Extreme Value Analysis of Urban Air Quality using Internet of Things

Anurag Barthwal and Debopam Acharya
Computer Science and Engineering, Shiv Nadar University, Gautam Buddha Nagar, U.P., India - 201314

Extremely poor levels of air quality are often experienced in high pollutant concentration regions like Delhi. The analysis and forecasting of such extreme events is suitably performed using probability distributions. In this work, an end-to-end IoT based system has been developed to collect, visualize and analyze air pollution data in Delhi and NCR. Central fitting and extreme value distributions, Lognormal, Gumbel, Frechet and Weibull have been compared using goodness of fit criteria to select the best fit model for computation of exceedance probabilities and return periods for air pollution extremes. The exceedance probabilities and return periods have been compared to actual occurrences of extreme events. The results indicate that the Gumbel distribution based air quality model is best suited for forecasting air quality in this high pollutant concentration region.

Keywords: IoT, AQI, smart city, probability distribution, extreme value distribution, goodness of fit, coefficient of determination.

1. INTRODUCTION

Rapid and unplanned growth in urban areas has led to over-crowdedness, poor sanitation, insufficient or contaminated water, loss of green cover and a surge in the number of vehicles. There is a huge rise in energy demands with surging human population for cooking, air conditioning, ventilation and lighting. Increase in number of vehicles, construction activities, diesel generator sets for electricity and disposal of waste by burning generate dust and gaseous pollutants which contaminate the outdoor air. Industries that come up to support a huge urban population are another major source of air pollutants. Air pollution is a global cause for millions of deaths and respiratory diseases every year. Prolonged exposure to particulate matter (PM) of diameter less than $2.5\mu m$ increases risk of mortality from heart disease, respiratory infections, lung cancer and chronic obstructive pulmonary disease [Amann et al. 2017; Bhanarkar et al. 2018]. Quality of air in Delhi and surrounding National Capital Region (NCR) has worsened due to vehicular emissions, exhaust from diesel generators, dust from construction sites, burning garbage and agricultural waste, thermal power plants and industrial activities [Zhang et. al. 2017; Cohen et. al. 2017]. Delhi is the sixth most populous urban city in the world, and the most populous if the entire NCR is included. Accurate forecast of urban air quality is useful in enhancing the scientific understanding of air pollution, development of optimal pollution control strategies and providing alerts to sensitive population such as children, elderly and people with respiratory ailments. This work aims to collect and analyze air quality data using an IoT based sensing system and use this data to develop predictive models of air quality in real time.

Air quality information is usually monitored and made available to citizens using fixed air quality monitoring stations. These air quality stations deploy expensive instruments and are located at fixed points. Air pollution is a context aware phenomenon and on several occasions, the air quality stations fail to provide context aware fine grained data about air quality at a particular location. This limitation motivated us to take the crowd-sourcing approach and create a vehicle mounted mobile IoT system to sense and monitor air quality on the travel path. Our IoT system consists of various sensors connected to a micro-controller which transmits sensed data to an Android application in a smartphone which, in turn, uploads it to the cloud. The collected information is analyzed and modeled for predicting future trends, which could be used by the population for their daily travel plans.

Statistical analysis and probabilistic distributions have been extensively used in recent years to estimate the current state of air quality and forecast future air quality. Events of occurrence of extremely poor air quality are of most interest, given their potential for substantial impact on health, well-being and productivity of citizens. However, central fitting distributions such as *normal*, *log-normal* and *gamma* distributions do not provide good solutions for air pollutant data if the pollutant concentration is high. Hence, extreme value distributions have also been explored in this work to select the best fit distribution for forecasting extreme events of poor air quality in this region. Two locations in Delhi and the National Capital Region (NCR) have been identified for sensing and modeling air quality data.

Contribution of this work: Events of occurrence of extremely poor air quality are of most interest, given their potential for substantial impact on health, well-being and productivity of citizens. In this work, we show that for a high pollutant concentration region, occurrence of events of extremely poor air quality can be better forecasted using extreme value distributions as compared to central fitting distributions. We have compared the performance of one central fitting distribution- lognormal distribution and three types of extreme value distributions- Gumbel, Frechet and Weibull to estimate the risk of occurrence of extremely poor air quality. The most suitable model for forecasting has been selected using goodness of fit tests. Coefficient of determination (R^2) and index of agreement (IA) have been used for this purpose. Subsequently, the best fit model has been used to forecast the exceedance probabilities and the return periods of extreme values of air quality at two locations in Delhi-NCR. The *exceedance probability* is the probability that a threshold value of air quality index (AQI) will be breached. The *return period* is the estimated time of recurrence of an event, such as exceedance of a certain value of AQI. This information is useful to estimate the actual state of air quality of a place or geographical area. The forecasted exceedance probabilities for different AQI levels are then compared with actual exceedances to evaluate the accuracy of the model.

2. RELATED WORK

The monitoring, analysis and modeling of outdoor air quality has been discussed in some of the recent works. Bechir Raggad [2018] has used extreme value theory to analyse extreme temperature events in Riyadh, capital of Saudi Arabia. The generalized extreme value (GEV) distribution and the generalized Pareto (GPD) distribution models have been used for modelling, analyzing and forecasting daily maximum temperatures.

Chen et al. [2018] have developed a *Multivariate Long Short-Term Memory (MuLSTM)* model to learn the temporal dependency and spatial correlation of traffic patterns at different base stations, to accurately forecast the traffic in future. A weighed graph of the base stations is built afterwards, according to their traffic patterns, and an algorithm to find the optimal base station clustering scheme is proposed.

Wang et al. [2017] have developed a vehicular sensor network (VSN) to monitor urban air quality and a data gathering and estimation system on *VSN*. The data gathering algorithm also uses *Delaunay triangulation* to infer AQIs of the locations without any sensed data. Zhu et al. [2017] have developed a system with the help of which they intend to determine city wide air quality, using sensed data from few air quality monitoring stations, that are spaced far apart. Granger causality has been used to analyze all causality relations responsible for generation of air pollution.

Martins et al. [2017] have analyzed 16 years of hourly air pollutant data of *Sao Paulo* and 7 years of hourly air pollutant data of *Rio de Jeneiro* to provide information about extreme pollution events and their return period to government agencies, decision makers and urban citizens. Generalized Extreme Value (GEV) and Generalized Pareto Distribution (GPD) were applied to investigate the behavior of pollutants in these two regions. Probability of occurrence of extreme values and return periods for SO_2, NO_2, NO, O_3, $PM_{2.5}$ and PM_{10} was computed.

Ercelebi et al. [2009] have used extreme value distribution (EVD) type I and II to analyze SO_2 and NO_2 concentrations from two fixed air quality stations in Istanbul. Future largest SO_2 and NO_2 concentrations for the next 12 months have been forecasted using EVD. Exceedance probabilities and return periods for largest pollutant concentrations has also been calculated. To the best of our knowledge, there is no single recent work that has involved end-to-end process of building IoT systems to sense, analyze and model

air quality in real time in a large metropolitan region like the Delhi-NCR region. This is our motivation to take up this comprehensive work of developing an IoT system to sense, visualize, model and analyze outdoor air quality data.

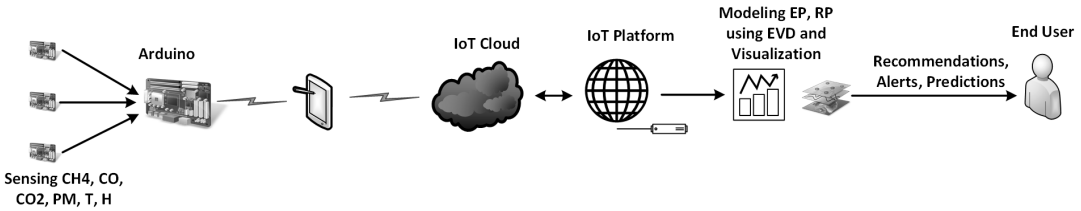

Fig.1: The IoT Sensing, Visualization and Analysis System

3. OUR IOT SYSTEM ARCHITECTURE

Our IoT sensing system consists of multiple sensors connected to a micro-controller that transmits the sensed data to an IoT cloud. The sensing system is placed in a vehicle. Arduino Uno is used for collection and transmission of sensed data. It is based on ATmega328P 8-bit micro-controller. It has 6 analog input pins, 14 digital input/output pins, a USB connection, a power jack, a 16 MHz quartz crystal, as ICSP header and a reset button. The architecture of our sensing system is shown in Figure 1.

The concentration of air pollutants can be obtained by using metal-oxide semiconductor (MOS), optical sensors or electro-chemical sensors, each having unique features. Electro-chemical sensors have high sensitivity, low energy consumption and are less sensitive to environmental changes, hence we have used these sensors for the detection of carbon dioxide (CO_2), carbon monoxide (CO) and methane (CH_4) concentrations in air [Kim et al. 2014].

Optical sensor *GP2Y1010AU0F* has been used in our sensing system for determination of particulate matter (PM) concentrations as it is low-cost, utilizes less power, possesses small size and is fairly inexpensive. *GP2Y1010AU0F* optical sensor has been preferred for determination of $PM_{2.5}$ and PM_{10} concentrations over other sensors because of acceptable sensitivity, as it is capable of measuring PM density up to 0.5 mg/m^3 [Barthwal et al. 2018].

The gas sensors have been calibrated before using them to record actual pollutant concentration. The sensor output is a function of the ratio of sensor resistance (R_s) to load resistance (R_L). To calculate R_L, we burn the sensor in fresh, clean air at a temperature of 22-25 °C and a relative humidity of 60-70% for 3-4 hours. The average output resistance during this period is recorded and R_L is fixed using the knob in the sensor. After calibration, the output voltage across the ground and V_{cc} pins of the sensor is a function of (R_s / R_L). The sensor output $\propto$ (R_s / R_L) obtained in this way is converted to ppm or $\mu g/m^3$ value using the sensor data-sheet provided by the manufacturer.

3.1 Hardware setup

The details of sensors used in the sensing system are provided in Table I. MQ2 gas sensor has been used for detection of carbon monoxide (CO), MQ4 for methane (CH_4), MQ135 for CO_2, Sharp optical sensor *GP2Y1010AU0F* for $PM_{2.5}$ and PM_{10} and DHT22 sensor for temperature and humidity respectively. As shown in Figure 1, the sensors transmit the sensed data to *Arduino Uno*, which in turn transmits the data to an Android application in a smartphone using Bluetooth module (HC-05). The smartphone application transmits the sensed data to IoT cloud, where it is stored for visualization and analysis. The stored air quality data in IoT cloud is used to compute exceedance probability and return period of AQI levels for 2 locations, and is made available to an end user through an Android application in the user's smartphone or a website.

Table I: Details of Sensors of IoT Setup

S.No.	Environmental Parameter	Unit	Sensor
1	CO_2	ppm	MQ135
2	CO	ppm	MQ2
3	CH_4	ppm	MQ4
4	Noise	dBA	KY-038 Sound sensor
5	Particulate Matter	$\mu g/m^3$	GP2Y1010AU0F sensor
6	Temperature	$^{\circ}$C	DHT22
7	Relative Humidity	%	DHT22
8	Latitude, Longitude	degrees	Adafruit Ultimate GPS

3.2 Software Setup

The software setup of our IoT system consists of (a) a program stored in an Arduino Uno micro-controller, which is used for calibration, collection and transmission of the sensed data using Bluetooth to an Android application installed in a smartphone; (b) An Android application that receives the sensed data and stores this data in the phone memory in the form of a *.csv* file; (c) A Cloud that receives data from the Android application. In our IoT system, the cloud is developed using IBM Cloud. The stored data at IoT cloud can be availed by a user with the help of an Android app in the user's smartphone or by using a web-browser. The details of software system are described in Table II.

Table II: Software setup

S.No.	Application	Software
1	Arduino Uno	C++ and AVR C
2	Android	Java
3	IoT Cloud	IBM cloud

4. DATA COLLECTION SETUP

Hourly data of air pollutants for the months of August, 2018 to December, 2018 has been used in this work to calculate the exceedance probabilities and the return periods for AQI levels. Data has been collected using our IoT system which is placed in a vehicle. Figure 2 shows our vehicular sensing system in the vehicle with the windows open. Open windows allow the outside polluted air to freely flow inside the vehicle chamber so that actual outdoor pollution can be captured. The sensed parameters are used to calculate the air quality index (AQI).

AQI is an index for reporting air quality. The higher the AQI value, the poorer the quality of air and the greater the health concern. The National Air Quality Index [CPCB. 2014] has divided AQI into 6 categories , each category corresponding to different level of health concern. Separate colours have been assigned to each category. Different colour for each category is helpful in conveying to the general public, whether air quality is reaching unhealthy levels in their area or not. AQI bands and the corresponding levels of health concern has been depicted in Table III.

Table III: Details of AQI values and levels of health concern

AQI Value	Level of Health concern	Colour
0 to 50	Good	Green
51 to 100	Moderate	Yellow
101 to 150	Unhealthy for sensitive groups	Orange
151 to 200	Unhealthy	Red
201 to 300	Very unhealthy	Purple
301 to 500	Hazardous	Maroon

In our work, carbon monoxide, $PM_{2.5}$ and PM_{10} have been taken into consideration for calculation of AQI. As PM_{10} was found to be the most significant pollutant throughout the duration of our study,

Fig.2: Data collection setup: The IoT sensing system placed inside a vehicle to collect air quality data

it has been used to calculate the AQI. Two locations have been selected as test-bed for conducting this experiment in the Delhi-NCR region: AnandVihar in Delhi and Dadri in NCR. AnandVihar has been chosen as a test-bed because it is one of the most polluted places of this region, characterized by heavy vehicular traffic and emissions from nearby industrial areas. On the other hand, Dadri is a semi-urban area which is less populated and polluted as compared to AnandVihar.

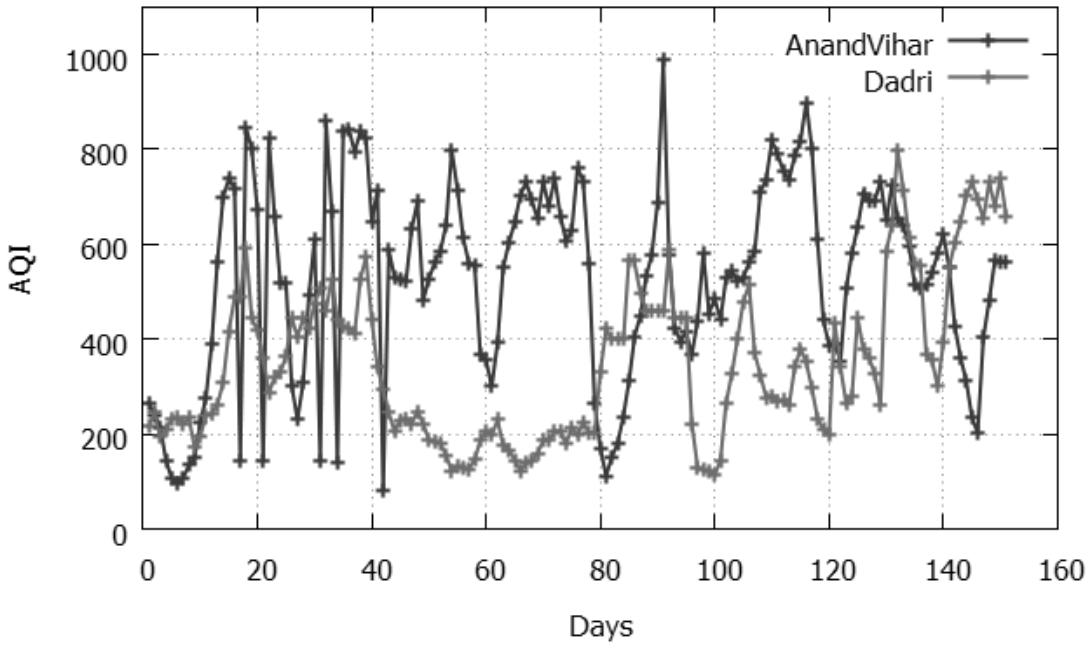

Fig.3: Variation of daily average AQI during the months of August, 2018 to December, 2018 at Anand Vihar and Dadri respectively

Figure 3 shows the variation of daily average AQI at AnandVihar and Dadri during the period from August, 2018 to December, 2018, as recorded by our vehicular sensing system. These plots indicate that occurrences of poor air quality are frequent and AQI level is often hazardous.

5. OUR AIR QUALITY MODELS

Statistical analysis and probability distributions have been extensively used in recent years to estimate the current state of air quality and forecast future air quality [Wang et al. 2004; Lu et al. 2014]. Central fitting distributions (CFDs) such as *normal*, *log-normal* and *gamma* distributions do not provide good solutions for air pollutant data if the pollutant concentration is high [Karaca et al. 2005; Sharma et al. 1999]. EVD has been used worldwide to forecast the occurrence of floods, storms, sea waves, droughts, wind, earthquakes, huge fluctuations in exchange rates and market crashes [Mdyusof et al. 2011]. We believe that to model the high pollution concentration air of this region, EVDs are better suited than CFDs. This work proposes to investigate central fitting and extreme value distributions to identify the most suitable probability distribution and model the high concentration air pollutant data of this region. We propose to use the best fit model which could subsequently be applied to forecast the occurrence of events of extremely poor air quality. The two indicators used for that purpose are exceedance probability and return period. Exceedance probability is the probability that a certain value of AQI will be breached. Return period is the estimated time of recurrence of an extreme event, such as exceedance of certain level of AQI. The probability distributions that we are going to evaluate to identify the best fit distribution for our data are (1) Lognormal, (2) Gumbel, (3) Frechet and (4) Weibull.

Hourly AQI data constitutes a positive, discrete random process. This random process has values X_0, X_1, .. , X_N where X_0 is the value at an arbitrary starting point and X_N is the value of the N^{th} observation. The change in concentration of a pollutant from one observation to another due to many diverse processes that act simultaneously, may be defined as

$$X_j - X_{j-1} = p_j X_{j-1} \tag{1}$$

where X_j and X_{j-1} are the values of the process at times *j* and *j - 1* respectively, p_j is a random variable of proportionality and it represents the effect of the diverse random processes on the pollution concentration observed at time *j - 1*. Rewriting equation 1, we get

$$\frac{(X_j - X_{j-1})}{X_{j-1}} = p_j, and \tag{2}$$

$$\sum_{j=1}^{N} \frac{(X_j - X_{j-1})}{X_{j-1}} = \sum_{i=1}^{N} p_j \tag{3}$$

We assume that the change in pollutant concentration at each point in time is small

$$\sum_{j=1}^{N} \frac{(X_j - X_{j-1})}{X_{j-1}} \cong \int_{X_0}^{X_N} \frac{dx}{x} = logX_N - logX_0 \tag{4}$$

Thus,

$$logX_N = logX_0 + p_1 + p_2 + .. + p_N \tag{5}$$

The central limit theorem states that logX_N is normally distributed regardless of the distribution of X_N and therefore, lognormally distributed. So, anything that is formed, grows or changes according to law of proportionate effect, follows lognormal distribution. For a large number of observations of AQI, the probability density function (pdf), f(x) and cumulative density function (cdf), F(x) of the pollutant, using Lognormal distribution is given as

$$f(x) = \frac{1}{x\alpha\sqrt{2\pi}} exp\Big[-\frac{1}{2} exp\Big(\frac{(ln(x) - \beta)}{\alpha} \Big)^2 \Big], x > 0, \alpha > 0, \beta > 0 \tag{6}$$

$$F(x) = \frac{1}{2\pi} \int_{-\infty}^{\frac{(lnx-\beta)}{\alpha}} exp^{\frac{-x^2}{2}}, x > 0, \alpha > 0, \beta > 0 \tag{7}$$

The Gumbel distribution (Extreme Value Type I distribution) was extensively developed and applied to flood flows by *Emil Julius Gumbel* in 1950s and 60s. For a large number of observations of AQI, the probability density function (pdf) and cumulative density function (cdf) of the pollutant, using Gumbel distribution is given as

$$f(x) = \frac{1}{\beta} exp\Big[- \frac{(x-\delta)}{\beta} - exp\Big(\frac{(x-\delta)}{\beta}\Big)\Big], \quad -\infty < x < \infty, -\infty < \delta < \infty, \beta > 0 \tag{8}$$

$$F(x) = exp\Big[- exp(\frac{(x-\delta)}{\beta}\Big], x > 0, -\infty < \delta < \infty, \beta > 0 \tag{9}$$

where, x is the variation of the air pollutant over a period of time, δ is the location parameter, which is the mode of the AQI distribution in this case.

$$\frac{df(x)}{dx} = 0, \text{for} x = \delta \tag{10}$$

For a large number of pollutant observations, the parameter β is the measure of dispersion, which depends on variance of random variable, X. The mean, E(X) and variance, V(X) of the random variable, X are given as

$$E(X) = \mu = \delta + n_c \beta, \tag{11}$$

$$V(X) = \sigma^2 = \frac{\pi^2 \beta^2}{6}, \tag{12}$$

where n_c = 0.5772 (Euler's constant).

Using equations (11) and (12), the values of β and δ are found to be

$$\beta = \frac{\sqrt{6}}{\pi}.\sigma, \tag{13}$$

$$\delta = \mu - n_c \beta = \mu - \frac{n_c \sqrt{6}}{\pi}.\sigma, \tag{14}$$

The third distribution that we are evaluating is the Frechet distribution (Extreme value distribution type II). It was developed by the French Mathematician *Maurice Rene Frechet* in 1927, who applied it to flood flows. For a large number of observations of AQI, the pdf and cdf of the pollutant, using Frechet distribution is given as

$$f(x) = \frac{\alpha}{\beta}\Big(\frac{\beta}{x}\Big)^{\alpha+1} exp\Big[- \Big(\frac{\beta}{x}\Big)^{\alpha}\Big], x > 0, \alpha > 0, \beta > 0 \tag{15}$$

$$F(x) = exp\Big[- \Big(\frac{\beta}{x}\Big)^{\alpha}\Big], x > 0, \alpha > 0, \beta > 0 \tag{16}$$

where α and β are the scale and shape parameters of Frechet distribution. The value of α and β is obtained by using the coefficient of variation (CV)

$$CV = \sqrt{\frac{\Gamma\Big(1 - \frac{2}{\alpha}\Big)}{\Gamma^2\Big(1 - \frac{1}{\alpha}\Big)}}, \alpha > 2 \tag{17}$$

$$\beta = \frac{\bar{x}}{\Gamma\left(1 - \frac{1}{\alpha}\right)} \tag{18}$$

The Weibull distribution (Extreme value distribution, type III) is the fourth probability distribution that is being considered in this work for forecasting extreme events. It was named after the Swedish mathematician *Waloddi Weibull*. For a large number of observations of AQI, the pdf and cdf of the pollutant, using Weibull distribution is given as

$$f(x) = \frac{\alpha}{\beta}\left(\frac{(x}{\beta}\right)^{\alpha-1} exp\left(-\frac{x}{\beta}\right)^{\alpha}, x > 0,\ \alpha > 0, \beta > 0 \tag{19}$$

$$F(x) = 1 - exp\left[-\left(-\frac{x}{\beta}\right)^{\alpha}\right], x > 0,\ \alpha > 0, \beta > 0 \tag{20}$$

where α and β are the scale and shape parameters of Weibull distribution.

6. SELECTION OF BEST FIT DISTRIBUTION

To find the distribution that fits our AQI data the best, we use hourly air quality data for the months of August to December 2018 to first calculate scale (α), shape (β) and location (δ) parameters for the four types of distributions. The scale parameter provides an estimate of the scale on the x-axis. It squeezes or stretches the distribution. The location parameter gives the location where the distribution is centered on the x-axis. The shape parameter defines the skewness and kurtosis. The parameters for AQI data at AnandVihar and Dadri are calculated using equations in Section 5 and are shown in Table IV.

Table IV: Parameters for Lognormal, Gumbel, Frechet and Weibull distributions for AQI data

Distributions	Lognormal		Gumbel		Frechet		Weibull	
Parameters	α	β	β	δ	α	β	α	β
AnandVihar	0.293	4.31	26.4	134.3	2.86	363.47	2.94	540.41
Dadri	0.273	5.31	32.62	147.61	1.93	167.21	1.99	300.31

The histogram of AQI data with lognormal, Gumbel, Frechet and Weibull distributions are shown in Figures 4, 5, 6 and 7. As seen from these plots, Lognormal and Gumbel distributions seem to fit our AQI data quite well. To identify the distribution that has best forecasting performance, 'goodness of fit tests' are performed using, (1) coefficient of determination (R^2) and (2) index of agreement (IA).

R^2 is the proportion of the variance in the dependent variable that can be predicted from independent variable. It is used to estimate how well a model is able to explain and forecast future outcomes. It is defined as

$$R^2 = \frac{\sum_{i=1}^{N}(F_i - \bar{F})^2}{\sum_{i=1}^{N}(F_i - \bar{F})^2 + \sum_{i=1}^{N}(F_i - F)^2} \tag{21}$$

here, F is the estimated cumulative probabilities that are derived from the cumulative probability distribution function of the proposed model, $\bar{F}$ is the mean of distribution function ($\bar{F} = \frac{\sum_{i=1}^{N}}{N}$). R^2 ranges between 0 and 1, the value 0 implying that the model explains none of the variability of the response data. A value of 1 indicates that the model explains and predicts future outcomes perfectly. Another measure of goodness of fit that we have used in our work is the Index of Agreement (IA). It is defined as

$$IA = 1 - \frac{\sum_{i=1}^{N}(O_i - P_i)}{\sum_{i=1}^{N}\left(|P_i - \bar{O}| + |O_i - \bar{O}|\right)} \tag{22}$$

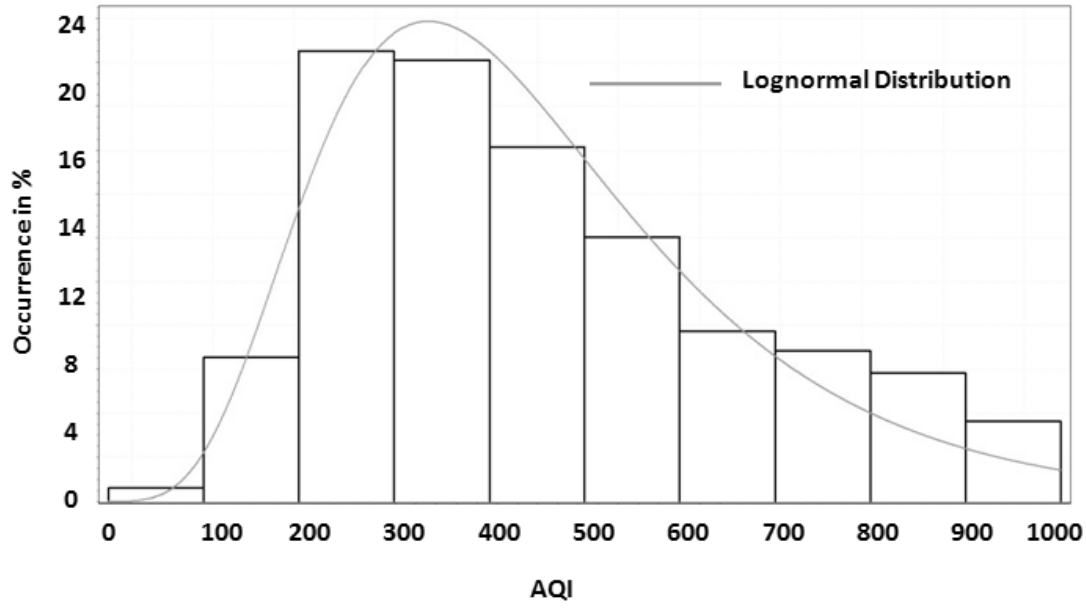

Fig.4: Histogram of hourly AQI at AnandVihar over the period from August, 2018 to December, 2018 with Lognormal Distribution

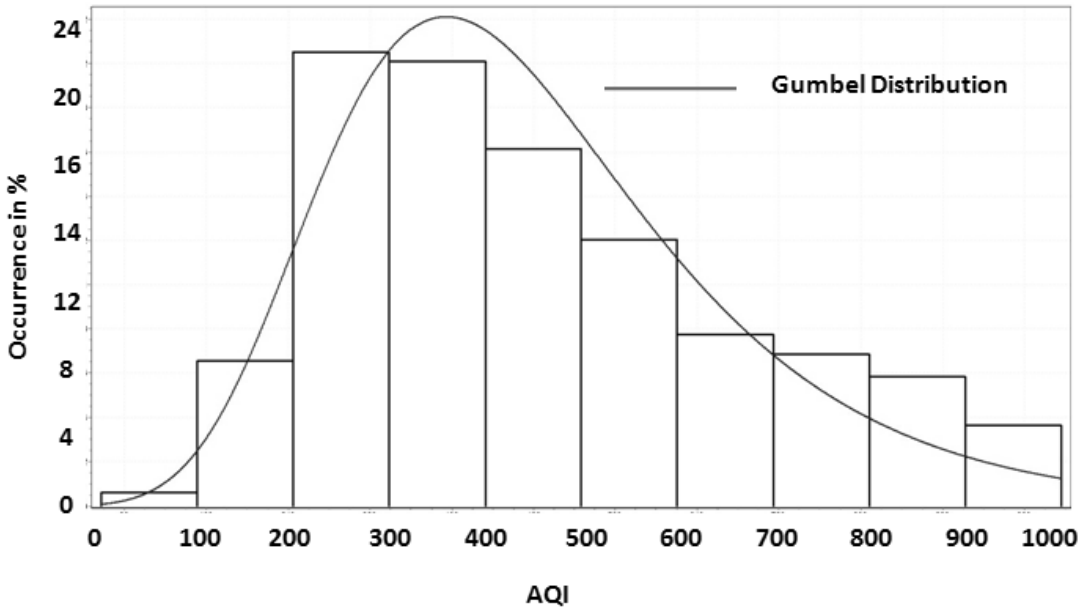

Fig.5: Histogram of hourly AQI at AnandVihar over the period from August, 2018 to December, 2018 with Gumbel Distribution

here, O_i and P_i are the i^{th} observed and predicted values and N is the number of observations of AQI data. The index of agreement is also defined as the ratio of mean square error (MSE) and potential error (PE), multiplied by the number of observations (N), and then subtracted from one.

$$IA = 1 - N.\frac{MSE}{PE} \tag{23}$$

IA lies between 0 and 1, the value of 1 indicating a perfect agreement between modeled values, P_i and the observations, O_i. A value of 0 indicates no agreement between modeled and observed values. Table V shows the results of goodness of fit test using R^2 and IA.

Table V: Results of Goodness of Fit Tests

Year	Distribution	R^2	IA
AnandVihar	Lognormal	0.79	0.75
	Gumbel	**0.97**	**0.91**
	Frechet	0.84	0.81
	Weibull	0.88	0.83
Dadri	Lognormal	0.79	0.81
	Gumbel	**0.94**	**0.94**
	Frechet	0.81	0.89
	Weibull	0.88	0.90

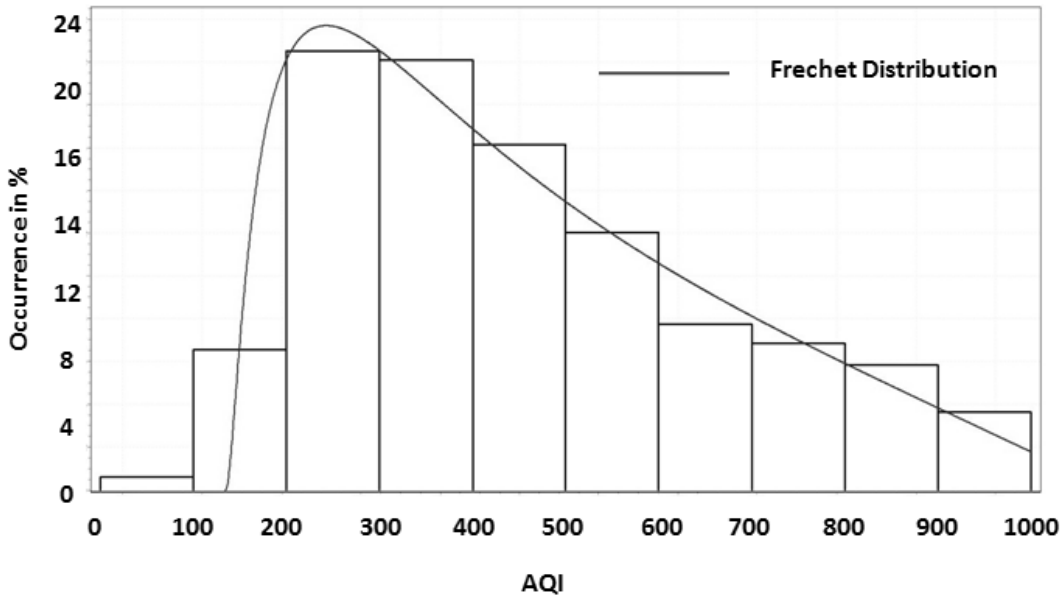

Fig.6: Histogram of hourly AQI at AnandVihar over the period from August, 2018 to December, 2018 with Frechet Distribution

The results show that EVD type I, Gumbel has highest values of R^2 and IA, and hence is best suited for forecast of extreme events of poor air quality and return periods.

7. COMPUTATION OF EXCEEDANCE PROBABILITY AND RETURN PERIOD OF AQI

Exceedance probability is the probability that a certain level of AQI will be breached. Return period is the estimated time of recurrence of an event, such as exceedance of certain level of AQI. The occurrence of peaks of extreme values of AQI, for a large number of observations at a location could be treated as a random variable. Let y_1, y_2, y_3, y_4, .. , y_n be a set of AQI observations of an independent random variable Y, with y_1 being the largest value, and y_n being the smallest. The experiment of generating the sequence- $y_1 > y_2 > y_3 > y_4 >$, .. , $> y_n$ is repeated N times, resulting in N such values of X_m, where X_m is the m^{th} largest value of X from a sample size of n. The observed x_m for each value 'm' are arranged in descending order resulting in the sequence x_m(r); r=1 to N. Here, r is the rank of a particular x_m within the

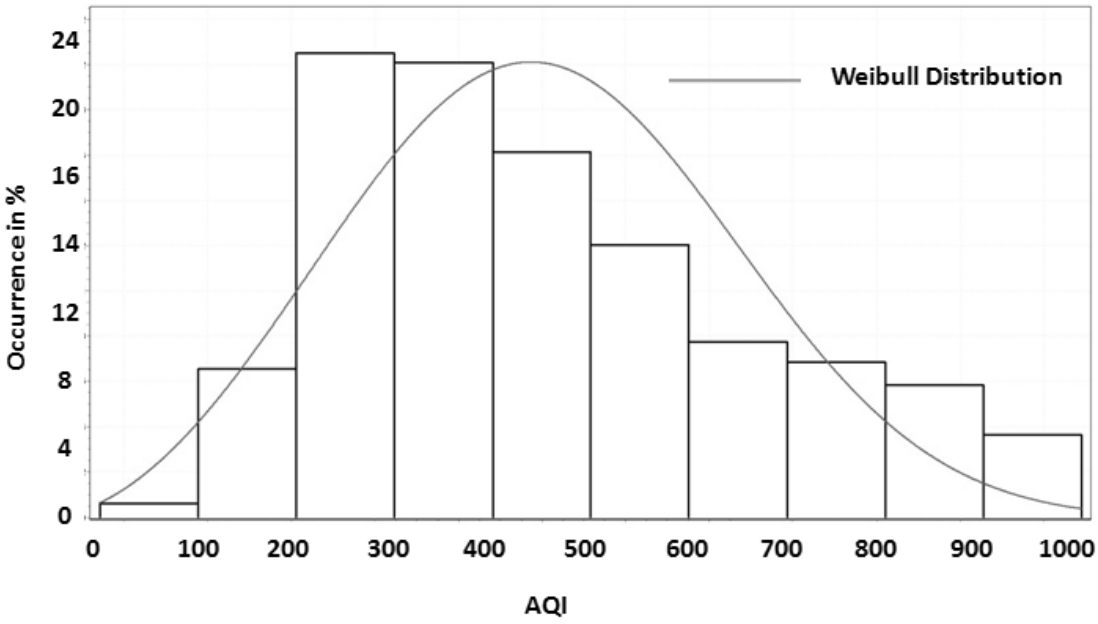

Fig.7: Histogram of hourly AQI at AnandVihar over the period from August, 2018 to December, 2018 with Weibull Distribution

X_m sequence. The probability $P_{1n}(x)$ that the largest AQI observation X_1 in a sample of size n is less than or equal to x if and only if the 'n' independent AQI observations are all less than or equal to x

$$P_{1n}(x) = Pr\{X \leqslant x\} = \Pi Pr(X \leqslant x) = \{F(x)\}^n \tag{24}$$

$P_{1n}(x)$ approaches an extreme value distribution as 'n' or number of AQI observations, increases indefinitely. The Extreme Value Type I Distribution or the *Gumbel* distribution is obtained from an exponential type of initial distribution. Examples of such distributions are exponential, normal, log-normal, *Weibull*, gamma and beta. For large values of the variables, right tails of their pdf converge to exponential form. The pdf, p(x) or the probability of exceedance of AQI, which is the derivative of the *cdf* $P_{1n}(x)$ in equation (24), is of the form $\lambda.e^{-\lambda x}$, λ being the rate parameter of the distribution. Hence, the probability of non-exceedance of the AQI value is given by

$$F(x) = 1 - \lambda.e^{-\lambda x} \tag{25}$$

From equation (24) and (25), we have

$$P_{1n}(x) = Pr\{X \leqslant x\} = (1 - \lambda.e^{-\lambda x})^n \tag{26}$$

Location and dispersion parameters β and α are added to the equation (26), to obtain equation (27)

$$P_{1n}(x) = \left[1 - \frac{\exp\{-\alpha(x-\beta)\}}{n}\right] \tag{27}$$

For large 'n' or large number of AQI observations, $P_{1n}(x)$ is approximated as

$$\lim_{n\to\infty} P_{1n}(x) = \tilde{P}_{1n}(x) = \exp\{-e^{-\alpha(x-\beta)}\} \tag{28}$$

The equation (28) represents the double exponential Gumbel distribution. In form of reduced random variate, equation (28) can be written as

$$\tilde{P}_{1n}(y) = \exp\{-e^{-y}\} \tag{29}$$

Here, $\tilde{P}_{1n}(y)$ is the exceedance probability for a large number of AQI observations, y = $\alpha(x-\beta)$, $\beta = \bar{x} - 0.45\sigma_x$ and $\alpha = 1.2825/\sigma_x$

The Return Period (T_y) is the average interval between the occurrence of AQI, of magnitude equal to or greater than a given 'x'. The probability that the time interval 'T' between two exceedances of X of magnitude x equals N, is given by,

$$\begin{aligned} Pr(\bar{T} = n) &= \Pi Pr(Y_i < y).Pr(Y_n < y) \\ &= \{Pr(Y < y)\}^{N-1}.Pr(Y < y) \end{aligned} \tag{30}$$

The equation (30) represents a geometric distribution, where the experiment is performed until threshold value of y is exceeded once. The expected value of time period E($\bar{T}$) is calculated using the geometric distribution where 'N' can take any value between 1 and ∞,

$$\begin{aligned} E(\bar{T}) &= \sum_{N=1}^{\infty} N.Pr(\bar{T} = N) \\ &= \sum_{N=1}^{\infty} N.\{1 - Pr(Y > y)\}^{N-1}.Pr(Y > y) \\ &= 1/Pr(Y > y) \end{aligned} \tag{31}$$

But, Pr($Y > y$) = 1 - $P_{1n}(y)$, where
$P_{1n}(y)$ = Pr($Y < y$) = $\exp\{-\exp(-y)\}$. Therefore,

$$Return\,Period\,(T_y) = [1 - P_{1n}(y)]^{-1} \tag{32}$$

where $P_{1n}(y)$ is the exceedance probability for a certain value of AQI.

Relation between Exceedance Probability and Return Period: Equation (32) is re-arranged to give

$$P_{1n}(y) = 1 - \frac{1}{T_y} = \frac{T_y - 1}{T_y} \tag{33}$$

where $P_{1n}(y)$ is the exceedance probability and T_y is the return period.

8. PERFORMANCE EVALUATION

Hourly AQI data of two locations, AnandVihar in Delhi and Dadri in the National Capital Region (NCR), has been used to forecast the exceedance probability and return period for AQI levels. Exceedance probabilities and return periods have been forecasted for AQI values as per the AQI bands shown in Table III. Using Gumbel distribution, forecasted exceedance probabilities and return periods for AQI values of 50, 100, 150, 200 and 300 at two locations- AnandVihar and Dadri have been displayed in form of plots in Figure 8 and 9.

Table VI shows the forecasted exceedance probabilities for an AQI of 50, 100, 150, 200 and 300 at AnandVihar (A.Vihar) and Dadri respectively. The graph between exceedance probability and AQI has been plotted in Figure 8.

AQI value of more than 150 is considered unhealthy for humans. The above table indicates that there is a high probability of exceedance of AQI extremes of 150, 200 and 300 at AnandVihar, while such a probability is lower at Dadri.

Figure 9 shows the return period for different AQI extremes at AnandVihar and Dadri respectively. The return period for AQI values of 200 and 300 is considerably lower at AnandVihar, as compared with

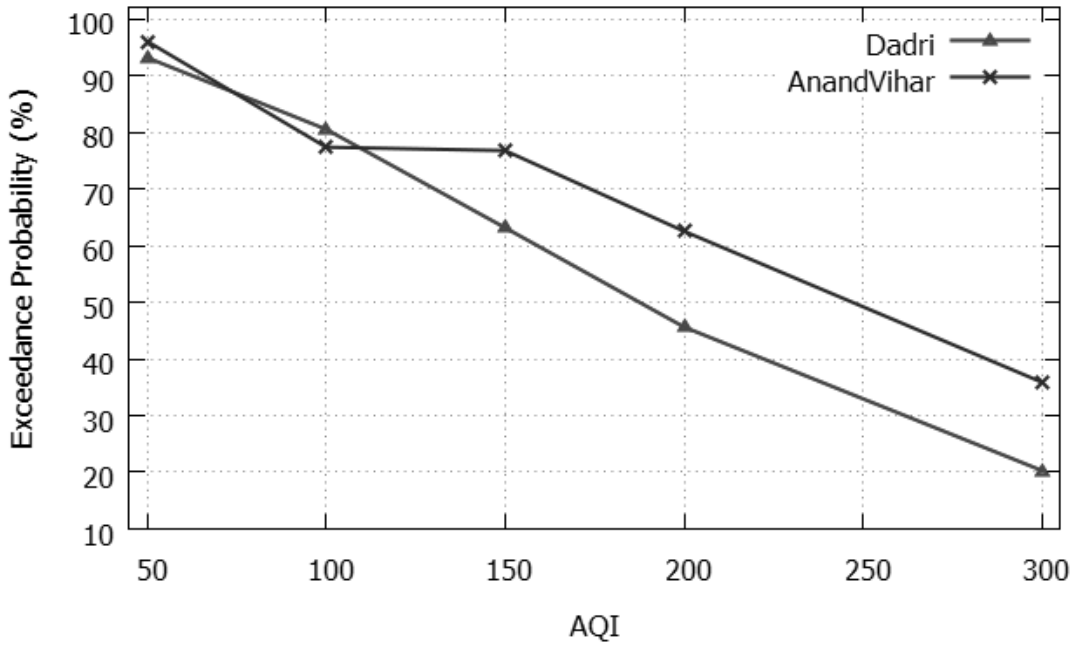

Fig. 8: Exceedance probabilities for different threshold values of AQI at AnandVihar and Dadri respectively

Table VI: Exceedance Probability (%)

AQI (>)	Exceedance Prob., A.Vihar (%)	Exceedance Prob., Dadri (%)
50	96.00	93.12
100	77.34	80.49
150	76.73	63.13
200	62.51	45.62
300	35.87	20.31

Dadri, indicating that the possibility of occurence of an event of extremely poor air quality is higher at AnandVihar in comparison to Dadri.

Table VII: Return Period (hours)

AQI (>)	Return Period, A.Vihar (hours)	Return Period, Dadri (hours)
50	1.04	1.07
100	1.29	1.24
150	1.30	1.58
200	1.56	2.19
300	2.78	4.92

8.1 Comparison between Forecasted and Actual Exceedance Probabilities

In this section, the exceedance probabilities and return periods of AQI levels forecasted using Gumbel distribution have been compared with actual values for the period of August, 2018 to December, 2018. The actual exceedance probability is given as

$$P\{(AQI) > (AQI_{TH})\} = \frac{days\,on\,which\,AQI\,exceeds}{total\,no.\,of\,days} \tag{34}$$

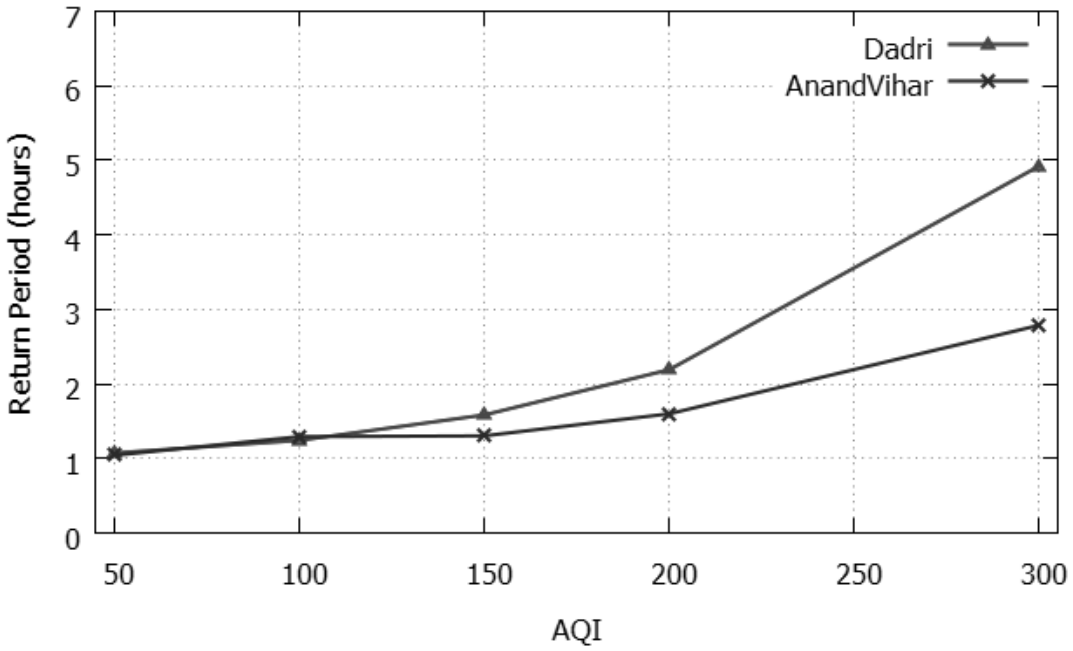

Fig.9: Return Period for different threshold values of AQI at Anand Vihar and Dadri respectively

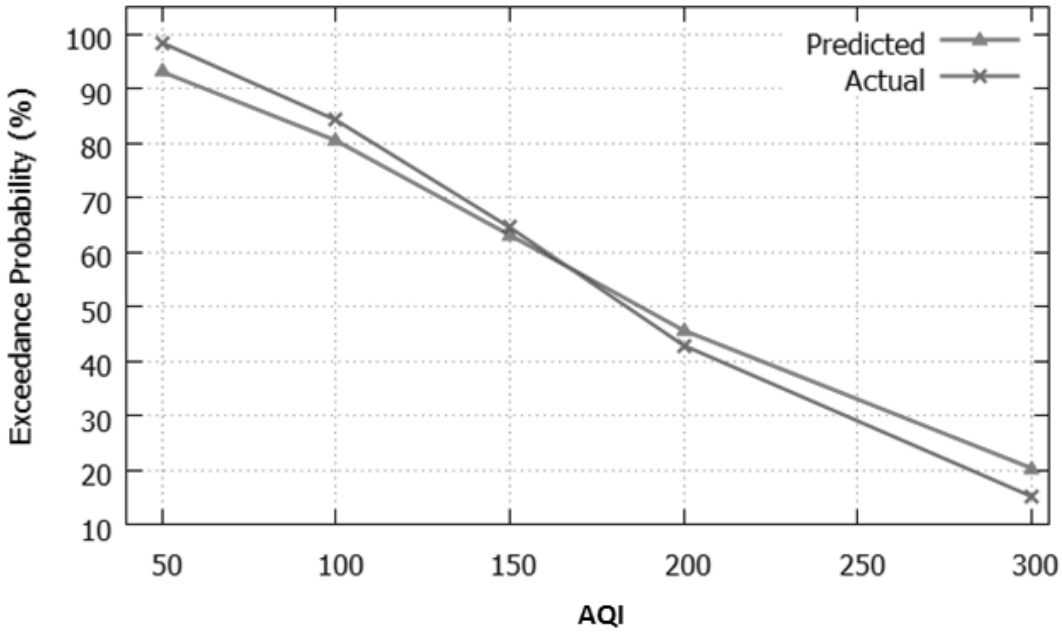

Fig.10: Forecasted and Actual exceedances at Dadri over the period between August, 2018 and December, 2018

The actual exceedance probabilities and the ones calculated using the EVD type I at AnandVihar and Dadri have been plotted in Figures 10 and 11.

Table VIII and IX present the forecast error in exceedance probabilities at AnandVihar and Dadri respectively. The difference between actual and forecasted exceedance probabilities (forecast error) at

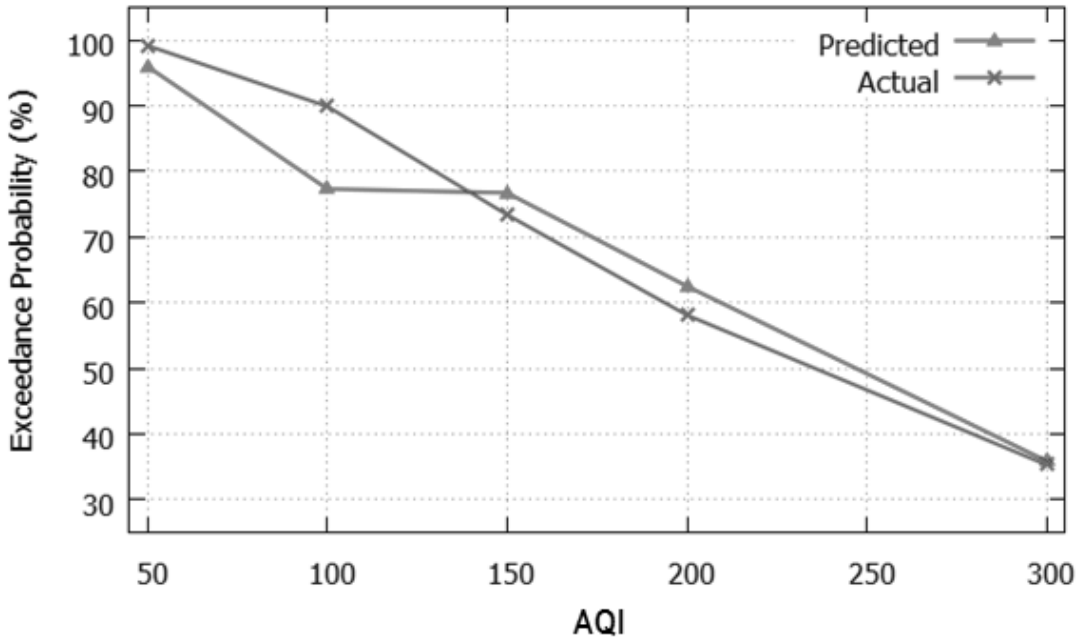

Fig.11: Forecasted and Actual exceedances at AnandVihar over the period between August, 2018 and December, 2018

AnandVihar decreases sharply as the AQI value increases, as shown in Table VIII. The average absolute forecast error at AnandVihar is 6.18%.

Table VIII: Comparison of Actual and Forecasted Probabilities at AnandVihar

AQI (>)	Actual (%)	Forecasted (%)	Error (%)
50	96	99.18	3.20
100	77.34	88.91	13.91
150	76.73	73.41	-4.52
200	62.51	58.11	-7.57
300	35.87	35.30	-1.61

As shown in Table IX, the average absolute forecast error at Dadri is 6.67%. These results at Anand-Vihar and Dadri validate that the Gumbel distribution model to forecast exceedances of AQI levels is suitable for these regions.

Table IX: Comparison of Actual and Forecasted Probabilities at Dadri

AQI (>)	Actual (%)	Forecasted (%)	Error (%)
50	98.45	93.12	5.41
100	84.26	80.49	4.47
150	64.54	63.13	2.19
200	42.86	45.62	-6.43
300	17.51	20.11	-14.61

9. RESULTS AND CONCLUSION

An Internet of Things system has been developed to collect, visualize, analyze and model urban air quality in this work. Hourly air pollutant data has been collected at two locations in the Delhi-NCR region, i.e. AnandVihar and Dadri, for the months of August, 2018 to December, 2018. Three types of EVD models and one central fitting distribution was considered for analyzing and forecasting events of extremely poor air quality. Two goodness of fit criteria were used to determine the best fit model for analysis of air quality of this high pollutant region. The results with coefficient of determination and index of agreement indicate that EVD type I, Gumbel distribution based model is indeed better suited for analysis and forecasting of air quality in this region. Forecast model using Gumbel distribution has been used to forecast exceedance probabilities and return periods, which is useful in understanding how frequently extreme values of AQI are going to occur. The forecast error has been computed at both locations. Average forecast error has been found to be 6.18% and 6.67% at AnandVihar and Dadri respectively, validating the use of extreme value distributions in these high pollutant concentration regions.

REFERENCES

AMANN, M., PUROHIT, P., BHANARKAR, A. D., BERTOK, I., BORKEN-KLEEFELD, J., COFALA, J., VARDHAN, B. H. 2017. Managing future air quality in megacities: A case study for Delhi. *Atmospheric Environment* 161, 99 - 111.

BHANARKAR, A. D., PUROHIT, P., RAFAJ, P., AMANN, M., BERTOK, I., COFALA, J., KUMAR, R. 2018. Managing future air quality in megacities: Co-benefit assessment for Delhi. *Atmospheric Environment* 186, 158 - 177.

ZHANG, QIANG ET. AL. 2017. Transboundary health impacts of transported global air pollution and international trade. *Nature* Vol - 543, SP - 705, Macmillan Publishers Limited, Springer Nature.

COHEN, AARON J. ET. AL. 2017. Estimates and 25-year trends of the global burden of disease attributable to ambient air pollution: an analysis of data from the Global Burden of Diseases Study 2015. *Lancet* Volume 389, Issue 10082, 2017, Pages 1907-1918, ISSN 0140 - 6736,

RAGGAD, B.. 2018. Stationary and Non-stationary Extreme Value Approaches for Modelling Extreme Temperature: the Case of Riyadh City, Saudi Arabia. *Environ Model Assess* 23: 99, Elsevier.

CHEN, L., YANG, D , ZHANGDAQIN,G , WANG C. , LI, J., NGUYEN, THI-MAI-TRANG. 2018. Deep mobile traffic forecast and complementary base station clustering for C-RAN optimization. *Journal of Network and Computer Applications, Elsevier* Volume 121, Pages 59 - 69, ISSN 1084 - 8045.

WANG, Y. AND CHEN, G. 2017. Efficient Data Gathering and Estimation for Metropolitan Air Quality Monitoring by Using Vehicular Sensor Networks. *IEEE Transactions on Vehicular Technology* Vol. 66, no. 8, pp. 7234-7248, Aug. 2017.

ZHU, J. Y., SUN, C. AND LI, V. O. K. 2017. An Extended Spatio-Temporal Granger Causality Model for Air Quality Estimation with Heterogeneous Urban Big Data. *IEEE Transactions on Big Data* Vol. 3, no. 3, pp. 307 - 319, Sept. 1, 2017.

MARTINS, LEILA DROPRINCHINSKI ET AL. 2017. Extreme value analysis of air pollution data and their comparison between two large urban regions of South America. *Weather and Climate Extremes*, Volume 18, Pages 44 - 54, ISSN 2212 - 0947 (Elsevier).

ERCELEBI, SELAMET, TOROS AND HSEYIN. 2009. Extreme Value Analysis of Istanbul Air Pollution Data. *CLEAN Soil, Air, Water,* 37. 122 - 131(2009).10.1002/clen.200800041.

KIM, J., CHU, C. AND SHIN, S. 2014. ISSAQ: An Integrated Sensing Systems for Real-Time Indoor Air Quality Monitoring. *IEEE Sensors Journal*, vol. 14, no. 12, pp. 4230 - 4244, Dec. 2014.

BARTHWAL, A. AND ACHARYA, D. 2018. An Internet of Things System for Sensing, Analysis & Forecasting Urban Air Quality," 2018 *IEEE (CONECCT)*, Bangalore, 2018, pp. 1 - 6.

WANG, X. MAUZERALL, D. L. 2004. Characterizing Distributions of Surface Ozone and Its Impact on Grain Production in China, Japan and South Korea: 1990 and 2020. *Atmospheric Environment*. 38, p. 4383 - 4402.

LU, H. C. 2004. The Statistical Characters of PM10 Concentration in Taiwan Area. *Atmospheric Environment*. 36, p. 491 - 502.

CENTRAL POLLUTION CONTROL BOARD, INDIA. 2014. `http://cpcb.nic.in/National-Air-Quality-Index/`.

KARACA, F., ALAGHA, O., ERTURK, F. 2005. Statistical characterization of atmospheric PM10 and PM2.5 concentrations at a non-impacted suburban site of Istanbul, Turkey. *Chemosphere*. 59, p.1183 - 1190.

SHARMA, P., KHARE, M. AND CHAKRABARTI, S.P. 1999. Application of extreme value theory for predicting violations of air quality standards for an urban road intersection. *Transportation Research Part D: Transport and Environment, 4.* 201 - 216.

MDYUSOF, N. F. F., RAMLI, N. A., YAHAYA, A. S., SANSUDDIN, N., GHAZALI, N. A. AND AL MADHOUN, W. A. 2011. Central fitting distributions and extreme value distributions for prediction of high PM10 concentration. *2011 International Conference on Multimedia Technology*, Hangzhou, 2011, pp. 6263 - 6266.

[18] KAHN, HENRY D. 1973. Note On The Distribution of Air Pollutants. *Journal of the Air Pollution Control Association*, 23:11, 973 - 973.

Anurag Barthwal is a Ph.D. candidate in the Department of Computer Science and Engineering at Shiv Nadar University. His research interests are in the areas of Wireless Communications, Internet of Things and Mathematical Modeling. His research is supported by HumanSense grant funded by Information Technology Research Academy (ITRA), Ministry of Electronics and IT, Government of India.

Dr. Debopam Acharya Dr. Debopam Acharya is currently Associate Professor of Computer Science and Engineering at Shiv Nadar University, India. He is the Founding Head of the Department of Computer Science and Engineering at Shiv Nadar University and Chaired the department till 2018. His research interests are in the areas of Internet of Things, Mobile Sensing and Smart Environment. Prior to joining SNU, he has worked in the Location Based Services group at Garmin Inc., Olathe, Kansas,USA and tenure track faculty at Georgia Southern University, Statesboro, GA, USA. His work on SAVE (Sun-java based automatic vehicular accident reporting system), a sensor based technology to prevent and report an automobile accident has featured in leading international publications. His research has been funded by several public and private sector agencies like BlackBerry, Dell India, and Ministry of Electronics and IT, Govt. of India. Dr. Acharya earned his PhD in 2006 from the University of Missouri-Kansas City, USA.