# Source details

## Scientific Reports

*Open Access*

Scopus coverage years: from 2011 to Present

Publisher: Springer Nature

ISSN: 2045-2322

Subject area: Multidisciplinary

Source type: Journal

CiteScore 2021
6.9

SJR 2021
1.005

SNIP 2021
1.389

View all documents > | Set document alert | Save to source list | Source Homepage

CiteScore | CiteScore rank & trend | Scopus content coverage

**Improved CiteScore methodology**

CiteScore 2021 counts the citations received in 2018-2021 to articles, reviews, conference papers, book chapters and data papers published in 2018-2021, and divides this by the number of publications published in 2018-2021. Learn more >

### CiteScore 2021

$$6.9 = \frac{564{,}351 \text{ Citations 2018 - 2021}}{81{,}511 \text{ Documents 2018 - 2021}}$$

Calculated on 05 May, 2022

### CiteScoreTracker 2022

$$7.5 = \frac{643{,}744 \text{ Citations to date}}{86{,}191 \text{ Documents to date}}$$

Last updated on 05 April, 2023 • Updated monthly

### CiteScore rank 2021

| Category | Rank | Percentile |
|---|---|---|
| Multidisciplinary<br>Multidisciplinary | #11/120 | 91st |

View CiteScore methodology > | CiteScore FAQ > | Add CiteScore to your site

# Source details

## Scientific Reports

*Open Access*

Scopus coverage years: from 2011 to Present

Publisher: Springer Nature

ISSN: 2045-2322

Subject area: Multidisciplinary

Source type: Journal

CiteScore 2021
6.9

SJR 2021
1.005

SNIP 2021
1.389

View all documents > | Set document alert | Save to source list | Source Homepage

CiteScore | CiteScore rank & trend | Scopus content coverage

Export content for category

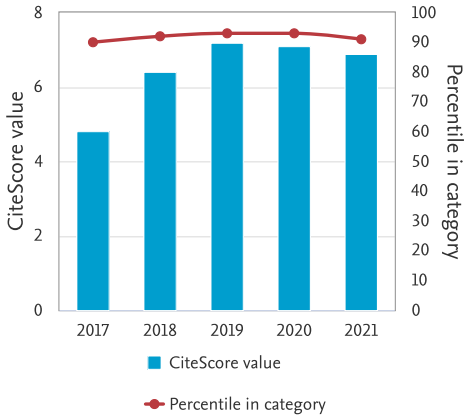
## CiteScore rank 2021

In category: Multidisciplinary

#11/120 Scientific Reports 6.9 91st percentile

| Rank | Source title | CiteScore 2021 | Percentile |
|---|---|---|---|
| #1 | Nature | 70.2 | 99th percentile |
| #2 | Science | 57.8 | 98th percentile |
| #3 | Nature Communications | 23.2 | 97th percentile |
| #4 | National Science Review | 19.1 | 97th percentile |
| #5 | Science advances | 18.5 | 96th percentile |
| #6 | Proceedings of the National Academy of Sciences of the United States of America | 18.1 | 95th percentile |
| #7 | Science Bulletin | 17.2 | 94th percentile |
| #8 | Journal of Advanced Research | 17.1 | 93rd percentile |
| #9 | Research | 10.5 | 92nd percentile |
| #10 | The Innovation | 10.0 | 92nd percentile |
| #11 | Scientific Reports | 6.9 | 91st percentile |
| #12 | Tsinghua Science and Technology | 5.8 | 90th percentile |

## CiteScore trend



scientific reports

Check for updates

OPEN

# An integrated analysis of air pollution and meteorological conditions in Jakarta

Teny Handhayani

**Air pollution and climate change are general problems for society. This paper proposes an integrated analysis of the Air Quality Index (AQI) and meteorological conditions in Jakarta. The column-based data integration model is applied to create integrated data of the Air Quality Index and meteorological conditions. The integrated data is then used to generate a causal graph using the PC algorithm. The causal graph reveals that there exist causal relationships between pollutants and meteorological conditions, e.g, humidity, rainfall, wind speed, and duration of sunshine affect particulate matter 10 ($PM_{10}$); wind speed affects sulfur dioxide ($SO_2$); temperature affects ozone ($O_3$). The historical data records that the average wind speed is decreased and the number of unhealthy days has risen. Ozone and particulate matter are two pollutants that mainly influence poor air quality in Jakarta. The integrated data is also used to train Long Short-Term Memory (LSTM) and Gated Recurrent Unit (GRU) for forecasting. Experimental results show that LSTM using integrated data produces smaller errors for forecasting AQI and meteorological conditions.**

Poor air quality is dangerous to civilization and the environment[1–6]. Air pollution is the largest cause of non-communicable diseases in some countries and regions, for instance, in Southeast Asia[7]. In 2020, the government of Republic Indonesia establishes that air quality is measured from the concentration of 7 parameters: particulate matter 10 ($PM_{10}$), particulate matter 2.5 ($PM_{2.5}$), sulfur dioxide ($SO_2$), nitrogen dioxide ($NO_2$), carbon monoxide (CO), ozone ($O_3$), and hydrocarbons (HC).

Air pollution might have a linkage to meteorological conditions or vice versa. Some research has been conducted to analyze the linkage of air pollutants and meteorological conditions[8–13]. A study in Taiwan reveals that temperature was associated with the incidence of CO poisoning[14]. A Bayesian Network graphical model has been used to analyze the statistical dependencies between environmental parameters, air pollution variables, and health data[15]. A study found that the maximum aerosol optical depth (AOD) in Palangka Raya, Pontianak, and Jambi happened in the dry season from July to October[16].

The historical data of Air Quality Index (AQI) and meteorological conditions in Jakarta record some important information[17,18]. The increasing number of unhealthy days happened from 2010–2013, 2015–2018, and 2020–2021. It is understandable that in 2020, the air quality was getting better because of the limited activities during the Covid-19 outbreak. However, the number of unhealthy days raised in 2021. From 2010 to 2021, the number of unhealthy days is always higher than healthy days. This is an early warning for the society that poor air quality might worsen if it is not managed properly. The average temperature slightly increased around 0.55°C from 2013 to 2019 and the average wind speed decreased.

Correlation measures a relationship between variables. However, correlation does not imply causation[19]. It means that statistical properties alone do not determine causal structures. The causal learning methods are enable to analyze the dependence structures among variables. A study has been conducted to observe the performance of learning algorithms to learn Bayesian network structures from climate data[20]. Some studies have been done to analyze the causal effects between pollution and health. A research has revealed the causal effects between local air pollution on daily deaths[21]. Gaussian process model and information geometric causal inference criterion have been implemented to obtain the correct causal directions between air pollutants[22]. A causal inference approach named Total Events Avoided (TEA) has been used for evaluating the health impacts of an air pollution regulation[23].

Analyzing the trend of air pollution is beneficial for the government and society to find the important factors that contribute to air quality. This research is conducted to study the causal relationships between air pollutants and meteorological conditions in Jakarta. The problem of this research is how to analyze the causal effect of air pollution and meteorological conditions in Jakarta. This paper proposes an integrated analysis of AQI and

Fakultas Teknologi Informasi, Universitas Tarumanagara, Jakarta, Indonesia. email: tenyh@fti.untar.ac.id

meteorological conditions using a causal learning approach. It implements the PC algorithm to generate a causal graph from a dataset. The causal graph is then used to analyze the cause and effect relationships among variables. The proposed method is useful to analyze the linkage of air pollution and meteorological conditions in Jakarta. The integrated data is also applied to train models for forecasting. This paper implements Long Short-Term Memory (LSTM) and Gated Recurrent Unit (GRU) to forecast AQI and meteorological conditions. The research contribution is an integration model to analyze the dependency relationships among variables and prediction of the future values of AQI and meteorological conditions, the case study in Jakarta.

## Methods

**Air Quality Index (AQI).** The Ministry of Environment and Forestry in the Republic of Indonesia measures the Air Quality Index (AQI) using equation (1), where $I$, $I_a$, $I_b$, $L_a$, $L_b$, and $L_x$ represent AQI score, upper limit AQI, lower limit AQI, upper limit ambient concentration, lower limit ambient concentration, and measurement results of real ambient concentration, respectively[24]. Air Quality Index (AQI) standard values are categorized as good (1–50), moderate (51–100), unhealthy (101–200), very unhealthy (201–300), and hazardous (≥301).

$$I = \frac{I_a - I_b}{L_a - L_b}(L_x - L_b) + I_b \tag{1}$$

**PC algorithm.** A causal graph is a graphical model that represents cause and effect relationships among variables. Assume that causal information between variables can be represented by a directed acyclic graph (DAG) where the nodes represent random variables and the edges represent direct causal effects[25–28]. Each causal DAG implies a set of conditional independence relationships[25]. A simple graph $A \rightarrow B$ (i.e., $A$ is a parent of $B$) represents that A is a direct cause of B. $A$ is a (possibly indirect) cause of $B$ only if there is a directed path from $A$ to $B$ ($A$ is an ancestor of $B$). One of the algorithms for learning a causal graph from a dataset is the PC algorithm[26,28,29].

The PC algorithm applies conditional independence tests to generate a causal graph from a dataset[26]. Suppose $E$, $\hat{\rho}$, $\alpha$ , $n$, and $\Phi(.)$ denotes the separation set, the partial correlation, the significance level, the number of samples, and the cumulative distribution function (cdf) of $\mathcal{N}(0, 1)$, respectively. An equation (2) can be used to compute a conditional independence test for Gaussian data[30,31]. It tests a question 'is a variable $D_u$ conditionally independent $D_v$ of given $D_E$?'

$$D_u \perp D_v \mid D_E \Leftrightarrow \sqrt{n - \mid E \mid -3}\left|\frac{1}{2}\log\left(\frac{1+\hat{\rho}_{uv|E}}{1-\hat{\rho}_{uv|E}}\right)\right| \leq \Phi^{-1}(1-\alpha/2). \tag{2}$$

The correlation coefficient of two random variables $X$ and $Y$ is $\rho_{XY} = \frac{\sigma_{XY}}{\sigma_X \sigma_Y}$, where $\sigma$ is standard deviation[32]. The partial correlation can be computed from correlation matrix using Eq. (3), where $A$, $B$, and $C$ are random variables[33].

$$\hat{\rho}_{AB.C} = \frac{\rho_{AB} - \rho_{AC}\rho_{CB}}{\sqrt{(1-\rho_{AC}^2)(1-\rho_{CB}^2)}} \tag{3}$$

In general, the PC algorithm has two main steps: generating graph skeleton and orienting the edges[34]. Suppose a dataset consists of $v$ variables. The first step is generating a complete undirected network consisting of $v$ vertices. The conditional independence tests are run for every triplet vertices. The output of the first step is a skeleton. The information of the conditional independence test in the first step is used to orient the edges. The output of the PC algorithm is a graph represented by a Completed Partially Directed Acyclic Graph (CPDAG)[30]. The PC algorithm can be used to learn causal graphs by assuming there are no latent variables in the dataset.

**Long short-term memory (LSTM).** Long short-term memory (LSTM) is an efficient gradient-based method[35,36]. LSTM refers to a standard recurrent neural network (RNN) that has long-term memory and short-term memory. Suppose $\zeta$, $X_t$, $\tilde{S}_t$, $S_{t-1}$, $S_t$, $\circ$, $O_t$ denote the sigmoid function, the preprocessed data, the new state of memory cell, the previous state of the memory cell, the final state of memory cell, Hadamard product, and the final output of the memory unit, respectively. Let $i_t$, $f_t$, $o_t$ be the output of different gates and $W^{(i)}$, $W^{(f)}$, $W^{(o)}$, $W^{(c)}$, $U^{(i)}$, $U^{(f)}$, $U^{(o)}$, $U^{(c)}$ be coefficient matrices. The mathematical models related to the LSTM memory unit are defined by Eqs. (4–9)[37]. LSTM networks work well for making predictions based on time series data[38–42].

$$i_t = \zeta\left(W^{(i)}X_t + U^{(i)}S_{t-1}\right) \tag{4}$$

$$f_t = \zeta\left(W^{(f)}X_t + U^{(f)}S_{t-1}\right) \tag{5}$$

$$o_t = \zeta\left(W^{(o)}X_t + U^{(o)}S_{t-1}\right) \tag{6}$$

$$\tilde{S}_t = \tanh\left(W^{(c)}X_t + U^{(c)}S_{t-1}\right) \tag{7}$$

$$S_t = f \circ S_{t-1} + i_t \circ \tilde{S}_t \tag{8}$$

$$O_t = o_t \circ \tanh(S_t) \tag{9}$$

**Gated recurrent unit (GRU).** Gated recurrent unit (GRU) is recurrent neural networks (RNN) using gating mechanism[43,44]. Let $W_z, W_r, W, U_z, U_r, U, b_z, b_r$, and $b$ be model parameters. Suppose $\odot$ represents element-wise multiplication. For each $j$-th hidden unit, GRU has a reset gate $r_t^j$ and an update gate $z_t^j$ to control the hidden state $h_t^j$ at each time $t$ which are computed using Eqs. (10–13). GRU has been successfully implemented for forecasting the time series datasets[45–47].

$$r_t = \zeta(W_r x_t + U_r h_{t-1} + b_r) \tag{10}$$

$$z_t = \zeta(W_z x_t + U_z h_{t-1} + b_z) \tag{11}$$

$$\tilde{h}_t = \tanh(W x_t + U(r_t \odot h_{t-1}) + b) \tag{12}$$

$$h_t = (1 - z_t) \odot h_{t-1} + z_t \odot \tilde{h}_t \tag{13}$$

**Evaluation metric.** The evaluation metric for forecasting are mean absolute error (MAE), mean square error (MSE), and root mean square error (RMSE). MAE reflects the actual situation of the prediction error and RMSE evaluates the degree of change and accuracy of the data. Let $y'$, $y$, and $n$ be the predicted value, true value, and the number of samples. Equations (14–16) are used to compute MAE, MSE, and RMSE, respectively[47,48].

$$MAE_{y',y} = \frac{1}{n}\sum_{i=1}^{n} |y_i' - y_i| \tag{14}$$

$$MSE_{y',y} = \frac{1}{n}\sum_{i=1}^{n} (y_i' - y_i)^2 \tag{15}$$

$$RMSE_{y',y} = \sqrt{\frac{1}{n}\sum_{i=1}^{n} (y_i' - y_i)^2} \tag{16}$$

**Dataset.** This paper use AQI and meteorological conditions in Jakarta from public datasets. The AQI data owned by DKI Jakarta Provincial Government can be accessed at https://data.jakarta.go.id/organization/badan-pengelolaan-lingkungan-hidup-daerah[18]. The meteorological conditions dataset is obtained from an open dataset belonging to Indonesian Agency for Meteorological, Climatological and Geophysics (Badan Meteorologi, Klimatologi, dan Geofisika or simply BMKG) that is available at http://dataonline.bmkg.go.id/home[49].

The air quality dataset is daily records Air Quality Index (AQI) of $PM_{10}$, $PM_{2.5}$, $SO_2$, CO, $O_3$, and $NO_2$ from 2010 to 2021. $PM_{2.5}$ is only available from 2021. The meteorological conditions is a daily record of average temperature (°C), average relative humidity (RH) (%), average rainfall (mm), average duration of sunshine (hours), and average wind speed (m/s) from 2010 to 2021.

**The proposed method.** This paper proposes an integrated analysis of air pollution data and meteorological condition to analyze the air quality in Jakarta. The proposed method is illustrated in Fig. 1. The stages of the proposed method are data integration, causal graph generation, and forecasting. The integration process of meteorological data and AQI data use column-based integration. The datasets are time series data with numerical values. The idea of data integration has been used to learn simultaneously from multiple data sources[50,51]. The integration data requires not only the same date for each sample but also the same number of samples from all resources. In this paper, the integrated data is a single table containing variables from meteorological data and AQI data. This data is then used as input for generating a causal graph and forecasting. A causal graph is generated using the PC algorithm. LSTM and GRU are implemented for forecasting.

This paper uses the PC algorithm from *bnlearn* in the R package[52,53]. A causal graph is generated in R Studio. It also implements LSTM and GRU from TensorFlow Keras. The forecasting is run in Jupyter Notebook for Python.

LSTM and GRU are implemented to forecast the prediction of AQI and meteorological conditions. The LSTM and GRU models consist of stacked layers with 128 and 64 units, dropout layer and dense layer. LSTM and GRU are run for 50 epochs and they implement the Softmax activation function. This paper uses multivariate forecasting. The experiments use integrated and not integrated data. The letter *i* and *p* indicate that the algorithm is implemented for forecasting using integrated data and not integrated data, respectively. A not-integrated data refers to AQI data or a meteorological conditions dataset. An integrated dataset is a dataset containing AQI and meteorological conditions obtained from data integration process. This paper runs multivariate forecasting in 3 different scenarios:

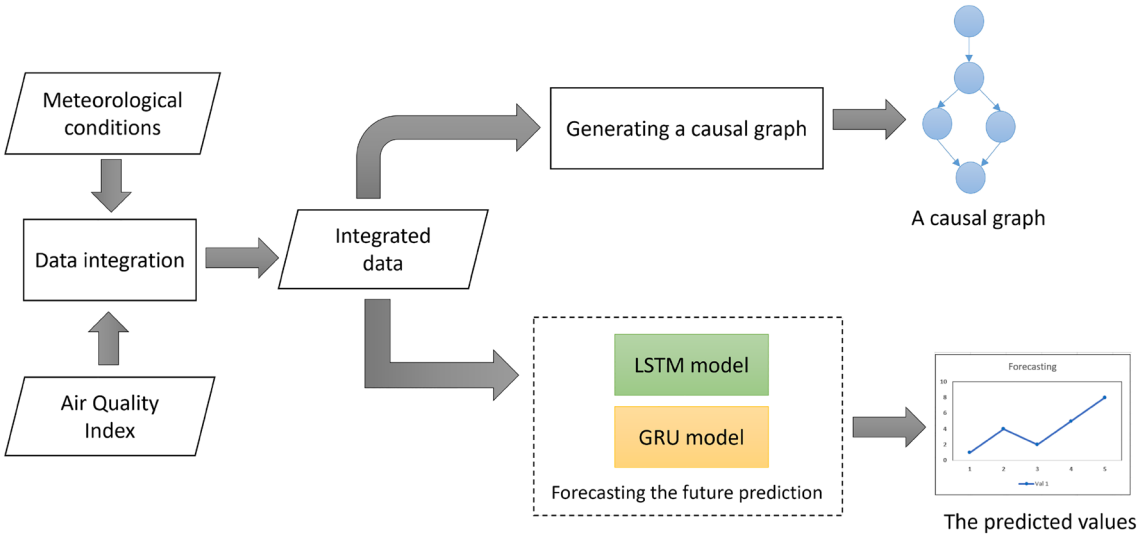



**Figure 1.** The proposed method of an integration analysis of air pollution and meteorological conditions.

- Experiment 1 using training set from 2010 to 2018 and testing set from 2019.
- Experiment 2 using training set from 2010 to 2019 and testing set from 2020.
- Experiment 3 using training set from 2010 to 2020 and testing set from 2021.

## Results and discussion

The datasets are containing less than 5% missing values. The missing values are filled up using an average value of the observed variables from 7 days before the observed date. After preprocessing phase, it implements column-based integration to create a single formed data from AQI and meteorological condition datasets. The integrated data is used to generate a causal graph and to train models for forecasting.

**Causality analysis.** This paper examines the dependence relationships between air pollutants represented by AQI and meteorological conditions. A graph is generated from an integrated data of AQI and meteorological conditions from 2010 to 2021. The dataset consists of 4383 samples and 10 variables (temperature, humidity, rainfall, sunshine, wind speed, $PM_{10}$, $SO_2$, CO, $O_3$, and $NO_2$). $PM_{2.5}$ is not included to the experiments due to the samples are only available from 2021. Figure 2 shows a causal graph generated using the PC algorithm at significance level of $\alpha = 0.05$. The graph finds some information that will be explained as follows.

- Humidity, rainfall, and duration of sunshine are causal parameters for $PM_{10}$. Those findings are corresponding to some previous studies. Humidity influences PM's natural deposition process; moisture particles adhere to PM and accumulate atmospheric PM concentration[9]. The increasing humidity reduces $PM_{10}$ concentrations in the atmosphere because moisture particles grow in size to a point where 'dry deposition' happens. $PM_{10}$ continually reduced with humidity rising[10]. The precipitation has a certain wet scavenging effect on $PM_{2.5}$ and $PM_{10}$[11]. Precipitation scavenging refers to the cleaning of gases and particles by cloud and precipitation elements. A study of ambient air quality in Jakarta found that the concentration of suspended particulate matter is decreased in the wet season (October–March) and increased in the dry season (April–September) because rainfall removes the pollutant in the atmosphere[54].
- CO has a dependent relationship to humidity. The previous study shows that higher humidity has a negative effect on the adsorption of carbon monoxide[55].
- Wind speed has dependence relationships to $SO_2$, $NO_2$, and $PM_{10}$.
- Temperature has a causal relationship to $O_3$. The chemical reactions in the formation or destruction of $O_3$ are influenced by temperature, solar radiation, and wind speed[56]. A study found that diurnal temperature range, precipitation, and wind speed had the largest impact on $SO_2$ in Shandong, China[57].
- CO causes $O_3$ and it is similar to a study in Kota Bharu, Malaysia that discovers CO as a causal parameter for $O_3$[58].
- $PM_{10}$, $SO_2$, and CO affect $O_3$. $O_3$ is an air pollutant that is formed in the atmosphere from a combination of nitrogen oxides, volatile organic compounds, CO, and methane in the presence of sunlight[59].
- $NO_2$ affects $PM_{10}$ and $SO_2$.
- Sunshine affects $PM_{10}$.

The causal graph in Fig. 2 explains the connection of certain parameters from meteorological conditions to air pollution. Those relationships are not revealed when the analysis is done separately.

**Correlation analysis.** This paper highlights correlation coefficient ($\rho$) with the values $\rho \geq \pm 0.2$. The correlation coefficient between variables except for $PM_{2.5}$ is computed from samples of 2010–2021. The correlation

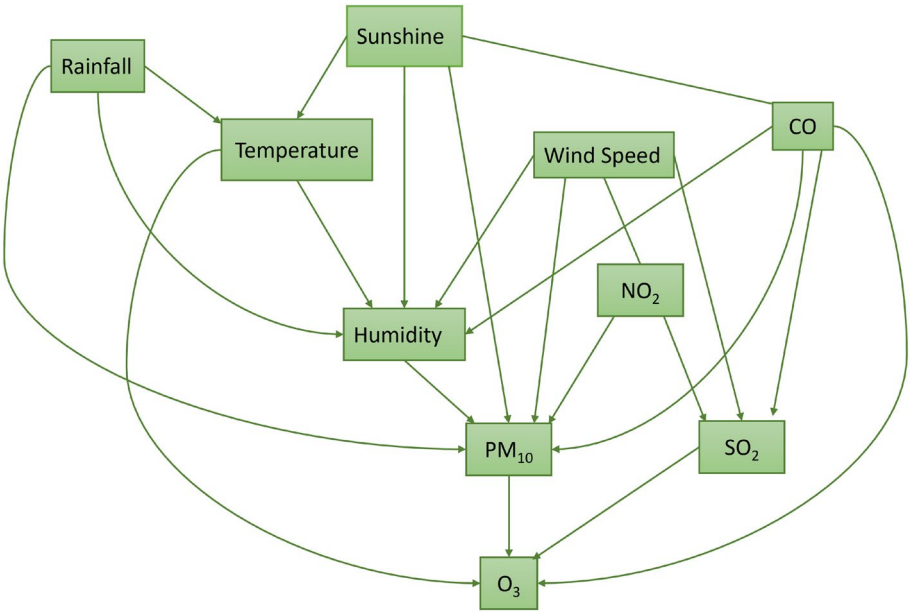



**Figure 2.** A causal graph is generated from AQI and meteorological conditions.

coefficient involving $PM_{2.5}$ is obtained from the dataset of 2021. Table 1 shows correlation coefficient between two variables computed using Pearson correlation. The longer the sunshine duration makes the higher temperature, lower humidity, and lower rainfall. The temperature and duration of sunshine have a positive correlation to $PM_{10}$ and $PM_{2.5}$. The more concentration of PM the higher temperature will be. Humidity has a negative correlation to $PM_{10}$ and $PM_{2.5}$. This is one of the possible ways to decrease PM concentration by increasing humidity. Higher rainfall increases humidity. Weather modification to create artificial rain is useful to decrease PM concentration. Humidity and CO have a positive correlation. Meanwhile, wind speed and $SO_2$ have a negative correlation.

The annual average AQI from 2010 to 2021 is illustrated in Fig. 3A. The highest exposure to $O_3$ happened in 2012. Figure 3B shows the monthly average AQI in Jakarta from 2010 - 2021. The top 3 air pollutant are $O_3$, $PM_{2.5}$ and $PM_{10}$. AQI score of $O_3$ is always higher than 50 and it reaches over 100 in October to November which is categorized as an unhealthy condition.

Sunrise and sunset in Jakarta are not significantly different every day throughout the year because it lies on a latitude of − 6°12′ 52.63′ ′ S and a longitude of 106°50′ 42.47′ ′ E. The length of daylight remains the same every day, so the duration of sunshine is mostly affected by clouds. In the last 10 years, the low average rainfall happens from May to September, and the lowest is around 1.8 mm in August. Meanwhile, the longest average sunshine duration occurs in August, September, and October at 5.7, 6.4, and 5.3 hours, respectively. The month of June to October has a high average level of $PM_{10}$ over 73 and the highest is 76.79 in August. The lowest average of $PM_{10}$ is 50.24 in January. This finding is closed to the previous study[54] which is states that the highest concentration of $PM_{10}$ occurs in September 2015 and the lowest one is in February 2017. In 2021, the two highest average AQI for $PM_{2.5}$ are 80.56 in June and 86.32 in July. In May and October, the average temperature is around 29.1 °C

| | Temperature | Humidity | Rainfall | Sunshine | Wind | $PM_{10}$ | $SO_2$ | CO | $O_3$ | $NO_2$ | $PM_{2.5}$ |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Temperature | 1 | − 0.69 | − 0.38 | 0.47 | − 0.02 | 0.31 | 0.12 | − 0.12 | 0.17 | 0.03 | 0.41 |
| Humidity | − 0.69 | 1 | 0.40 | − 0.49 | − 0.06 | − 0.35 | − 0.12 | 0.25 | − 0.16 | − 0.02 | − 0.34 |
| Rainfall | − 0.38 | 0.40 | 1 | − 0.21 | − 0.03 | − 0.20 | − 0.07 | 0.07 | − 0.07 | − 0.02 | − 0.36 |
| Sunshine | 0.47 | − 0.49 | − 0.21 | 1 | − 0.02 | 0.27 | 0.09 | − 0.18 | 0.13 | 0.02 | 0.19 |
| Wind | − 0.02 | − 0.06 | − 0.03 | − 0.02 | 1 | − 0.05 | − 0.32 | 0.02 | 0.01 | − 0.10 | − 0.32 |
| $PM_{10}$ | 0.31 | − 0.35 | − 0.20 | 0.27 | − 0.05 | 1 | 0.05 | 0.19 | 0.35 | 0.12 | 0.73 |
| $SO_2$ | 0.12 | − 0.12 | − 0.07 | 0.09 | − 0.32 | 0.05 | 1 | − 0.12 | − 0.06 | 0.60 | 0.33 |
| CO | − 0.12 | 0.25 | 0.07 | − 0.18 | 0.02 | 0.19 | − 0.12 | 1 | 0.15 | 0.05 | 0.19 |
| $O_3$ | 0.17 | − 0.16 | − 0.07 | 0.13 | 0.01 | 0.35 | − 0.06 | 0.15 | 1 | 0.00 | 0.32 |
| $NO_2$ | 0.03 | − 0.02 | − 0.02 | 0.02 | − 0.1 | 0.12 | 0.60 | 0.05 | 0.00 | 1 | 0.14 |
| $PM_{2.5}$ | 0.41 | − 0.34 | − 0.36 | 0.17 | − 0.32 | 0.73 | 0.33 | 0.19 | 0.33 | 0.14 | 1 |

**Table 1.** Correlation coefficient.

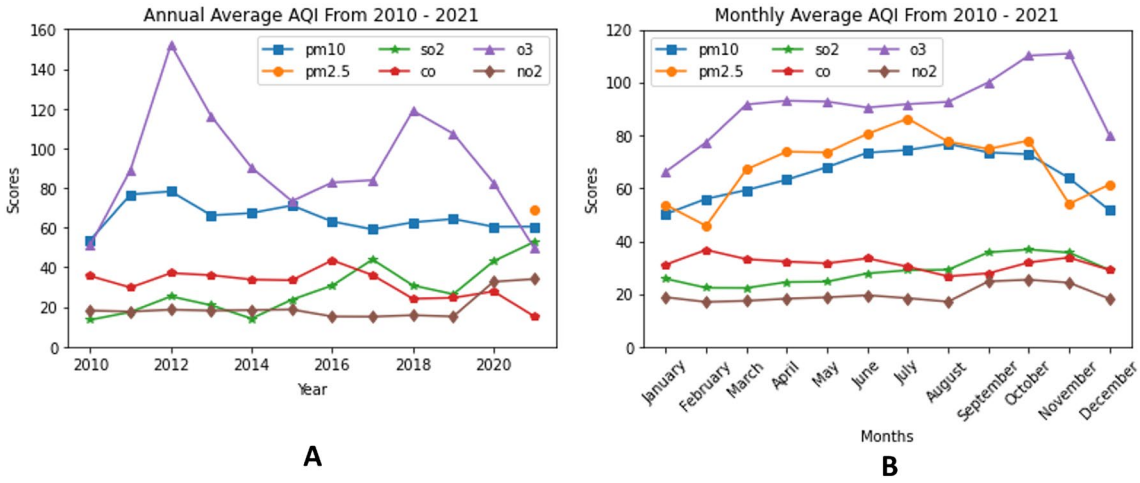



**Figure 3.** The (**A**) annual and (**B**) monthly average AQI in Jakarta from 2010 to 2021.

which is higher than the overall average temperature of 28.49 °C. The average humidity during July–October is around 70–74%. Since 2015, the average wind speed decreases around 1 m/s than that in 2010. In 2021, the correlation between wind speed and $PM_{2.5}$ $\rho$(wind speed and $PM_{2.5}$) is − 0.32. The decrement in wind speed contributes to increasing $PM_{2.5}$. Wind speed and $SO_2$ have a negative correlation, so decreasing wind speed rises $SO_2$. A positive correlation is obtained between $SO_2$ and $NO_2$ as 0.6, indicating that the concentration of those pollutants rises together. $O_3$ has a positive correlation to $PM_{10}$ and $PM_{2.5}$.

**The historical data and forecasting models.** A record of the number of unhealthy (U) and very unhealthy days (VU) in the year 2010–2021 is presented in Table 2. The historical data shows that $O_3$ is the pollutant that mostly causes unhealthy and very unhealthy days. There are 108 days where on the same day two pollutants have AQI scores over 100 but only 22 days were labeled as very unhealthy because they only pay attention to a pollutant that has the highest AQI scores on that days. In 2020, on three consecutive days, the three pollutants together ($SO_2$, $O_3$, and $NO_2$) have AQI scores of more than 100 and those are categorized as unhealthy. It needs further study for a case when more than two pollutants have AQI scores over 100 in a day. It is possible to be more hazardous when the concentration of multiple pollutants reaches the unhealthy limit at the same time, so the categories of air pollution levels need to be evaluated.

The previous studies reveal various effects of the pollutants. The ambient temperature increased acute cardiovascular-respiratory mortality effects of $PM_{2.5}$[60]. Exposure to $PM_{10}$, $NO_2$, and $O_3$ generates a relative risk to human health[61]. The effect of humans inhaling $O_3$ possibly leads to acute lung function changes and inflammation[62]. $PM_{2.5}$ may contribute to the development of diabetes mellitus, increase cardiopulmonary morbidity and mortality, and cause adverse birth outcomes[63]. Epidemiological evidence shows that $PM_{2.5}$ damage the human respiratory system[64]. The accumulating of exposure to low concentrations of carbon monoxide can affect a number of organ systems[65].

The actual data and forecasting of AQI from 2010 to 2021 are described in Fig. 4 A and B, respectively. The performance of LSTM and GRU are evaluated using MAE and RMSE. According to the experimental results, LSTM using integrated data produces the smaller error. In general, LSTM and GRU show a good performance in forecasting $PM_{10}$, CO, and $O_3$.

| Year | $_UO_3$ | $_U$CO | $_UPM_{2.5}$ | $_UPM_{10}$ | $_USO_2$ | $_UNO_2$ | $_{VU}O_3$ |
|---|---|---|---|---|---|---|---|
| 2010 | 18 | 0 | 0 | 0 | 0 | 0 | 4 |
| 2011 | 105 | 0 | 0 | 41 | 0 | 0 | 5 |
| 2012 | 123 | 0 | 0 | 25 | 0 | 0 | 116 |
| 2013 | 177 | 0 | 0 | 4 | 0 | 0 | 27 |
| 2014 | 83 | 1 | 0 | 6 | 0 | 0 | 8 |
| 2015 | 43 | 0 | 0 | 21 | 0 | 0 | 0 |
| 2016 | 48 | 0 | 0 | 0 | 0 | 0 | 1 |
| 2017 | 105 | 0 | 0 | 5 | 0 | 0 | 0 |
| 2020 | 57 | 25 | 0 | 1 | 8 | 10 | 3 |
| 2021 | 2 | 0 | 136 | 2 | 1 | 1 | 0 |

**Table 2.** Pollutants that affect unhealthy (U) and very unhealthy (VU) days in Jakarta.

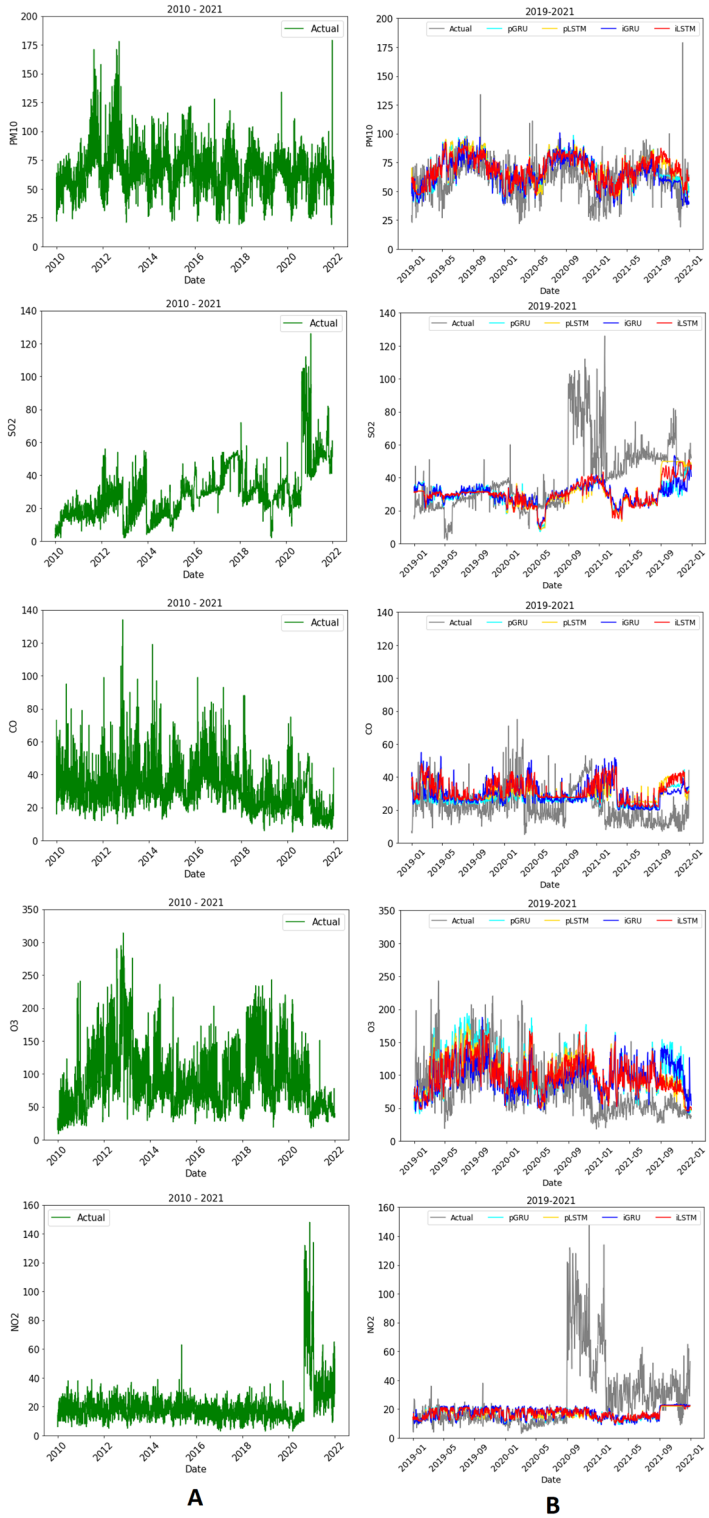



**Figure 4.** The actual data (**A**) and forecasting (**B**) of AQI in Jakarta.

The actual data and forecasting meteorological conditions are described in Fig. 5A and B, respectively. LSTM and GRU work well to forecast temperature, humidity, sunshine duration and wind speed. However, they are less accurate to predict rainfall.

The two highest AQI of $PM_{10}$ were in 2011 and 2013 when the averages were 76.59 and 78.21. The AQI of $SO_2$ was consistently rising around 3 times higher than in 2010. The AQI of CO increased from 2010 to 2017, but it decreased from 2020 to 2021. The AQI of $O_3$ was also rising and the highest was in 2012–2013. Figure 6 shows the values of MAE and RMSE for forecasting results. LSTM using integrated data produces smaller errors.

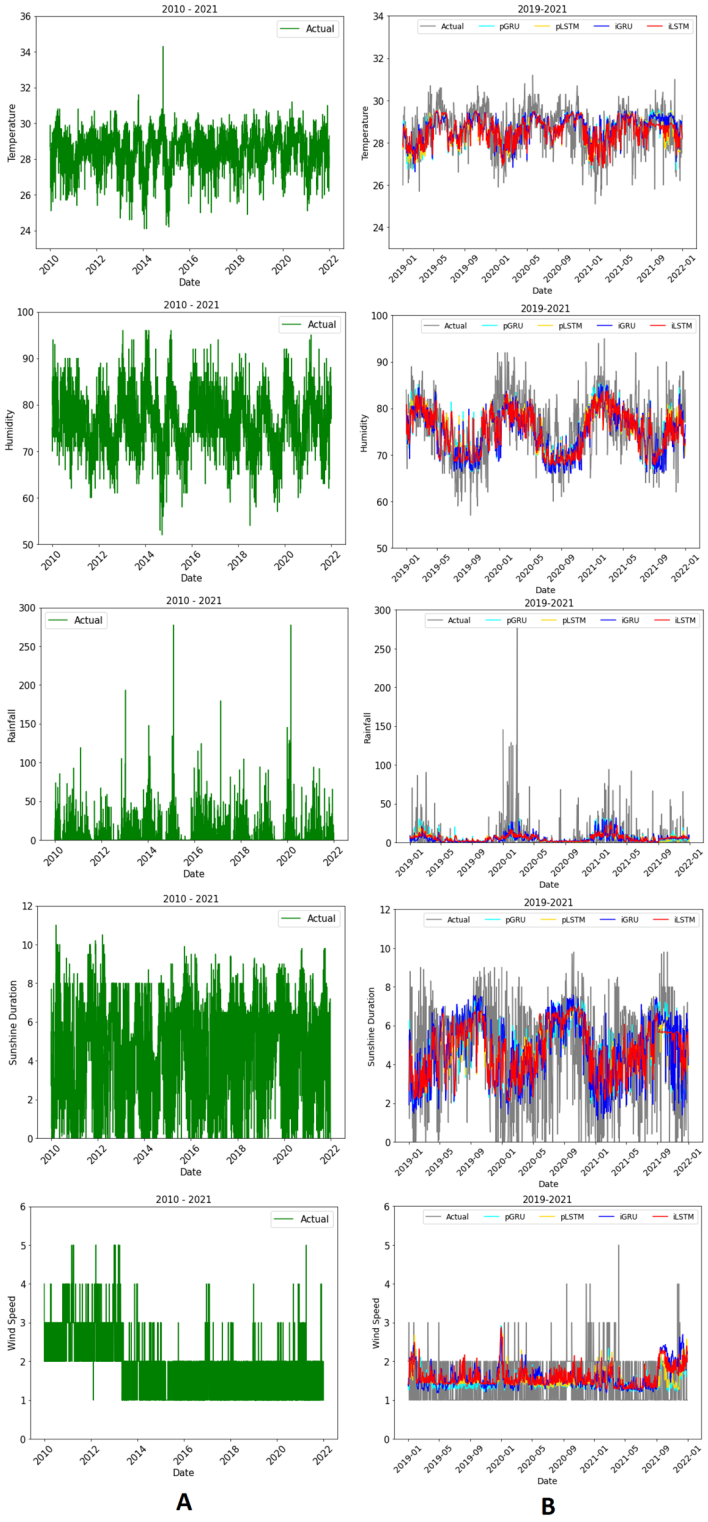



**Figure 5.** The actual data (**A**) and forecasting (**B**) of meteorological condition in Jakarta.

In general, the forecasting results of AQI data from 2020 to 2021 have higher errors than that from 2019. It is suspected that major restrictions in some activities during the Covid-19 outbreak influence that condition, for instance, the national or local lockdown reduces the use of motor vehicles which decreases the CO level. There was a huge increase in $SO_2$ and $NO_2$ from September 2020–January 2021 but the reason is unknown. It needs further study for investigation. Comparing to the other study which is forecasting the observed variables using not integration data[66], the forecasting using integration data produces slightly lower MAE and RMSE.

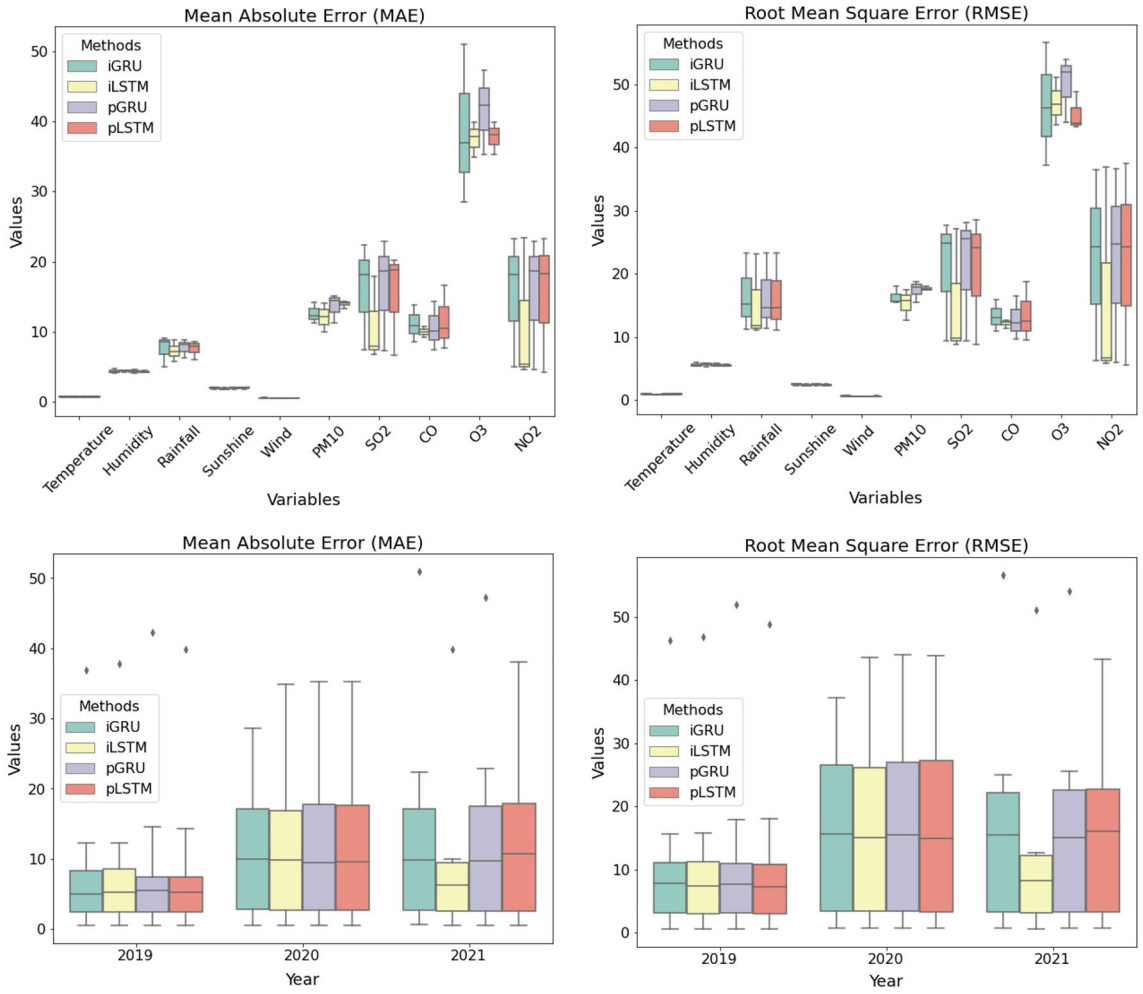



**Figure 6.** A comparison performance using integrated data (*i*) and not integrated data (*p*).

The findings in this paper are expected to enrich the knowledge of the linkage between air pollution and climate change. This contribution is beneficial to determining the proper handling of air pollution and climate change problems.

## Conclusion

In conclusion, the integration analysis successfully discovers the linkage between air pollution and meteorological conditions in Jakarta. The integrated data is used to generate a causal graph and to train models for forecasting. A causal graph shows that there exist dependence relationships between AQI and meteorological conditions. This information is beneficial for handling air pollution and climate change. LSTM and GRU work well as models for forecasting $PM_{10}$, CO, $O_3$, temperature, humidity, sunshine duration, and wind speed. However, those models show less accurate to predict $SO_2$, $NO_2$, and rainfall. LSTM using integrated data produces a smaller error. The forecasting results of air pollution before the Covid-19 outbreak are more accurate. The Covid-19 outbreak influences human activities that probably affect air quality, e.g, decreasing CO, and increasing $NO_2$ and $SO_2$. The future work is implementing machine learning approach for an integrated analysis to find the connection of population growth, industries, human activities and air pollution to the climate change in Indonesia.

## Data availibility

The datasets are available from the corresponding author by request for strong reasons.

Received: 16 October 2022; Accepted: 3 April 2023
Published online: 09 April 2023

## References

1. Kan, H. *et al.* Part 1 a time-series study of ambient air pollution and daily mortality in Shanghai, China. *Res. Rep. Health Eff. Inst.* **154**(1), 17–78 (2010).
2. Qian, Z. *et al.* Part 2 association of daily mortality with ambient air pollution, and effect modification by extremely high temperature in Wuhan, China. *Res. Rep. Health Eff. Inst.* **154**(1), 91–217 (2010).

3. Tramuto, F. *et al.* Urban air pollution and emergency room admissions for respiratory symptoms: A casecrossover study in palermo, Italy. *Environ. Health* **10**(31), 1–11 (2011).
4. Zhang, J., Wei, Y. & Fang, Z. Ozone pollutan a major health hazard worldwide. *Front. Immunol.* **10**(1), 1–10 (2019).
5. Holm, S. M. & Balmes, J. R. Systematic review of ozone effects on human lung function, 2013 through 2020. *Chest* **161**(1), 190–201 (2022).
6. Peng, H. *et al.* Relationship between meteorological factors, air pollutants and hand, foot and mouth disease from 2014 to 2020. *BMC Public Health* **22**(1), 1–10 (2022).
7. Asian development bank. Air quality in Asia: Why is it important, and what can we do? (2022; accessed 20 Sept 2022); https://www.adb.org/sites/default/files/publication/780921/air-quality-asia.pdf.
8. He, J. *et al.* Air pollution characteristics and their relation to meteorological conditions during 2014–2015 in major chinese cities. *Environ. Pollut.* **223**, 484–496 (2017).
9. Hernandez, G., Berryand, T.-A., Wallis, S. L. & Poyner, D. Temperature and humidity effects on particulate matter concentrations in a sub-tropical climate during winter. In *International Proceedings of Chemical, Biological and Environmental Engineering* 41–49 (2017).
10. Lou, C. *et al.* Relationships of relative humidity with pm2.5 and pm10 in the yangtze river delta, china. *Environ. Monit. Assess.* **189**(582), 1–16 (2017).
11. Yansui, L., Zhou, Y. & Lu, J. Exploring the relationship between air pollution and meteorological conditions in china under environmental governance. *Sci. Rep.* **10**, 1–11 (2020).
12. Liu, Z. *et al.* Analysis of the influence of precipitation and wind on $PM_{2.5}$ and $PM_{10}$ in the atmosphere. *Adv. Meteorol.* **2020**(1), 1–13 (2020).
13. Hou, K. & Xu, X. Evaluation of the influence between local meteorology and air quality in Beijing using generalized additive models. *Atmosphere* **13**(24), 1–14 (2021).
14. Wang, C. H. *et al.* Quantifying the effects of climate factors on carbon monoxide poisoning a retrospective study in Taiwan. *Front. Public Health* **9**(1), 1–7 (2021).
15. Vitolo, C., Scutari, M., Ghalaieny, M., Tucker, A. & Russell, A. Modeling air pollution, climate, and health datavusing bayesian networks: A case studyvof the English regions. *Earth Space Sci.* **5**, 76–88 (2018).
16. Kusumaningtyas, S. D. A. *et al.* Aerosols optical and radiative properties in Indonesia based on AERONET version 3. *Atmos. Env.* **282**, 119174 (2022).
17. Dinas Lingkungan Hidup Provinsi DKI Jakarta: Laporan Kualitas Udara Jakarta (2022; accessed 10 Jul 2022); https://lingkunganhidup.jakarta.go.id/files/14477-2022-06-24-07-45-08.pdf.
18. Portal Data Terpadu Pemprov DKI Jakarta: Dataset Indeks Standar Pencemaran Udara (2022, accessed 24 Jun 2022); https://data.jakarta.go.id/group/lingkungan-hidup.
19. Peters, J., Janzing, D. & Schölkopf, B. *Elements of Causal Inference* (The MIT Press, 2017).
20. Scutari, M., Graafland, C. E. & Gutiérrez, J. M. Who learns better bayesian network structures: Accuracy and speed of structure learning algorithms. *Int. J. Approx. Reason.* **115**, 235–253 (2019).
21. Schwartz, J., Bind, M. A. & Koutrakis, P. Estimating causal effects of local air pollution on daily deaths effect of low levels. *Environ. Health Perspect.* **125**(1), 23–29 (2017).
22. Zhang, Y., Gen, Y. & Luo, G. Causal direction inference for air pollutants data. *Comput. Electr. Eng.* **68**, 404–1411 (2018).
23. Nethery, R. C., Mealli, F., Sacks, J. D. & Dominici, F. Evaluation of the health impacts of the 1990 clean air act amendments using causal inference and machine learning. *J. Am. Stat. Assoc.* **16**(1), 1–12 (2020).
24. Kementrian Lingkungan Hidup dan Kehutanan: Indeks Standar Pencemaran Udara (ISPU) Sebagai Informasi Mutu Udara Ambien di Indonesia (2022, accessed 10 Jul 2022); https://ditppu.menlhk.go.id/portal/read/indeks-standar-pencemar-udara-ispu-sebagai-informasi-mutu-udara-ambien-di-indonesia.
25. Pearl, J. *Causality Models, Reasoning and Inference* (Cambridge University Press, 2000).
26. Spirtes, P., Glymour, C. & Scheines, R. *Causation, Prediction, and Search* (The MI Press, 2001).
27. Pearl, J. Causal inference in statistics an overview. *Stat. Surv.* **3**, 96–146 (2009).
28. Colombo, D., Maathuis, M. H., Kalisch, M. & Richardson, T. S. Learning high dimensional directed acyclic graphs with latent and selection variables. *Ann. Stat.* **40**(1), 294–321 (2012).
29. Kalisch, M., Mächler, M., Colombo, D., Maathuis, M. H. & Bühlmann, P. Causal inference using graphical models with the R package pcalg. *J. Stat. Softw.* **47**(11), 1–26. https://doi.org/10.18637/jss.v047.i11 (2012).
30. Kalisch, M. & Buhlmann, P. Estimating high dimensional directed acyclic graphs with the PC algorithm. *J. Mach. Learn. Res.* **8**, 613–636 (2007).
31. Cui, R., Groot, P. & Heskes, T. Copula PC algorithm for causal discovery from mixed data. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases* (2016).
32. Walpole, R. E., Mayers, R. H. & Myers, S. L. *Probability and Statistics for Engineers and Scientists* (Prentice Hall, New Jersey, 2011).
33. Meloun, M. & Militký, J. *Statistical Data Analysis A Practical Guide* (India PVT LTD, 2011).
34. Colombo, D. & Maathuis, M. H. Order independent constraint based causal structure learning. *J. Mach. Learn. Res.* **14**(2014), 3921–3962 (2016).
35. Hochreiter, S. & Schmidhuber, J. Long Short Term Memory. *Neural Comput.* **9**(8), 1735–1780 (1997).
36. Gers, F. A. & Schmidhuber, J. LSTM recurrent networks learn simple context free and context sensitive languages. *IEEE Trans. Neural Netw.* **12**(6), 1333–1340 (2001).
37. Zhao, Z., Chen, W., Wu, X., Chen, P. C. Y. & Liu, J. LSTM network a deep learning approach for short term traffic forecast. *IET Intel. Transport Syst.* **11**(2), 68–75 (2017).
38. Tsai, Y.-T., Zeng, Y.-R. & Chang, Y.-S. Air pollution forecasting using RNN with LSTM. In *2018 IEEE 16th Intl Conf on Dependable, Autonomic and Secure Computing, 16th Intl Conf on Pervasive Intelligence and Computing, 4th Intl Conf on Big Data Intelligence and Computing and Cyber Science and Technology Congress(DASC/PiCom/DataCom/CyberSciTech)* 1074–1079 (2018).
39. Belavadi, S. V., Rajagopal, S., Ranjani, R. & Mohan, R. Air quality forecasting using LSTM RNN and wireless sensor networks. In *The 11th International Conference on Ambient Systems, Networks and Technologies (ANT) April 6–9, 2020, Warsaw, Poland* 241–248 (2020).
40. Poornima, S. & Pushpalatha, M. Prediction of rainfall using intensified LSTM based recurrent neural network with weighted linear units. *Atmosphere* **10**, 1–18 (2019).
41. Alhirmizy, S. & Qader, B. Multivariate time series forecasting with LSTM for Madrid, Spain pollution. In *2019 International Conference on Computing and Information Science and Technology and Their Applications ICCISTA* 1–5 (2019).
42. Ghanbari, R. & Borna, K. Multivariate time series prediction using LSTM neural networks. In *2021 26th International Computer Conference, Computer Society of Iran CSICC* 1–5 (2021).
43. Cho, K., Merrienboer, B. V., Bahdanau, D. & Bengio, Y. On the properties of neural machine translation encoder decoder approaches. In *Eighth Workshop on Syntax, Semantics and Structure in Statistical Translation SSST-8* 103–111 (2014).
44. Che, Z., Purushotham, S., Cho, K., Sontag, D. & Liu, Y. Recurrent Neural Networks for multivariate time series with missing values. *Sci. Rep.* **8**, 1–12 (2018).
45. Zhou, X., Xu, J., Zeng, P. & Meng, X. Air pollutant concentration prediction based on GRU method. In *IOP Conf. Series: Journal of Physics: Conf. Series* 1–6 (2019).

46. Athira, V., Vinayakumar, R. & Kumar, P. S. Deepairnet applying recurrent networks for air quality prediction. In *International Conference on Computational Intelligence and Data Science ICCIDS 2018* 1394–1403 (2018).
47. Tao, Q., Li, Y. & Sidorov, D. Air pollution forecasting using a deep learning model based on 1d Convnets and Bidirectional GRU. *IEEE Access* **7**, 76690–76698 (2019).
48. Iglewicz, B. & Myers, R. H. Comparisons of approximations to the percentage points of the sample coefficient of variation. *Technometrics* **12**(1), 166–169 (1970).
49. Pusat Database BMKG: Data Harian (2022; accessed 23 June 2022). http://dataonline.bmkg.go.id/home?language=indonesia.
50. Pavlidis, P., Weston, J., Cai, J. & Grundy, W. N. Gene functional classification from heterogeneous data. In *Proceedings of the Fifth Annual International Conference on Computational Biology* 249–255 (2001).
51. Daemen, A., Gevaert, O. & Moor, B. D. Integration of clinical and microarray data with kernel method. In *Proceedings of the 29th Annual International Conference of the IEEE EMBS* 5411–5415 (2007).
52. Scutari, M. bnlearn—an R package for Bayesian Network learning and inference (2010, accessed 2 October 2022); https://www.bnlearn.com/.
53. Scutari, M. Learning Bayesian Networks with the bnlearn R Package. *J. Stat. Softw.* **35**(3), 1–22 (2010).
54. Kusumaningtyas, S. D. A., Aldrian, E., Wati, T. & Atmoko, D. Sunaryo: The recent state of ambient air quality in Jakarta. *Aerosol Air Qual. Res.* **18**(9), 2343–2354 (2018).
55. Eslamian, M., Nadimi, E. & Salehi, A. Effect of humidity on gas sensing properties of tin dioxide toward carbon monoxide: A first principle study. In *2017 Iranian Conference on Electrical Engineering (ICEE)* 276–278 (2017).
56. Alvim-Ferraz, M. C. M., Sousa, S. I. V., Pereira, M. C. & Martins, F. G. Contribution of anthropogenic pollutants to the increase of tropospheric ozone levels in the oporto metropolitan area, portugal since the 19th century. *Environ. Pollut.* **140**, 516–524 (2006).
57. Wu, H., Hong, S., Hu, M., Li, Y. & Yun, W. Assessment of the factors influencing sulfur dioxide emission in Shandong, China. *Atmosphere* **13**, 1–14 (2022).
58. Raffee, A. F., Hamid, H. A., Rahmat, S. N. & Jaffar, M. I. The cause-and-effect analysis of ground level ozone ($O_3$), air pollutants and meteorological parameters using the causal relationship approach. *J. Eng. Res.* **1**, 1–21 (2022).
59. Schneidemesser, E. V. *et al.* Chemistry and the linkages between air quality and climate change. *Chem. Rev.* **115**(10), 3856–3897 (2015).
60. Li, Y., Ma, Z., Zheng, C. & Shang, Y. Ambient temperature enhanced acute cardiovascular-respiratory mortality effects of $PM_{2.5}$ in Beijing, China. *Int. J. Biometeorol.* **59**, 1761–1770 (2015).
61. Khaniabadi, Y. O. *et al.* Exposure to $PM_{10}$, $NO_2$, and $O_3$ and impacts on human health. *Environ. Sci. Pollut. Res.* **24**, 2781–2789 (2016).
62. Bromberg, P. A. Mechanisms of the acute effects of inhaled ozone in humans. *Biochem. Biophys. Acta* **12**, 2771–2781 (2016).
63. Feng, S., Gao, D., Liao, F., Zhou, F. & Wang, X. The health effects of ambient $PM_{2.5}$ and potential mechanisms. *Ecotoxicol. Environ. Saf.* **128**, 67–74 (2016).
64. Xing, Y.-F., Xu, Y.-H., Shi, M.-H. & Lian, Y.-X. The impact of $PM_{2.5}$ on the human respiratory system. *J. Thorac. Dis.* **8**(1), 69–74 (2016).
65. Townsend, C. L. & Maynard, R. L. Effects on health of prolonged exposure to low concentrations of carbon monoxide. *Occup. Environ. Med.* **59**(10), 708–711 (2022).
66. Handhayani, T., Lewenusa, I., Herwindiati, D. E. & Hendryli, J. A comparison of LSTM and BiLSTM for forecasting the air pollution index and meteorological conditions in jakarta. In *5th International Seminar on Research of Information Technology and Intelligent Systems (ISRITI)* (eds Kartadie, R. & Wibowo, F.W.) 334–339 (2022).



## Acknowledgements
The author thanks Dr. Siti Syuhaida Mohamed Yunus for the meaningful discussions and suggestions for this research.


## Author contributions
The author contributes to collecting the dataset, running the experiments, and preparing the manuscript.

## Competing interests
The author declares no competing interests.

## Additional information
**Correspondence** and requests for materials should be addressed to T.H.

**Reprints and permissions information** is available at www.nature.com/reprints.

**Publisher's note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.