Extreme and Records Value Analysis for Evaluating Air Quality in Bekaa Valley, Lebanon

Alya Atoui, Sami Abbad Andaloussi, Kamal Slim, Régis Moilleron, and Zaher Khraibani*

***Abstract*—Air pollution is a major public health problem that affects economic development, agriculture, and the ecosystem. Lebanon is one of the most polluted countries in the Middle East region due to the increase in the concentrations of atmospheric air pollution which exceed the required levels according to the global guidelines. The main objective of this paper is to identify the extreme concentrations of air pollutants in order to minimize their adverse effects. The peak of concentration of the pollution which is measured within a specific period could be described by using the extreme value theory mainly as one kind of the three different types of the extreme value theory and the record theory as a second kind. These two approaches will be applied to predict the expected extreme concentration in the future and the probability of occurrence of a new record. Whereas daily measurements of SO_2, CO, NO, PM_{10}, $PM_{2.5}$ throughout the period 2016-2019 in the Bekaa Valley. The findings indicate that the concentrations of the 15, 30, 40, and 50-year return pollutant levels are continuously increasing. The percentage of the change of SO_2, CO, NO, PM_{10}, $PM_{2.5}$ after 50 years is 64%, 5%, 3%, 29%, 20% and 12%, respectively. The records have been observed at the beginning of a time series and an interval point prediction was given for each measure. The future record values of SO_2, CO, NO, PM_{10}, $PM_{2.5}$ were increased by 0.2%, 0.7%, 1%, 0.5%, 6%, and 5.7%, respectively, over 1 year. There was a 66% chance to observe a new record-breaking pollutant level that exceeded the guidelines after two years.**

***Index Terms*—Air pollution, extreme, prediction, records.**

I. Introduction

Atmospheric air quality has emerged as a significant issue in the last century and a half, with effects on people varying according to the level of pollution and duration of the exposure, from respiratory illness to heart attack to death. The effects of air pollution on humans vary depending on the level of air quality and the length of time of exposure, from respiratory problems to heart failure or death. Many studies have been initiated to control the adverse effects of industrial pollution, deforestation, hazardous disposal, land degradation ,and other urgent environmental issues. Atmospheric air pollution has increased to an alarming level and premature mortality from exposure to air pollution in the world increases each year (WHO) [1]. There are numerous people who are already being exposed to air pollutants, and pollution is still rising due to a number of different factors, among them the burning of fossil combustibles. The landscape in the Middle East is varied and consists of high ground, deserts and the Sahara, coastal zones, and large areas of plain lands. Climate differs across regions, it experiences strong episodes of heavy pollution, high levels of particulates, and heavy levels of acid deposits [2] in several regions due to large industrialized areas, lack of any effective public transportation system, dense traffic areas, and high population densities [3]. Located on the Mediterranean Sea, Lebanon is a relatively compact state with an increasing population, especially in the urban areas, where there is no system of public transit (MoE) available altogether [4]. Moreover, the country is exposed to steady winds from eastern European countries and to Saharan desert windstorms, which make Lebanon among the dirtiest countries in the region of the Middle East. This makes it among the most environmentally polluted in the Middle East region. The major environmental challenges are caused by a failure to develop appropriate policies and environmental facilities. Atmospheric air pollution from numerous different sources was identified as being one of the most pressing health issues the country confronts, particularly in urban zones [5]. In recent several decades, atmospheric air pollution has been considered a significant problem in Lebanon, which affects public health and the environment generally. Consequently, surveillance and analyses of the extreme pollutant concentrations will be critical in order to control environmental danger on one hand and to reduce the human exposition and the health hazards involved on the other. Lebanon's urban environment is affected by high levels of air pollutants from industrial and transportation sectors in the urban areas, as well as high levels of air quality [6]. Climate change, especially increased environmental temperature, is projected to severely impact biodiversity [7]. The average number of deaths due to air pollution in Lebanon is noted to be 2700 cases in 2018, or an average of 4 deaths for every 10000 people. This is considered to be the largest in the Mediterranean region [8]. Research showed that the increasing number of cars, the uncontrolled industry, and the privately operated generator which are located in the residential areas are mainly the cause of the persistence of air pollution [9]. Regional traffic is an important factor in the high concentrations of ozone, the diesel engines generate large quantities of particulate pollutants, and pollutant concentrations are above WHO guidelines in most Lebanese areas [10] thus requiring emergency measures. So far, few research have been conducted to investigate air pollution in Lebanon. Consequently, atmospheric air pollution in Lebanon should be studied and controlled through statistical prediction approaches. Numerous researchers have investigated temporal distributions of atmospheric pollution through the use of probabilistic models [11], [12], and statistical descriptive methods [13]. Time series analysis

Manuscript received April 27, 2022; revised May 9, 2022; accepted June 27, 2022.

Alya Atoui, Sami Abbad Andaloussi, and Régis Moilleron are with the Univ Paris Est Creteil, LEESU, F-94010, Creteil, France.

Kamal Slim was with the National Council for Scientific Research, Lebanon

Zaher Khraibani was with the Lebanese University, Faculty of Sciences, Department of Applied Mathematics, Hadat, Lebanon.

*Correspondence: Zaher.khraibani@ul.edu.lb

doi: 10.18178/ijesd.2023.14.1.1408

methods [14], [15] where the aim is to demonstrate trends in concentrations of the pollutants over time, and hence impact of air pollutants on health, have also been employed [16]-[18]. Some other researchers used statistical approaches such as Principal Component Analysis (PCA) and Hierarchical Agglomerative Clustering (HAC) [19]. Consequently, this study, being the first in Lebanon, investigates pollutant performance through record-based analysis of pollutant concentrations extremes and the theory of the most important ones applied to air pollutant records. This article seeks an introduction to a new extreme values pattern that is more appropriate for record-based environment surveys. The theory allows efficient problem recognition of air pollutants. The presented findings may be used as an air quality control management instrument providing policymakers the means to identify actions necessary for pollution abatement and create a warning mechanism for very high peaks. Results suggest that both extremes and especially record theory are easily applied and can reach an even greater precision than other models in case of a restricted observational database. The extreme phenomena of atmospheric air quality are of special relevance. They are a risk to human beings and may also affect the ecosystems of a city, a nation, and even a planet. The Extreme Value Theory (EVT) offers a valuable framework to analyze the air pollution quality data and predict the future of extreme events while detecting the level and return period of the extremes. The main purpose of the study (EVT) is to develop strategies to reduce criteria air pollutants. Numerous research investigations have used the theory of EVT and records for such environmental purposes as extreme temperature, extreme precipitation, extreme wavelength, etc. [20]-[24]. This approach is advantageous as it shows extreme values and their return period, which in turn assists with creating a system of warnings about the future. Furthermore, it is different from the other statistics approaches that look at the averages of data, as when trying to analyse the environmental data, the focus must be on extreme values that are alarming. It should be noticed that record keeping has been applied in several fields. From Climate change to the environmental sector, to finance and sports. This approach has proven to be relevant to the temporal evolution of a data set over time [25]-[27]. The application of these extreme approaches to modeling the air quality will assist to model a pattern of pollution extreme or peak values, to predict future peak pollution, and identify the time of year when pollution level is the greatest. In this article, we are interested in the study of air pollutants like carbon monoxide ($CO\ mg|m^3$), nitrogen dioxide ($NO_2\ \mu g|m^3$), and sulfur dioxide ($SO_2\ \mu g|m^3$) which are now recognized as indicators of anthropogenic emissions (vehicles, industry, construction, ports, airports, road networks, etc.), and particulate matter ($PM_{10}\ \mu g|m^3$ and $PM_{2.5}\ \mu g|m^3$) which can be very harmful to human health. The study was made to encompass Bekaa region, which is an elaborate area with a variety of human activities. The Bekaa region is witnessing a catastrophic environmental scenario affecting the air and water quality in the region, and thus the public health and the ecosystem in general. Signs of climate change such as storms, floods, and heat waves are becoming usual and periodic [20]. These extreme events which are a result of high pollution are affecting the Lebanese economy, especially in the Bekaa region which depends primarily on agricultural activity [28].

II. Background

In general, the theory of extremes corresponds to the study of the occurrence of events with a low probability whose objective is to obtain reliable estimates of the probabilities of occurrence of rare events. This theory is based on the theory of probability to study the tails of the distribution of a sequence of independently and identically distributed random variables. This paper aims to introduce a new model of extremes that is based on the theory of records to predict a new record of pollutant concentrations. Record theory is related to extreme value theory. This theory involves studying the limiting distribution of the maximum $M_n = max(X_1, \ldots, X_n)$ of a series of independent and identically distributed random variables. Record theory deals with the exact distribution of the number of records N_n and record times L_n). These two distributions do not depend on the distribution of observations, so it takes into account the temporal structure of $\{X_1, \ldots\ldots, X_n\}$ insofar as we study the instants of the L_n jumps of M_n that correspond to the instants of records.

A. Extreme Value Analysis

Studies of extremes began in the 1920s (Fisher and Tippett, 1928 [29]), and developed rapidly (Gumbel, 1942 [30]); Leadbetter *et al*., 1983 [31]. Two main methods are developed in the study of extreme values. The first is the block maxima method based on the limit distribution of the maximums (Fisher-Tippett theorem): Assume that $X_1, \ldots, X_n$ be a sequence of independent and identically distributed random variables and $M_n = \max(X_1, \ldots, X_n)$. If a sequence of pairs of real numbers (a_n, b_n) exists such that $a_n > 0$ and $\lim_{x \to +\infty} P\left(\frac{M_n - b_n}{a_n} \leq x\right) = G(x)$, where G is a non degenerate distribution function. The limit distribution G belongs to one of the three distributions below called the generalized extreme value distributions (GEV) which defined respectively by Gumbel, Frechet, Weibull:

- $\Omega(x) = \exp[-\exp(-x)]$, where $x \in \mathbb{R}$
- $\phi_\theta(x) = \exp[-x^{-\theta}]$, where $x \geq 0$ and $\theta > 0$
- $\psi_\theta(x) = \exp[-(-x)^{\theta}]$, where $x \geq 0$ and $\theta > 0$.

The general form of (GEV) distribution is given by:

$$G_{\gamma,\Gamma,\sigma}(x) = G_\gamma\left(\frac{x-\Gamma}{\sigma}\right) = \exp\left[-\left(1+\gamma\left(\frac{x-\Gamma}{\sigma}\right)\right)^{-\frac{1}{\gamma}}\right] \quad (1)$$

$1 + \gamma\left(\frac{x-\Gamma}{\sigma}\right) > 0$, $\gamma \in \mathbb{R}$, $\Gamma \in \mathbb{R}$, $\sigma > 0$, Γ the location parameter, σ the scale parameter and γ the shape parameter.

The second approach is the peak over threshold (POT) method, where we select a threshold u, and every value exceeding this threshold is considered an extreme value.

The extreme value theory consists in studying the observations which exceed a threshold u. The excess Y of the variable X above the threshold is given by $X - u$ if $X > u$. The distribution function F_u of the observations above the threshold u is given for $y > 0$:

$$F_u(y) = \frac{F(u+y) - F(u)}{1 - F(u)} \quad (2)$$

For a large enough threshold u, the survival function is approximated by a Generalized Pareto Distribution (GPD), (Pickands theorem 1975 [32], Balkema and de Haan 1974 [33]):

$$\bar{H}_{\gamma,\sigma}(y) = \begin{cases} \left(1+\gamma\frac{y}{\sigma}\right)^{-\frac{1}{\gamma}} & \text{if } \gamma \neq 0 \\ exp\left(\frac{-y}{\sigma}\right) & \text{if } \gamma = 0 \end{cases}$$

The convergence between the distribution of the maximum to a (GEV) and that of the threshold exceedances to a (GPD) :

$$\lim_{n\to\infty} P\left(\frac{X_{n,n}-a_n}{b_n} \leq x\right) = G_\gamma(x)$$

$$\rightarrow \lim_{u\to x_{F[0,x_F-u]}} \sup \left|\bar{F}_u(y) - \bar{H}_{\gamma,\sigma(y)}(y)\right| = 0 \quad (3)$$

where x_F is the endpoint of the cumulative distribution function F. Both methods provide the $T-year$ return period of a given variable, with a probability of $1/T$.

B. Threshold Selection

There is no precise method for choosing the threshold. In general, the threshold chosen should be small so that the number of observations that exceed the threshold provides an accurate estimate of the model parameters. In the literature, the "Mean Excess Function" (MEF) is used to choose the threshold u [34].

$$MEF(u) = E(X-u|X>u),$$

If the adjustment of exceedance with a valid (GDP) with a certain threshold u_0 and for $\gamma < 1$ then:

$$E(X-u_0|X>u_0) = \frac{\sigma_{u_0}}{1-\gamma}$$

For every $u > u_0$:

$$E(X-u|X>u) = \frac{\sigma_u}{1-\gamma} = \frac{\gamma}{1-\gamma}u + \frac{\sigma_{u_0}-\gamma u_0}{1-\gamma}$$

MEF is a linear function in terms of u. The smallest value of u is given such that the (MEF) becomes linear for $\gamma < 1$.

C. Probability and Return Period

The probability of occurrence of pollutant concentrations exceeding a given threshold as well as the return period is calculated based on the limits recommended by the World Health Organization (WHO) [35]. The return period is represented by the time of occurrence of high peaks of pollutant concentrations that exceed the thresholds in the station "Memshieh Garden". In the following, the probability and the return period are calculated in both cases (GEV) and (BM).

$$p = 1 - \exp\left\{-\left[1+\gamma\left(\frac{x-\mu}{\sigma}\right)\right]^{-\frac{1}{\gamma}}\right\} \quad (4)$$

p is the probability and $T = 1/p$ is the return period.

Z_p the return level of the $1/p$ observation. will be exceeded every $1/p$ observations.

$$Z_p = F^{-1}(p)$$

$F^{-1}(\mathrm{p})$ is the approximation of the tail survival function:

$$F^{-1}(p) \approx \begin{cases} u + \frac{\sigma}{\gamma}\left(\left[\frac{p}{F(u)}\right]^{-\gamma} - 1\right) & \text{if } \gamma \neq 0 \\ u - \sigma\log\left(\frac{p}{F(u)}\right) & \text{if } \gamma = 0 \end{cases}$$

where p is small enough in order to have $Z_p > u$.

D. Records Theory

Records theory began in (1952) with Chalender [36] who introduced a basic properties of records. The main results were given in the period 1952-1983. Assume a sequence of observations $(X_n)_{n\geq 1}$ independent and identically distributed random variables observed successively from a cumulative distribution function F. The record process is composed by (R_n, L_n). The record-breaking index $\{L_n, n > 0\}$ is defined by:

$$L_0 = 1 \text{ with probability } 1$$

for $n > 1$,

$$L_n = min\{j: X_j > X_{L_{n-1}}\}$$

The record value $\{R_n\}$ is then defined by:

$$R_n = X_{L_n}, \qquad n = 0,1,2,\ldots$$

R_0 is the record trivial. The number of records N_n is defined by:

$$N_n = \{\text{number of records among } X_1,\ldots,X_n\}.$$

where $N_1 = 1$, X_1 is a trivial record. The number of record is defined by: $N_n = \sum_{i=1}^{n} \delta_i$; with δ_i is a sequence of record indicator:

$$\delta_i = \begin{cases} 1 & \text{if i} = 1 \\ \mathbb{I}\{X_i > max(X_1,\ldots,X_{i-1})\} & \text{if i} > 1 \end{cases}$$

By symmetry the δ_i has a Bernoulli distribution, $Ber\left(\frac{1}{i}\right)$.

E. Extreme Records Value

It is necessary to know the distribution of extreme values to study the records.. Based on the subsection A, the (GEV) distribution is defined by:

$$F(x) = 1 - exp\left[-e^{\left(\frac{x-\mu}{\sigma}\right)}\right],$$

where $\mu \in \mathbb{R}$, $\sigma > 0$. The upper records values be observed from an extreme value distribution with density function:

$$f(x,\mu,\sigma) = \frac{1}{\sigma}e^{(x-\mu)/\sigma)}e^{-e^{(}x-\mu)/\sigma)}, x \in \mathbb{R} \quad (5)$$

For such a random variable we have

$$X \overset{d}{=} \mu + \sigma\log X^*,$$

where $X^* \sim Exp(1)$. The corresponding record value sequence can be described by

$$R_n \overset{d}{=} \mu + \sigma \log\left(\sum_{i=0}^{n} X_i^*\right).$$

Arnold [37] give the expected value and the variance of R_n:

$$E(R_n) = \mu - \sigma\gamma + \sigma \sum_{j=1}^{n} \frac{1}{j},$$

$$V(R_n) = \sigma^2 \left(\frac{\Pi^2}{6} - \sum_{j=1}^{n} \frac{1}{j^2}\right).$$

The expected value and the variance of the number of records are given by:

$$E(N_n) = \sum_{j=1}^{n} \frac{1}{j} \approx \ln(n) + \gamma, \quad (6)$$

$$V(N_n) = \sum_{j=1}^{n} \frac{1}{j}\left(1 - \frac{1}{j}\right) \approx \ln(n) + \gamma - \frac{\pi^2}{6}. \quad (7)$$

where γ is Euler's constant 0.5772. The joint probability distribution, $f(r_0, \ldots, r_n)$ of the record values $R_0, R_1, \ldots, R_n$ from a continuous cumulative distribution function $F(r)$, is defined by:

$$f_{R_0,\ldots,R_n}(r_0, \ldots, r_n) = f(r_n) \prod_{i=0}^{n-1} h(r_i),$$

$$for \ -\infty < r_0 < r_1 < \ldots < r_n \quad (8)$$

where $h(r) = f(r)/1 - F(r)$ is the hazard rate function.

III. Records Inference

In this section, the statistical inference, point estimation, interval estimation, and prediction of air pollutant concentrations are discussed. The maximum likelihood estimation of the distribution parameters (MLE) is discussed. In addition, the best linear unbiased parameter estimate is shown. The point prediction of the future records in which the best linear unbiased prediction and the best linear invariant prediction are described.

A. Maximum Likelihood Estimation

Assume that $R_0, R_1, \ldots, R_n$ are the upper record values observed from any location-scale distribution with cumulative distribution function $F(x, \mu, \sigma) = F\left(\frac{x-\mu}{\sigma}\right)$.

The joint probability distribution of $R_0, R_1, \ldots, R_n$ is:

$$f(r_0, r_1, \ldots, r_n; \mu, \sigma) = \frac{1}{\sigma^{n+1}} f\left(\frac{r_n - \mu}{\sigma}\right) \prod_{i=0}^{n-1} \left[\frac{f\left(\frac{r_i-\mu}{\sigma}\right)}{1-F\left(\frac{r_i-\mu}{\sigma}\right)}\right] \quad (9)$$

From equation (8), the likelihood function is given by:

$$L = \frac{1}{\sigma^{n+1}} \prod_{i=0}^{n-1} \left(\frac{f(r_i^*)}{1-F(r_i^*)}\right) f(r_n^*), \quad (10)$$

$$-\infty < r_0^* < r_1^* < \ldots < r_n^* < \infty$$

where $r_i^* = \frac{r_i - \mu}{\sigma}$ for $i = 0,1, \ldots, n$.

The log-likelihood function is obtained from equation (10) [38]:

$$\log L = -(n+1)\log\sigma - \sum_{i=0}^{n-1} \log(1 - F(r_i^*)) + \sum_{i=0}^{n} \log f(r_i^*), \quad (11)$$

The Best Linear Unbiased Estimator (BLUE) of μ and σ is given by (David, 1981), [39] and (Balakrishnan, 1991), [40]:

$$\mu^* = R_n - \frac{\alpha_n}{n} \sum_{i=0}^{n-1} (R_n - R_i) \quad (12)$$

$$\sigma^* = \frac{1}{n} \sum_{i=0}^{n-1} (R_n - R_i) \quad (13)$$

where $\alpha_n = -\gamma + \sum_{i=0}^{n-1} \frac{1}{i}$ with γ is Euler's constant.

B. Future Records Prediction

We focus on the prediction of record values of pollutants to evaluate the extreme records. We first introduce the point prediction of a future record and then present the conditional prediction intervals for future records. Based on the BLUE's defined above (μ^*, σ^*), the Best Linear Unbiased Predictor (BLUP) R_{n+1}^* of the records values $R_n$in global case do to Goldberger (Goldberger, 1962):

$$R_{n+1}^* = \mu^* + \alpha_m \sigma^* + \omega^T \Sigma^{-1}(R - \mu^* \mathbf{1} - \sigma^* \alpha) \quad (14)$$

where R is the vector of observed record values, **1** is a vector of **1**'s, α is the vector of means records values, Σ is the variance-covariance matrix of the standard record values and $\omega^T = (\sigma_{0,n}, \sigma_{1,n}, \ldots, \sigma_{n,m})$. The BLUE of σ_n^* is given by the equation (13). Similarly, the BLUE of σ based on $(R_0, R_1, \ldots, R_{n+1})$is given by:

$$\sigma_{n+1}^* = \frac{1}{n+1} \sum_{i=0}^{n-1} (R_{n+1} - R_i) \quad (15)$$

From the two equations (13), and (15), we get the point prediction of the future record

$$\begin{aligned} R_{n+1}^* &= R_n - \frac{1}{n} \sum_{i=0}^{n-1} R_i + \frac{1}{n+1} \sum_{i=0}^{n} R_i \\ &= R_n + \frac{1}{n(n+1)} \sum_{i=0}^{n-1} (R_n - R_i). \end{aligned} \quad (16)$$

Let us take $EV(\mu, \sigma)$ distribution considered in section E, the $100(1-\alpha)\%$ conditional prediction interval for R_{n+1} is given by Chan (1998) [38]:

$$[R_n + \sigma^* z_{3,\alpha/2}, R_n + \sigma^* z_{3,1-\alpha/2}] \quad (17)$$

where $z_{3,\alpha}$ the $\alpha - th$ quantile of the conditional distribution:

$$z_{3,\alpha} = \left(\frac{n}{n+1}\right)(1-\alpha)^{-1/n} - 1 + \frac{1}{n+1}.$$

By Chandler (1952) [36], the waiting time T_2^* to observe a new record increases proportionally to the number of records.

The probability to observe a new record in the coming n_2 years given observations for n_1 years is defined by:

$$P(T_2^* > n_2) = \frac{n_1}{n_1+n_2} \quad (18)$$

IV. Surveillance Station and Data Collection

The Environmental Resources Monitoring in Lebanon (ERML) project began in 2013 with the support of the United Nations Environment Program (UNEP) and the United Nations Development Program (UNDP). Under the project, the Lebanese Ministry of Environment has installed Air Quality Surveillance Stations (AQMS) in various sites in major Lebanese cities (Hadath, Beirut, Baalbeck, Zahleh, Saida), it was the first real-time air monitoring system aimed at tracking the exposure of the population from industries, electric plants, and traffic on the road in addition to urban sources of air pollutants. The monitoring stations include web-based analyzers for monitoring criteria pollutants: $SO_2\ \mu g|m^3$, $NO\ \mu g|m^3$, $NO_2\ \mu g|m^3$, $CO\ mg|m^3$, $PM_{10}\ \mu g|m^3$, $PM_{2.5}\ \mu g|m^3$. These settings were recorded hourly between 01/01/2016 and 19/06/2019 in the station "Memshieh Garden - Zahleh" located in Bekaa valley at an altitude of 1150 m with a latitude of $33°\ 51'3.01"\ \mathrm{N}(34{,}270\ )$ and a longitude of $35°53'45.54"$ E (Fig. 1). The dataset is composed of a matrix of size (30370×6). It should be emphasized that this study is made on one specific site which is located close to Litani river, therefore results may differ from other stations or data of different time periods.

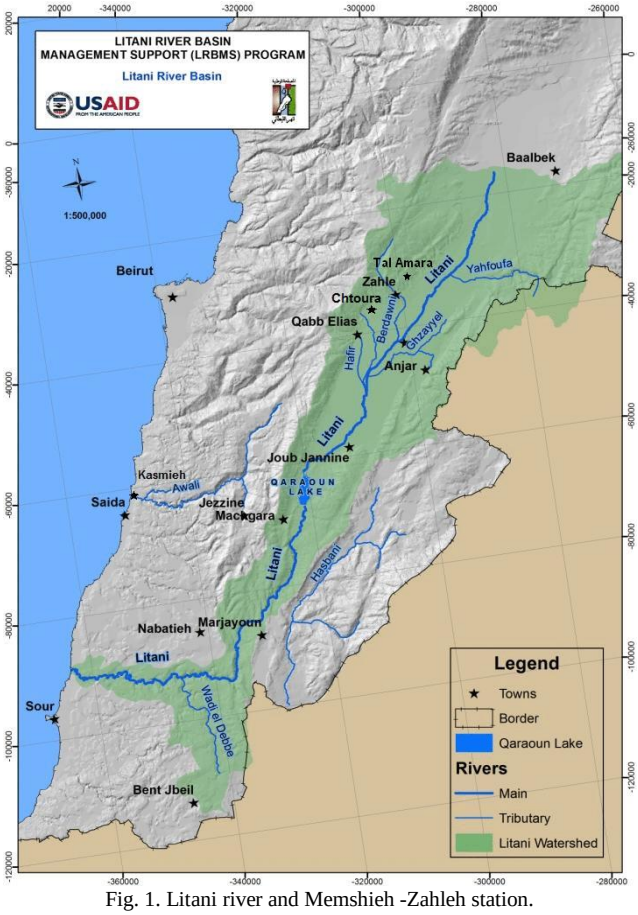


Fig. 1. Litani river and Memshieh -Zahleh station.

Within the framework of the monitoring of environmental resources in Lebanon the monitoring of air quality, through the design and implementation of a national air quality monitoring network several monitoring stations are settled in Lebanon. Fig. 1 shows the Memshieh Garden station, the choice of this station to study the air quality in the Bekaa region which is characterized by agriculture and tourism and its geographical location. In addition, there are no previous studies to study the air quality and the extreme concentrations of these measurements in this region.

V. Results and Discussion

The following section presents an application of the extreme values theory and the records theory on the complete data of six pollution parameters using the statistical software "R". The principal objective is to predict the extreme quantiles and calculate the return level of each pollution parameter to forecast future levels of atmospheric pollution.

A. The Data

Data collected from the Lebanese Ministry from 2016 to 2019 are registered on each hour (MoE: National Environmental Action Plan, 2020). The initial sample size equals n=30370 observations. We encountered missing data (NA). This was caused due to either technical issues, the monitoring station being out of order for many months, or mistakes in data input. The measurements are registered hourly and data is validated on a daily, a monthly, and a yearly basis as a final validation step. The data validation is generally carried out with the use of software that automatically checks the data transmission and the possibility of a technical problem. If there is a reported issue, a service technician will check the monitoring sites and either replace or correct the monitoring equipment. Nevertheless, some data gaps must be completed for use in the survey. To address the missing data problem, we use the R software package "missForest" which provides an imputing algorithm for the missing value that is random forest (RF) method based [41]. We use (RF) method since it can deal with high-dimensional datasets and provides robust results. It is a nonparametric approach, meaning it can deal with complex and non-linear data interaction and does not consider probability distributions of data. In addition, (RF) is able to approximate the out-of-bag (OOB) error without requiring a testing package. An explanation for ((OOB) is described in the article (James *et al*. (2013)) [42]. Some discrepancy is registered for the interquartile range, the median, and the mean and also for the measure of kurtosis. The OOB error rate is 3.46%. Nevertheless, in this article, we are interested in the study of extreme pollutants. The maximum 24-hour maximum was used for the air pollutant. The maximum sampling size to which we are applying the methods of the extremes and the records is *n*=1265 observations. A summary of missing data is shown in the table that follows.

Table I represents the missing value for each parameter of the initial data set. The percentage rate of missing values in the entire data set is 14.013%. This percentage is relatively small compared to the available data, but it varies from one parameter to another, we notice the presence of high percentages for SO_2, PM_{10} and $PM_{2.5}$. To avoid bias in the analysis processes, the data are supplemented by using the

"missForest" algorithm. We proceed in the following with some classical descriptive statistics to get information about the air quality parameters. It should be noted that the presence of high variability in the air quality parameters affects some descriptive statistics [15]. The following table is the summary of the new dataset produced by the " missForest " algorithm. The descriptions of the maximum daily observations are given in the table that follows:

TABLE I: NUMBER AND PERCENTAGE OF MISSING VALUES FOR EACH PARAMETERS

Parameters	SO_2 $\mu g\|m^3$	NO $\mu g\|m^3$	NO_2 $\mu g\|m^3$	CO $mg\|m^3$	PM_{10} $\mu g\|m^3$	$PM_{2.5}$ $\mu g\|m^3$
NA	7765	2631	2398	589	6593	6597
%NA	22.57%	7.66%	6.9%	7.52%	19.71%	19.72%

TABLE II: DESCRIPTIVE STATISTICS FOR THE MAXIMUM HOURLY DURING THE 24 HOURS FOR AIR POLLUTION

Parameters	SO_2 $\mu g\|m^3$	NO $\mu g\|m^3$	NO_2 $\mu g\|m^3$	CO $mg\|m^3$	PM_{10} $\mu g\|m^3$	$PM_{2.5}$ $\mu g\|m^3$
Min	0.19	2.02	11.88	0	0	0
1st Qu.	5.836	20.3	55.29	1.72	43.52	33.42
Median	12.22	46.78	77.07	3.44	53.6	43.10
Mean	19.18	67.36	81.84	35.86	84.81	69.32
3st Qu.	22.25	93.67	100.9	76.33	88.10	72.10
Max.	318.58	493.64	283.0	313.120	1553	1454.8
C.V.	149.0	94.32	41.08	133.89	132.14	144.82

The coefficient of variation (C.V.) is a measure of variation over time [43], which allows us a comparison of pollutants in which $SO_2 \mu g|m^3$ has the highest variation over time while NO_2 has the lowest variation. The variations of these parameters can be due to several factors a meteorological conditions SO_2 $\mu g|m^3 < PM_{2.5} \mu g|m^3 < PM_{10} \mu g|m^3 < CO$ $mg|m^3 < NO$ $\mu g|m^3 < NO_2 \mu g|m^3$. The values in Table II represent the statistical description of the maximum values of each 24h over three years of hourly measurements. With respect to WHO guidelines, the yearly average values of all the pollutants exceed the permissible limits, with the exception of SO_2 $\mu g|m^3$, which is still within the permissible range even though it is very close to the upper limit.

TABLE III: CORRELATION COEFFICIENT BETWEEN THE DIFFERENT POLLUTANTS

Parameters	SO_2 $\mu g\|m^3$	NO $\mu g\|m^3$	NO_2 $\mu g\|m^3$	CO $mg\|m^3$	PM_{10} $\mu g\|m^3$	$PM_{2.5}$ $\mu g\|m^3$
SO_2	1					
NO	0.23	1				
NO2	0.218	0.67	1			
CO	0.43	-0.24	-0.25	1		
PM_{10}	0.29	0.35	0.41	-0.025	1	
$PM_{2.5}$	0.23	0.39	0.44	-0.09	0.91	1

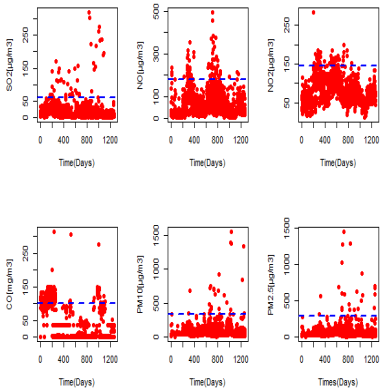


Fig. 2. Maximum hourly per day for air pollution parameters in Memshieh garden station.

Also we present in Table III the matrix of correlation coefficients between air pollutants parameters, which shows us that most of the air pollutants are positively correlated with the exception correlation coefficient between PM_{10} $\mu g|m^3$ and respectively NO $\mu g|m^3$, NO_2 $\mu g|m^3$, CO $mg|m^3$ and $PM_{2.5} \mu g|m^3$. The positive correlation indicates coincident time fluctuations of air pollutants. The higher positive correlations between $NO \mu g|m^3$, $NO_2 \mu g|m^3$, CO $mg|m^3$ were detected.

In Fig. 2, the time series for the maximum air pollution concentration recorded per day between 2016 and 2019 is presented. The concentrations of both CO $mg|m^3$ and NO $\mu g|m^3$ seem to trend. Also evident is the fact that SO_2 $\mu g|m^3$ and NO $\mu g|m^3$ do not have a consistent significant time trend. The peak $PM_{2.5}$ $\mu g|m^3$ shows a trend of increase with respect to time. The blue dashed line is the threshold u, which determines the "average exceedance plot". In the following section, we explore the pollutant exceedance concentration over this threshold by using the (POT) approach.

B. Univariate Extreme Values

This subsection describes the theory of extreme values for the air quality monitoring data. In order to adjust the extreme value distribution, in general, two methods are provided: Block Maxima (BM) and the Peak Over Threshold (POT). The statistical results were carried out with the R software.

The (BM) involves breaking down data into a number of blocks (monthly blocks) and selecting the peak of each one to find a distribution that fit the peak of the data.

The distribution adjusted (GEV) (equation (1)) for the data is shown as follows:

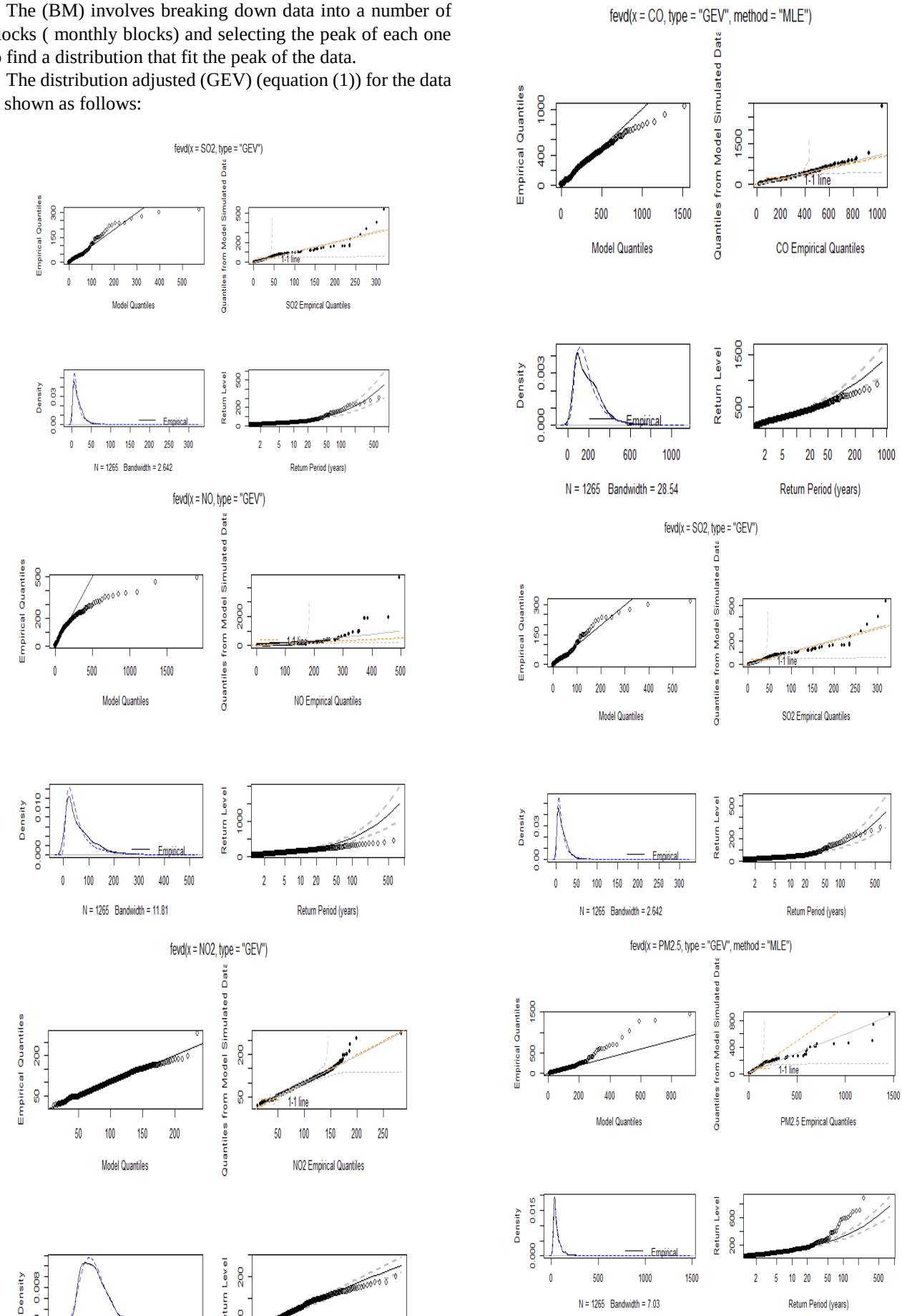


Fig. 3. Diagnostic plot for the fitted GEV model for each maximum pollutant.

We note that the probability plot has a linear pattern (close

to straight line) and that the two curves of the estimated and the empirical densities seems to be coherent and fitting to each other, indicating that the adjusted EGV gives a reasonable fit to data.

However the quantiles plot exhibits some deviation from the linearity, which is probably related to the higher uncertainty level. Furthermore, empirical estimations in the return level graph is close to the model -based linear line, which is nearly straight line with about 95% of confidence interval level, as the estimated form of distribution parameters (GEV) is near to zero for pollutants concentration. It should also be noted that, although the estimates of the return levels appear to be convincing, the biases in confidence intervals for large return periods also indicate that uncertainty affects this model to high levels. As illustrated in Fig. 3, the diagnostics for the maximum hours per day of pollutant concentrations shows that they both yield a strong good fit to a Fréchet distribution that has a positive shape parameter γ. The observations were generated from a (GEV) distribution. The estimation of this parameter can be carried out with two distinct approaches: the method of maximum likelihood and the weighted moments method. [44]. The parameters of location, scale and shape of the (GEV) were estimated from the data. The table of the parameters estimates for each concentration of the pollutant :

From the Table IV, the probability distribution of the pollutants as a function of the form parameter (γ) sign is specified. All the parameters (γ) are positive, thus the (GEV) tends towards the Frechet distribution, with the exception that for $NO_2\ \mu g|m^3$ where (γ) is negative, and it shows that Weibull is the best distribution to adjust $NO_2\ \mu g|m^3$. Now let us verify different future return values for pollutant concentration in next 15, 30, 40 and 50 years.

TABLE IV: ESTIMATED PARAMETERS OF THE (GEV) DISTRIBUTION FOR EACH POLLUTANT CONCENTRATION

Variables	Parameters	Estimate (Std error)	95% C.I.
$SO_2 \mu g\|m^3$	Location(μ)	8.42 (0.24)	[7.94 ; 8.90]
	Scale(σ)	7.32 (0.23)	[6.86 ; 7.79]
	Shape(γ)	0.52 (0.032)	[0.45 ; 0.58]
$NO \mu g\|m^3$	Location(μ)	33.15 (0.95)	[31.28 ; 35.02]
	Scale(σ)	28.36 (0.9)	[26.58 ; 30.14]
	Shape(γ)	0.49 (0.033)	[0.42 ; 0.55]
$NO_2 \mu g\|m^3$	Location(μ)	67.24 (0.84)	[65.58 ; 68.9]
	Scale(σ)	26.98 (0.61)	[25.78 ; 28.18]
	Shape(γ)	-0.03 (0.019)	[-0.08 ; -0.005]
$CO\ mg\|m^3$	Location(μ)	3.36 (0.15)	[3.06 ; 3.65]
	Scale(σ)	5.04 (0.27)	[4.51 ; 5.57]
	Shape(γ)	1.41 (0.037)	[1.34 ; 1.48]
$PM_{10} \mu g\|m^3$	Location(μ)	0.36 (0.017)	[0.33 ; 0.4]
	Scale(σ)	0.58 (0.033)	[0.52 ; 0.65]
	Shape(γ)	1.52 (0.039)	[1.44 ; 1.6]
$PM_{2.5} \mu g\|m^3$	Location(μ)	5 (0.08)	[4.84 ; 5.16]
	Scale(σ)	2.67 (0.07)	[2.53 ; 2.81]
	Shape(γ)	0.34 (0.017)	[0.31 ; 0.38]

C. Return Level

We noted that data are broken into different daily blocks which are modeled by the (GEV) distribution. The accuracy of the return level estimates generated from the (GEV) distribution is investigated. The return levels may be valuable for evaluating the trends of the pollutant parameters and for predicting their future pollutant extremes. However, the extreme quantiles of order 1-q of the maximum distribution over a given period of time is of particular concern in this context. Our interest is to find x_q where [20]:

$$P(Max(X_1, \ldots, X_n) \leq x) \approx G_{\mu,\gamma,\sigma}(x_q) = 1 - q$$

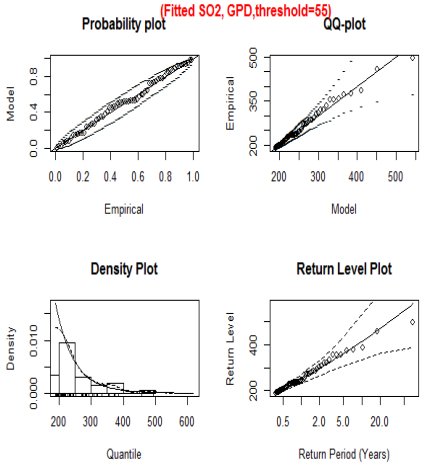
TABLE V: RETURNING PERIOD FOR EACH CONCENTRATION OF THE POLLUTANTS ACCORDING TO THE GEV-APPROACH

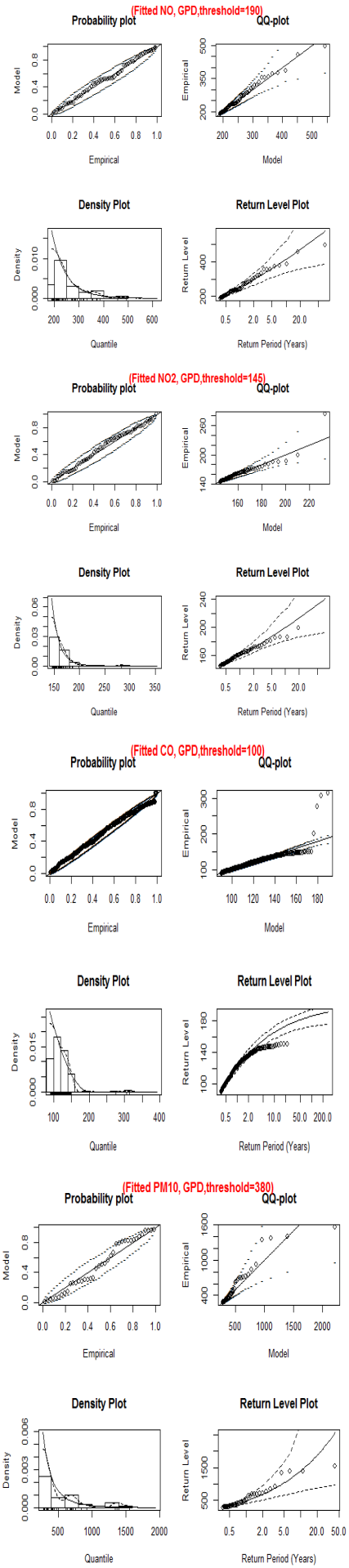
		SO_2 $\mu g\|m^3$	NO $\mu g\|m^3$	NO_2 $\mu g\|m^3$	CO $mg\|m^3$	PM_{10} $\mu g\|m^3$	$PM_{2.5}$ $\mu g\|m^3$
RETURN PERIOD	15	50.9	190	135.8	431	227.8	1471.4
	30	76.2	279.5	152.5	543.5	674.1	1516.7
	40	89.6	326.4	159.4	597.8	1052.5	1527
	50	101.5	367.6	164.6	636.8	1485.3	1532.9

According to Fig. 3, the return value of the fitted (GEV) to the maximum likelihood approach data (solid black line) and the associated 95% confidence interval (dotted black line) are shown. The prediction for each peak concentration of the pollutants over 15, 30, 40, and 50 years is provided in Table V.

Below, the approach (POT) for return period computation will be discussed, such that we fit the concentrations of the pollutants through the distribution (GPD). The thresholds for the parameters $SO_2\ \mu g|m^3$, $NO\ \mu g|m^3$, $NO_2\ \mu g|m^3$, $CO\ mg|m^3$, $PM_{10}\ \mu g|m^3$, $PM_{2.5}\ \mu g|m^3$ after being calculated by the MEF method are respectively equal to 55, 140, 150, 100, 400, 350.

The probability plots show the empirical versus observed probabilities for the model. It is expected that the model should fit linearly as shown in the above figures. We note that the (GPD) fit of the actual data exhibits a strong agreement with the approach (POT). Figures also shows return periods (in years) for the concentrations of pollutants. We used the (POT) approach and the generalized Pareto distribution (GPD) introduced by Pickands (1975) [32], which is widely applied. The advantage of this approach is that it studies the concentration that exceed some threshold level, which makes possible to investigate every possible available data exceeding the threshold, instead of selecting a maximum set of data, like in the (GEV) theory.





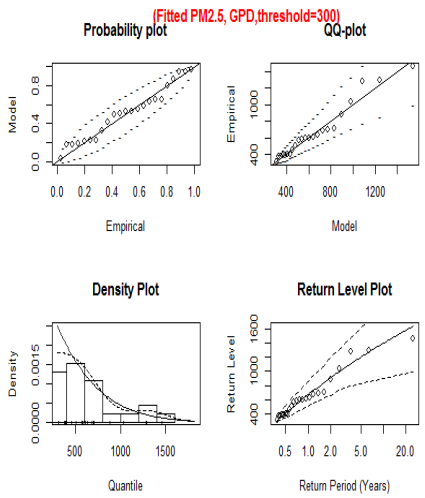


Fig 4. A diagnostic graph for the adjusted POT pattern for each peak air pollutant.

TABLE VI: THE RETURN PERIOD FOR EACH CONCENTRATION OF POLLUTANT ACCORDING TO THE (POT) APPROACH

		SO_2 $\mu g\|m^3$	NO $\mu g\|m^3$	NO_2 $\mu g\|m^3$	CO $mg\|m^3$	PM_{10} $\mu g\|m^3$	$PM_{2.5}$ $\mu g\|m^3$
RETURN PERIOD	15	440.3	500.8	279.3	389.5	1933.8	1588.8
	30	487.7	517.3	3228	455.2	2159.1	1702.7
	40	507.4	523.3	343.3	482.2	2252.5	1745.6
	50	522.7	527.6	361.5	503.1	2325.1	1777.3

With the POT approach, return levels for the 6 parameters were computed at 15, 30, 40 and 50 years. It can be noticed that concentrations for all the parameters continuously increased. The percentage of change of $SO_2 \mu g|m^3$, $NO\ \mu g|m^3$, $NO_2 \mu g|m^3$, $CO\ mg|m^3$, $PM_{10} \mu g|m^3$, and $PM_{2.5} \mu g|m^3$ after 50 years is respectively 64%, 5%, 3%, 29%, 20%, and 12%. The greatest change in concentration is expected to happen after 15 years, then the change concerning time will be smaller. SO_2 has the highest percentage of change, followed by CO and the particulate matter.

The results of this research indicate that extreme values theory involves investigating a maximum of a random sequence of variables. These results yield insights on the future extreme impacts of the pollutants on Lebanon. The results of (POT) are more coherent with the historic pollutants concentrations compared with the (BM) and the (POT) approaches. We further remark that all return value of the (GEV) approach are lower than the maximal pollutant concentration shown in Table II. Consequently, the (POT) approach is most adequate.

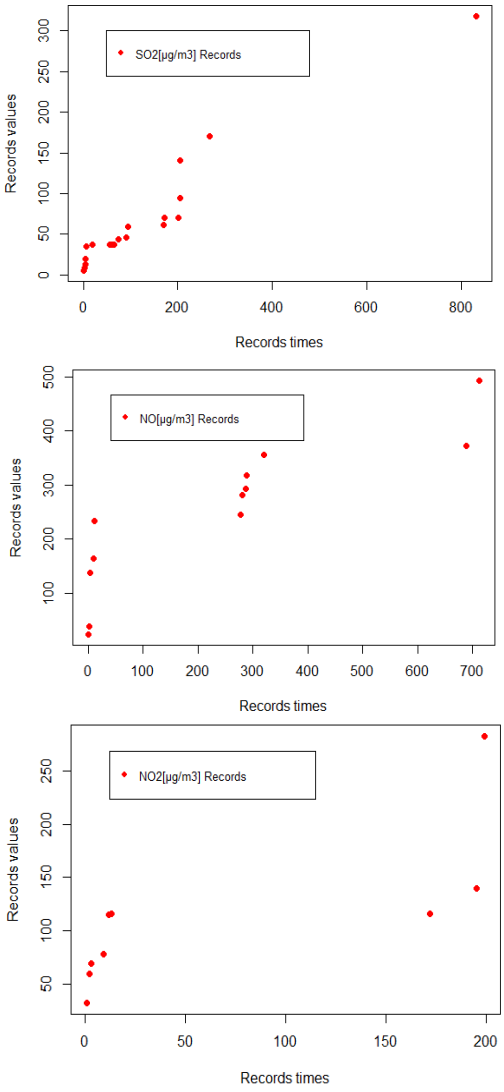
D. Extreme Records

An analysis of record occurrence values for daily pollutant parameters is provided herein. The aim was to gain insight and to quantify statistical records for these parameters within the framework of climate change and atmospheric pollution in Lebanon. Similar to the extremes approach, a simple mathematical pattern of identically distributed random variables has been used to forecast increases in the peak values of the atmospheric pollutants. The statistics of these parameters are investigated further based on a new significant contribution from record theory. Within this framework, however, we also consider the record values and times, which will be required for the prediction of the future

observed records.

1) Records results

As a baseline, a plot of the upper record values for pollutant parameters throughout 2016 to 2019 is given in the figures as shown below:



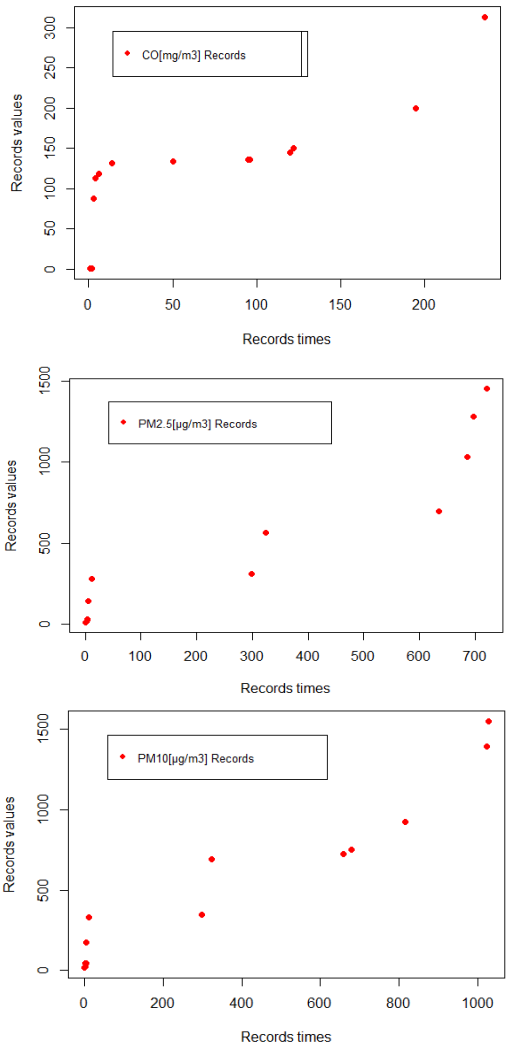


Fig. 5. Record values for the various parameters of each pollutant.

From Fig. 5, we remark that records are clustered amongst the first observations. The record rate asymptotically converges towards to zero $P(\delta_i) = 1/i$. Likewise $E(N_n) \approx \ln n$ which implies that the records have a tendency to get further away over time. Finally, the record values R_n, the record times L_n and the number of observed records N_n for the pollutant parameters are derived. Results obtained are shown in the table below:

TABLE VII: RECORD TIMES, RECORD VALUES AND NUMBER OF OBSERVED RECORDS SO_2, NO, AND NO_2

$SO_2\ \mu g/m^3$			$NO\ \mu g/m^3$			$NO_2\ \mu g/m^3$		
Record Dates	L_n	R_n	Record Dates	L_n	R_n	Record Dates	L_n	R_n
01/01/2016	1	5.11	01/01/2016	1	23.73	01/01/2016	1	31.79
03/01/2016	3	8.85	02/01/2016	2	38.79	02/01/2016	2	59.75
04/01/2016	4	12.65	03/01/2016	3	137.58	03/01/2016	3	68.66
05/01/2016	5	19.56	10/01/2016	10	164.73	09/01/2016	9	77.8
06/01/2016	6	34.76	12/01/2016	12	234.44	12/01/2016	12	114.53
13/01/2016	19	36.83	04/10/2016	278	245.91	13/01/2016	13	115.83
28/02/2016	56	36.97	07/10/2016	281	281.03	20/06/2016	172	116.10
05/03/2016	59	37.01	13/10/2016	287	293.73	13/07/2016	195	140.03
11/03/2016	65	37.42	14/10/2016	320	317.96	18/07/2016	199	283
21/03/2016	74	43.47	18/11/2017	688	372.67			
10/04/2016	91	45.87	12/12/2017	712	493.64			
14/04/2016	95	58.91						
15/05/2016	169	61.2						

17/05/2016	171	69.95			
16/06/2016	201	70.03			
19/06/2016	204	94.99			
20/06/2016	205	140.64			
24/08/2016	268	170.96			
09/03/2018	832	318.58			
$N_n(SO_2) = 19$			$N_n(NO) = 11$		$N_n(NO_2) = 9$

TABLE VIII: RECORD TIMES, RECORD VALUES AND NUMBER OF OBSERVED RECORDS CO, $PM_{2.5}$, AND PM_{10}.

$CO\ mg\|m^3$			$PM_{2.5}\mu g\|m^3$			$PM_{10}\mu g\|m^3$		
Record Dates	L_n	R_n	Record Dates	L_n	R_n	Record Dates	L_n	R_n
01/01/2016	1	0.92	01/01/2016	1	11	01/01/2016	1	16
02/01/2016	2	1.20	02/01/2016	2	17.6	02/01/2016	2	21.20
03/01/2016	3	87.30	03/01/2016	3	28.8	03/01/2016	3	42.344
04/01/2016	4	113.32	04/01/2016	4	32.67	04/01/2016/	4	44.963
06/01/2016	6	118.14	05/01/2016	5	145.6	05/01/2016	5	174.6
01/02/2016	14	131.93	12/01/2016	12	280.6	12/01/2016	12	332.3
19/02/2016	50	133.76	25/10/2016	299	313.5	25/10/2016	299	347.8
04/04/2016	95	135.6	19/11/2016	324	565.3	19/11/2016	324	690.1
05/04/2016	96	135.86	18/09/2017	634	659.9	20/10/2017	659	722.4
29/04/2016	120	145.24	16/11/2017	686	1032.1	09/11/2017	679	754.2
01/05/2016	122	149.97	28/11/2017	697	1279.3	25/03/2018	815	923.8
13/07/2016	195	200	21/12/2017	721	1454.8	17/10/2018	1021	1396.2
23/08/2016	236	313.12				23/10/2018	1027	1553
$N_n(CO) = 13$			$N_n(PM_{2.5}) = 12$			$N_n(PM_{10}) = 13$		

From the above tables, we observe that records occur early in the time series and the high occurrence of the observed records of the pollutant parameters occurs in the year 2016. While the pollution would be exceeded during summer because of the high temperature, the findings indicate that record values are mostly achieved during the colder weather months. This is explained the fact that some of these pollutants like $NO\ \mu g|m^3$, $NO_2\ \mu g|m^3$, and $SO_2\ \mu g|m^3$ are a result of incomplete combustion of fossil fuels. In the cold season, the combustion process becomes much more unlikely to be completed due to low temperature, which leads to increased incomplete products of combustion. The probability of a new record can be computed for every pollutant and any nearby time by using $P(\delta_i)$ for an independent identical distribution, e.g., the probability of a new record being observed in 2020 is equal to 0.018. Therefore, according to the formula (6), there is a direct relationship between the number of expected records theoretically derived by $E(N_n)$ and observed record values of concentration pollutants. Results are shown in more detail in Table IX below.

TABLE IX: MEASUREMENT OF PREDICTION ERROR OF EACH PARAMETERS

Parameters	SO_2 $\mu g\|m^3$	NO $\mu g\|m^3$	NO_2 $\mu g\|m^3$	CO $mg\|m^3$	PM_{10} $\mu g\|m^3$	$PM_{2.5}$ $\mu g\|m^3$
Obs. Records Val.	19	11	9	13	12	13
Exp. Nb. Records	17.21	10.62	10.08	12.2	11.62	12.15
Exact relative error	9.42%	3.4%	12%	6.76%	3.16%	6.53%

Observed record values closely approximate expected values, which indicates that the record patterns being considered were well attributed to the concentration pollutants. Furthermore, the minimal accurate relative error is recorded across all parameters, which indicates a good fit of the recording model. Moreover, we can notice that, when n increases, the records have a tendency to be further away in time. The majority of the records may be observed during the first several years, however, when n grows sufficiently large, then the number of records will behave as ln(n).

2) Records prediction

We will predict the next record value for the pollutant parameters based on the record values collected in previous years. Using equation (16) we predict the record point values and equation (17) gives the confidence interval of prediction for the pollutants.

TABLE X: PREDICT RECORDS VALUES FOR EACH PARAMETERS

Parameters	R_n^*	C.I.(95%)
$SO_2 \mu g\|m^3$	319.42	[318.51,328.83]
$NO\ \mu g\|m^3$	497.37	[493.52,537.91]
$NO_2\ \mu g\|m^3$	286.13	[282.84,310.97]
$CO\ mg\|m^3$	314.84	[312.99,332.22]
$PM_{10}\ \mu g\|m^3$	1464.12	[1454.18,1501.62]
$PM_{2.5}\ \mu g\|m^3$	1561.53	[1552.39,1597.68]

Table X gives point prediction and the upper 95% confidence limit prediction for the (BLUP) of each parameter. After one year of monitoring, the values of records of $SO_2\ \mu g|m^3$, $NO\ \mu g|m^3$, $NO_2\ \mu g|m^3$, $CO\ mg|m^3$, $PM_{10}\ \mu g|m^3$, and $PM_{2.5}\ \mu g|m^3$ increased by 0.2%, 0.7%, 1%, 0.5%, 6%, and 5.7% respectively. Thus, pollution is expected to become more of a problem with time, particularly with regard to particulates, which show the largest increases after year one, and which are known to cause a number of effects to human health in the short and long term. The Bekaa region is a complex region with a variety of human activities. This region is witnessing a catastrophic environmental scenario affecting air quality. These extreme events, which result from heavy pollution, affect the Lebanese economy, particularly in the Bekaa region, which depends mainly on agricultural activity. Atmospheric factors such as carbon monoxide $CO\ mg|m^3$)nitrogen dioxide ($NO_2\ \mu g|m^3$), and sulfur dioxide ($SO_2\ \mu g|m^3$) are expected to reach future high records. They are emitted by vehicle engines and larger high-temperature combustion plants, industries, and road networks. Similarly, fine particles $PM_{10}\ \mu g|m^3$ and

$PM_{2.5}\ \mu g|m^3$ also reached record high values mainly because of heating and road fuels which can be very harmful to human health. Nitrogen oxides are pollutants mostly linked to emissions from road traffic.

Now we compute the probability to expect this record by using the formula (18). Hence, for $n_1 = 4$ years (number of years in the database), the probability that we will have to wait for more than 2 years for example before observing a new record is $\frac{4}{4+2} = 0.66$.

VI. Conclusion

In this article, extreme and record theories have been used to analyze the air data. The fundamentals of extreme value and record theories were reviewed. Those theories have been applied on six different pollutants. The data were sufficient to investigate the extreme value and make an accurate prediction of the return level of the pollutant concentration after 15, 30, 40 and 50 years using the peak-on-threshold approach of the extreme value theory. The extreme records, the values and the time interval between them can also be followed according to the theory of extremes. We found that extreme pollutant concentration levels occur throughout the observation, but the record extremes are mainly concentrated at the beginning of the observation, and then move farther away from each other over time. This implies that pollutant concentration should be increasing, but the time between records is widening. In addition, most surveys have been performed during the winter season, when there is heavy fossil fuel burning and incomplete combustion as a result of the low temperatures, which leads to more harmful by-products being generated. Note that results on extreme pollutant prediction may differ among various geographies since the pollution concentration is affected by many factors like emission levels, meteorological conditions as well as the geography. Eventually, extremes and record theories may be used as a managerial device for detecting the pollution levels in order to establish a warning system for the future. Another point that should be emphasized is the record values which are higher than the WHO guidelines. These elevated levels may endanger population health when frequently reoccurring.

Conflict of Interest

The authors declare no conflict of interest.

Author Contributions

ZK, RM, SA, KM conducted the research; AA analyzed and collected the data; wrote the paper, conceived and designed the analysis; all authors had approved the final version.

Acknowledgment

The authors would like to thank the Lebanese Ministry of Environment for providing the data preferred.

References

[1] World Health Organization (WHO), *Air Quality Guidelines*, Copenhagen, Denmark: WHO Regional Office for Europe, 2021.

[2] N. A. Saliba, H. Kouyoumdjian, and M. Roumie, "Effect of local and long-range transport emissions on the elemental composition of PM_{10} and $PM_{2.5}$ in Beirut," *Atmos. Environ*, vol. 41, pp. 6497–6509, 2007,

[3] J. Lelieveld, P. Hoor, P. Jöckel, A. Pozzer, P. Hadjinicolaou, J.-P. Cammas, and S. Beirle, "Severe ozone air pollution in the Persian Gulf region," *Atmos. Chem. Phys*, vol. 9, pp. 1393–1406, 2009,

[4] MoE: National environmental action plan, *Ministry of Environment*, Beirut, Lebanon : s.n., 2020.

[5] N. A. Saliba, S. Moussa, H. Salame, and M. El-Fadel, "Variation of selected air quality indicators over the city of Beirut, Lebanon: Assessment of emission sources," *Atmos. Environ*., vol. 40, pp. 3263–3268. 2006,

[6] C. Abdallah, C. Afif, N. El Masri, F. Öztürk, M. Keleş, and K. Sartelet, "A first annual assessment of air quality modeling over Lebanon using WRF/Polyphemus," *Atmos. Pollut. Res*., vol. 9, pp. 643-654, 2018.

[7] R. Ochoa-Hueso, S. Munzi, R. Alonso, M. Arroniz-Crespo, A. Avila, V. Bermejo, R. Bobbink, C. Branquinho, L. Concostrina-Zubiri, and C. Cruz, "Ecological impacts of atmospheric pollution and interactions with climate change in terrestrial ecosystems of the Mediterranean Basin: current research and future directions," *Environ. Pollut*., vol. 227, pp. 194-206, 2017.

[8] A. Farrow, K. A. Miller, and L. Myllyvirta, "Toxic air: The price of fossil fuels," *Greenpeace Middle East and North Africa*, June 2020, p. 35.

[9] W. Farah, M. M. Nakhle, M. Abboud *et al*., "Time series analysis of air pollutants in Beirut, Lebanon," *Environ Monit Assess*, vol. 186, pp. 8203–8213, 2014.

[10] C. Abdallah, K. Sartelet, and C. Afif, "Influence of boundary conditions and anthropogenic emission inventories on simulated O_3 and $PM_{2.5}$ concentrations over Lebanon," *Atmospheric Pollution Research*, ISSN 1309-1042, pp. 971-979, 2016.

[11] K. E. Bencala and J. H. Seinfeld, "On frequency distribution of air pollutant concentrations," *Atmospheric Environment*, vol. 10, pp. 941–950, 1976.

[12] C. K. Lee, D. S. Ho, C. Yu, C. Wang, and Y. Hsiao, "Simple multifractal cascade model for air pollutant concentration (APC) time series," *Environmetrics*, vol. 14, no. 2, pp. 255–269, 2003.

[13] K. Voigt, G. Welzl, and R. Bruggemann, "Data analysis of environmental air pollutant monitoring systems in Europe," *Environmetrics*, vol. 15, pp. 577–596, 2004.

[14] J. Schwartz and A. Marcus, "Mortality and air pollution in London: A time series analysis," *American Journal of Epidemiology*, vol. 31, pp. 85–194, 1990.

[15] R. L. R. Alvim, M. Ferraz, C. Alves, and F. Martins, "Time series analysis of air pollution data," *Salcedo Atmos. Environ*., vol. 33, pp. 2361-2372, 1999.

[16] C. A. Pope *et al*., "Respiratory health and PM_{10} pollution-a daily time series analysis," *American Review of Respiratory Disease*, vol. 144, pp. 668–674, 1991.

[17] N. Gouviea and T. Fletcher, "Time series analysis of air pollution and mortality: effects by cause, age and socioeconomic status," *Journal of Epidemiology and Community Health*, vol. 54, pp. 750–755, 2000.

[18] G. Touloumi, R. Atkinson, and A. L.Terte, "Analysis of health outcome time series data in epidemiological studies," *Environmetrics*, vol. 15, pp. 101–117, 2004.

[19] C. Afif, A. L. Dutot, C. Jambert *et al*., "Statistical approach for the characterization of NO_2 concentrations in Beirut," *Air Qual Atmos Health*, vol. 2, pp. 57-67, 2009.

[20] A. Hayek, Z. Khraibani, D. Radwan, N. Tabaja, S. A. Andaloussi, J. Toufaily, and E. G.-Z. T. Hamieh, "Analysis of the extreme and records values for temperature and precipitation in Lebanon," *Communications in Statistics: Case Studies, Data Analysis and Applications*, vol. 6, no. 4, pp. 411-428,2020.

[21] M. Aline and Z. Khraibani, "Oil rig protection against wind and wave in Lebanon," *Communications in Statistics: Case Studies, Data Analysis and Applications*, vol. 6, pp. 191-214, 2020.

[22] H. M. Badran, Z. Khraibani, H. Khraibani, H. Zeineddine, A. Mefleh, and H. Hamie, "Dependency modelling of natural rare phenomena: Application on oil rigs," *International Journal of Probability and Statistics*, vol. 5, p. 25, 2016.

[23] Z. Khraibani, H. M. Badran, and H. Khraibani, "Records method for the natural disasters application to the storm events," *Journal of Environmental Science and Engineering*, vol. 5, pp. 643–651, 2011.

[24] S. Gulati and F. G. B. M. G. Kibria, "Analysis of hurricane extremes and record values in the Atlantic," *Communications in Statistics: Case Studies, Data Analysis and Applications*, vol. 5, no. 2, pp. 101-110, 2019.

[25] G. Wergen and J. Krug, *Record-Breaking Temperatures Reveal a Warming Climate*. s.l. : Europhys Lett 92:30008., 2010.

[26] G. Wergen, A. Hense, and J. Krug, "Record occurrence and record values in daily and monthly temperatures," *Clim. Dynam*., 2013.
[27] A. S. Hoayek, G. R. Ducharme, and Z. Khraibani, "Distribution-free inference in record series," *Extremes*, vol. 20, pp. 585–603, 2017.
[28] Z. Khraibani, H. Trabulsi, A. Atoui, A. Hayek, and D. Radwan, "Climate change, agriculture and economic growth in Lebanon: A VAR approach," *American Journal of Economics*, vol. 10, pp. 126-131, 2020.
[29] R. A. Fisher and L. H. C. Tippett, "Limiting Forms of the frequency distribution of the largest or smallest member of a sample," *Mathematical Proceedings of the Cambridge Philosophical Society*, vol. 24, pp. 180-190, 1928.
[30] E. J. Gumbel, "On the frequency distribution of extreme values in meteorological data," *Bull. Amer. Meteor. Soc*, vol. 23, pp. 96-105, 1942.
[31] M. R. Leadbetter, G. Lindgren, and H. Rootzén, *Extremes and Related Properties of Random Sequences and Processes*, New York: Springer Verlag, 1983.
[32] J. I. Pickands, "Statistical inference using extreme value order statistics," *The Annals of Statistics*, vol. 3, pp. 119–31, 1975.
[33] A. A. Balkema and L. Haan, "Residual life time at great age," *The Annals of Probability*, vol. 2, pp. 792–804, 1974.
[34] S. Ghosh and S. Resnick, "A discussion on mean excess plots," *Stochastic Processes and their Applications*, vol. 120, pp. 1492-1517, ISSN 0304-4149, 2010.
[35] World Health Organization (WHO). (2021). [Online]. Available: https://www.who.int/health-topics/air-pollution.
[36] K. N. Chandler, "The distribution and frequency of record values," *Journal of the Royal Statistical Society*, vol. 14, pp. 220-228, 1952.
[37] B. C. Arnold, N. Balakrishnan, and H. N. Nagaraja, *Records*, New York: Wiley, 2011.
[38] C. P. Shing, "Interval estimation of location and scale parameters based on record values," *Statistics Probability Letters*, vol. 37, pp. 49-58, 1998.
[39] H. A. DAVID, *Order Statistics*, Second edition, New York: John Wiley L Sons, 1981.
[40] N. Balakrishnan and A. C. Cohen, *Order Statistics and Inference: Estimation Methods*, San Diego: Academic Press, 1991.
[41] D. J. Stekhoven and P. Bühlmann, "MissForest—non-parametric missing value imputation for mixed-type data," *Bioinformatics*, vol. 28, pp. 112-118, January 2012.
[42] J. Gareth *et al*., *An Introduction to Statistical Learning with Applications in R*, 2013.
[43] K. C. Lee, "Multifractal characteristics in air pollutant concentration time series," *Water Air Soil Poll*., vol. 135, pp. 389-409, 2002.
[44] H. L. F. Ana, *Extreme Value Theory: An Introduction*, Springer, 2007.

Alya Atoui received the bachelor of petroleum structures with honours from Lille University and Lebanese University. She is currently pursuing a PhD in environmental science with the Water Environment and Urban Systems Laboratory (LEESU), Paris Est-Creteil University.