Methodology To Develop City-Level, Science-Based, Air-Pollution Action Plans

Working Paper Series

ESCAP 75
MOVING FORWARD TOGETHER

The shaded areas of the map indicate ESCAP members and associate members.*

Acknowledgement: This policy brief on "Methodology to Develop City Level, Science-Based Air Pollution Action Plans" has been issued in January 2023. Preparation of this paper is coordinated by the Economic and Social Commission for Asia and the Pacific, through its Environment and Development Division, under the supervision of Sangmin Nam, Director, Environment and Development Division, Curt Garrigan, Chief, Environment and Development Policy Section and authorship by Matthew Perkins, Environmental Affairs Officer, and Worasom Kundhikanjana.

This document also benefitted from contributions by Abigail Smith and Mervin Chin, whose inputs were integral to the production of this report. Financial support by the Republic of Korea is gratefully acknowledged.

For further information on this policy brief, please address your inquiries to: Environment and Development Division United Nations Economic and Social Commission for Asia and the Pacific (ESCAP)

Email: ESCAP-EDD@un.org

** The designations employed and the presentation of material on this map do not imply the expression of any opinion whatsoever on the part of the Secretariat of the United Nations concerning the legal status of any country, territory, city or area or of its authorities, or concerning the delimitation of its frontiers or boundaries.*

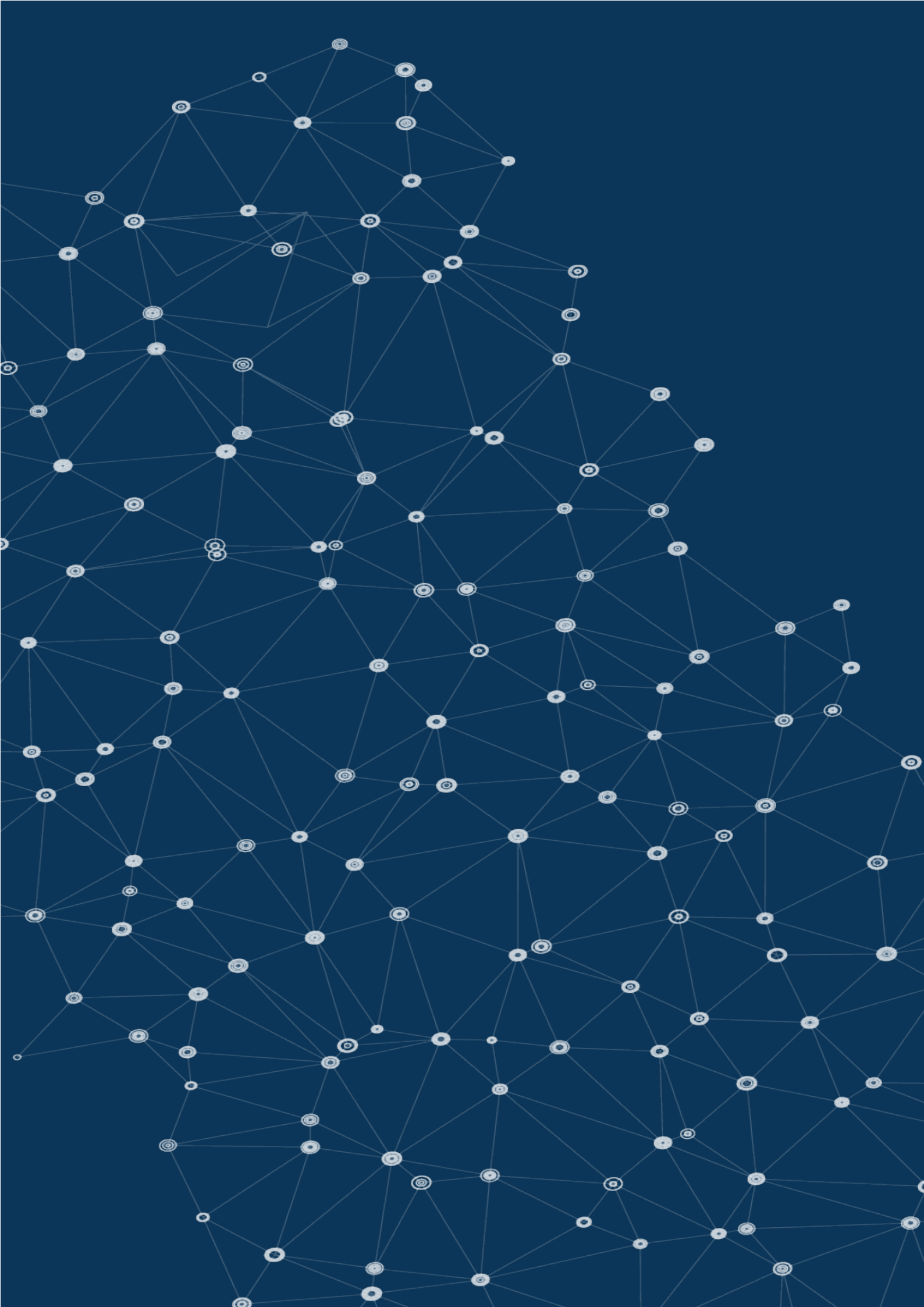


Table of Contents

Abstract

It is universally recognized that air pollution is a pressing environmental challenge that has increased considerably in recent years, leading to a rise in premature deaths, threatening livelihoods and the sustainable development of the region, in particular in many cities in Asia and the Pacific where air pollution is a major public health hazard to an increasing urban population.

However, many rapidly urbanizing cities in developing countries lack the resources and data to clearly identify the causes of air pollution impacting local conditions. In order to address the pervasive problems of low data quality and availability, this paper describes methodologies using machine learning to enable better decision making. Through the use of these innovative techniques, decision makers and stakeholders can have a more insight through these scientific findings on which to create city level air pollution action plans which focus resources on the solutions which can make the most difference.

unsplash

Introduction

It is universally recognized that air pollution is a pressing environmental challenge that has increased considerably in recent years, leading to a rise in premature deaths, threatening livelihoods and the sustainable development of the region, in particular in many cities in Asia and the Pacific where air pollution is a major public health hazard to an increasing urban population.

ESCAP research has shown that the resulting damages from air pollution disproportionately impact the low-income and marginalized communities, and this can be a crucial factor that exacerbates income inequalities. According to the World Health Organization, more than 60 per cent of all premature deaths from household air pollution in 2012 were among women and children.

However, many rapidly urbanizing cities in developing countries lack the resources and data to clearly identify the causes of air pollution impacting local conditions. In order to address the pervasive problems of low data quality and availability, this paper describes methodologies using machine learning to enable better decision making.

Through the use of these innovative techniques, decision makers and stakeholders can have a more insight through these scientific findings on which to create city level air pollution action plans which focus resources on the solutions which can make the most difference. Subsequent working papers will illustrate how to apply and utilize these scientific findings in the creation of customized, urban-level air pollution action plans.

Examples of key areas for attention include increased energy consumption and use of inefficient energy technologies in households and industrial processes, unmanaged and inefficient construction work (buildings/roads, etc.) in the region intensify air pollution, especially in urban areas, with high concentrations of particulate matter in cities. Therefore, better city planning with an emphasis on environmental priorities such as clean air is needed.

This focus would encourage clean energy, energy efficiencies, local industry standards, and other low-carbon development, transport policies, etc. By using these innovative research techniques, city planners in developing countries will be better informed of which interventions would be most suitable to improve the air quality for their citizens.

Predicting PM2.5 with Machine Learning

In order to demonstrate the usefulness of machine learning approaches to understanding the causes of air pollution, the city of Chiang Mai, Thailand will be used as a case study. This city has been selected because it regularly suffers from seasonal, severe air pollution combined with low levels of data availability, particularly over time.

Additionally, because such conditions are experienced in several other of the cities in the Southeast Asian region with the worst air pollution, the findings are likely to be of general use. As such, this city can usefully serve as a test case for this type of data science.

This methodology will need to be able to analyses the history and primary air pollution sources for Chiang Mai, together with transboundary aspects. However, it is necessary to examine the causal factors driving the seasonal pollution functions to put together action plans and policies that could have tangible impacts to solve the issue.

This analysis has created a data-science based model that can accurately predict different the outcome of different scenarios, specifically with reference to reducing the number of open burning hotspots at varying distance from the city center. The development of this statistical model involved building a system with the following criteria:

- The ability to identify the primary air pollution sources in Chiang Mai and determine their influence on AQI levels.

- Manipulate the data so that air pollution in Chiang Mai is predicted at an hourly rate. The R2-score (a statistical measure of how close the data are to the fitted regression line) should be in line with state-of-the-art standards to support confidence in the findings. In line with the analysis provided in Annex 1, a model accuracy of .75R2 is considered a satisfactory result.

- Predict the air pollution level by implementing the various scenarios. For instance, demonstrating how Chiang Mai's seasonal air pollution would change if agricultural burning were decreased by 40% in a 200km radius instead of a 20% reduction in a 400km radius and understand which scenario would lead to better outcomes.

The most common atmospheric modeling types used to measure and predict these data are: 1) Atmospheric Chemistry, 2) Dispersion, 3) Machine Learning. However, these models are designed around multi-processing approaches involving real-time and historic emission records and meteorological data.

Their effectiveness of models such as Hybrid Single-Particle Lagrangian Integrated Trajectory (HYSPLIT) can be limited, as it is difficult to calculate air mass movement and source apportionment when there are multiple sources of pollution, leading to scaling difficulties. This is one aspect in which current approaches to Atmospheric Chemistry and Dispersion struggle to sufficiently capture the nonlinearity relationship between air pollution and source emissions.

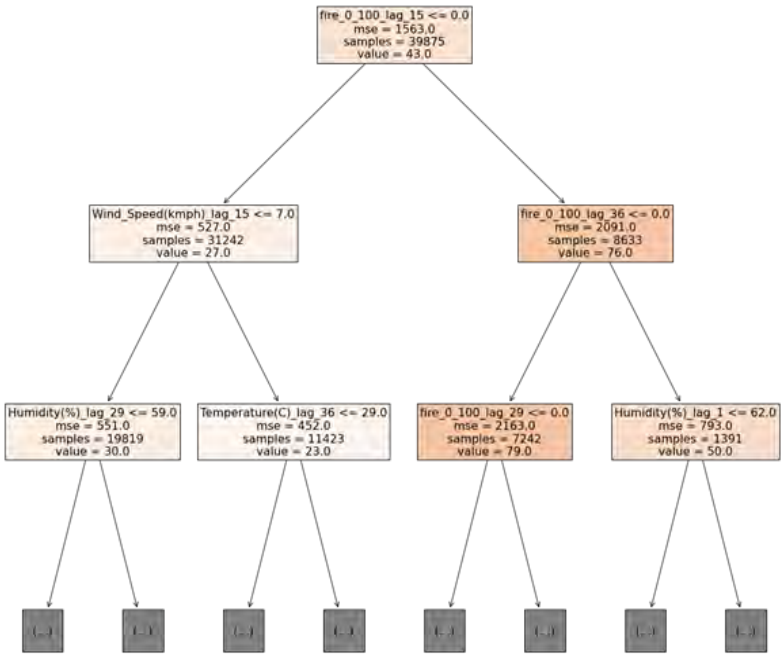
For facing this issue, machine learning models based on statistical algorithms seem favorable. Instead of focusing on physical and chemical processes, statistical models are rigidly based on historical data predictions. Machine learning is ideal for modeling air pollution levels because of its ability to handle many variables and non-linear relationships between independent and dependent variables.

A statistical approach is applied, which allows the Machine Learning Model, through supervised learning, to predict possible outcomes in different scenarios. For this step, a Random Forest Regressor (RF) demonstrates the most outstanding suitability because of its capacity to manage complex relationships. This RF model is composed of an ensemble of decision trees that predict the average behavior for a given set of conditions.

For example, a decision trees of which shoes to where for a given day is given by the condition of whether today is a weekend and the outside temperature. A similar idea is applied to predict the pollution level given the condition today and previous days. The figure below illustrates an example of one tree.

The model decides whether to increase or decrease the pollution level from level = 43 based on the number of hotspots in 100 km radius in the past 15 hours, then the decisions is split further down the node using the wind speed (left) or the number of hotspots in 100 km radius in the past 36 hours (right). This progress continues until the model used all the input data.

Figure 1: Random Forest Regressor Model for Chiang Mai



After obtaining a reasonable model, we can use the model in two ways. First, the capacity of the model to correctly predict the level of $PM_{2.5}$ air pollution is assessed, including the ability to rank the input feature by order of importance, which allow us to calculate contribution from various input (pollution sources). Second, the model can then be used simulate the relationship between meteorological data and air pollution with a high level of accuracy.

The process of applying this approach to predict the levels of $PM_{2.5}$ based on reduced levels of agricultural burning reduction for Chiang Mai was designed and implemented by building a data warehouse, data visualization to understand relationship between features and the pollution level, and model building, as described in the next sections.

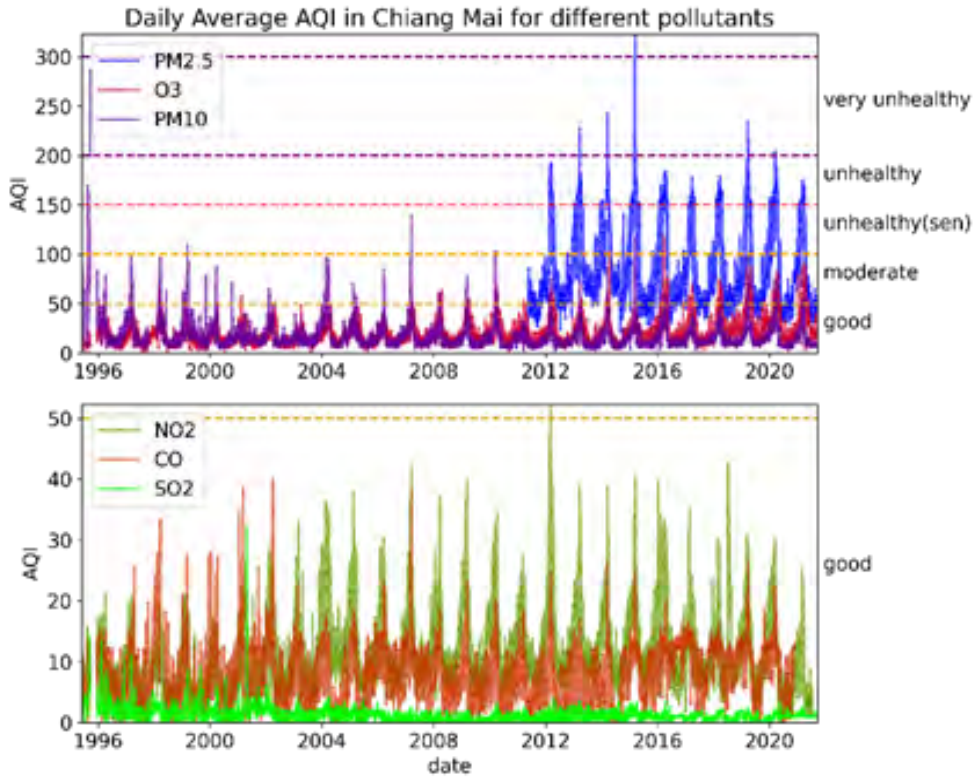
Data Warehousing

Data was gathered on chemical pollutant types, hotspots, weather, and other variables in constructing this data warehouse for Chiang Mai. Data sources include historical records, satellite readings, scrubbing from other sources, and even crowdsourcing. Some data was straightforward to obtain because of proper record-keeping, while other aspects were more complicated to obtain – but adjusted for with machine learning via feature engineering. For details of the data source, please see the appendix section.

An example of insufficient data that was adjusted for with machine learning is missing $PM_{2.5}$ historical data. As discussed in the section on monitoring air pollution in Chiang Mai, $PM_{2.5}$ was not tested there until 2012, as seen in the top graph of Figure 2. However, there is data since 1995 on PM_{10} and O_3 – two chemicals known to demonstrate a very close correlation to $PM_{2.5}$.

To increase the accuracy of $PM_{2.5}$ predictions, the model could be enhanced to use predict PM_{10} and O_3 first and infer the predicted levels of $PM_{2.5}$ (as seen in the lower graph of Figure 2). ESCAP will continue to pursue this line of analysis and publish further results in subsequent working papers.

Figure 2: Pollution Level Measures in Aggregate (United States AQI Index), Analysis; ESCAP 2021



Accounting for Meteorological Conditions

Wind speed has a considerable effect on the pollution level because it changes how the air circulates for a region, particularly for a mountainous area like Chiang Mai. High wind speeds help clean up the air; however, in the middle of the burning season, the wind could also bring the pollutants into the city.

For Chiang Mai, wind speed in the winter months is at its lowest level of about a 5 km/hr speed – it is at its fastest in the monsoon season (April to July). Figure 3 compares the annual levels of $PM_{2.5}$ in Chiang Mai against the wind speed. As seen in this graph, the annual pattern negatively correlates with the air pollution level throughout the year.

In Figure 4 the hourly value of the wind speed and hourly density of $PM_{2.5}$ in Chiang Mai were compared. It shows that the wind speed peaks in the late afternoon, and the pollution level drops simultaneously. Meteorological effects are notoriously difficult to simulate accurately, compounded by the natural geographic complexity of the area around Chiang Mai.

For example, low winds are observed in winter under influence of high pressure ridge when subsidence and radiative inversion are enhanced and the weather is dry, leading to thermal inversions which concentrate $PM_{2.5}$ particles in the area. Therefore, it is important to consider the totality of meteorological effects in order to make precise predictions. In this context, the wind can have an impact on how and when pollution moves through the region. The model considers these factors as part of air pollution's predictive factors, including specific weather pattern data to help forecast scenarios accurately.

Figure 3: $PM_{2.5}$ (Solid Blue Line) Hourly Patterns Over A 24 Hour Period.

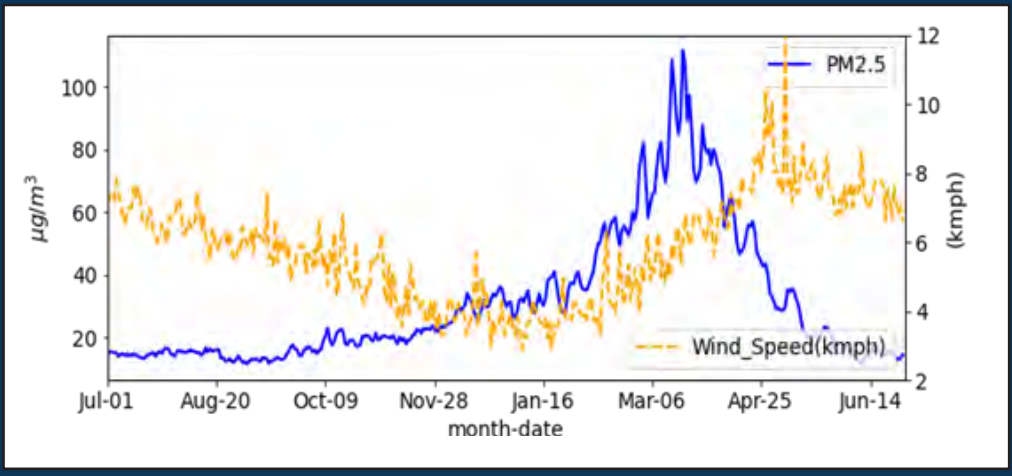
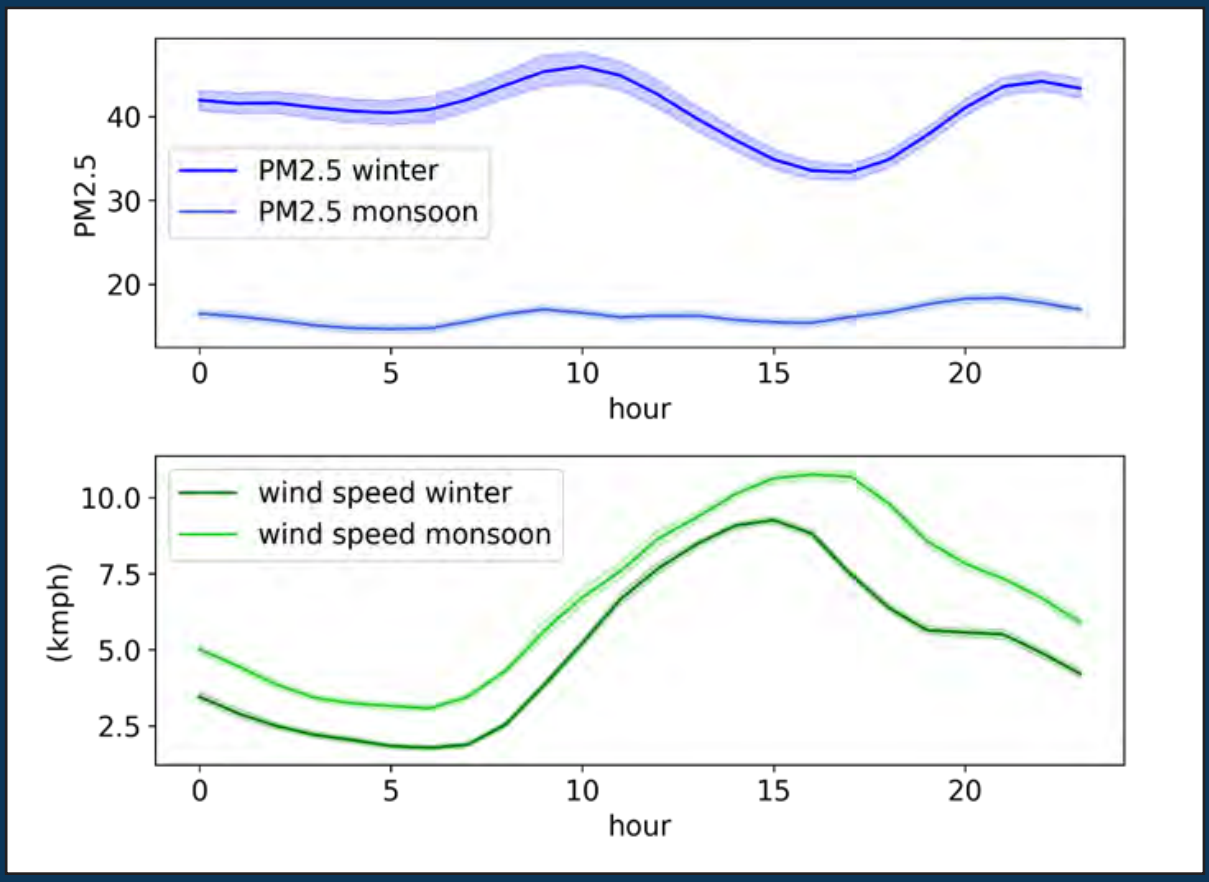
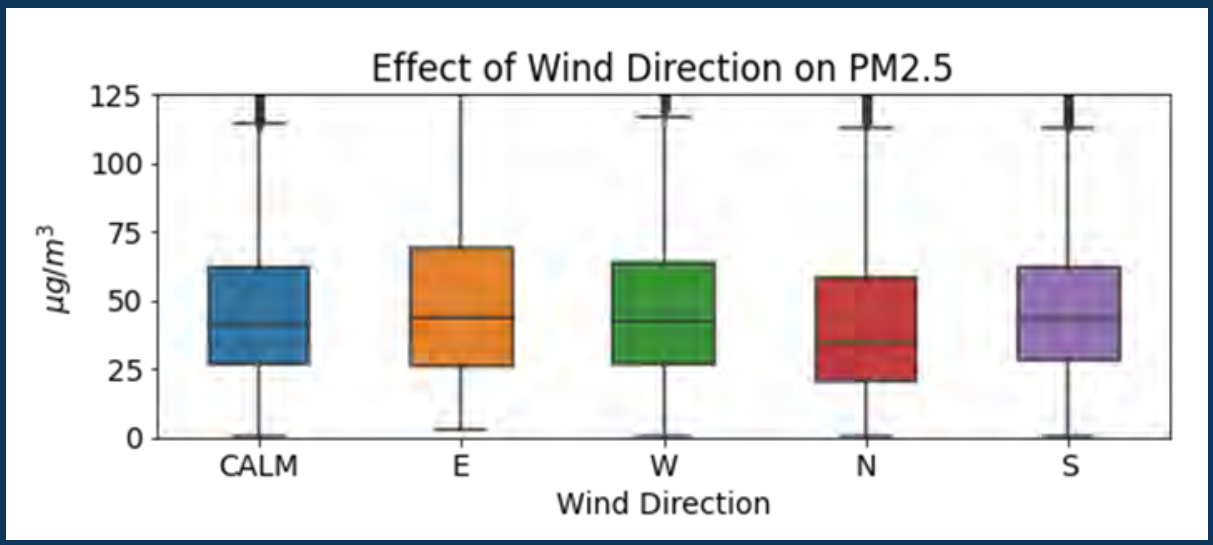


Figure 4: Wind speed (solid green line) hourly patterns over a 24-hour period, by season



In this plot, the hourly behavior for winter and monsoon season are plotted in separate lines, highlighting the relationship between these data points in winter (pollution) vs monsoon seasons. The hourly behaviors are moderated by windspeed, which reduces the pollution to be lowest around 3pm. Therefore, it can be seen that wind speed is a mitigating factor, but not a primary determinate of pollution levels.

Figure 5: Effect of Wind Direction on $PM_{2.5}$



Building the Air Pollution Model

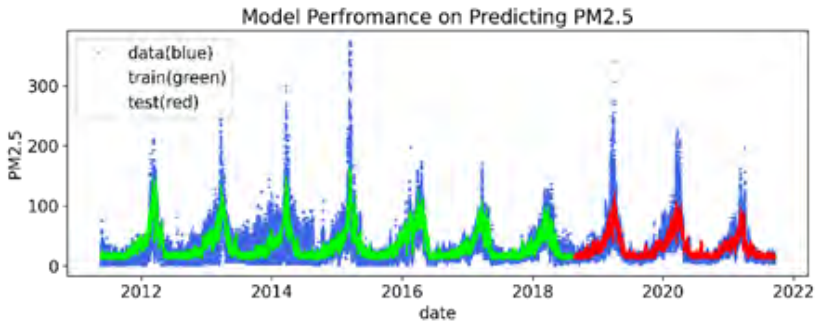
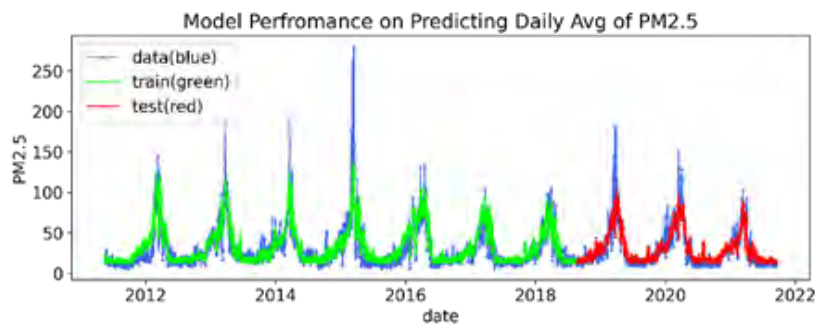
We build random forest regressor model to predict hourly $PM_{2.5}$ level using weather, burning hotspots in 900 km radius, and datetime information. The $PM_{2.5}$ level were averaged from two PCD station (35t ad 36t) located in downtown Chiang Mai. Categorical data such as wind directions and holiday are one-hot encoded. The model's performance is evaluated by measuring its objective performance and testing it against previously unseen data.

Model training allows us to make a prediction correctly as often as possible. The data is split into train, validation, and prediction datasets by order of occurrence using a splitting ratio of 4:3:3. This gives the training dataset between 2012 and mid-2015, the validation dataset between mid-2015 to 2019 and the prediction dataset from 2019 – current.

The validation dataset is the used to optimize the model parameters, and later merged with the training dataset for final model training. The prediction set is used for model prediction and creating simulated scenarios. Prediction and scenario simulation is undertaken with daily average values to provide higher model accuracy.

The figure below shows the model hourly and daily prediction accuracy for Chiang Mai. The daily prediction was obtain by averaging the hourly prediction, which slightly improve the model performance. The blue lines are the actual data, the green lines are training data, and the red lines are predictions.

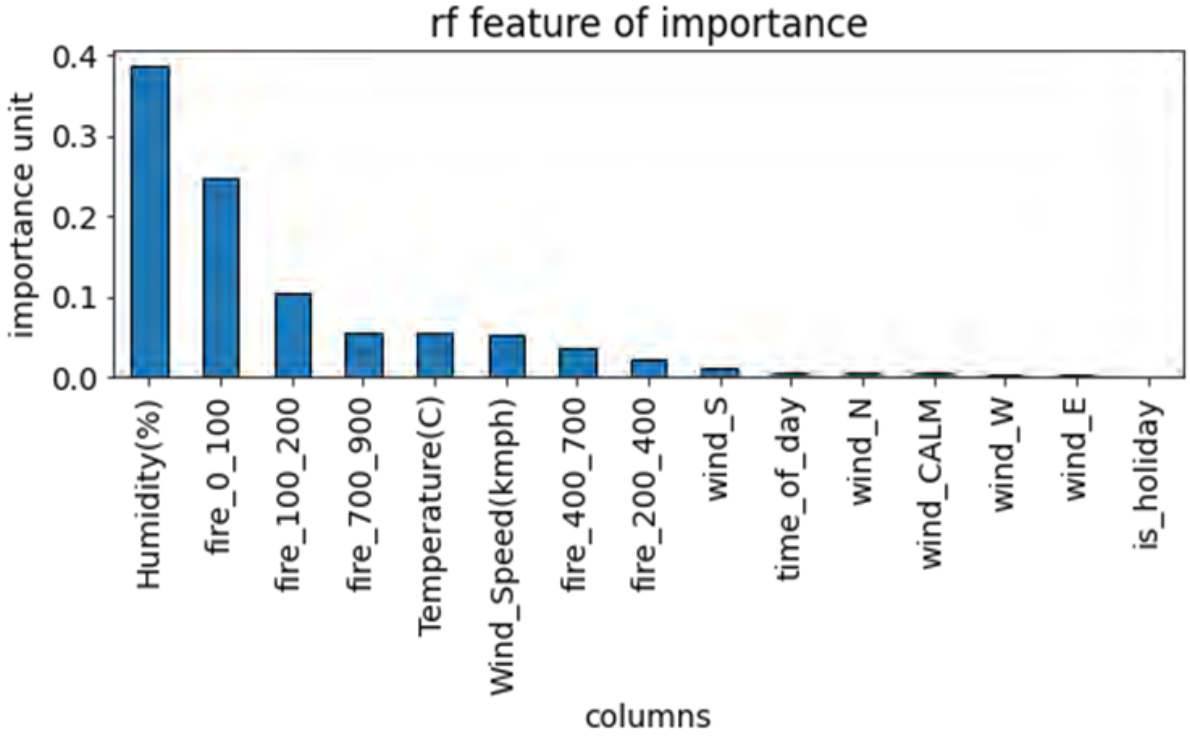
Figure 6: Hourly and Daily Prediction Accuracy for Chiang Mai

Optimization	R2-score	Error	Model Performance
Hourly Prediction	0.71	±15	
Daily Prediction	0.78	±12	

When the model was not allowed to see specific data sets, it still had a reliable prediction score (R2 = 0.66). The prediction error is +/- 16 m3, which is absorbed in the confidence band's error. This outcome gives the model confidence to begin making and analyzing predictions that can guide the region towards clearer skies.

The figure below shows the model feature of importance. In addition to metrological condition, the feature clearly indicates that hotspots are importance features, which indicates high contribution from agricultural burning activities. This is in agreement with various studies.

Figure 7: Feature of Importance



To quantify the effect of agricultural burning toward $PM_{2.5}$ level in Chiang Mai, we perform what we called a statistical prediction in order. The statistical prediction is performed by obtaining a range of possible weather data and fire activities from the training dataset. The range of possible independent variables is used to predict a range of possible hourly $PM_{2.5}$ levels. The values from the same date-time are averaged to produce final prediction values.

For example, to predict the range of possible $PM_{2.5}$ values on December 20, 2022, this approach will use the conditions around December 20, during previous years, feed this behavior into the model and predict the $PM_{2.5}$ level. The weakness of this approach is that the Random Forrest regressor can only predict values which it is has seen before.

If there is a year where the weather and fire patters are very different from the historical data, the prediction will be inaccurate. It is expected that events such as the impact of COVID19 mitigation measures will present analytical challenges for training and development of air pollution models using machine learning.

The figure below compares the daily average of actual $PM_{2.5}$ (blue) and the statistical prediction (red). The statistical prediction is very similar to the actual data values – meaning that our model can accurately predict $PM_{2.5}$based on multiple variables and situations.

Figure 8: Actual $PM_{2.5}$ Level In The Prediction Dataset (Blue) Between 1 October 2017 And 15 June 2020 And The Statistical Prediction Value (Red)

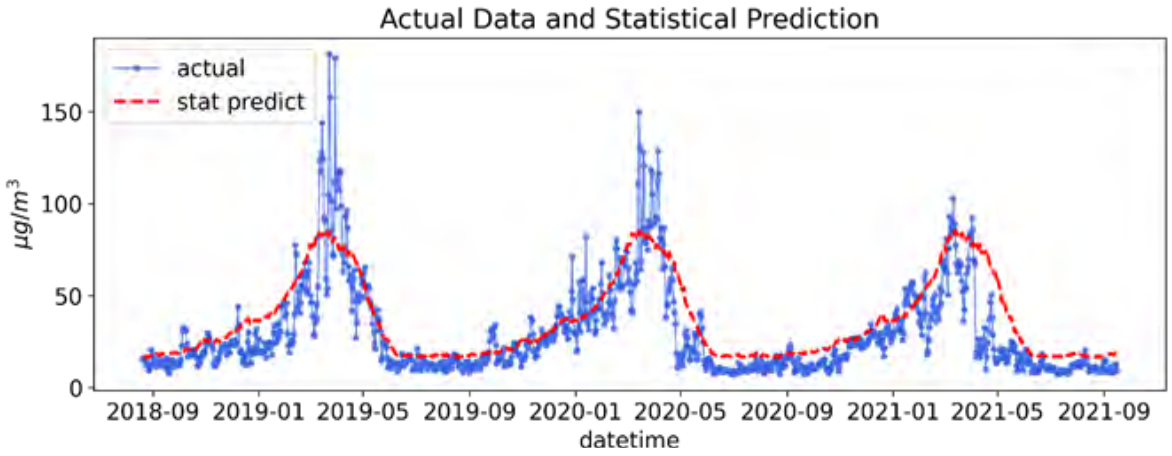


Figure 9: Seasonal Pattern of Data and Statistical Prediction

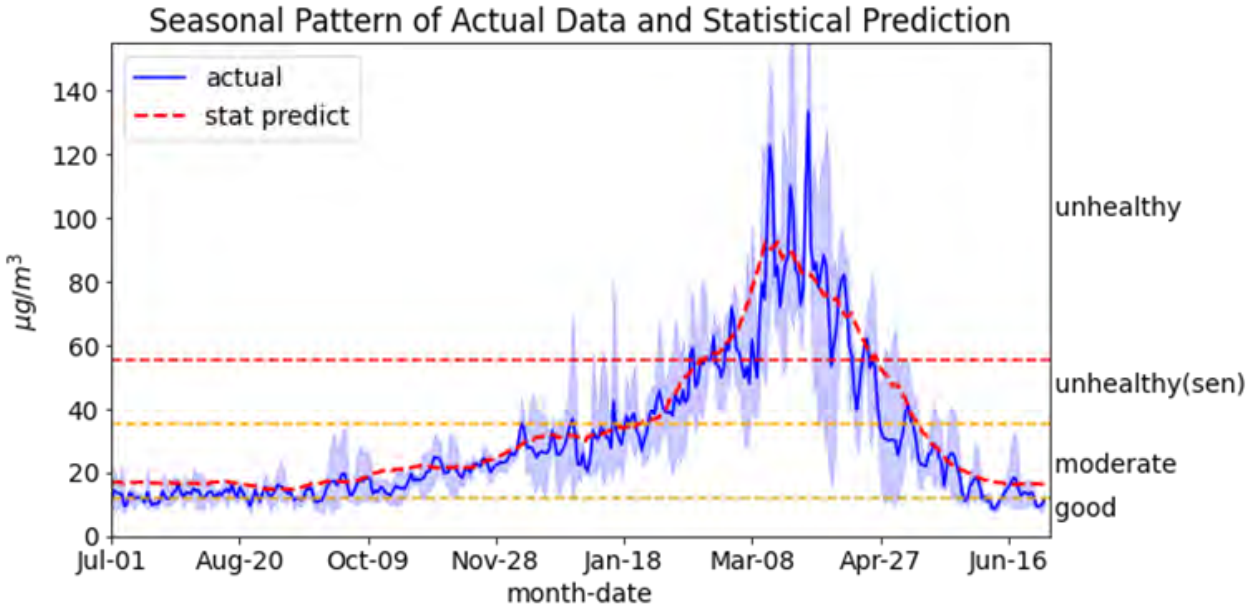
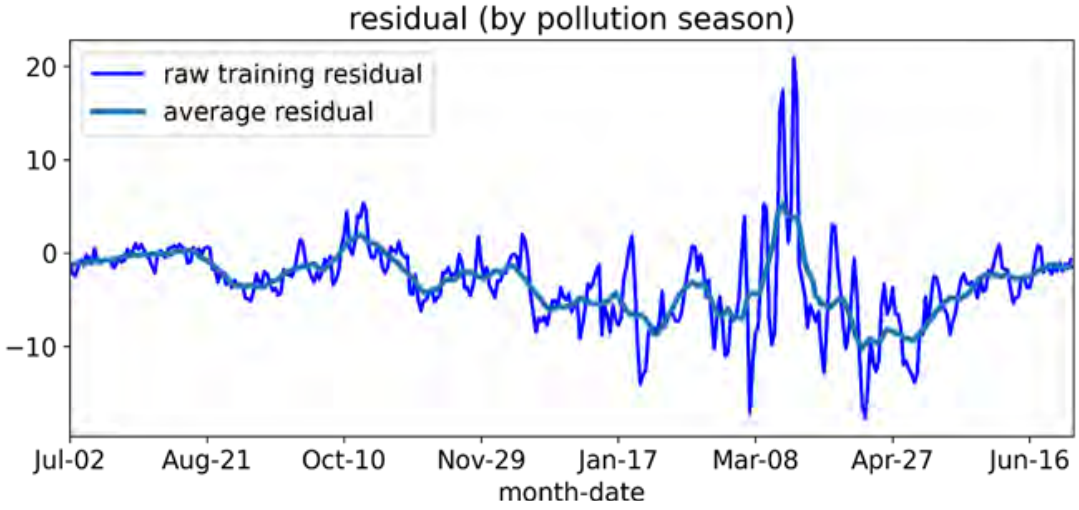


Figure 9: Raw Training Residual and Average Residual



Conclusion

The results of this research indicate that the use of machine learning can serve as viable mechanism for supporting reliable decision making for cities facing local air quality challenges. By utilizing machine learning, existing data sets can be used in innovative ways to identify the principal components of local air pollution and thereby enable targeted local action.

These approaches work best when air pollution factors follow predictable trends, enabling prediction within historically observed parameters. However, these methodologies struggle when large disruptive events occur which the machine learning model has not previously observed in the data. Policy makers and all stakeholders should be fully aware of these strengths and limitations as they seek to utilize these innovative approaches.

References

Araki, S., Shima, M. and Yamamoto, K. (2018) 'Spatiotemporal land use random forest model for estimating metropolitan NO2 exposure in Japan', Science of The Total Environment, 634, pp. 1269–1277. doi: 10.1016/j.scitotenv.2018.03.324.

Castelli, M. et al. (2020) A Machine Learning Approach to Predict Air Quality in California, Complexity. Hindawi. doi: https://doi.org/10.1155/2020/8049504.

Delavar, M. R. et al. (2019) 'A Novel Method for Improving Air Pollution Prediction Based on Machine Learning Approaches: A Case Study Applied to the Capital City of Tehran', ISPRS International Journal of Geo-Information, 8(2), p. 99. doi: 10.3390/ijgi8020099.

Gómez-Losada, Á. et al. (2019) 'A data science approach for spatiotemporal modelling of low and resident air pollution in Madrid (Spain): Implications for epidemiological studies', Computers, Environment and Urban Systems, 75, pp. 1–11. doi: 10.1016/j.compenvurbsys.2018.12.005.

Iskandaryan, D., Ramos, F. and Trilles, S. (2020) 'Air Quality Prediction in Smart Cities Using Machine Learning Technologies Based on Sensor Data: A Review', Applied Sciences, 10(7), p. 2401. doi: 10.3390/app10072401.

Liu, J., Weng, F. and Li, Z. (2019) 'Satellite-based PM2.5 estimation directly from reflectance at the top of the atmosphere using a machine learning algorithm', Atmospheric Environment, 208, pp. 113–122. doi: 10.1016/j.atmosenv.2019.04.002.

Ma, Z. et al. (2019) 'Effects of air pollution control policies on PM 2.5 pollution improvement in China from 2005 to 2017: A satellite-based perspective', Atmospheric Chemistry and Physics, 19, pp. 6861–6877. doi: 10.5194/acp-19-6861-2019.

Martınez-Espana, R. et al. (2018) 'Air-Pollution Prediction in Smart Cities through Machine Learning Methods: A Case of Study in Murcia, Spain', Journal of Universal Computer Science, 24(3), p. 16.

Masih, A. (2019) 'Machine learning algorithms in air quality modeling', Global Journal of Environmental Science and Management, 5(4), pp. 515–534. doi: 10.22034/GJESM.2019.04.10.

Stafoggia, M. et al. (2019) 'Estimation of daily PM10 and PM2.5 concentrations in Italy, 2013–2015, using a spatiotemporal land-use random-forest model', Environment International, 124, pp. 170–179. doi: 10.1016/j.envint.2019.01.016.

Xiao, Q. et al. (2018) 'An Ensemble Machine-Learning Model To Predict Historical PM2.5 Concentrations in China from Satellite Data', Environmental Science & Technology. doi: 10.1021/acs.est.8b02917.

Xue, T. et al. (2019) 'Spatiotemporal continuous estimates of PM2.5 concentrations in China, 2000–2016: A machine learning method with inputs from satellites, chemical transport model, and ground observations', Environment International, 123, pp. 345–357. doi: 10.1016/j.envint.2018.11.075.

Xue, Y. et al. (2020) 'Hourly PM2.5 Estimation over Central and Eastern China Based on Himawari-8 Data', Remote Sensing, 12(5), p. 855. doi: 10.3390/rs12050855.

Zani, N. B. et al. (2020) 'Long-term satellite-based estimates of air quality and premature mortality in Equatorial Asia through deep neural networks', Environmental Research Letters, 15(10), p. 104088. doi: 10.1088/1748-9326/abb733.

Annex 1: Review of Methodologies for Use Data Science in Air Quality Applications

Aims:

I) Review the existing literature to establish a benchmark based on consensus within peer reviewed air pollution modelling research for a measure of the strength of association between hindcasted pollution levels generated by machine learning models and historical data, as described by the Pearson's correlation coefficient or '*r*' value, that is considered meaningful.

II) Review the existing literature to gain an understanding of the best practices and methodological lessons learned regarding the use of machine learning models and data science for predicting air quality.

Objectives:

I) To draw upon the existing literature to make case for whether our model is producing results that can be considered meaningful/significant.

II) Incorporate and learn from methodological best practices in our own research so as to improve the validity of the results and identify areas for further investigation/data collection.

An overview of the strength of association (r2 value) between modelled air quality levels and historical levels in the literature

Methodological best practices when using data science and machine learning models for air quality predictions

Study	Aim	Geographic Area	Time	Pollutants Modelled/Measured	Input parameters	Machine learning methods	Prediction methods, errors and *r* value	Other key findings
(Delavar *et al.*, 2019)	Assess prediction models to determine PM_{10} and $PM_{2.5}$ pollution concentrations in Tehran.	Tehran, Iran		PM_{10}, $PM_{2.5}$	Day of week, month of year, topography, meteorology, pollutant rate of two nearest neighbours	Regression support vector machine, geographically weighted regression, artificial neural network,	"The most reliable algorithm for the prediction of air pollution was autoregressive nonlinear neural network with external input using the proposed prediction model, where its one-day prediction error reached 1.79 $\mu g/m^3$."	"a genetic algorithm was used with data for day of week, month of year, topography, wind direction, maximum temperature and pollutant rate of the two nearest neighbours which were identified as the most effective parameters in the prediction of air pollution."
(Martınez-Espana *et al.*, 2018)	Analyse the performance of several machine learning methods for predicting the O_3 levels in the region of Murcia.	Murcia, Spain		O_3	Nitrogen Monoxide (NO), Nitrogen Dioxide (NO_2), Sulfur Dioxide(SO_2), Total Nitrogen Oxides (NOX), Particulate matter in suspension < 10µ g(PM_{10}), Benzeno (C_6H_6), Toluene (C_7H_8), Xileno (XIL), Temperature (TMP °C), Relative Humidity (HR) measured in %, wind direction (DD) in grades, Wind speed (VV) in meters per second (m/s), Atmospheric pressure (PRB) in bar and Solar Radiation (RS) in watts per square meter (w/m^2).	Bagging, Random Committee, Random Forest, a decision tree and an instance-based technique.	The quality and reliability of the models are measured by the Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE). Finally, the robustness and suitability are evaluated through the determination coefficient R^2.	"The technique that obtains the best fit in general is Random Forest, being this assertion validated by statistical tests. *The results indicate an R^2 setting between 80% and 90% overall and an O_3 prediction error less than 11 μ/m^3.* It is also important to note that among the parameters that most influence the ozone prediction we have found climatic variables related to temperature, humidity and wind. In addition, hierarchical clustering indicates that the air-pollution monitoring areas in the Region of Murcia can be divided into two zones only so as to create two general O_3 prediction models for the entire Region."
(Masih, 2019)	Highlight the underlying principles of machine learning techniques and their role in				The paper assesses the following aspects of the selected studies; 1) motivation of the work; 2) type of	Machine learning algorithms categorised into; Regression, ANNs, SVM,	The most popular evaluation criteria are the correlation coefficient (R^2), Mean Absolute Error (MAE), Root Mean Square Error	Machine learning techniques are mainly conducted in Europe and America. A multicomponent analysis also showed that pollution estimation is generally performed with ensemble

	prediction performance based on a review of 38 studies.				modeling i.e. forecast or estimation; 3) historical data of predictive features; 4) type of the machine learning algorithms employed e.g. Regression, ANNs, SVM, ensemble learning techniques, or hybrid models; 5) nature of prediction i.e. if a specific pollutant (PM_{10}, $PM_{2.5}$, NOx, O_3, SO_2 etc.) is predicted or air quality index (AQI) in general is calculated to learn pollution level; 6) geographic location where the study is performed; 7) time span and the number of data stations used; 8) evaluation methods to assess the model performance. The assessment is based on a comparison between model accuracy and the prediction of the actual value.	ensemble learning techniques, and hybrid models. By 2018 the three most popular machine learning algorithms (in order) were 'ensmble', NN and SVM.	(RMSE) and Relative Absolute Error (RAE). The various findings of the studies in relation to *r* value are as follows: i) Result obtained from one study recommend that the performance of proposed model (R^2=0.69) is better than that of obtainable by CTMs. ii) The average accuracy of another study calculated in terms of correlation coefficient (R^2) at all 31 stations was found to be 0.62. iii) The model of a further study performed well to estimate the concentration of CO and NO with R^2 =0.95 when FIS ensemble with RF and ANN respectively. iv) The performance of RF in another study has been superior having R^2 value equal to 0.85 as compared to Random Committee (0.83), Bagging (0.82) and Decision Tree (0.82). !SEE *Group 3: Satellite image and sensor-based monitoring*	learning and linear regression based approaches, whereas forecasting tasks tend to implement neural networks and support vector machines based algorithms.

							techniques to enhance pollution prediction accuracy for further values	
(Iskandaryan, Ramos and Trilles, 2020)	Comparative review of 41 studies related to air pollution prediction that use machine learning algorithms that are based on sensor data in the context of smart cities				Investigated the research strategies and process of 41 papers using the following questions: Which machine learning techniques are used to predict air quality in the smart city domain? How do the proposed methods handle different types of data in terms of air pollution? What temporal resolutions were analysed with the proposed techniques? The main features extracted and compared from each paper were; Year, Case Study, Methods, Algorithms, Evaluation Metrics, Prediction Target, Time Granularity, Data Rates, Dataset Types, Open Data, Advantages and	Regarding machine learning techniques, the studies used neural networks (38%), regression (24%), ensemble (22%), hybrid (11%) models	Overall, 29 metrics were applied, from which MAE and RMSE were the most used metrics, each of them being applied in 24 papers. The following observations of R^2 were made in relation to one study: "above 0.75R^2 was considered a satisfactory result and all the methods obtained higher from this threshold."	1. Authors are applying sophisticated rather than simple machine learning techniques 2. Most case studies are coming from China 3. $PM_{2.5}$ is the most common prediction target 4. In 41% of the studies the authors predicted air quality for the following day 5. 66% of the studies used hourly data 6. 49% of the studies used open source data and this has increased since 2016 7. External factors such as weather conditions, spatial characteristics and temporal features are key factors in the prediction of air quality

					Limitation/Future Work !!SEE: Table 2 'Features of selected papers' for these features in tabular form, including which papers used R^2 as an evaluation metric			
(Castelli *et al.*, 2020)	Use Support Vector Regression (SVR) to forecast pollutant and particulate levels and to predict the air quality index (AQI)	California, USA		O_3, CO, SO_2, and AQI	The dataset was extracted from EPA's Air Quality, containing hourly data separated by the pollutant or parameter being measured under the categories 'meteorological conditions', 'criteria gases', and 'particulates' —CO, SO_2, NO_2, ozone, $PM_{2.5}$, temperature, humidity, and wind from the state of California. The hourly events were collected between January 1, 2016, and May 1, 2018.	SVR with radial basis function (RBF) was the type of kernel that allowed the most accurate predictions. Using the whole set of available variables revealed a more successful strategy than selecting features using principal component analysis.	SVR with RBF kernel allows us to accurately predict hourly pollutant concentrations, like carbon monoxide, sulfur dioxide, nitrogen dioxide, ground-level ozone, and particulate matter 2.5, as well as the hourly AQI for the state of California. Classification into six AQI categories defined by the US Environmental Protection Agency was performed with an *accuracy of 94.1% on unseen validation data.* MAE, RMSE, nRMSE, and R^2 were used for checking the error matrics of the forecasting model for each pollutant. *For $PM_{2.5}$ the R^2 value for the SVR-RBF forecasting model was 0.937 for the training set and 0.647 for the validation set.*	The studied method produced a suitable model of the hourly atmospheric pollution, allowing us to obtain, generally, good accuracy in modeling pollutant concentrations like O_3, CO, and SO_2, as well as the hourly AQI for the state of California.
(Gómez-Losada *et al.*, 2019)	Use hourly time series of four key pollutants to estimate the exposure to background pollution in Madrid from 2001-2017, accounting for the temporal and spatial variability of exposure.	Madrid Spain	2001-2017	NO_2, O3, PM_{10}, SO_2	Hourly time series for each year from 2001 to 2017, of NO_2, O3, PM_{10} and SO_2 obtained from 38 monitoring stations.	The background air pollution concentration was estimated independently on annual time series of NO_2, O_3, PM_{10} and SO_2 pollutants at hourly resolution and summarized as an annual average concentration, using Hidden Markov Models (HMM). The spatial distribution of the background air pollution was estimated after averaging the estimates of one non-geostatistical (inverse distance weighting –IDW-) and one geostatistical (ordinary kriging - OK-) spatial interpolation methods.		It has been seen that these models provide a comprehensive overview, and probably a robust approach, of the complex estimation of background air pollution, which represents a chronic level of exposure to which the population is permanently exposed in cities.
(Xiao *et al.*, 2018)	Propose and test a new ensemble machine learning approach to provide reliable	China		$PM_{2.5}$		The modeling domain, China, was divided into seven regions using a spatial clustering	The ensemble prediction characterized the spatiotemporal distribution of daily $PM_{2.5}$ well with the cross-validation (CV) R^2 (RMSE) of 0.79 (21 $\mu g/m^3$).	Our hindcast modeling system allows for the construction of unbiased historical PM2.5 levels.

	$PM_{2.5}$ hindcast capabilities					method to control for unobserved spatial heterogeneity. A set of machine learning models including random forest, generalized additive model, and extreme gradient boosting were trained in each region separately. Finally, a generalized additive ensemble model was developed to combine predictions from different algorithms.	The cluster-based subregion models outperformed national models and improved the CV R^2 by ~0.05. Compared with previous studies, our model provided more accurate out-of-range predictions at the daily level (R^2 = 0.58, RMSE = 29 μg/m^3) and monthly level (R^2 = 0.76, RMSE = 16 μg/m^3).	
(Ma *et al.*, 2019)	Assess the effects of air pollution control policies from 2005 – 2017 on $PM_{2.5}$ levels using satellite remote sensing data, assimilated meteorology and land use data with statistical models	China	2005-2017	$PM_{2.5}$	Data used includes: ground-monitored $PM_{2.5}$ concentrations (μg m^{-3}), Aqua MODIS Collection 6 Dark Target (DT) AOD and Deep Blue (DB) AOD data, planetary boundary layer height (PBLH, 100 m), wind speed (WS, m s^{-1}) at 10 m above the ground, mean relative humidity in	A two-stage statistical model was developed for each year separately from 2014 to 2017. The first-stage linear mixed effects (LME) model included day-specific random intercepts and slopes for AOD, season-specific	For the monthly mean concentrations calculated from at least six daily $PM_{2.5}$ predictions, *the validation R^2 values range from 0.75 to 0.81*. The results show that *the overall prediction accuracy of the models from 2014 to 2017 is satisfying.*	The uneven spatial distribution of ground $PM_{2.5}$ monitors is a source of uncertainty in statistical models, however "high model performances" have been achieved in this study and previous similar studies e.g. Geng et al. (2015) estimated long-term $PM_{2.5}$ concentrations in China using a scaling method and found *the validation R^2 of $PM_{2.5}$ predictions was 0.72* compared to the 5-month averaged ground $PM_{2.5}$ concentrations for January–
					surface pressure (PS, hPa), precipitation of the previous day (Precip_Lag1; mm), MODIS active fire spots, urban cover (%), and forest cover (%) A previously developed two stage statistical model using MODIS Collection 6 AOD and assimilated meteorology, land use data, and ground monitored $PM_{2.5}$ concentrations with a cross validation R^2 *of 0.79 for daily estimates and 0.73 for monthly levels* was used to hindcast missing data (2004-2012). Where sufficient ground-monitored $PM_{2.5}$ data was available (2014-2017) a separate $PM_{2.5}$–AOD statistical model for each year to estimate the spatially resolved (0.1° resolution) $PM_{2.5}$ concentrations was developed.	meteorological variables, and fixed slope for precipitation and fire spots. The second-stage generalized additive model (GAM) established the relationship between the residuals of the first stage model and smooth terms of geographical coordinates, forest and urban cover. To evaluate the model over-fitting, 10-fold cross-validation (CV) was used. Monthly mean $PM_{2.5}$ concentrations for each grid cell were calculated to perform the time series analysis.		estimates combining scaling and statistical methods shows that their *validation R^2 of long-term average $PM_{2.5}$ was 0.67* for their first-stage scaling method (van Donkelaar et al., 2016).
(Xue *et al.*, 2019)	Develop a new machine learning model with high-dimensional expansion (HD-	China	2000-2016	$PM_{2.5}$	$PM_{2.5}$ monitoring data from 1497 sites across China, satellite AOD and auxiliary variables - (MODIS	To derive the final estimates of historical $PM_{2.5}$ a two-stage method was	The model was trained with data from 2013 to 2016 and its performance evaluated using annually-iterated cross-validation, which	This study produced AOD-based estimates of historical $PM_{2.5}$ with complete spatiotemporal coverage, which were evidenced as accurate, particularly in middle

	expansion) of numerous predictors (including AOD and other satellite covariates, meteorological variables and CTM simulations) to predict daily $PM_{2.5}$ concentrations during 2000-2016 across China and estimate long-term trends.				level 2 products of AOD at a spatial resolution of 3 km (MOD04_3K and MYD04_3K) from the earth observing satellites Terra (2000–2016) and Aqua (2002–2016), WRF and CMAQ simulations - simulated maps of meteorological variables including temperature, wind speed and direction, relative humidity, pressure, and planet boundary layer height using the WRF model - Driven by the outputs of the WRF model, we also simulated concentrations of $PM_{2.5}$ and their five major components, i.e., NO_3^-,SO_4^{2-}, elemental carbon (EC), organic carbon (OC) and NH_4^+, based on the 2000–2016 inventories from the Multi-resolution Emission Inventory of China model (http://meicmodel.org/) and using the CMAQ model.	designed and applied to the whole study domain. In stage I, we developed two separate ML models: with satellite AOD (ML:CMAQ + AOD) and without satellite AOD (ML: CMAQ). Estimates from the former model ($PM_{2.5}^{ML: CMAQ+AOD}$) were more accurate, but the latter ($PM_{2.5}^{ML: CMAQ}$) had complete spatiotemporal coverage. In stage II, to interpolate the missing values of $PM_{2.5}^{ML: CMAQ+AOD}$, we developed a generalized additive model (GAM) with an offset of $PM_{2.5}^{ML:CMAQ}$.	iteratively held out the in-situ observations for a whole calendar year (as testing data) to examine the predictions from a model trained by the rest of the observations. *Estimates were found to be in good agreement with in-situ observations, with R^2 of 0.61, 0.68, and 0.75 for daily, monthly and annual averages, respectively.* To interpolate the missing predictions due to incomplete AOD data, we incorporated a generalized additive model into the ML model. *The two-stage estimates of $PM_{2.5}$ sacrificed the prediction accuracy on a daily timescale (R^2= 0.55), but achieved complete spatiotemporal coverage and improved the accuracy of monthly (R^2= 0.71) and annual (R^2= 0.77) averages.*	and long term. The products could support large-scale epidemiological studies and risk assessments of ambient $PM_{2.5}$ in China and can be accessed via the website (http://www.meicmodel.org/dataset-phd.html).
					All satellite data, except nightlight data, were downloaded from https://search.earthdata.nasa.gov. Nightlight data were downloaded from the NCEI website https://ngdc.noaa.gov/eog/dmsp.html.			
(Zani *et al.*, 2020)	Use long-term satellite-based observations of chemical and physical parameters integrated with ground-based data of PM concentrations to develop a machine learning approach for continuous PM_{10} monitoring in the region and estimate long-term premature mortality.	Equatorial Asia	2005-2015	PM_{10}, $PM_{2.5}$	Long-term observations of PM_{10} concentrations from a network *comprising 52 ground-level monitoring stations* across Peninsular Malaysia and Malaysian Borneo. *Satellite retrievals of aerosol properties, trace gases and land use* are used to develop a satellite-based proxy able to capture the variability of ground-level PM_{10}; *AOD data* from MODIS 1x1km resolution from the multiangle implementation of atmospheric correction (MAIAC)	DNNs were trained using satellite data extracted with a 20 km averaging radius around each station and aggregated with a 7-d moving average. DNNs are trained on a randomly extracted 80% subset of all available data; then, validation is performed on the remaining 20%. The overall performance of DNN is evaluated with the Pearson (r) and Spearman (ρ) correlation	Across the seasons and across moving averages (MA) and monthly mean (MM) models *DNN performance ranged between R^2 0.643 (spring 7-d moving average PM_{10}) and R^2 0.905 (fall monthly means).*	DNNs show enhanced predictive skills and lower bias compared to the LM approach due to their ability to capture significant non-linear mechanisms and variable interactions dictating PM_{10} concentrations. Seasonally, the dry period brought by the Southwest monsoon, possibly enhanced by ENSO, is associated with higher PM_{10} and lead to an improved model performance during summer and fall. Higher PM_{10} in the fall is associated with the less frequent wet deposition processes and enhanced wildfires occurrence, which determined the extreme haze events recorded in 2006 and 2015. Future research should focus on including additional meteorological variables (e.g. wind speed/direction, ground temperature and planetary

					algorithm, *Column water vapor (CWV)* retrieved as a daily 1 km × 1 km resolution data from MODIS on Terra and Aqua and corrected through MAIAC, *Normalized difference vegetation index (NDVI)* retrieved as monthly Level 3 (L3) 1 km × 1 km resolution quantity from MODIS onboard Aqua, *Carbon monoxide (CO)* tropospheric amount derived from the Measurements of Pollution in the Troposphere (MOPITT) sensor onboard Terra, *Urban fraction (UF)* from the Consensus Landcover dataset, *Tropospheric amounts of trace gases and ultraviolet (UV) irradiance* measured by the ozone monitoring instrument (OMI) including; *NO_2, SO_2, HCHO and UV, and finally population data* retrieved from the Socioeconomic Data and Application	coefficients between observed and modeled values. To predict annual mean PM_{10} spatial fields, other DNNs are trained on monthly aggregated data. PM_{10} patterns are thus predicted from the monthly aggregated satellite variables homogenized to the reference grid at 0.25° × 0.25° resolution.		boundary layer height), which may enable description of aerosol vertical profiles and transport and dispersion processes on the local scale. While this study shows the skill of DNN in predicting surface PM concentrations, future investigations could quantify the predictive skills of a different machine learning approaches (e.g. random forest, gradient boosting machine or mixed models). *The annual PM_{10} and $PM_{2.5}$ maps reveal significant spatial and inter-annual patterns related to both anthropogenic drivers and wildfires*. The estimated health impacts indicate that *metropolitan areas remain the most affected, due to the combined effect of numerous anthropogenic emissions and high population density*. Conversely, *the effect of wildfires dominates on the regional scale*, as indicated by the strong inter-annual variability in the number of premature deaths over the region, which are significantly higher during fire years than adjacent non-fire years.
					Center (SEDAC) census.			
(Araki, Shima and Yamamoto, 2018, p. 2)	To develop a spatiotemporal land use random forest (LURF) model of the monthly mean NO_2 concentrations in a metropolitan area of Japan.	Amagasaki, Osaka, and Kobe Cities, Japan		NO_2	Air quality measurements (2011-2014) are obtained from monitoring stations, some located near heavy traffic to monitor automobile exhaust and some in locations that are not directly affected by specific emission sources – hourly mean concentrations were used to calculate monthly mean values. Data sets were chosen accounting for factors that affect the spatial distribution of air pollutants such as emission, advection and deposition. Green area ratio, road length, emission intensity and meteorological parameters, and satellite-derived NO_2 data were calculated for each grid cell.	*A spatiotemporal LURF* was constructed using a variable selection method proposed by Genuer et al. (2015). *A spatiotemporal land use regression model (LUR)* was also constructed based on a supervised stepwise selection procedure previously used to develop LUR models for NO_2 in Europe (Beelen et al., 2013).	The prediction accuracy of the LURF model is evaluated through a leave-one-monitor-out cross validation. *A high R^2 value of 0.79 is obtained, which is better than that of the conventional land use regression model using linear regression (R^2 of 0.73)*. Also evaluate the LURF model via a temporal and *overall cross validation and obtain R^2 values of 0.84 and 0.92, respectively*.	The study successfully integrates temporal and spatial components into the model, which exhibits higher accuracy than spatial models constructed individually for each month. The findings illustrate the advantage of using a LURF to model the spatiotemporal variability of NO_2 concentrations.
(Stafoggia *et al.*, 2019)	To estimate daily PM_{10} and $PM_{2.5}$ concentrations at 1-km^2 grid for	Italy	2013-2015	PM_{10}, $PM_{2.5}$	*24 hr mean PM_{10} and $PM_{2.5}$ from monitoring stations* at sites across the	Separate RF models were defined to: predict $PM_{2.5}$ and	The models were able to capture most of PM variability, with *mean cross validation (CV) R^2 of 0.75*	"We developed a five-stage approach where we merged multiple sources of spatial and temporal data, we predicted

	2013-2015 using a Random Forest (RF) machine learning approach				country. Multi-Angle Implementation of Atmospheric Correction (MAIAC) *Aerosol Optical Depth (AOD) data* (higher quality data at 1-km^2 spatial resolution than standard MODIS products). *Meteorological parameters* (daily mean air temperature, sea-level barometric pressure, precipitations, relative humidity, wind speed and direction) and planetary boundary layer height were retrieved by the ERA-Interim reanalysis project (Dee et al., 2011), the latest global atmospheric reanalysis produced by the ECMWF. Vegetation index (NDVI) from the MODIS NDVI product. Substantial spatial data at the grid cell level ranging from geo-climatic zones, point emission sources, road density	$PM_{2.5-10}$ concentrations in monitors where only PM_{10} data were available. (stage 1); impute missing satellite AOD data using estimates from atmospheric ensemble models (stage 2); establish a relationship between measured PM and satellite, land use and meteorological parameters (stage 3); predict stage 3 model over each 1-km^2 grid cell of Italy (stage 4); and improve stage 3 predictions by using small-scale predictors computed at the monitor locations or within a small buffer (stage 5).	*and 0.80 (stage 3) and 0.84 and 0.86 (stage 5) for PM_{10} and $PM_{2.5}$, respectively*. Model fitting was less optimal for $PM_{2.5-10}$, in summer months and in southern Italy. Finally, predictions were equally good in capturing annual and daily PM variability, therefore *they can be used as reliable exposure estimates for investigating long-term and short-term health effects.*	satellite AOD from atmospheric ensemble models, and we took full advantage of machine learning methods to obtain finely resolved PM predictions over large spatial and temporal domains. We also applied a local model (stage 5) with the aim of proving the validity of our approach for future epidemiological applications with individual data on residential addresses. We believe that machine learning methods, in combination with extensive data collection on multiple parameters, can be valid tools for predicting ground level air pollutants concentrations at fine spatial and temporal resolution."
					data, imperviousness surface areas etc.			
(Xue *et al.*, 2020)	Apply an improved geographically and temporally weighted regression (IGTWR) model to geostationary satellite (Himawari-8 AHI) data to estimate hourly $PM_{2.5}$ concentration data over central and eastern China in 2017	Central and Eastern China	2017	$PM_{2.5}$	In situ ground-level *$PM_{2.5}$ concentrations* from observation network, *Himawari-8 Advanced Himawari Imager (AHI) AOD data* (resolutions of 2 and 5 km), *land use type* showing the distribution of arable land, woodland, grassland, and desert at 10km resolution from the east to the inland north-western area, *relative humidity* at temporal resolution of 30 min – 3 hrs depending on site (linear interpolation to obtain relative humidity at each time, spatial interpolation for each time, and inverse distance-weighted interpolation for surface relative humidity of each time performed), boundary layer height (BLH) at 3hr temporal resolution	A generalized distance based on the longitude, latitude, day, hour, and land use type was constructed. AHI aerosol optical depth, surface relative humidity, and boundary layer height (BLH) data were used as independent variables to retrieve the hourly $PM_{2.5}$ concentrations at 1:00, 2:00, 3:00, 4:00, 5:00, 6:00, 7:00, and 8:00 UTC (Coordinated Universal Time).	The model fitting and cross-validation performance were satisfactory. For the model fitting set, the correlation coefficient of determination (R^2) between the measured and predicted $PM_{2.5}$ concentrations was 0.886, and the root-mean-square error (RMSE) of 437,642 samples was only 12.18 μg/m^3. The tenfold cross-validation results of the regression model were also acceptable; the correlation coefficient R^2 of the measured and predicted results was 0.784, and the RMSE was 20.104 μg/m^3, which is only 8 μg/m^3 higher than that of the model fitting set. The spatial and temporal characteristics of the hourly $PM_{2.5}$ concentration in 2017 were revealed. The model also achieved stable performance under haze and dust conditions.	"There are significant regional pollution characteristics over central and eastern China in 2017. Fine particulate matter pollution events are concentrated mainly in the North China Plain, Guanzhong Plain, Fenwei Plain, middle and lower reaches of the Yangtze River, South China, and Sichuan Basin. The ground-level $PM_{2.5}$ concentration is negatively correlated with the solar radiation. The predicted and measured values of the $PM_{2.5}$ concentration are highly consistent throughout the study area. The model results represent the actual distribution characteristics of $PM_{2.5}$ in central and eastern China. The model designed in this study can monitor the occurrence, development, and termination of extreme weather events. The predicted $PM_{2.5}$ concentration is very similar to the observed $PM_{2.5}$ concentration. Satellite-based hourly $PM_{2.5}$ monitoring will play an important role in predicting extreme weather events and providing warnings in advance."

(Liu, Weng and Li, 2019)	Develop an ensemble machine-learning algorithm for estimating $PM_{2.5}$ concentrations directly from Advanced Himawari Imager satellite measured top-of-the-atmosphere (TOA) reflectances integrated with meteorological parameters across China in 2016	China	2016	$PM_{2.5}$	Satellite products (TOA reflectances from the AHI onboard the Himarwari-8 to measure AOD), hourly averaged surface $PM_{2.5}$ concentration from ~1500 sites over mainland China, meteorological variables (surface atmospheric pressure (P, hPa), total column water (TCW, kg m^{-2}), 10-m u-wind (U_{10}) and v-wind (V_{10}) component, air temperature at an altitude of 2 m (T, K), total column ozone (kg m^{-2}), relative humidity (RH, %), and planetary boundary layer height (PBLH, m) were obtained from the ERA-Interim reanalysis)	First collocated AHI measurements, surface $PM_{2.5}$ concentrations, and meteorological variables to generate the training dataset. Then the random forest model was fitted by applying it to the training dataset. In model fitting, the modeling dataset was used in both the model fitting and the model validation. The best parameters of the model were determined by adjusting them until the best prediction accuracy was achieved. Ten-fold cross validation (CV) were then used to further adjust the model to avoid the over-fitting problem. The resulting model was finally used to estimate hourly surface $PM_{2.5}$ concentrations at the AHI pixel level.	The algorithm is demonstrated to perform well across China with high accuracies at different temporal scales. *The model has an overall cross-validation coefficient of determination (R^2) of 0.86 and a root-mean-square error (RMSE) of 17.3 μg m^{-3} for hourly $PM_{2.5}$ concentration estimation.* Such accuracies of the estimation on $PM_{2.5}$ concentration by using TOA reflectance directly are comparable with those of the common methods on estimating $PM_{2.5}$ concentration by using satellite derived AODs, but the former has a relatively stronger predictive power relating to spatial-temporal coverages than the latter.	Annual and seasonal variations of $PM_{2.5}$ concentration over three major the developed regions in China are estimated using the model and analysed. The relatively stronger predictive ability of developed model in this study may help provide information about the diurnal cycle of $PM_{2.5}$ concentrations as well as aid in monitoring the processes of regional pollution episodes and the evolution of $PM_{2.5}$ concentration.

Annex 2: Data Sources

- US AQI standard is from https://airnow.gov/index.cfm?action=airnow.calculator

- Thailand AQI standard is from http://air4thai.pcd.go.th/webV2/aqi_info.php

- Hourly air pollution in Chiang Mai from year 1996 – 2019 was partially obtained by submitted an official request to Thailand Pollution Control Department and the data from 2019 – 2021 was scraped from http://air4thai.pcd.go.th/webV2/aqi_info.php

- Hourly weather data from year 2010 – 2021 was scraped from https://www.wunderground.com/history/daily/th/mueang-chiang-mai/VTCC

- Daily hotpots data from year 2010 to 2021 was download from https://earthdata.nasa.gov/earth-observation-data/near-real-time/firms/c6-mcd14dl

- Holiday information from year 2010 – 2021 in Thailand is from https://www.timeanddate.com/holidays/thailand/

- El Niño index from year 2010 – 2021 is from https://origin.cpc.ncep.noaa.gov/products/analysis_monitoring/ensostuff/ONI_v5.php