Atmospheric Environment 234 (2020) 117544

Contents lists available at ScienceDirect

Atmospheric Environment

journal homepage: http://www.elsevier.com/locate/atmosenv

A framework for setting up a country-wide network of regional surface $PM_{2.5}$ sampling sites utilising a satellite-derived proxy – The COALESCE project, India

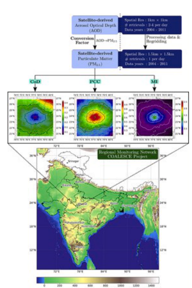
Nirav L. Lekinwala [a], Ankur Bhardwaj [c], Ramya Sunder Raman [c,d,*], Mani Bhushan [a,b], Kunal Bali [e], Sagnik Dey [e,f]

[a] Department of Chemical Engineering, Indian Institue of Technology Bombay, Mumbai, 400076, Maharashtra, India
[b] Interdisciplinary Programme in Climate Studies, Indian Institute of Technology Bombay, Mumbai, 400076, Maharashtra, India
[c] Department of Earth and Environmental Sciences, Indian Institute of Science Education and Research Bhopal, Bhopal Bypass Road, Bhauri, Bhopal, 462 066, Madhya Pradesh, India
[d] Center for Research on Environment and Sustainable Technologies, Indian Institute of Science Education and Research Bhopal, Bhopal Bypass Road, Bhauri, Bhopal, 462 066, Madhya Pradesh, India
[e] Center for Atmospheric Science, Indian Institute of Technology Delhi, 110016, Delhi, India
[f] Center of Excellence for Research on Clean Air, Indian Institute of Technology Delhi, 110016, Delhi, India

HIGHLIGHTS

- Novel method for identifying regional background sites for $PM_{2.5}$ measurement developed.
- Methodology includes use of physical criteria and satellite-derived $PM_{2.5}$ proxy.
- Satellite-derived $PM_{2.5}$ using AOD and conversion factors from GEOS-Chem utilized.
- Sites identified using metrics such as CoD, PCC, and MI to assess homogeneity in $PM_{2.5}$.
- Identification of 11 regionally representative $PM_{2.5}$ sampling sites for COALESCE.

GRAPHICAL ABSTRACT



ARTICLE INFO

Keywords:
Regionally representative $PM_{2.5}$ sampling sites
MODIS aerosol Optical depth
Coefficient of divergence (CoD)
Pearson correlation coefficient (PCC)
Mutual information (MI)
COALESCE project

ABSTRACT

Air quality management and assessment of aerosol climate effects over a large area require the strategic placement of regionally representative monitoring sites (RRMS) to capture the required information. Ground-based, fine particulate matter ($PM_{2.5}$) concentrations measured for a long duration at high spatial resolution i.e. at several potential locations in a region help identify an optimal regionally representative site for the monitoring network. However, in the absence of long-term $PM_{2.5}$ concentrations with high spatial resolution, identification of RRMS is a challenge. To identify for such situations, a novel methodology utilising satellite-derived $PM_{2.5}$ is presented in this study. High spatial resolution ($1\,km \times 1\,km$) daily aerosol optical depth (AOD) over several years (2004–2011) is used to derive surface $PM_{2.5}$ using appropriate conversion factors. $PM_{2.5}$ concentrations thus derived for a potential site and its nearby cells in the region are subjected to statistical

* Corresponding author. Department of Earth and Environmental Sciences, Indian Institute of Science Education and Research Bhopal, Bhopal Bypass Road, Bhauri, Bhopal, 462 066, Madhya Pradesh, India.
E-mail address: ramyasr@iiserb.ac.in (R. Sunder Raman).

https://doi.org/10.1016/j.atmosenv.2020.117544
Received 11 January 2020; Received in revised form 24 March 2020; Accepted 19 April 2020
Available online 4 May 2020
1352-2310/

analysis, to quantify linear/non-linear relationships between the respective $PM_{2.5}$ time-series. Metrics such as coefficient of divergence (CoD), Pearson correlation coefficient (PCC) and Mutual information (MI) are calculated to assess the regional representativeness of a site. These criteria combined with physical criteria such as proximity of the site to local sources, local meteorology analysis and clustered air parcel back-trajectory analysis, are used in a weight-of-evidence approach to establish site adequacy. The selection and validation of eleven sites for the COALESCE project, is used as an illustrative example, to demonstrate the effectiveness of the site-selection methodology developed in this study.

1. Introduction

Atmospheric aerosols consist of a mixture of sulphates, nitrates and carbonaceous particles and are produced from various natural and anthropogenic sources. These aerosols play an important role in regional air quality, atmospheric chemistry and climate change (e.g., Hansen et al., 2000; Pachauri et al., 2013). An increase in concentration of atmospheric aerosols has both health and environmental implications (e.g., Dockery et al., 1993; Shah et al., 2013). Thus, it is essential to identify the major regional sources of atmospheric aerosols to curb their increasing concentration. This, in turn, requires adequate monitoring of aerosols at a regional level which can be achieved by developing an effective monitoring network consisting of several sampling locations with appropriate sampling duration and sampling frequency. Each sampling site in the network should be representative of regional aerosols characteristics but not unduly influenced by nearby sources (Chow et al., 2002). A regional aerosol monitoring network helps apportion and identify the sources of aerosol emissions, thereby enabling appropriate policy-related decisions and development of suitable control strategies/technologies. Thus, setting up a monitoring network consisting of regionally representative monitoring sites (RRMS) is vital.

Regional aerosol networks are currently deployed in several countries. In particular, the US Environmental Protection Agency (USEPA) has set-up the Chemical Speciation Network (CSN) for ambient aerosol monitoring sites to enable control of emission of aerosols (US EPA, 1999). The sites selected for monitoring are objective-based and serve a particular purpose. Additionally, to improve the visibility in national parks in the US, National Park Service (NPS) introduced IMPROVE program which identified sites based on a particular set of conditions. These conditions mostly check the influence of local sources on the site by checking its distance from the nearest roads, checking for obstruction of airflow by trees and logistical availability of the site (IMPROVE, 2016). Other studies include short term measurements at several sites in a region and identifying the most RRMS based on statistical analysis using Coefficient of Divergence (CoD) and Pearson Correlation Coefficient (PCC) metrics (e.g., Kim et al., 2005; Wongphatarakul et al., 1998). In contrast to short-term sampling, long term sampling helps understand the characteristics of a site based on seasonal differences and specific events.

To set-up sites which represent the spatio-temporal characteristics of the region around it, i.e., RRMS, knowledge of the current $PM_{2.5}$ concentration levels and patterns within a region is essential. Further, carbonaceous aerosols often constitute a significant mass fraction of fine particles (particulate matter whose aerodynamic diameter is less than 2.5 μm, $PM_{2.5}$). India only has a limited number of operational ground-based $PM_{2.5}$ monitoring stations, and most of them are located in urban areas with inadequate rural coverage (Balakrishnan et al., 2015). Some of the significant PM monitoring network across India include, (a) The National Air Quality Monitoring Programme (NAMP) which monitors SO_2, NO_2 and PM_{10} at 703 air quality (AQ) stations across 307 cities and towns and; (b) Continuous Automatic Air Quality Monitoring Station (CAAQMS) which monitors SO_2, NO_2, PM_{10}, CO, O_3 and BTEX (Benzene, Toluene, Ethylbenzene and Xylene) at 130 CAAQM stations spread across 65 cities in India (CPCB, 2014). While data from both the networks are made available online for public use, it is often plagued with incomplete or missing data (Pant et al., 2019).

To the best of our knowledge, approaches to design an effective regional aerosol monitoring network in the absence of comprehensive long-term in-situ $PM_{2.5}$ measurements are not available in literature. This work addresses the problem of designing a regional aerosol monitoring network in the absence of comprehensive ground-based in-situ $PM_{2.5}$ measurements. While obtaining high spatial resolution $PM_{2.5}$ data using conventional ground-based measurement techniques may not be always feasible, aerosol products (such as Aerosol Optical Depth, AOD) from satellites can be used as a proxy for $PM_{2.5}$ measurements, which have high spatial coverage (global) over a number of years (2002-present). In this work, we utilise a widely available satellite aerosol product, AOD, instead of in-situ $PM_{2.5}$ measurements, coupled with other ancillary information, to build a weight-of-evidence approach for site selection. Ancillary information includes scaling factors developed from modelling to estimate surface $PM_{2.5}$ from AOD, information about the local sources and local meteorology along with clustered back-trajectory analysis and statistical analysis using complimentary metrics such as CoD, PCC and Mutual Information (MI) to check for the regional representativeness of a sampling site.

The usefulness of the framework developed in this work is demonstrated by its application to design a regional aerosol monitoring network in India, as a part of the COALESCE (Carbonaceous Aerosols Emission, Source apportionment and Climate Impacts) project. The main objectives of the COALESCE project are to understand, (1) regional carbonaceous aerosols – their primary source sectors and influence, (2) carbonaceous aerosols – atmospheric transformation and processes including those that influence regional air quality and climate. The project design and its specific objectives are discussed elsewhere (Venkataraman et al., 2020). This multi-institutional project seeks to operate eleven sites (number of sites dictated by budgetary and logistical constraints) sampling $PM_{2.5}$ every other day for two years (2019, 2020). It was desired to strategically locate these sites such that they capture not only regional $PM_{2.5}$ but also the spatial variability across regions of India.

2. Methodology

Understanding of the local and physical features of each site is essential to ensure its regional representativeness. Blanchard et al. (1999) defined spatial/regional representativeness of a monitoring site as the area within which the pollutant concentration is approximately constant. The site is termed as spatially representative if the concentrations at the adjoining area is within 20% of that recorded at the site. The methodology developed in this study evaluates a site and its surrounding area based on a weight-of-evidence approach. In addition to compliance with the physical site guidelines (see section 2.1), this approach consists of the use of trajectory-ensemble cluster analysis method along with statistical analysis on satellite-derived data for fine PM. Complementary metrics such as Coefficient of Divergence (CoD), Pearson Correlation Coefficient (PCC) and Mutual Information (MI) are calculated for fine PM concentrations to select regionally representative sites. Fig. 1 shows a summary of the steps involved in the site selection process with a discussion of each item in the following sub-sections.

2.1. Site information and meteorology

Local information plays an essential role in identifying key drivers of $PM_{2.5}$ concentration at a site. Specifications of local information criteria used to identify the potential measurement sites are discussed in detail in supplemental text S1. In a nutshell, the identified site should satisfy the following criteria:

1) Site should be physically distant from local sources such as paved and unpaved roads, agricultural field burning and residential biomass burning.
2) Site should not be downwind of a large point source to avoid its dominance on $PM_{2.5}$ concentration at the site: The influence of large local sources can be assessed by analysing local meteorolog- ywind speed and wind direction (see Section 2.1.1).
3) Site should not be affected by local topography such as lakes, hills, valleys and other features that may affect or alter the wind pattern in the nearby area (see supplemental text S1.2).
4) Site should be representative of the regional long-range transport characteristics: These characteristics can be assessed by using clustered back-trajectories (see Section 2.1.2).

2.1.1. Local meteorology

Shenfeld (1970) extensively studied the effect of meteorological parameters on the amount of pollution in the atmosphere. Meteorological parameters such as wind speed, wind direction together with the level of emission, affect the dispersion and transport of particulate matter and influence spatio-temporal variability. High wind speed hinders the rise of pollution in elevated sources, thus ground level concentration increases, while for sources at a lower height, the greater the wind speed, higher the dilution which results in low ground level concentration. Wind direction and its persistence is a significant factor in predicting the air pollution potential of an area when the primary source of the pollutants is elevated sources. A wind-rose plot can visualise the wind speed and wind direction data. This information along with a map of point sources are used to assess the potential for $PM_{2.5}$ sampled at the site to be locally dominated. Sites affected by local sources may be inhibited in their ability to capture background concentration and may not represent the region around it well.

2.1.2. Back trajectory ensembles

In-situ parameters and meteorology, while providing information about potential local source influence, fail to address the issue of air parcels transported over long distances. Air parcels while passing from one region to another accumulate and deposit pollutants. Air parcel back-trajectory analysis is widely used to determine the origin of the air masses and their transport pathways to establish a source-receptor relationship (e.g., Occhipinti et al., 2008; Han et al., 2005). These trajectories may not represent the exact path of the air parcel but are suitable to identify a particular synoptic-scale transport (e.g., Delcloo and De Backer, 2008; Stohl et al., 2002, 2001). Back-trajectory models such as HYSPLIT can be used to generate back trajectories. To obtain meaningful information, back-trajectories should be generated for different seasons. However, this can lead to a large number of trajectories over a span of a year, making it difficult to directly interpret the origins of the trajectories. To address this issue, this study uses *k*-means clustering to group the trajectories to obtain broad categories of origins. k-means is a multivariate clustering technique and can be used to cluster the trajectories into a user-specified number of groups such that the

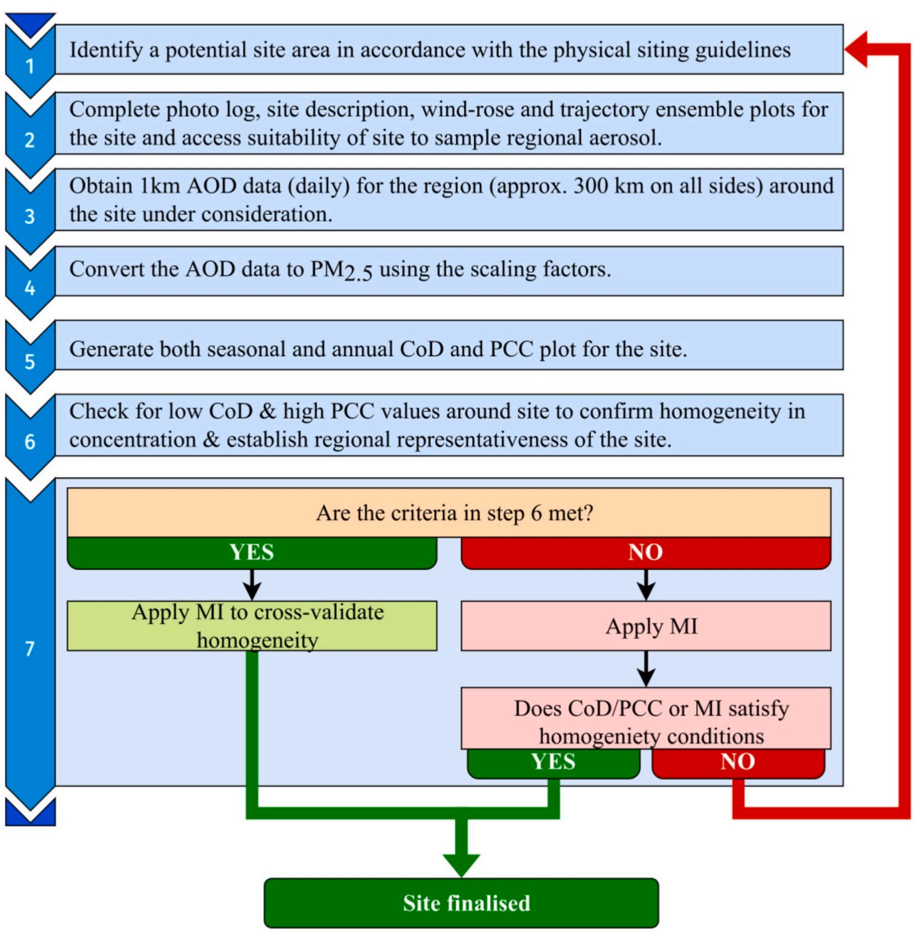


Fig. 1. Stepwise weight-of-evidence approach for site selection.

intra-cluster variability is minimized. The change in objective function with a change in a number of clusters is observed and the point beyond which there is no significant reduction in the objective function value, is used to decide the number of clusters. The resulting clustered trajectories help identify significant source regions and probable sources that contribute to the concentration of particulate matter at the site under consideration. Details of the approach are discussed in supplemental text S2.

2.2. Utilising a satellite derived proxy for daily surface $PM_{2.5}$

While it is essential to analyse the site based on local information, for a site to be regionally representative, the $PM_{2.5}$ concentrations at the site should be representative of the $PM_{2.5}$ concentrations in the neighbourhood of the site. To assess this representativeness, information about the $PM_{2.5}$ concentrations in the neighbouring regions is required. In the absence of such information from extensive ground-based surface measurements, this study utilises a satellite-derived proxy.

Satellites such as NASA's Aqua and Terra, equipped with moderate resolution imaging spectrometer (MODIS) sensor, pass over the Indian region daily at 10:30 IST and 13:30 IST (and 01:30 IST – night time) respectively (NASA, 2000 and NASA, 2002). The satellite-derived aerosol product is processed using the Multi-Angle Implementation of Atmospheric Correction (MAIAC) algorithm which generates high spatial resolution (1 km × 1 km) AOD data (Lyapustin et al., 2018, 2011). This study uses the MAIAC AOD data to estimate the $PM_{2.5}$ concentration by using the conversion factor (CF) (available offline) developed by van Donkelaar using the product of GEOS- Chem Chemical Transport Model (van Donkelaar et al., 2010, 2015) on a daily scale with 10 km × 10 km resolution. The conversion factor represents the ratio of the 24 h model derived $PM_{2.5}$ to AOD (Liu et al., 2004) as shown in equation (1);

$$\text{satellite derived } PM_{2.5} = \left(\frac{PM_{2.5}}{AOD}\right)_{\text{model}} \times \text{satellite derived AOD} \tag{1}$$

The daily value of model-derived 24-h average $PM_{2.5}$ calculation is based on two different relative humidity values such as at 35% for US/Canada and 50% for European Surfaces (van Donkelaar et al., 2010). So, some uncertainties could be present in our $PM_{2.5}$ estimation over Indian region, because of the different relative humidity used in the conversion factor. Literature also reports $PM_{2.5}$ estimation using the MAIAC AOD product (Just et al., 2015; Lee et al., 2016) with several advantages and disadvantages (Chu et al., 2016). In this study, we estimated the $PM_{2.5}$ product by downscaling the conversion factor's resolution to 1 km (according to MAIAC resolution) and then multiplying it (CF) with the MAIAC AOD product. Despite the uncertainties in the specific values of $PM_{2.5}$ estimated by Equation (1), this approach facilitates analysis of long-term trends and inter-comparisons between $PM_{2.5}$ in various grid cells.

The high-resolution data obtained from the satellite can be used to assess the representativeness of the potential monitoring site by comparing its satellite-derived $PM_{2.5}$ values with satellite-derived $PM_{2.5}$ values in the neighbouring cells. Satellite data is available for several years (2002–2019), and thus, data at cells of interest can be compared in terms of magnitude and changes in trends. Pearson Correlation Coefficient (PCC) and Coefficient of Divergence (CoD) are complementary metrics which can be used for site assessment and both have found application in the field of air quality monitoring (AQM), for comparing the concentration of pollutants at two monitoring sites (Krudysz et al., 2009; Xie et al., 2012a). However, their use with satellite data derived $PM_{2.5}$ for network site selection presented in this work is novel. For subsequent discussions in this section, let vectors X and Y indicate the time series of satellite derived $PM_{2.5}$ values at the potential monitoring site, and a neighbouring cell under consideration, respectively. In particular,

$$X = [x_1\ x_2\ \dots\ x_n]' \tag{2}$$

$$Y = [y_1\ y_2\ \dots\ y_n]' \tag{3}$$

where x_t, y_t, for all $t = 1, 2, \dots n$, are the satellite-derived $PM_{2.5}$ values at time instant t at the potential monitoring site, and the neighbouring cell under consideration.

2.2.1. Coefficient of divergence

Coefficient of Divergence (CoD) (e.g., Kim et al., 2005; Wongphatarakul et al., 1998) is used to evaluate the similarity of $PM_{2.5}$ values between the potential sampling site and the neighbouring cells and it is defined as:

$$\text{CoD} = \sqrt{\frac{1}{n}\sum_{t=1}^{n}\left(\frac{x_t - y_t}{x_t + y_t}\right)^2} \tag{4}$$

where, $x_t$and y_t are values of $PM_{2.5}$ at the reference cell and a neighbouring cell at t^{th} time instant respectively, n is the total number of time instants.

CoD values close to zero represent uniformity between a pair of samples and in turn represent similarity in the cell values, while the values close to unity represents complete divergence thereby indicating that the $PM_{2.5}$ values in the two cells are quite different. Studies in the literature suggest that if the CoD value is less than 0.2, the cells may be considered spatially homogeneous while when the value is above 0.2, the cells are spatially heterogenous (eg., Krudysz et al., 2009; Wilson et al., 2005). CoD compares the pair of values at a certain time instant, and fails to take trends in time series into consideration; thus, PCC is used alongside CoD to check if there exist a linear relationship between the $PM_{2.5}$ values in the two cells under consideration.

2.2.2. Pearson Correlation Coefficient

Pearson Correlation Coefficient (PCC or r) is used to show the degree of correspondence of satellite-derived $PM_{2.5}$ between the potential sampling site and the neighbouring cells. It is defined as:

$$r_{x,y} = \frac{\frac{1}{n}\sum_{t=1}^{n}(x_t - \mu_x)(y_t - \mu_y)}{\sqrt{\frac{1}{n}\sum_{t=1}^{n}(x_t - \mu_x)^2}\sqrt{\frac{1}{n}\sum_{t=1}^{n}(y_t - \mu_y)^2}} \tag{5}$$

where, $r_{x,y}$ is the PCC between vector at the reference cell and neighbouring cell over all n time instants, $x_t$and y_t are values of $PM_{2.5}$ at the reference cell and neighbouring cell at t^{th} time instant respectively, while μ_x and μ_y are mean values of $PM_{2.5}$ at the reference cell and neighbouring cell over all n time instants respectively.

PCC between cells can be used to check the uniformity of pollutant in a particular region when long time series data is available (Wongphatarakul et al., 1998). In particular, higher magnitude of PCC, i.e., a value close to unity (both positive and negative), indicates that the $PM_{2.5}$ values at the two sites are proportional. While PCC is a measure of strength and direction of the linear relationships between the two variables under consideration, it will have low values (close to zero) in case the variables are not linearly related but nevertheless have a strong nonlinear relationship. In such cases, PCC will not be adequate to assess representativeness and other metrics which can quantify similarity in the presence of nonlinear relationships are needed.

CoD and PCC are used as complementary tools because, while a high CoD (indicating heterogeneity) could be the result of concentration changes, it may still lead to high PCC if the two cells captured the same air mass since in this case the change in concentration will not affect their co-variance. On the other hand, a high CoD accompanied by a low PCC is a strong indicator of the two cells being heterogeneous and low CoD with high PCC indicates homogeneity between two cells (Krudysz et al., 2008; Wongphatarakul et al., 1998; Xie et al., 2012b). Thus, CoD and PCC plots are used together for regional background site identification to minimize the likelihood of obtaining a false positive (locations identified as homogenous in concentration when they are not) or a false negative (locations identified as heterogeneous in concentrations when they are actually not so); especially when the potential sampling site is being compared with the neighbouring cells which are not in close proximity to this site. The false-positives may arise due to the stagnant air mass while false-negatives may arise due to transport/diffusional, and depositional losses (Siudek and Frankowski, 2017). Additionally, the vectors X and Y may be related non-linearly which can be attributed to both meteorology and topographical feature of the site and nearby region. In such cases, despite being regionally representative, the site will fail the CoD and PCC based test, and hence additional metrics which can assess the representativeness for such situations are needed. In this work, we use Mutual Information (MI) for this purpose.

2.2.3. Mutual information

Used extensively in probability and information theory, mutual information (MI) is used to measure the mutual dependence between two variables. This dependence can be explained in terms of the amount of information about one variable that can be obtained from the other variable. $PM_{2.5}$ values used for the analysis are continuous variables. For continuous random variables, the MI is defined as (Cover and Thomas, 1991):

$$I(X,Y) = \int_Y \int_X P(x,y) \log\left[\frac{P(x,y)}{P(x)P(y)}\right] dxdy \tag{6}$$

where, $P(x,y)$ is the joint probability density function of X and Y and;$P(x)$ and $P(y)$ are the marginal probability density functions of X and Y, respectively.

An alternative way to define MI is (Cover and Thomas, 1991):

$$I(X,Y) = h(X) - h(X|Y) = h(Y) - h(Y|X) \tag{7}$$

where, $h(X)$ is the differential entropy of variable X, and $h(X|Y)$ is the conditional differential entropy of X given Y.

Thus, MI captures the reduction in entropy of X due to knowledge of Y. It should be noted that MI is a symmetric measure, i.e. $I(X,Y) = I(Y,X)$. Intuitively, the amount of information that X and Y share constitutes the MI. Hence, for cases where X and Y do not share any information, i.e., the variables are independent, MI is zero. Note that variables which are dependent but uncorrelated will have PCC value of zero, but MI will still be high. Thus, MI can be used as an additional metric for assessing regional representativeness when the variables may have a nonlinear relationship. Given vectors X and Y of satellite-derived $PM_{2.5}$ values, the proposed MI algorithm shown in Fig. 2, can be divided into two parts.

1. **Kernel density estimation (KDE):** KDE is a non-parametric method to estimate the probability density function (PDF) of the variable under consideration. To obtain the joint density, a 2D kernel density estimation procedure needs to be used. For our problem, we obtained ranked and normalized (Steuer et al., 2002) vectors $\tilde{X}$, $\tilde{Y}$ from the original vectors X and Y. A 2D Gaussian kernel density is then fit through $\tilde{X}$ and $\tilde{Y}$ data to estimate the joint probability density function (joint-PDF). Once joint-PDF is obtained, the marginal-PDF is

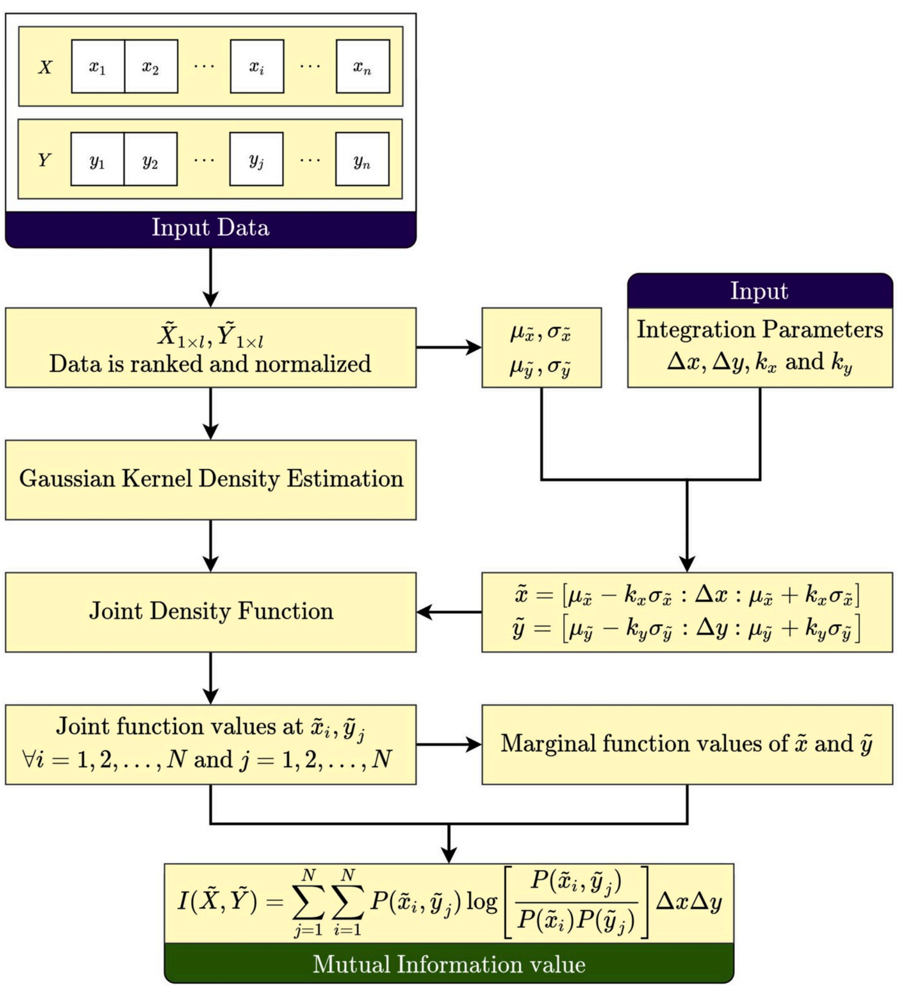


Fig. 2. Schematic for computing Mutual Information.

obtained by numerically integrating the joint-PDF. These joint and marginal-PDFs are used to compute MI.

2. **Numerical Integration:** Since the analytical form of the joint-PDF and the marginal PDFs is not available; the integral in Equation (6) needs to be computed numerically. Towards this end, we approximated Equation (6) as:

$$I(\tilde{X},\tilde{Y}) = \sum_{j=1}^{N}\sum_{i=1}^{N} P\left(\tilde{x}_i,\tilde{y}_j\right)\log\left[P\left(\tilde{x}_i,\frac{\tilde{y}_j)}{P(\tilde{x}_i)P(\tilde{y}_j)}\right]\Delta x\Delta y \quad (8)$$

where, Δx and Δy are user-specified interval length parameters for x and y, while the number of points (N) used in Equation (8) in the x (or y) dimension are considered to be equispaced with grid spacing Δx (or Δy) in the range $\mu_{\tilde{x}} - k_x\sigma_{\tilde{x}} : \Delta x : \mu_{\tilde{x}} + k_x\sigma_{\tilde{x}}$ (similarly for $\mu_{\tilde{y}} - k_y\sigma_{\tilde{y}} : \Delta y : \mu_{\tilde{y}} + k_y\sigma_{\tilde{y}}$) with k_x (or k_y) being a user-defined parameter.

In addition to meeting physical siting and meteorology-based criteria, a proposed site can be considered regionally representative for $PM_{2.5}$ sampling, if low CoD, high PCC and high MI values for $PM_{2.5}$ concentration are obtained between the selected location and grid cells in the region (typically 200 km–500 km) around it.

3. Results and discussion

3.1. Site selection for the COALESCE project – illustrative example

It was required to identify eleven regionally representative $PM_{2.5}$ sampling sites across India under the aegis of the COALESCE project (Venkataraman et al., 2020). The weight-of-evidence approach discussed in the previous section was utilized to finalize the sampling sites (Fig. 3). Details of site selection are presented in the following sections.

3.1.1. Local information and meteorology for site selection

Each consortium institute of the COALESCE project identified several sites and performed a detailed evaluation using the local and meteorological information. Bhauri site proposed by IISER Bhopal is used to elucidate the details about the local and meteorological data collected for each site; details of all other sites are presented in supplemental text S4.

Bhauri (Bhopal) site is about 20 km from the city of Bhopal and has an average altitude of about 500 m above the mean sea level. The site is surrounded by grass, trees, pavements and some buildings within 200 m. The site does not have any continuous pollution source and is about 200 m away from a roadway with moderate traffic. The site is at least 500 m away from crops, barren land and loose soil, unpaved road and any domestic or agricultural burning sites. The site is also at least 500–1000 m away from other sources of pollution such as diesel generators, water treatment plant, and construction sites, making the site an ideal choice for ambient sampling. The site is easy to access throughout the year and has continuous availability of electricity. The site is adequately away from most of the local sources and is not downwind of any large point sources.

To assess the regional long-range transport characteristics, an offline version of the HYSPLIT model (Draxler and Hess, 1998) is used to obtain 5-day (120 h) daily back trajectories at arrival heights of 100, 500 and

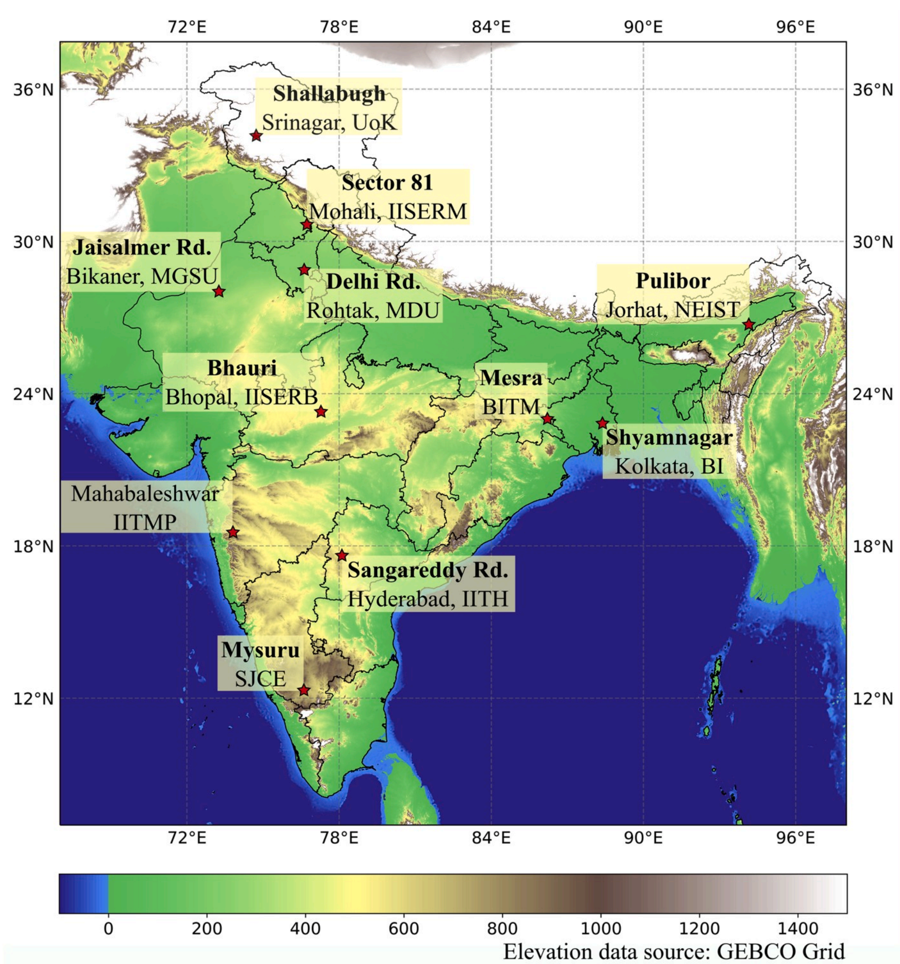


Fig. 3. Site locations identified for the COALESCE project (adapted from Venkataraman et al., 2020).

1000 m. A total of 1095 trajectories were clustered using *k*-means clustering into five clusters. The annual clustered back-trajectories are shown in Fig. 4. It shows that the clusters 2 and 4 (contributing 52% of the trajectories) originate over North-Western and North-Eastern India (Indo-Gangetic Plain (IGP)). Further, clusters 1 and 3 (contribute about 32% of the total trajectories) originate over the Arabian Sea and 17 per cent of the total trajectories originating over the Arabian Peninsula constitute cluster 5 (Fig. 4).

The clusters of trajectories received at the receptor site contribute differently during the different seasons, as shown in Fig. 5. Table 1 summarises the seasonal information provided by the trajectories. Cluster 2 and 4 originating over the Indian landmass, with few trajectories originating over the Arabian Peninsula and Northern Africa dominate during the winter season. Cluster 1 and 5 originating over the Arabian Sea and Arabian peninsula and cluster 4 originating over North-western India contribute to the majority of the trajectories during the pre-monsoon season. Cluster 1 and 3, both originating over the Arabian Sea contribute to the trajectories during the monsoon season, and both the local clusters, cluster 4 originating over northern India and parts of Pakistan and cluster 2 originating over the Indo-Gangetic Plain (IGP) contribute to the trajectories during the post-monsoon season. Both winter and post-monsoon season share similar long-range transport characteristics. This analysis helps identify sites that are likely to be influenced by sources located in various geographical locations, thus ensuring that the samples collected are regionally representative.

While back trajectory analysis helps understand the long-range transport of air parcel received at the site, meteorological information from the nearest weather station helps understand the local wind patterns. Indian Space Research Organisation's (ISRO) automatic weather station (AWS) station located at Samardha Village about 40 km North-East-East of Bhopal was used to obtain daily meteorological data (from 2011 to 2016, but with approximately 43% data missing in 2014). It was observed that most of the air parcels received at the site were from the NNE to SE direction (Fig. 6) suggesting sources located in this wind sector may affect the concentrations at the site. It can be concluded that local sources in the above directions may have greater influence as compared to source in other directions. It is however pertinent to note that there are no major point sources in the immediate vicinity of the chosen sampling site.

3.1.2. Surface $PM_{2.5}$ derived from satellite AOD

For the eleven sites identified based on criteria discussed in Section 2.1, their regional representativeness was assessed using high-resolution satellite-derived AOD to obtain surface $PM_{2.5}$ concentrations. The $PM_{2.5}$ data used is 1.5 km × 1.5 km spatial resolution (1 km × 1 km is regridded to 1.5 km × 1.5 km) daily data for the Indian landmass for 8 years (2004–2011). While the AOD data is available from 2000 onwards, the conversion factors used to convert AOD to $PM_{2.5}$ were only available from 2004 to 2011. Different metrics are used to check the spatial representativeness of the reference cell (X) with the neighbouring cell (Y). There is a possibility that the data is not available at all-time instants at both the cells simultaneously due to cloud cover and different satellite swath; such data is discarded and only data available at the same time instant for both the cells are used for computations. Details about stepwise download and processing of the data are explained in supplemental text S3.

Based on CoD and PCC metrics, most of the sites comply with the set criteria of CoD (<0.2 in the neighbourhood (at least 80–100 km on all sides of the RRMS) and PCC (>0.5 in the neighbourhood) for regional representativeness, except for Pulibor site at Jorhat. Results for Bhauri (Bhopal) site and Pulibor (Jorhat) site are used to illustrate the use of CoD, PCC and MI in assessing the regional representativeness. Results for all other sites are discussed in supplemental text S4.

Pulibor (Jorhat) site showed a small region where the CoD values are less than 0.2 while the PCC values were high over a vast region. The high CoD values imply spatial heterogeneity while high PCC values suggest

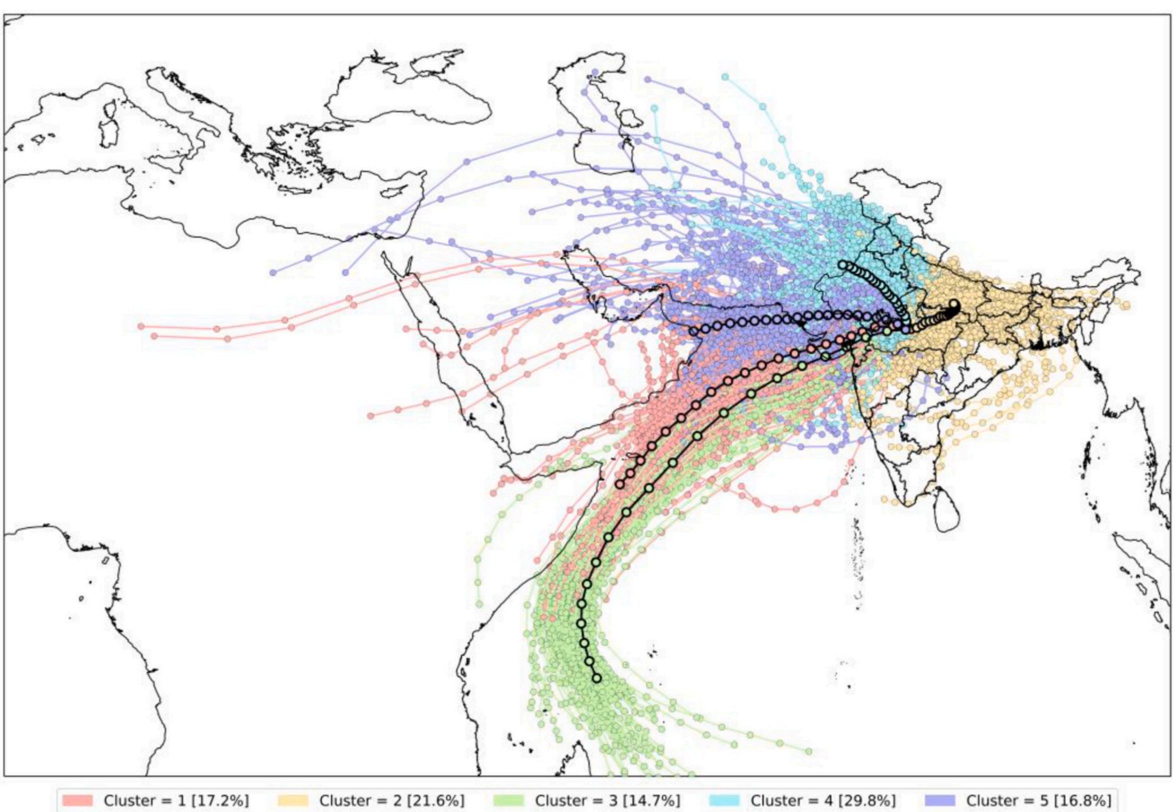


Fig. 4. Bhauri (Bhopal) site's annual HYSPLIT back trajectories for the year 2017, classified into five clusters using *k*-means clustering.

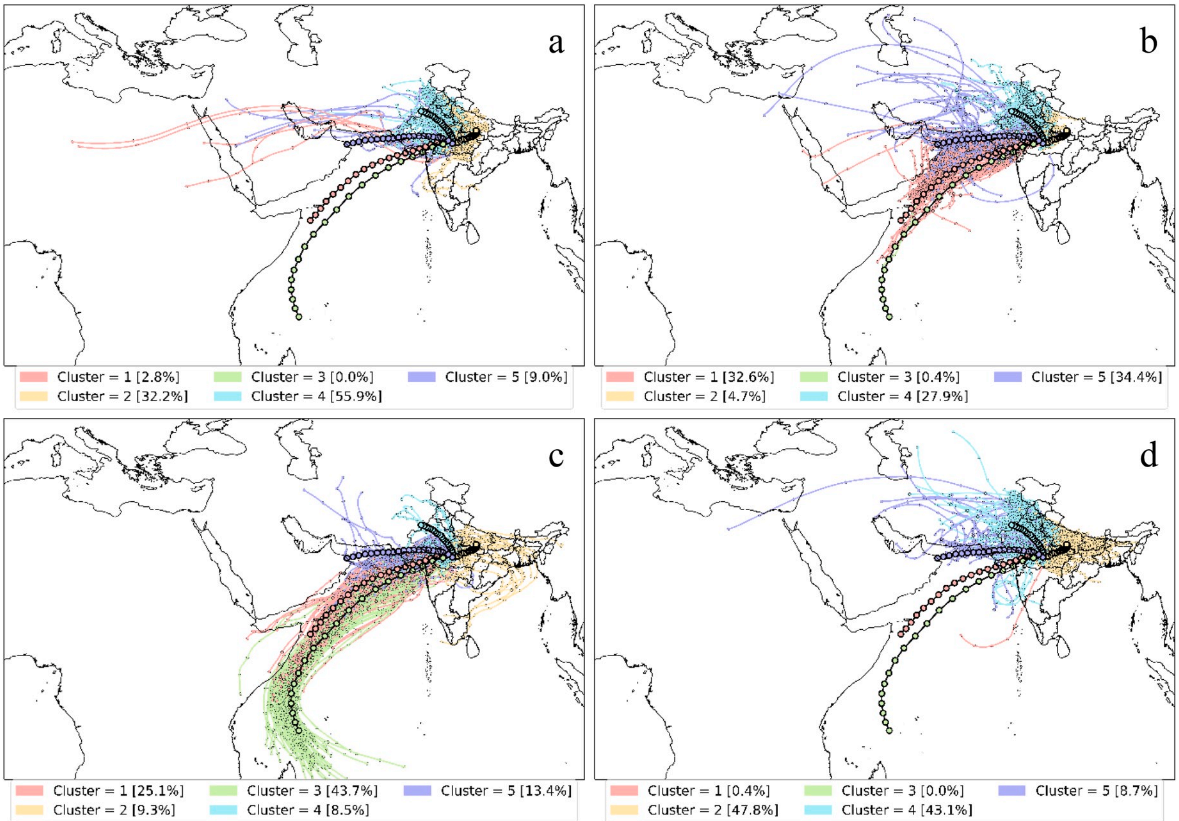


Fig. 5. Bhauri (Bhopal) site's seasonal HYSPLIT back trajectory for year 2017 **a.** Winter (JF), **b.** Pre-Monsoon (MAM), **c.** Monsoon (JJAS) and **d.** Post-Monsoon (OND).

Table 1
Bhauri (Bhopal) site's annual and seasonal cluster contribution in percent for the year 2017.

C#	Dominant Direction	Annual Contribution	Seasonal Contribution			
			Winter	Pre-Monsoon	Monsoon	Post-Monsoon
1	SW	17.2	2.8	**32.6**	**25.1**	0.4
2	E	21.6	**32.2**	4.7	9.3	**47.8**
3	SW	14.7	0.0	0.4	**43.7**	0.0
4	NW	29.8	**55.9**	27.9	8.5	**43.1**
5	W	16.8	9.0	**34.4**	13.4	8.7

good linear relationship temporally around the site as shown in Figs. 7 and 8, respectively. Pulibor (Jorhat) site does not qualify as a regionally representative site based on the CoD and PCC because of a small region (approx. 40 km in all directions) with CoD values less than 0.2 along with high PCC values in a larger region (approx. over 100 km in all direction) around the site. The use of MI captures the non-linear relationships in the data over time. The region with MI values more than 0.5 covers a vast area (approx. over 100 km in all direction) correspondingly, the region with MI values more than 0.25 covers most of North-East India (Fig. 9). Thus, it can be concluded that despite not meeting the CoD criteria, the site represents the region well based on PCC and MI and measurements at the sites can be used proxy for the region.

Bhauri (Bhopal) another proposed site exhibited both low CoD and high PCC in the site's neighbourhood as shown in Figs. 10 and 11, respectively and confirms to the criteria for good regional representativeness, which is also verified by the MI values (Fig. 12). Thus, for sites where the CoD and PCC fails to establish regional representativeness which essentially captures the linear relationship in the data, the sites/cells may be related non-linearly due to meteorology, topography and other atmospheric factors. These non-linear relationships are captured by MI. Good agreement of CoD, PCC and MI suggest the utility of MI to

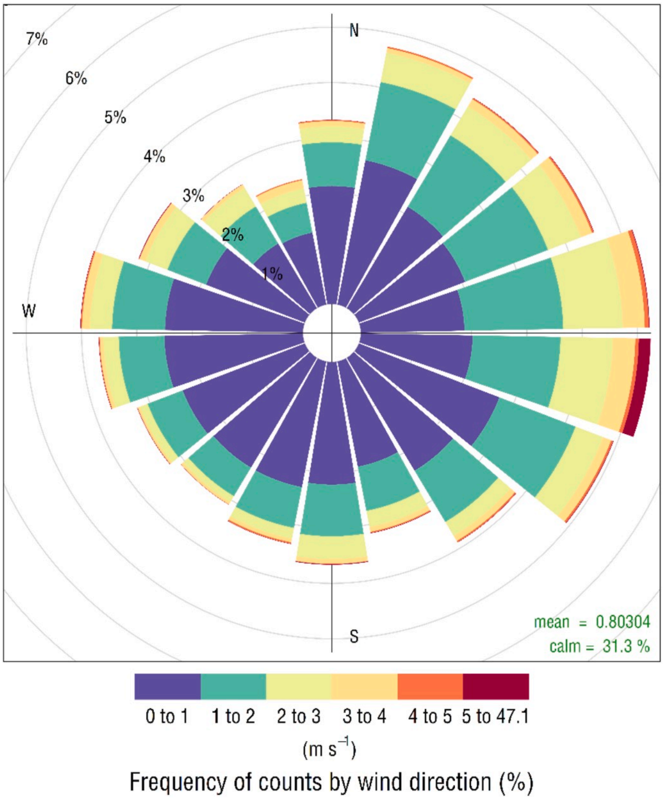


Fig. 6. Wind-rose plot for Bhauri (Bhopal) Site (2011–2016) (MOSDAC, 2020).

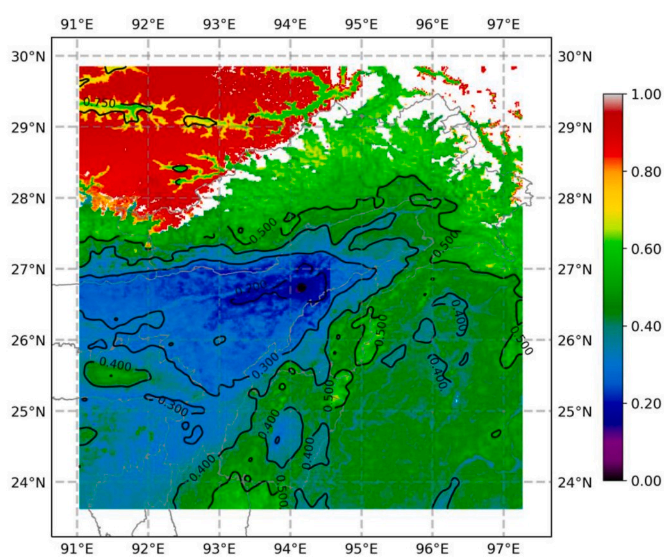


Fig. 7. $PM_{2.5}$ Coefficient of Divergence plot for Pulibor (Jorhat) site.

validate the results of CoD and PCC.

3.2. Summary of findings

The $PM_{2.5}$ sampling sites finalized for the COALESCE project utilising the methodology developed in this study are geographically well spread out across India, with Mahabaleshwar and Shallabaug (Kashmir) sites being high-altitude sites. These high-altitude sites help capture background regional aerosol concentrations. All the eleven sites were subjected to compliance checks with the multiple criteria for acceptance. It was found that most of the sites were away from the major sources such

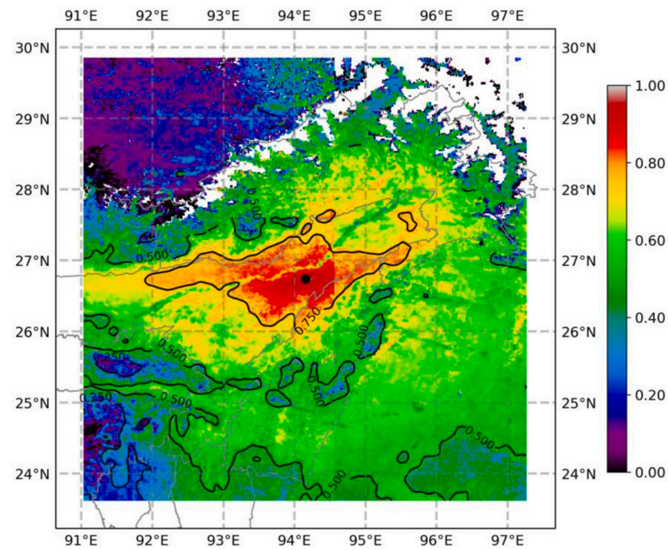


Fig. 8. $PM_{2.5}$ Pearson correlation coefficient plot for Pulibor (Jorhat) site.

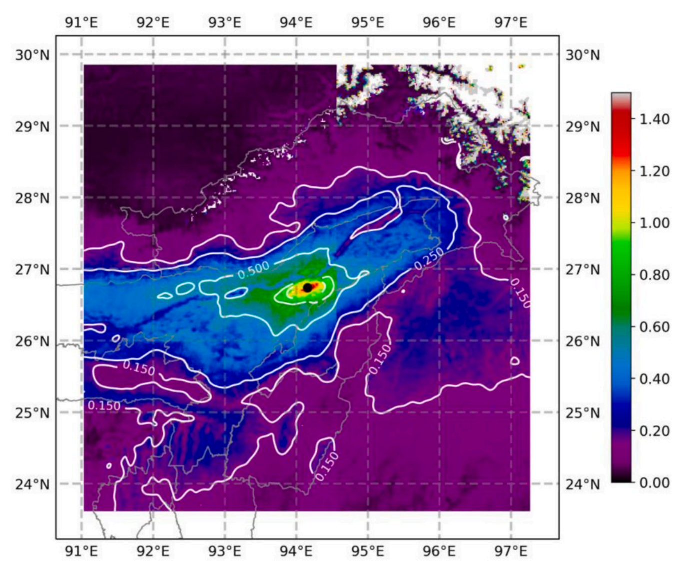


Fig. 9. $PM_{2.5}$ Mutual Information plot for Pulibor (Jorhat) site.

as traffic, agricultural field burning, domestic biomass burning and unpaved roads. For some sites such as Sangareddy Rd, Hyderabad and Mesra where these sources are within 200 m, the emissions from these sources were very small in magnitude and are unlikely to affect the concentrations at the site. Each site was also assessed for influence from the large point sources and all sites except Delhi Rd., Rohtak were not subject to such influences. The Delhi Rd., Rohtak site was found to be downwind of an industrial area constituting mainly plastic and films plants, sugar mills, cattle feed plant and other medium and small-scale industries. However, this site was selected both for logistical reasons and due to the fact that this region of the country is seldom uninfluenced by industrial centres. Additionally, clustered back-trajectory analysis performed for all the sites suggest that the aerosol transported across the Arabian Peninsula, Arabian sea, Bay of Bengal and by some fast-moving westerlies carrying air parcels from the Mediterranean Sea are likely to be captured by some sampling sites. Further, most of the sites except for Shallabaug (Kashmir), had a significant influence from trajectories that originate/pass over the Indian landmass, which is desirable to understand the influence of different regions on other sites. For most of the sites except, Pulibor (Jorhat) and Shallabaug (Kashmir), both CoD and

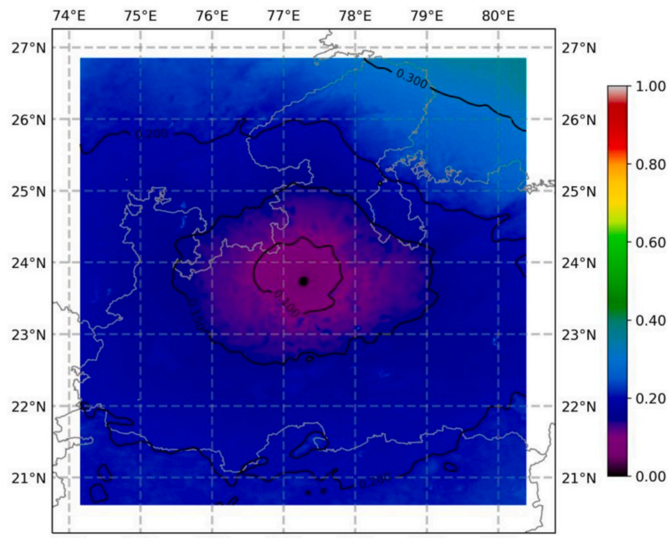


Fig. 10. $PM_{2.5}$ Coefficient of Divergence plot for Bhauri (Bhopal) site.

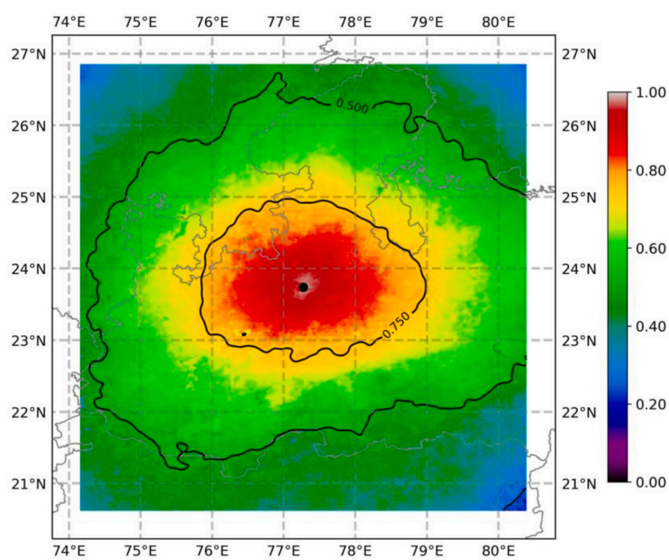


Fig. 11. $PM_{2.5}$ Pearson correlation coefficient plot for Bhauri (Bhopal) site.

PCC were useful to establish spatial representativeness of the site. Details about Pulibor (Jorhat) site were discussed in Section 3.1.2, while Shallabaug (Kashmir) site is affected mainly by the topography of the region around it. The site lies on the high-altitude plains trapped between the high hills and mountains in both the north and south directions. Logistics and socio-political conditions dictated the choice of this site as a sampling location. Despite making an exception, measurements made at this location will provide the first long term inputs on the air quality status in the Kashmir valley and will be useful in driving policy interventions for this region.

4. Discussions

The study focuses on setting up an air quality monitoring network which measures the $PM_{2.5}$ concentrations at regionally representative monitoring sites. It was proposed that regional representativeness of a site can be established using different metrics namely CoD, PCC and MI which enable comparison of a site with its neighbourhood. In this section, we discuss a few issues related to use of AOD-derived $PM_{2.5}$ data, quantifying representativeness and assessing adequacy of the overall

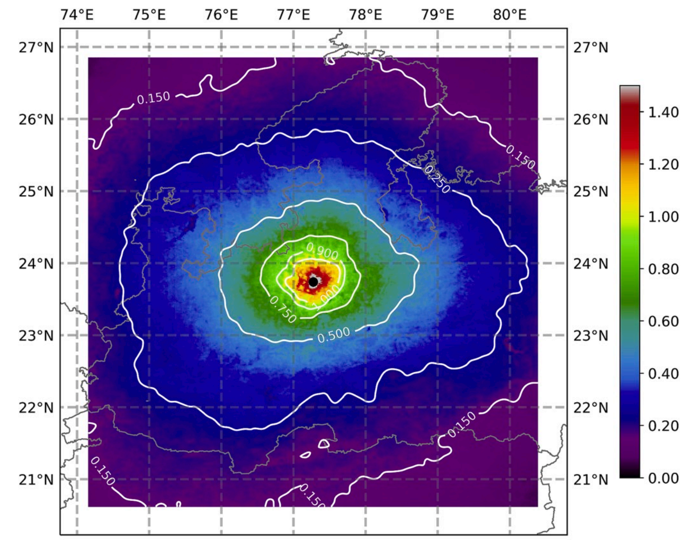


Fig. 12. $PM_{2.5}$ Mutual Information plot for Bhauri (Bhopal) site.

monitoring network.

4.1. AOD based analysis

Ideally, in-situ $PM_{2.5}$ measurements at high spatio-temporal resolution are desired for computing representativeness metrics (CoD, PCC and MI) proposed in this work but in the COALESCE study discussed in the current work such data was not available. Thus, $PM_{2.5}$ obtained from widely available satellite aerosol product-AOD, was used. This involved estimation of surface $PM_{2.5}$ from AOD using conversion factors developed from modelling studies (van Donkelaar et al., 2010). In literature, Chu et al. (2016) has reviewed various studies for predicting ground $PM_{2.5}$ concentration using satellite aerosol optical depth. While this conversion enabled analysis with AOD-derived $PM_{2.5}$, the conversion from AOD to $PM_{2.5}$ step introduces uncertainties in the analysis. Additionally, this conversion may not be possible if the conversion factors are not available. Thus, for the potential sites in the COALESCE project, we additionally performed the representativeness analysis directly with AOD data. The spatial plots for CoD, PCC and MI generated using AOD data are presented along with the corresponding plots obtained using AOD converted $PM_{2.5}$ data in Supplemental Figs. S4, S6, S8, S10, S12, S14, S16, S18, S20, and S22. The visual assessment of these figures leads to similar conclusions as that obtained from the analysis of the AOD converted $PM_{2.5}$ data.

4.2. Quantifying representativeness

In this work, generation of CoD, PCC and MI spatial maps for a potential site was proposed to visually assess the representativeness of the site. While these maps provide a visual and qualitative evaluation of representativeness, it may be desirable to come up with point values to assess the representativeness and compare across multiple potential sites. In this work, we propose use of distance weighted metrics for this purpose. To illustrate, consider CoD. For a given potential sampling site, the distance weighted metrics (X_{d^*}) can be computed as:

$$X_{d^*} = \sum_{i,j \in N(S)} \frac{X_{i,j}}{d_{i,j}^2}$$

where.

$d_{i,j}$ is the distance of the cell (i,j) from the site,

$N(S)$ is the neighbourhood of site S consisting of all the cell (i, j) whose distance $d_{i,j} \leq d^*$,
$d^* = 100$, 200 and 300 km is used in the current work.

Such distance weighted metrics were computed for CoD, PCC and MI for all the eleven sites in the COALESCE project for 100 km, 200 km and 300 km distances (also see supplemental text S5). The resulting values are presented in Supplemental Table S14.

To facilitate comparison across various sites, the MI, PCC and CoD values were scaled as follows:

CoD	PCC	MI
$\frac{CoD_{S_i} - \min_j(CoD_{S_j})}{CoD_{S_i}}$	$\frac{PCC_{S_i}}{\max_j(PCC_{S_j})}$	$\frac{MI_{S_i}}{\max_j(MI_{S_j})}$

The scaled values for different sites of the COALESCE project are presented in Table 2. It can be seen that almost all the sites have high MI and PCC values along with low CoD values (Table 2), thereby indicating that these sites are regionally representative when compared to the representativeness of the best site in the network.

However, for Srinagar and Jorhat sites, it is seen that MI (both sites) and PCC (Srinagar) are on the lower side while CoD is on the higher side of the threshold values (Table 2). This observation indicates relative lack of representativeness of these sites which may be because of their topography as discussed in Sections 3.1.2 and 3.2, respectively. However, these sites were chosen taking into account the feasibility and logistics of operating a sampling site in these areas.

4.3. Adequacy of the overall monitoring network

In any monitoring network, it is essential to strike a balance between the number of sites and the spatial coverage of the network. Network with a higher number of sites ensures a broader spatial coverage but adds to the cost of the installation and operation without really generating additional meaningful information due to high redundancy. In contrast, network with sparse sites may be low on operational cost but may not be able to detect critical regional contributing sources. In the current work, while we focussed on testing adequacy of an individual site, it may be desirable to be able to comment on the adequacy of the overall monitoring network. Towards this end, we propose comparison of the selected sites in a pair-wise manner using the metrics CoD, PCC and MI, which were used to establish regional representativeness of an individual site. A pair-wise comparison across sites helps highlight the similarities and differences between different sites. In particular, if the CoD values are very high (close to 1), and PCC and MI values are very low, the sites are very dissimilar and the number of sites in the network may not be adequate. On the other hand, low CoD and high PCC and MI values for site pairs may indicate redundancy in the network.

Fig. 13 shows a pair-wise comparison of the eleven sites using CoD, PCC and MI metrics for the COALESCE project. All site-pairs have CoD values greater than 0.2 (Fig. 13a) indicating that the values at these sites are not similar. Similarly, PCC plot (Fig. 13b) shows that most of the values lie in [-0.5, 0.5] band indicating the absence of a strong linear relationship in the values across the sites. The weak linear relationship that exists may be attributed to the seasonal variations in the $PM_{2.5}$ values across the region. The MI plot (Fig. 13 c) shows low values for all site pairs except for Delhi Rd. (Rohtak) site and Sector 81 (Mohali) site which may be due to their proximity to each other (they are about 240 km apart). However, based on the overall behaviour of CoD and PCC it is concluded that there is no redundant site in the network and hence all the sites were retained as part of the network. Additionally, Jorhat's Pulibor site has high CoD values, low magnitude of PCC values and low MI values with other sites which indicates that the site is significantly different from other sites. They may be attributed to the significantly different topography of this region as compared to other sites. This analysis indicates that additional sites in this is region could be investigated but this was not possible in the COALESCE project due to logistical constraints.

5. Conclusions

Identification of appropriate sites is an essential part of setting up an air quality monitoring network. While the approaches available in literature use short term sampling at different candidate sites, long term characteristics of the sites may vary based on the seasonal differences and the frequency of occurrence of episodic events. This study has demonstrated the usefulness of a weight-of-evidence approach in identifying sites for the set-up of a regional $PM_{2.5}$ network. With advancing technologies, satellite data is readily available for long time-durations at high spatial resolution compared to what can be obtained from ground-based $PM_{2.5}$ measurements. Various metrics (both capturing linear and nonlinear relationships) have been proposed in this work to assess the site representativeness based on the satellite-derived measurements. Use of satellite-derived $PM_{2.5}$ measurements enable saving precious manpower and financial resources for site identification. With air quality standards becoming more stringent across the world, the approach proposed in this work can be used to design and operationalize regional $PM_{2.5}$ monitoring networks. This approach is especially valuable in countries where background $PM_{2.5}$ measurements may not be

Table 2
Site-wise distance weighted MI, CoD and PCC values for 100, 200 & 300 km (scaled).

Site	Mutual Information			Coef. Of Divergence			Pearson Corr. Coef.		
	100 km	200 km	300 km	100 km	200 km	300 km	100 km	200 km	300 km
Sangareddy Rd., Hyderabad	0.75	0.75	0.75	0.06	0.06	0.05	0.96	0.96	0.95
Mahabaleshwar	0.62	0.61	0.61	0.15	0.17	0.17	0.92	0.90	0.89
Delhi Rd., Rohtak	0.90	0.89	0.89	0.13	0.17	0.19	0.97	0.96	0.95
Sector 81, Mohali	0.86	0.85	0.85	0.18	0.25	0.29	0.94	0.91	0.90
Jaisalmer Rd., Bikaner	0.92	0.92	0.91	0.03	0.06	0.07	0.97	0.96	0.95
Shallabaug, Srinagar	0.56	0.55	0.55	0.43	0.48	0.49	0.87	0.84	0.83
Shyamnagar, Kolkata	**1.00**	**1.00**	**1.00**	0.11	0.16	0.19	**1.00**	**1.00**	**1.00**
Pulibor, Jorhat	0.66	0.65	0.66	0.34	0.38	0.40	0.94	0.92	0.92
Mesra	0.73	0.73	0.72	0.10	0.12	0.13	0.96	0.95	0.94
Bhauri, Bhopal	0.82	0.82	0.82	**0.00**	**0.00**	**0.00**	0.99	0.99	0.98
Mysuru	0.65	0.64	0.64	0.12	0.15	0.15	0.93	0.91	0.89

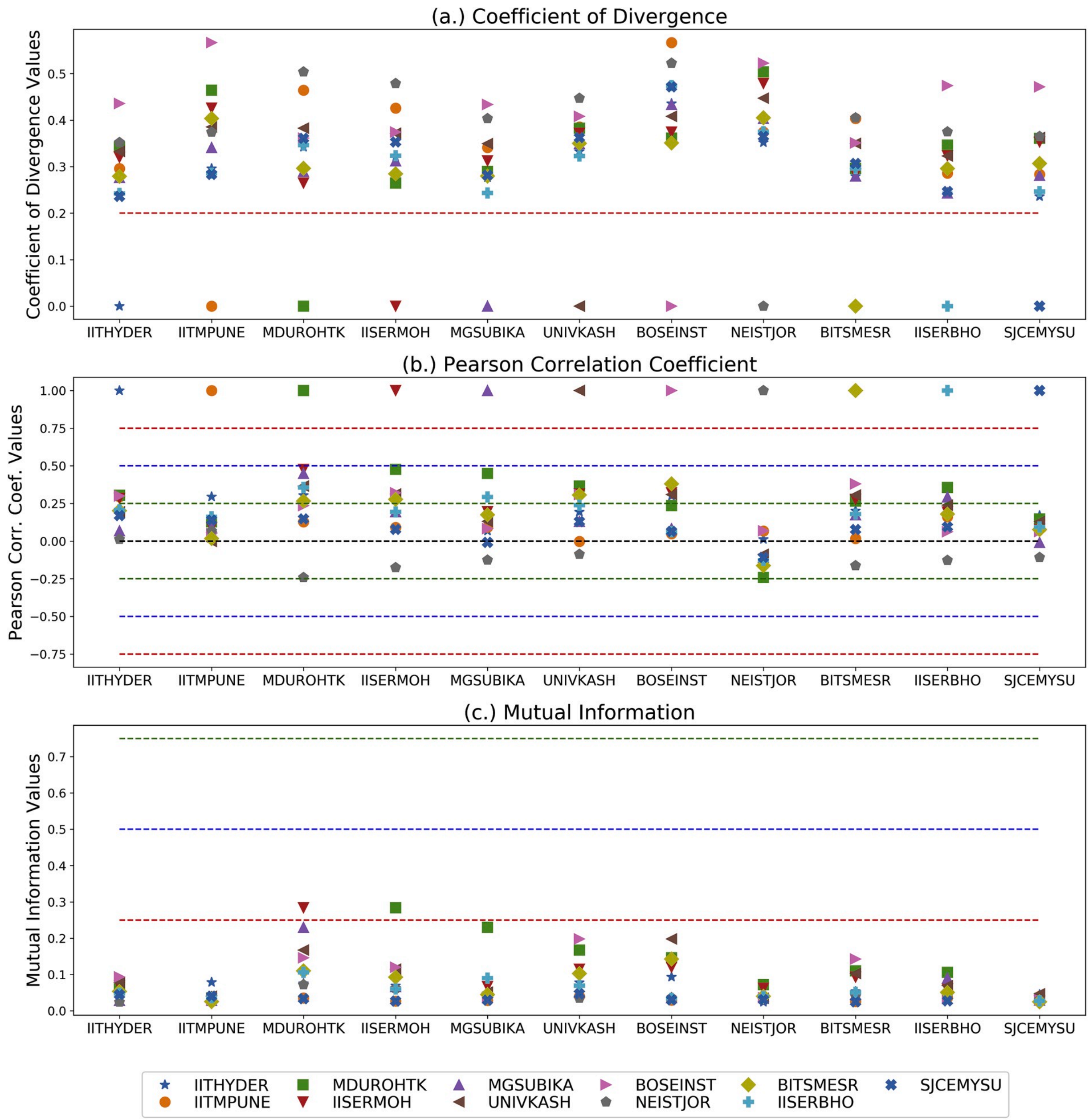


Fig. 13. Pair-wise CoD, PCC and MI plots for different sites (a) CoD plot, (b) PCC plot and (c) MI plot.

available at reasonable temporal and spatial resolutions, to assist with sampling network design.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

CRediT authorship contribution statement

Nirav L. Lekinwala: Data curation, Formal analysis, Visualization, Writing - original draft. **Ankur Bhardwaj:** Data curation, Formal analysis, Writing - original draft. **Ramya Sunder Raman:** Conceptualization, Methodology, Validation, Writing - review & editing. **Mani Bhushan:** Investigation, Validation, Supervision, Writing - review & editing. **Kunal Bali:** Data curation, Formal analysis. **Sagnik Dey:** Data curation, Resources, Supervision.

Acknowledgements

This work was supported by the Ministry of Environment, Forest and Climate Change (MoF&CC), Government of India, under the NCAP-COALESCE project {Grant No.14/10/2014-CC(Vol.II)}. The authors thank the internal review committee of the NCAP-COALESCE project for their comments and suggestions on this paper. The views expressed in this document are solely those of authors and do not necessarily reflect those of the Ministry. The Ministry does not endorse any products or commercial services mentioned in this publication. The authors thank the COALESCE consortium partners (https://ncapcoalesce.iitb.ac.in/partner-institutes/) for sharing the local information and other information about the site used in the study. The authors also gratefully acknowledge the NOAA Air Resources Laboratory (ARL) for providing the HYSPLIT transport and dispersion model (https://ready.arl.noaa.gov/HYSPLIT.php) used in this publication. Sagnik Dey acknowledges funding from IIT Delhi for Institute Chair position and DST-FIST program for computing support at IIT Delhi.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.atmosenv.2020.117544.

References

Balakrishnan, K., Sambandam, S., Ramaswamy, P., Ghosh, S., Venkatesan, V., Thangavel, G., Mukhopadhyay, K., Johnson, P., Paul, S., Puttaswamy, N., Dhaliwal, R.S., Shukla, D.K., 2015. Establishing integrated rural–urban cohorts to assess air pollution-related health effects in pregnant women, children and adults in Southern India: an overview of objectives, design and methods in the Tamil Nadu Air Pollution and Health Effects (TAPHE) s. BMJ Open 5 (6). https://doi.org/10.1136/bmjopen-2015-008090 e008090.

Blanchard, C.L., Carr, E.L., Collins, J.F., Smith, T.B., Lehrman, D.E., Michaels, H.M., 1999. Spatial representativeness and scales of transport during the 1995 integrated monitoring study in California's San Joaquin Valley. Atmos. Environ. 33 (29), 4775–4786. https://doi.org/10.1016/S1352-2310(99)00284-8.

Chow, J.C., Engelbrecht, J.P., Watson, J.G., Wilson, W.E., Frank, N.H., Zhu, T., 2002. Designing monitoring networks to represent outdoor human exposure. Chemosphere 49 (9), 961–978. https://doi.org/10.1016/S0045-6535(02)00239-4.

Chu, Y., Liu, Y., Li, X., Liu, Z., Lu, H., Lu, Y., Liu, F., 2016. A review on predicting ground PM2.5 concentration using satellite aerosol optical depth. Atmosphere 7 (10), 129. https://doi.org/10.3390/atmos7100129.

Cover, T.M., Thomas, J.A., 1991. Elements of Information Theory, ed. Wiley, ISBN 978-0-471-24195-9.

Delcloo, A.W., De Backer, H., 2008. Five day 3D back trajectory clusters and trends analysis of the Uccle ozone sounding time series in the lower troposphere (1969–2001). Atmos. Environ. 42 (19), 4419–4432. https://doi.org/10.1016/j.atmosenv.2008.01.072.

Dockery, D.W., Pope, C.A., Xu, X., Spengler, J.D., Ware, J.H., Fay, M.E., Ferris, B.G., Speizer, F.E., 1993. An association between air pollution and Mortality in Six U.S. Cities. N. Engl. J. Med. 329 (24), 1753–1759. https://doi.org/10.1056/NEJM199312093292401.

Draxler, R.R., Hess, G.D., 1998. An overview of the HYSPLIT_4 modelling system for trajectories, dispersion and deposition. Aust. Meteorol. Mag. 47 (4), 295–308.

Han, Y.-J., Holsen, T.M., Hopke, P.K., Yi, S.-M., 2005. Comparison between back-trajectory based modeling and Lagrangian Backward dispersion modeling for locating sources of Reactive Gaseous Mercury. Environmental Science & Technology 39 (6), 1715–1723. https://doi.org/10.1021/es0498540.

Hansen, J., Sato, M., Ruedy, R., Lacis, A., Oinas, V., 2000. Global warming in the twenty-first century: an alternative scenario. Proc. Natl. Acad. Sci. Unit. States Am. 97 (18), 9875–9880. https://doi.org/10.1073/pnas.170278997.

Just, A.C., Wright, R.O., Schwartz, J., Coull, B.A., Baccarelli, A.A., Tellez-Rojo, M.M., Moody, E., Wang, Y., Lyapustin, A., Kloog, I., 2015. Using high-resolution satellite aerosol optical depth to estimate daily PM2.5 geographical Distribution in Mexico city. Environmental Science & Technology 49 (14), 8576–8584. https://doi.org/10.1021/acs.est.5b00859.

Kim, E., Hopke, P.K., Pinto, J.P., Wilson, W.E., 2005. Spatial variability of fine particle mass, components, and source contributions during the Regional Air Pollution Study in St. Louis. Environ. Sci. Technol. 39 (11), 4172–4179. https://doi.org/10.1021/es049824x.

Krudysz, M.A., Froines, J.R., Fine, P.M., Sioutas, C., 2008. Intra-community spatial variation of size-fractionated PM mass, OC, EC, and trace elements in the Long Beach, CA area. Atmos. Environ. 42 (21), 5374–5389. https://doi.org/10.1016/j.atmosenv.2008.02.060.

Krudysz, M.A., Dutton, S.J., Brinkman, G.L., Hannigan, M.P., Fine, P.M., Sioutas, C., & Froines, J.R., 2009. Intra-community spatial variation of size-fractionated organic compounds in Long Beach, California. Air Quality, Atmosphere & Health 2 (2), 69–88. https://doi.org/10.1007/s11869-009-0035-1.

Lee, M., Kloog, I., Chudnovsky, A., Lyapustin, A., Wang, Y., Melly, S., Coull, B., Koutrakis, P., Schwartz, J., 2016. Spatiotemporal prediction of fine particulate matter using high-resolution satellite images in the Southeastern US 2003-2011. J. Expo. Sci. Environ. Epidemiol. 26 (4), 377–384. https://doi.org/10.1038/jes.2015.41.

Liu, Y., Park, R.J., Jacob, D.J., Li, Q., Kilaru, V., Sarnat, J.A., 2004. Mapping annual mean ground-level PM2.5 concentrations using Multiangle Imaging Spectroradiometer aerosol optical thickness over the contiguous United States. J. Geophys. Res.: Atmospheres 109 (D22). https://doi.org/10.1029/2004JD005025.

Lyapustin, A., Wang, Y., Laszlo, I., Kahn, R., Korkin, S., Remer, L., Levy, R., Reid, J.S., 2011. Multiangle implementation of atmospheric correction (MAIAC): 2. Aerosol algorithm. J. Geophys. Res.: Atmospheres 116 (D3). https://doi.org/10.1029/2010JD014986.

Lyapustin, Alexei, Wang, Y., Korkin, S., Huang, D., 2018. MODIS collection 6 MAIAC algorithm. Atmospheric Measurement Techniques Discussions 1–50. https://doi.org/10.5194/amt-2018-141.

Occhipinti, C., Aneja, V.P., Showers, W., Niyogi, D., 2008. Back-trajectory analysis and source-receptor relationships: particulate matter and Nitrogen Isotopic Composition in Rainwater. J. Air Waste Manag. Assoc. 58 (9), 1215–1222. https://doi.org/10.3155/1047-3289.58.9.1215.

Pachauri, Aparna, Singla, Vyoma, Lakhani, Anita, Kumari, K., Maharaj, T., S, 2013. Characteristics and sources of carbonaceous aerosols in PM2.5 during wintertime in Agra, India. Aerosol and Air Quality Research 13 (3), 977–991. https://doi.org/10.4209/aaqr.2012.10.0263.

Pant, P., Lal, R.M., Guttikunda, S.K., Russell, A.G., Nagpure, A.S., Ramaswami, A., Peltier, R.E., 2019. Monitoring particulate matter in India: recent trends and future outlook. Air Quality, Atmosphere & Health 12 (1), 45–58. https://doi.org/10.1007/s11869-018-0629-6.

Shah, A.S.V., Langrish, J.P., Nair, H., McAllister, D.A., Hunter, A.L., Donaldson, K., Newby, D.E., Mills, N.L., 2013. Global association of air pollution and heart failure: a systematic review and meta-analysis. Lancet 382 (9897), 1039–1048. https://doi.org/10.1016/S0140-6736(13)60898-3.

Shenfeld, L., 1970. Meteorological aspects of air pollution control. Atmosphere 8 (1), 3–13. https://doi.org/10.1080/00046973.1970.9676578.

Siudek, P., Frankowski, M., 2017. The effect of sources and air mass transport on the variability of trace element deposition in central Poland: a cluster-based approach. Environ. Sci. Pollut. Control Ser. 24 (29), 23026–23038. https://doi.org/10.1007/s11356-017-9932-2.

Steuer, R., Kurths, J., Daub, C.O., Weise, J., Selbig, J., 2002. The mutual information: detecting and evaluating dependencies between variables. Bioinformatics 18 (2), S231–S240. https://doi.org/10.1093/bioinformatics/18.suppl_2.S231.

Stohl, A., Haimberger, L., Scheele, M.P., Wernli, H., 2001. An intercomparison of results from three trajectory models. Meteorol. Appl. 8 (2), 127–135. https://doi.org/10.1017/S1350482701002018.

Stohl, Andreas, Eckhardt, S., Forster, C., James, P., Spichtinger, N., Seibert, P., 2002. A replacement for simple back trajectory calculations in the interpretation of atmospheric trace substance measurements. Atmos. Environ. 36 (29), 4635–4648. https://doi.org/10.1016/S1352-2310(02)00416-8.

van Donkelaar, A., V, M.R., Michael, B., Ralph, K., Robert, L., Carolyn, V., J, V.P., 2010. Global estimates of ambient fine particulate matter concentrations from satellite-based aerosol optical depth: development and application. Environmental Health Perspectives 118 (6), 847–855. https://doi.org/10.1289/ehp.0901623.

van Donkelaar, A., V, M.R., Michael, B., L, B.B., 2015. Use of satellite observations for long-term exposure assessment of global concentrations of fine particulate matter. Environmental Health Perspectives 123 (2), 135–143. https://doi.org/10.1289/ehp.1408646.

Venkataraman, C., Bhushan, M., Dey, S., Ganguly, D., Gupta, T., Habib, G., Kesarkar, A., Phuleria, H., Sunder Raman, R., 2020. Indian network project on carbonaceous aerosol emissions, source apportionment and climate Impacts (COALESCE). Bull. Am. Meteorol. Soc. https://doi.org/10.1175/BAMS-D-19-0030.1. Accepted.

Wilson, J.G., Kingham, S., Pearce, J., Sturman, A.P., 2005. A review of intraurban variations in particulate air pollution: implications for epidemiological research. Atmos. Environ. 39 (34), 6444–6462. https://doi.org/10.1016/j.atmosenv.2005.07.030.

Wongphatarakul, V., Friedlander, S.K., Pinto, J.P., 1998. A comparative study of $PM_{2.5}$ ambient aerosol chemical databases. Environ. Sci. Technol. 32 (24), 3926–3934. https://doi.org/10.1021/es9800582.

Xie, M., Coons, T.L., Dutton, S.J., Milford, J.B., Miller, S.L., Peel, J.L., Vedal, S., Hannigan, M.P., 2012. Intra-urban spatial variability of $PM_{2.5}$-bound carbonaceous components. Atmos. Environ. 60, 486–494. https://doi.org/10.1016/j.atmosenv.2012.05.041.

References

CPCB, 2014. National air quality Index. Control of urban pollution series CUPS 82/2014–15. https://app.cpcbccr.com/ccr/#/caaqm-dashboard/caaqm-landing last accessed November 29 2019.

Improve, 2016. SOP 126, Site Selection. Colorado State University, Boulder, CO. http://vista.cira.colostate.edu/Improve/wp-content/uploads/2016/04/SOP-126-Site-Selection.pdf. accessed May 29, 2018.

Meteorological & Oceanographic satellite data archival centre (MOSDAC) data for Samradha Village. https://www.mosdac.gov.in, last accessed January 03 2020.

National Aeronautics, Space Administration, 2000. Terra mission handbook. retrieved from. https://eospso.gsfc.nasa.gov/sites/default/files/mission_handbooks/Terra.pdf.

National Aeronautics, Space Administration, 2002. Aqua mission handbook. retrieved from. https://eospso.gsfc.nasa.gov/sites/default/files/mission_handbooks/Aqua.pdf.

U.S. Environmental Protection Agency (EPA), 1999. Particulate Matter ($PM_{2.5}$) Speciation Guidance Final Draft, first ed. US EPA Monitoring and Quality Assurance Group. Emission, Monitoring, and Analysis Division, Office of Air Quality Planning and Standards, Research Triangle Park, NC, USA https://www.epa.gov/sites/production/files/2017-01/documents/final_draft_pm2.5_speciation_guidance_1999.pdf.