley: West Sussex, UK, 2016.

**Publisher’s Note:** MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.



© 2020 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<http://creativecommons.org/licenses/by/4.0/>).