

Rethinking Machine Learning Weather Normalisation: A Refined Strategy for Short-term Air Pollution Policies

Yuqing Dai¹, Bowen Liu², Chengxu Tong¹, David C. Carslaw³, A. Rob MacKenzie¹, Zongbo Shi¹

¹School of Geography, Earth and Environmental Science, University of Birmingham, Birmingham, B15 2TT, U.K.

²Department of Management, Business School, University of Birmingham, Birmingham, B15 2TT, U.K.

³Wolfson Atmospheric Chemistry Laboratories, University of York, York, YO10 5DD, U.K.

Correspondence to: Zongbo Shi (z.shi@bham.ac.uk), A. Rob MacKenzie (a.r.mackenzie@bham.ac.uk)

Abstract. Air pollution causes millions of premature deaths annually, driving widespread implementation of clean air interventions. Quantitative evaluation of the efficacy of such interventions is critical in air quality management. Machine learning-based weather normalization (ML-WN) has been employed to isolate meteorological influences from emission-drive changes; however, it has its own limitations, particularly when abrupt emission shifts occur, e.g., after an intervention. Here we developed a logical evaluation framework, based on paired observational datasets and a test of ‘ML algebra’ (i.e., the ‘commutation’ of a normalisation step), to show that ML-WN significantly underestimates the immediate effects of short-term interventions on nitrogen oxides (NO_x), with discrepancies reaching up to 42% for one-week interventions. This finding challenges assumptions about the robustness of ML-WN for evaluating short-term policies, such as emergency traffic controls or episodic pollution events. We propose a refined approach (MacLeWN) that **can reduce such** underestimation biases by >90% in idealised but plausible cases studies. We applied both approaches to evaluate the impact of COVID-19 lockdown on NO_x as measured at Marylebone Road, London. For the one-week period after the lockdown, ML-WN estimates approximately 17% smaller NO_x reductions compared to MacLeWN, and such underestimation diminishes as policy duration extends, decreasing to ~10% for one-month and becoming insignificant after three months. Our findings indicate the importance of carefully selecting evaluation methodologies for air quality interventions, suggesting that ML-WN should be complemented or adjusted when assessing short-term policies. Increasing model interpretability is also crucial for generating trustworthy assessments and improving policy evaluations.

Formatted: Highlight

1 Introduction

Air pollution remains one of the most pressing global environmental challenges, responsible for an estimated 4.2 million premature deaths annually due to cardiovascular disease, stroke, lung cancer, and chronic respiratory diseases (Lee et al., 2020b; Fuller et al., 2022). In response, policymakers worldwide have enacted diverse strategies to mitigate air pollution, ranging from long-term emission reduction plans to short-term measures aimed at avoiding acute pollution episodes. Evaluating the effectiveness of these interventions is critical for ensuring cost-effective policy design and maintaining public

trust in governance. However, such evaluations are inherently complex due to the dynamic interplay of emission sources, atmospheric chemistry, deposition processes, and, importantly, meteorological variability (Seo et al., 2018). Meteorological conditions, in particular, exert a profound influence on observed pollutant concentrations, often masking or amplifying changes in emissions over time (Shi et al., 2021). Compounding this challenge, fluctuations in human activities, such as

seasonal industrial output or agricultural practices, introduce additional variability that can obscure the true impact of specific policy measures. Consequently, robust methodologies are needed to disentangle the confounding effects of meteorology and periodic anthropogenic activities from the signal of emission changes attributable to policy interventions. A promising approach to address this challenge is the machine learning-based weather normalisation (ML-WN) method developed by Grange et al. (2018). This data-driven strategy has gained traction for its ability to isolate meteorological influences from observed pollutant concentration trends, enabling clearer attribution of air quality changes to emission-related factors. For instance, ML-WN has been widely applied to assess the transient air quality improvements during COVID-19 lockdowns (Cole et al., 2020; Vu et al., 2019; Dai et al., 2021) and to investigate the impact of on ozone (O_3) concentrations and particulate matter compositions (Ding et al., 2023; Ding et al., 2021). Unlike traditional statistical techniques, which often rely on rigid assumptions about linear relationships between variables, ML-WN flexibly captures complex, non-linear interactions between meteorological parameters and emissions. This adaptability allows for more efficient decomposition of weather-driven variability from policy-driven changes in pollution time series. Furthermore, ML-WN circumvents the computational demands and inherent simplifications of chemistry-transport models (CTMs), positioning it as a pragmatic tool for rapid policy evaluation.

Despite its advantages, the ML-WN framework is not without limitations. Its reliability depends on the performance of the underlying machine learning model, which is susceptible to overfitting, especially when applied to sparse or noisy datasets. The selection of input variables, such as wind speed, temperature, or boundary layer height, introduces potential biases if critical predictors are omitted or redundant ones included. Additionally, while ML-WN excels at modelling non-linear relationships, its "black-box" nature complicates interpretability, raising concerns about whether the model genuinely captures causal mechanisms or merely correlates superficial patterns in its training data set. A more fundamental challenge lies in the absence of a definitive "ground truth" for validating weather-normalized pollution trends, as real-world systems are subject to concurrent socio-environmental changes that cannot be fully controlled. These limitations collectively hinder precise quantification of policy impacts, risking misinterpretations that could misguide public health strategies or resource allocation.

To address these gaps, we propose a logical benchmarking framework designed to evaluate the accuracy of weather normalisation methods in isolating policy-driven changes in air quality. Focusing on paired nitrogen oxides (NO_x) time series, a key pollutant influenced by both meteorology and anthropogenic emissions, we systematically test the ability of the ML-WN approach to recover known policy effects after removing meteorological noise. Our analysis reveals a critical shortcoming: as currently implemented, ML-WN could underestimate the short-term efficacy of interventions, particularly those with immediate impacts, such as traffic restrictions or industrial shutdowns. This underestimation arises from the

65 method’s tendency to over-smooth transient signals in the data, conflating abrupt policy-driven changes with stochastic meteorological variability. Left unaddressed, this bias could lead policymakers to undervalue the benefits of rapid-response measures or misallocate resources toward less effective long-term strategies. In response, we introduce an alternative weather normalisation strategy that explicitly accounts for transient policy signals by incorporating intervention-specific covariates into the model architecture.

70 **2 Materials and Methods**

2.1 Data Source

To establish a robust baseline of urban air pollution patterns unaffected by the anomalous atmospheric conditions during the COVID-19 pandemic, hourly nitrogen oxides (NO_x = nitrogen oxide (NO) + nitrogen dioxide (NO_2)) concentrations from 2017 to 2019 were analysed. These data were obtained from two Automatic Urban and Rural Network (AURN) sites in
75 London: Marylebone Road (MR, UKA00315) and North Kensington (NK, UKA00253). The MR site, situated within a typical street canyon, represents a high-traffic urban environment with elevated NO_x levels (Masson et al., 2020; Zhong et al., 2016). In contrast, the NK site serves as an urban background location within much less direct traffic influence, reflecting lower baseline pollution levels (Bigi and Harrison, 2010). Hourly surface meteorological variables, including ambient wind speed (w_s , in m s^{-1}) and wind direction (w_d , in degrees), air temperature (temp , in $^{\circ}\text{C}$), relative humidity (rh , in %), surface
80 pressure (sp , in hPa), and precipitation (precip , in mm) were obtained from the London Heathrow Airport weather station. These data were retrieved using the ‘worldmet’ package from the National Oceanic and Atmospheric Administration (NOAA) Integrated Surface Database (ISD), available at <https://CRAN.R-project.org/package=worldmet>. Consistent with
previous studies (Woolley et al., 2024; Betancourt et al., 2023), missing values in the meteorological and pollutant concentration datasets were handled by linear interpolation when gaps were less than three consecutive hours; longer gaps
85 were retained as missing to avoid introducing excessive bias through over-interpolation.

2.2 Machine Learning Weather Normalisation

The original machine learning weather normalisation (ML-WN) approach was introduced by Grange et al. (2018). Building upon this methodology, we developed an independent version of weather normalisation for air pollutants using a gradient boosting machine (GBM) model within H2O.ai’s Automated Machine Learning (AutoML) framework. AutoML is a
90 function within the H2O platform—an open-source R/Python package for data analysis developed by H2O.ai (Ledell and Poirier, 2020). AutoML automates the iterative process of hyperparameter selection and streamlines the machine learning pipeline, including preprocessing, feature engineering, model training, and model evaluation. It enables systematic comparison of multiple algorithms (e.g., generalized linear models, random forest, GBMs) within a predefined computational budget, and it provides a leaderboard ranking the models based on predefined metrics, such as model
95 performance and training time.

In this study, AutoML seeks the best function $f_*^{s,p}(\cdot)$ from an ensemble of 30 trained models $f^{s,p}(\cdot)$, for which:

$$\begin{cases} Y_t^{s,p} - f_*^{s,p}(\cdot) \rightarrow 0 \\ f_*^{s,p} = \operatorname{argmin}_{f^{s,p} \in \mathcal{F}_{\mathcal{E}_X \sim \mathcal{D}}(p, T, X_t^e, X_t^m)} \end{cases} \quad (1)$$

where $Y_t^{s,p}$ is the concentration of pollutant p (i.e., NO_x) at a given time point t and site s ; The function $f^{s,p}(\cdot)$ represents the machine learning models that have been trained for predicting pollution. The model output is p and input features include a time trend T and two matrices of regressors: X_t^e for temporal variables such as hour of the day and day of the week, which act as proxies for diurnal and weekly emission patterns; and X_t^m for meteorological variables (wd, ws, temp, rh, sp, and precip). Model training utilized 80% of the dataset, with the remaining 20% reserved for evaluation for each site (Vu et al., 2019; Grange et al., 2018; Goodfellow et al., 2016). Full model configurations and performance metrics are provided in Tables S1 and S2.

After the training process, the selected GBM model was applied to generate weather-normalised NO_x concentrations due to its strong predictive performance across both monitoring sites, achieving index of agreement (IOA) values of 0.84 (MR) and 0.82 (NK). In the ML-WN method, weather-normalised concentration $Y_t^{s,p}$ are derived by resampling meteorological variable X_t^m while fixing temporal emission proxies X_t^e :

$$Y_t^{s,p} = \frac{1}{n} \sum_{i=1}^n f_*^{s,p}(T, X_e^i, X_m^i) \quad (2)$$

where X_m^i is the i -th resampled meteorological dataset, and n is the total number of resampling (without replacement), determined by the robustness of model predictions and the practicality of computational costs (e.g., used 300 times here similar to previous studies (Shi et al., 2021; Vu et al., 2019)). The rationale behind the ML-WN approach is to construct a reliable machine learning model to capture pollutant concentrations under all possible weather conditions based on historical records. By repeatedly resampling the meteorological inputs and averaging the resulting predictions, ideally the method approximates the conditional expectation of concentration with meteorological variance removed; the residual signal is then interpreted as arising from changes in emissions.

Here, we also introduce an alternative strategy for weather normalisation, denoted as MacLeWN, to isolate emission-drive trends from meteorological variability. Unlike ML-WN, which averages out meteorological effects while holding temporal emission proxy's constant, MacLeWN firstly filters temporal variations (e.g., hourly, weekly cycles) that correlate with emission patterns:

$$Y_t^{s,p} = \frac{1}{n} \sum_{i=1}^n f_*^{s,p}(T, X_e^i, X_m) \quad (3)$$

Here, $Y_t^{s,p}$ represents time-trend-normalised concentrations, where temporal fluctuations are averaged out. Because the randomised emission proxies at each time step no longer encode emission levels such as rush-hour traffic peaks or weekend effects, the residual variability in the normalised output $Y_t^{s,p}$ reflects how the fixed meteorological conditions

Deleted: (e.g., rush-hour traffic peaks)

Deleted: T

influence averaged emission levels. Similarly, a baseline concentration $Y_t^{s,p}$ at each time step can be calculated by resampling both emission proxies and meteorology while retaining the long-term trend index T :

$$Y_t^{s,p} = \frac{1}{n} \sum_{i=1}^n f^{s,p}(T, X_e^i, X_m^i) \quad (4)$$

130 This unperturbed baseline contains only the slow, secular trend T , and no high-frequency fluctuations from either emissions or weather. The meteorological impact factor $met_factor_t^{s,p}$ is then derived as:

$$met_factor_t^{s,p} = 100\% \times \frac{Y_t^{s,p} - Y_t^{s,p}}{Y_t^{s,p}} \quad (5)$$

Subsequently, weather-normalised concentrations $\gamma_t^{s,p}$ are calculated as:

$$\gamma_t^{s,p} = \frac{observations}{1 - met_factor_t^{s,p}} \quad (6)$$

135 Because the numerator is the measured raw observations with impacts from emissions and weather and the denominator removes the quantified meteorological multiplier, the quotient represents the pollutant level under the “normalised” weather condition.

2.3 Evaluation Methodology

2.3.1 Assessment Using a Theoretical Framework

140 Evaluating the efficacy of weather normalisation in isolating emission-driven air quality changes requires scenarios with well-defined “ground truth” outcomes. However, real-world policy assessments are often confounded by overlapping variables, making it challenging to disentangle meteorological and anthropogenic effects. To address this, we designed idealised but plausible policy interventions targeting NO_x reductions at the MR site, a high-traffic location where NO_x concentrations typically exceed urban background levels (NK site) by a quantifiable “road increment” (Harrison et al., 2021; Bannister et al., 2021). After applying weather normalisation ML-WN, this increment represents the additional NO_x attributable to traffic emissions, calculated as the differences between weather-normalised MR (denoted as MR_wnNO_x) and NK (NK_wnNO_x) concentrations, as meteorological influences are expected to be minimized through the normalisation process.

150 In our simulations, interventions temporarily eliminated the road increment during predefined periods (e.g., 1 week to 6 months), after which concentrations reverted to baseline. Practically, we achieved this by replacing MR_wnNO_x concentrations during the intervention periods with the equivalent-but-lower NK_wnNO_x values, generating a synthetic time series (synth_wnNO_x) that isolates emission-driven changes, enabling direct comparison with observed or re-normalised data (Fig. S1). While this approach produces an idealized emission reduction that is unattainable in practice due to persistent traffic, street canyon effects, green infrastructure, and complexities of all kinds, it does provide a precisely defined benchmark time series against which to test the logic and sensitivity of weather normalisation methods.

Deleted: influences acting on

Deleted: To quantify these meteorological contributions, first we compute ...

Deleted: filters out all short-term variability,

Deleted: .

The synth_wnNO_x data can be used to quantify the ability of weather normalisation approaches during the policy window (see Fig. S1, illustrating the difference between synth_wnNO_x and MR_wnNO_x). To reintroduce meteorological variability into the idealised scenarios, we quantified the hourly meteorological contribution factors (MCF) at the MR site using the relative difference between observed NO_x and weather-normalised concentrations:

$$165 \quad MCF_t^{MR_NO_x} = 100\% \times \frac{C_t^{MR_wnNO_x} - C_t^{MR_NO_x}}{C_t^{MR_wnNO_x}} \quad (7)$$

$C_t^{MR_NO_x}$ and $C_t^{MR_wnNO_x}$ represent the observed and weather-normalised NO_x concentrations at a time point t , respectively. Here, a positive MCF indicates meteorological conditions enhancing pollutant dispersion (lower observed NO_x), while negative values reflect conditions exacerbating local accumulation (Fig. S2). These contributions were applied to synth_wnNO_x to simulate “observed” concentrations under policy interventions, denoted as reconstitute NO_x (recon_NO_x):

$$170 \quad C_t^{recon_NO_x} = C_t^{synth_wnNO_x} \times \left(1 - \frac{MCF_t^{MR_NO_x}}{100\%}\right) \quad (8)$$

Outside policy windows, recon_NO_x recovers precisely actual MR observations; during interventions, differences between recon_NO_x and NK_NO_x reflect both emission reductions and site-specific meteorological interactions (Fig. S3), such as reduced ventilation in the MR street canyon (Jeanjean et al., 2017; Dai et al., 2022).

175 One final step completes the assessment of the approaches to weather normalisation. By reapplying weather normalisation to recon_NO_x, we generated a re-normalised time series (wn-recon_NO_x). In theory, wn-recon_NO_x should closely match, if not exactly replicate, the synth_wnNO_x if the method perfectly isolates emission effects. That is to say, ideally, the weather normalisation operation should ‘commute’ algebraically. Discrepancies between the two normalised time series indicate systematic biases in the weather normalisation process, quantifying over- or underestimation of policy impacts. This comparison provides, to our knowledge, the first evaluation of the accuracy of machine learning weather normalisation approaches in assessing the influence of policy interventions on air quality. The framework described above is generalisable to any policy setting in which the policy impact is masked by weather-like ‘noise’. Figure 1 shows the schematic diagram for the whole evaluation process, and Table 1 provides a comprehensive list of terminology for NO_x time series used in this study.

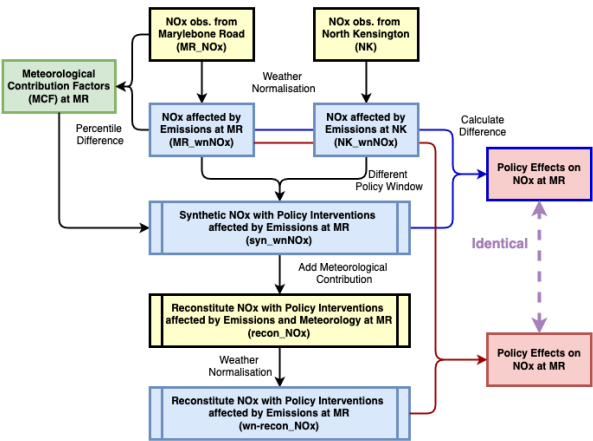


Figure 1: The schematic diagram showing the analytics pipeline to quantify the ability of weather normalisation approaches.

Table 1: A summary of the NO_x timeseries used in the study.

Name	Description
MR_NO _x	NO _x observations at Marylebone Road
NK_NO _x	NO _x observations at North Kensington
MR_wnNO _x	Weather normalised NO _x concentrations at Marylebone Road
NK_wnNO _x	Weather normalised NO _x concentrations at North Kensington
synth_wnNO _x	NO _x at Marylebone Road with “interventions” affected by only emissions
recon_NO _x	NO _x at Marylebone Road with “interventions” affected by emissions and meteorology
wn-recon_NO _x	NO _x at Marylebone Road with “interventions” affected by only emissions (reconstituted)

We designed scenarios mimicking diverse real-world policies, including sustained interventions (1 week–6 months), phased reductions (3–6 months), and cyclic interventions (1-month intervals) (Table 2, Fig. S1). These scenarios represent different types of policy interventions and help assess how the weather normalisation methods perform under different temporal patterns of emission changes. Although those sustained one-week to six-month cases are idealised “step” emission reductions, we also include phase-out and cyclic patterns specifically to emulate more gradual or heterogeneous real-world responses (e.g., staggered traffic bans or variable industrial curtailments), thereby spanning the continuum from abrupt to progressive interventions. To ensure that our conclusions are not model-specific, we replicated our analysis using different

machine learning approaches, including eXtreme Gradient Boosting (XGBoost) and distributed random forest (DRF) models. Performance metrics for these models (Tables S7–S8) and consistent results across these algorithms (Figs. S11–S12) indicates the generalizability of our findings. Following the methodological framework described above, we further extend the assessment to the MacLeWN. The weather normalised dataset, the meteorological factors, and the “original” dataset influenced by both emissions and meteorology were presented in Figs. S6–S8, respectively.

Table 2: Overview of air quality intervention durations and strategies for eight test scenarios (S1–S8). In the second part of each scenario name, “s” = sustained intervention, “p” = phased-out intervention, and “c” = cyclic intervention. Intervention durations are for a number (1, 2, 3, or 6) of weeks (‘w’), or months (‘m’).

Scenario	Intervention period	Description
S1_s1w	1 st Aug 2018 – 7 th Aug 2018	Sustained intervention for one week
S2_s2w	1 st Aug 2018 – 14 th Aug 2018	Sustained intervention for two weeks
S3_s1m	1 st Aug 2018 – 31 st Aug 2018	Sustained intervention for one month
S4_s3m	1 st Aug 2018 – 31 st Oct 2018	Sustained intervention for three months
S5_s6m	1 st Aug 2018 – 31 st Jan 2019	Sustained intervention for six months
S6_p3m	1 st Aug 2018 – 31 st Oct 2018	Phase-out intervention for three months
S7_p6m	1 st Aug 2018 – 31 st Jan 2019	Phase-out intervention for six months
S8_c1m	1 st Aug 2018 – 30 th April 2019	Cyclic intervention at one month interval

2.3.2 Application of Weather Normalisation to COVID-19 Lockdown Data

To further evaluate the performance of the ML-WN and MacLeWN approaches, we applied both techniques to analyse changes in NO_x concentrations during the COVID-19 lockdowns at London Marylebone Road, which will provide a basis to assess emission reductions under abrupt, real-world conditions. Prior studies have used weather normalisation to isolate lockdown effects, but the lack of a definitive “ground truth” for actual pollutant reductions during lockdowns makes it challenging to evaluate these methods absolutely. Our analysis introduces a direct comparison between the ML-WN and MacLeWN under identical scenarios, allowing us to assess their relative performance in isolating lockdown-induced emission changes from meteorological variability.

The lockdown in London was first announced on March 23rd, implemented on March 26th, and eased on June 23rd, 2020 (Davies et al., 2021). The COVID-19 lockdown measures led to an acute drop in NO_x levels at these sites (Lee et al., 2020a). We gathered hourly concentrations of NO_x from 2018 to 2021 for Marylebone from AURN. The corresponding meteorological data (i.e., wd, ws, RH, temp, sp, precip) were obtained from Heathrow Airport. The predictive variables for the machine learning models included temporal variables and meteorological variables as mentioned above. Weather-normalised daily NO_x concentrations around the time of the lockdowns are presented in Fig. S10. [Fig. S13 shows the variable-importance ranking based on mean absolute SHapley Additive exPlanations \(SHAP\) values and SHAP contribution plots, whereas Fig. S14 provides partial-dependence plots for the six most influential predictors.](#)

After weather normalisation, we adopted an analytical framework to evaluate the effects of the lockdowns on NO_x concentrations as used in Shi et al. (2021). The baseline period for the lockdown was defined as the one month-to-one week preceding each lockdown (i.e., excluding the final transitional week, Fig. S10), with post-lockdown impacts assessed during the first week of restrictions:

$$\Delta C_{NO_x,t} = C_{NO_x,t} - \overline{C_{NO_x}}_t \quad (9)$$

$$P_{NO_x,t} = 100\% \times \frac{\Delta C_{NO_x,t}}{\overline{C_{NO_x}}_t} \quad (10)$$

$\overline{C_{NO_x}}_t$ represents the average value of NO_x at a given hour during the baseline period, $C_{NO_x,t}$ represents the NO_x concentrations at a time t during the first week, one month, or three months after the lockdown, $\Delta C_{NO_x,t}$ and $P_{NO_x,t}$ represent the absolute and percentage changes in NO_x concentrations in 2020, respectively.

Emissions within a given year are subject to temporal trends, such as gradual policy shifts, economic fluctuations, or seasonal patterns like reduced heating fuel use during winter-to-spring transitions. These trends, unrelated to lockdown measures, risk conflating long-term or cyclical changes with short-term lockdown effects. To account for this, we detrended the data by comparing 2020 NO_x concentrations against the averaged values from the corresponding periods in 2018 and 2019. This two-year baseline (i.e., for “trend”) was selected to mitigate the influence of interannual variability. The lockdown-driven changes in NO_x concentration ($C_{NO_x,t}^*$) and their percentage equivalents ($\Delta P_{NO_x,t}^*$) were calculated as:

$$\Delta C_{NO_x,t}^* = \Delta C_{NO_x,t} - \Delta C_{NO_x,t}^{2018-2019} \quad (11)$$

$$\Delta P_{NO_x,t}^* = 100\% \times \frac{\Delta C_{NO_x,t}^*}{\overline{C_{NO_x}}_t} \quad (12)$$

3 Results

3.1 Comparison of two approaches in the theoretical framework

Figure 2(a) presents a comparative analysis of NO_x reductions, contrasting the “actual” intervention outcomes (red bars) with the predicted simulations derived from the ML-WN approach (blue bars) across eight idealised intervention scenarios (S1-S8, Table 1). The consistent discrepancies between the red and blue bars indicate that the ML-WN approach systematically underestimates the effectiveness of policy interventions, particularly in short-term scenarios. This underestimation occurs because the ML-WN method averages each time-step over meteorological samples drawn from the whole historical record; such averaging sometimes could be unrealistic that “blurs” the sharp drop introduced by the intervention, which is discussed further in the discussion. This underestimation has not been reported before, primarily because it is challenging to detect when comparing a single time series with and without weather normalisation. Importantly,

Deleted: may not fully capture abrupt changes in emission patterns due to it...

Deleted: s reliance on

Deleted: data

Deleted: and the smoothing effects inherent in machine learning models...

Deleted: Such

the same qualitative pattern (i.e., $MacLeWN > ML-WN$) holds also for both phase-out and cyclic scenarios, showing robustness even when the rebound signal after the intervention is not instantaneous.

Our results show that the underestimation is particularly significant for short-term interventions. For example, for the one-week sustained intervention (S1_s1w), the ML-WN approach underestimates the policy effects by $101.1 \mu\text{g m}^{-3}$, corresponding to a 42.2% discrepancy. Such substantial underestimations could lead to significant misjudgements of the short-term air quality improvements and the resultant health impacts of air pollutants (Meng et al., 2021). As the duration of the intervention increases to two weeks (S2_s2w) and one month (S3_s1m), the underestimation decreases to $54 \mu\text{g m}^{-3}$ (23.5%) and $35.2 \mu\text{g m}^{-3}$ (13.8%), respectively. This pattern continues for policy interventions sustained over longer periods, with the underestimation further decreasing to $17.9 \mu\text{g m}^{-3}$ (6.3%) for three months (S4_s3m) and $7.9 \mu\text{g m}^{-3}$ (less than 3%) for six months (S5_s6m). These results indicate that the ML-WN method becomes more accurate over longer intervention periods, possibly because the model adjusts to new emission patterns over time.

Furthermore, the degree of underestimation is more associated with the duration of policy interventions than with the type of policy (i.e., sustained, phased out, or cyclic). For example, both sustained and phased-out policies spanning three and six months show similar levels of underestimation. In contrast, cyclic policies with one-month intervals (S9_c1m) result in an underestimation of $16 \mu\text{g m}^{-3}$ (4.8%), which is smaller compared to the underestimations observed in other policy types. Although these discrepancies are smaller compared to short-term interventions, they are still significant when considering long-term average threshold values for public health impacts (Faustini et al., 2014).

Figure 2(b) illustrates the time-dependent discrepancies in the underestimation of policy impacts on NO_x concentrations, as estimated by the ML-WN approach under different intervention scenarios. For example, S6_p3m scenario, representing a phased-out policy implemented over three months, has an overall underestimation of $16.6 \mu\text{g m}^{-3}$ (6%, light green bar) but exhibits a $38.1 \mu\text{g m}^{-3}$ (14.4%) underestimation in the first week (light blue bar). These results show that, for policies with immediate effects (i.e., sustained policies), the initial underestimation of their efficacy by the ML-WN approach is most pronounced in the early stages but diminishes over time. The extent of this underestimation is inversely related to the duration of the policy implementation. As the duration of the policy extends beyond one month, the discrepancy between the anticipated and “actual” impact in the initial stages decreases significantly.

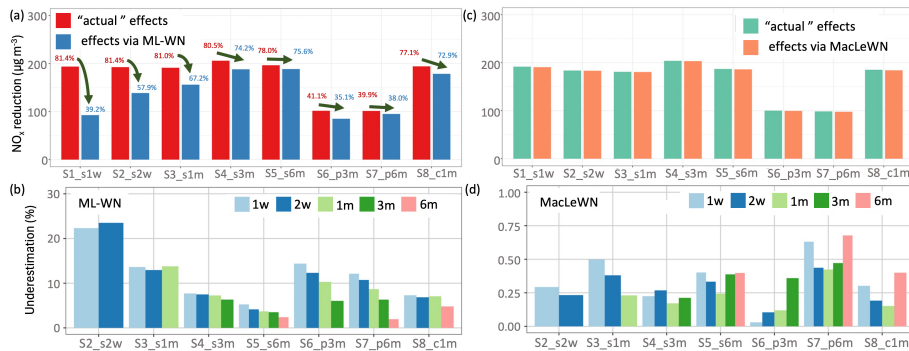


Figure 2: Average intervention effects on NO_x concentrations at Marylebone Road (MR) under different scenarios, including sustained interventions lasting from one week (S1_s1w) to six months (S1_s6m), phased-out interventions over three (S6_p3m) to six months (S7_p6m), and cyclic interventions with a one-month frequency (S8_c1m). (a) and (c): The red and green bars represent the “actual” intervention effects based on the theoretical evaluation framework, while the difference between red bars and effects estimated by weather normalisation approaches indicates the extent of underestimation; (b) and (d): The bars represent underestimated average intervention effects at different time periods as a percentage. Note the different y-axis ranges in (b) and (d).

Specifically, during the first week after policy implementation, the underestimation of NO_x reduction is 51.9 μg m⁻³ (13.6%) for a two-week policy (S2_s2w), diminishing to 13.4 μg m⁻³ (5.3%) for a policy sustained over six months (S5_s6m). Similarly, the net underestimations of NO_x reduction after two weeks are 54 μg m⁻³ (23.5%), 33.3 μg m⁻³ (13%), 19.4 μg m⁻³ (7.5%), and 11 μg m⁻³ (4.2%) for S2_s2w, S3_s1m, S4_s3m, and S5_s6m, respectively. In contrast, phased-out policies, which gradually reduce interventions over time, present a different pattern. Although their cumulative effects align with those of sustained policies, the initial weeks show a marked underestimation of impact. Figure 2(b) shows that a three-month phased-out policy (S6_p3m) results in an overall underestimation of 16.6 μg m⁻³ (6.0%), and a six-month policy (S7_p6m) at 6.4 μg m⁻³ (1.9%). However, the first-week underestimations are 38.1 μg m⁻³ (14.4%) and 29.9 μg m⁻³ (12.1%), respectively. Similarly, the cyclic policy (S8_c1m) shows a 7.3% underestimation in the first week, which is somewhat higher than the overall underestimation observed over the six-month duration (4.8%).

By employing the updated machine learning approach (MacLeWN), these underestimations were mitigated, as shown in Figures 2(c) and 2(d). The results show a strong agreement between the “known” NO_x reductions resulting from idealised policy interventions and those simulated by the MacLeWN approach. While minor discrepancies are observed in the time-dependent analysis in Fig. 2(d), these variations are consistently small, each remaining below one percent, and showing improvements of greater than an order of magnitude, often by a factor of 20 to 50, compared to the underestimations

- Deleted: A
- Deleted: C
- Deleted: B
- Deleted: D
- Deleted: B
- Deleted: D

observed with the ML-WN approach as in Fig. 2(b). Further details are provided in Tables S4 and S5, and Figs. S5 and S9 in Supporting Information.

3.2 Comparison of two approaches in the lockdown scenario

Figure 3 provides the impact of lockdown measures on NO_x concentrations near the roadside at London Marylebone Road air quality monitoring site. NO_x reductions were assessed over one week, one month, and three months after the lockdown implementation by using the direct observations (blue bars), the ML-WN (purple bars), and the MacLeWN (green bars) approaches, respectively. The bars depict the average detrended concentrations changes at the MR site over the policy implementation period following lockdown, and the error bars denote the standard error of the mean. For one-week lockdown effects, the observed NO_x decreased from 153.9 to 27.5 µg m⁻³ (-82.1%); the ML-WN estimates a decrease from 158.1 to 76.0 µg m⁻³ (-51.9%); and the MacLeWN estimates a decrease from 153.2 to 47.7 µg m⁻³ (-68.9%). In the case of one month intervention effect, direct observations indicate a NO_x decrease from 134.3 to 43.6 µg m⁻³ (-67.5%), the ML-WN indicates a decrease from 150.9 to 71.9 µg m⁻³ (-52.3%), and the MacLeWN points to a -62.3% reduction from 141.3 to 53.2 µg m⁻³. For three-month lockdown effects, NO_x reductions through the observations, the ML-WN, and the MacLeWN are -58.1±6.6 µg m⁻³ (-53.7%), -71.9±6.0 µg m⁻³ (-51.3%), and -72.8±8.1 µg m⁻³ (-57.4%), respectively (Table 3). These results are consistent with the theoretical evaluation presented above, as the ML-WN approach estimates around 17% lower policy impact compared to the MacLeWN approach but decrease with longer policy implementation time (i.e., about 10% lower for one-month lockdown and is insignificant for three-month effect).

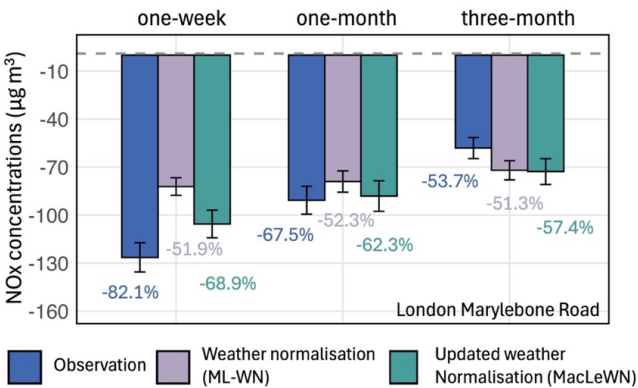


Figure 3: Intercomparison of detrended NO_x concentration changes at London Marylebone Road using direct observations, the weather normalisation (ML-WN), and the updated weather normalisation (MacLeWN) approaches. The bar graphs in the panel show concentration changes (C_p) in NO_x over one week, one month, and three months average following lockdown

335 implementation, respectively. Each graph contrasts the raw observation data (dark blue), ML-WN changes (light purple), and MacLeWN changes (dark green). For each bar, the average percentage changes (P_p^*) has been calculated and presented.

Table 3: Absolute and percentile COVID-19 lockdown impacts on NO_x concentrations at London Marylebone Road, analysed through direct observations, the weather normalisation (ML-WN), and the refined weather normalisation (MacLeWN) after measures implemented one week, one month, and three months, respectively.

Intervention period	Method	Concentration ($\mu\text{g m}^{-3}$)	Percentage (%)
One week	Observation	-126.0±9.1	-82.1±5.9
	ML-WN	-82.1±5.5	-51.9±3.4
	MacLeWN	-105.0±8.7	-68.9±5.6
One month	Observation	-90.7±8.7	-67.5±6.5
	ML-WN	-79.0±6.6	-52.3±4.4
	MacLeWN	-88.1±9.5	-62.3±6.7
Three months	Observation	-58.1±6.6	-53.7±6.1
	ML-WN	-71.9±6.0	-51.3±4.3
	MacLeWN	-72.8±8.1	-57.4±6.4

Note: In each case, the data are detrended following the method in Sect. 2.3.2. ~~the uncertainties are expressed as the standard error.~~

Deleted: .

340 **4 Discussion**

Accurately assessing the impact of policy interventions on air quality remains a critical challenge, as meteorological variability often obscures the signal of emission-driven changes. While machine learning-based weather normalisation (ML-WN) has emerged as a powerful tool to disentangle these effects, our findings reveal its limitations in evaluating short-term interventions. For example, we showed underestimations of up to 42% in the quantified effectiveness of one-week policies (Fig. 2 and Figs. S4 and S5), indicating that the ML-WN may not fully capture abrupt, non-linear changes in emissions following rapid policy implementation. This discrepancy highlights a fundamental tension: while ML-WN works well at isolating long-term meteorological influences, its reliance on historical patterns may render it less sensitive to transient disruptions in emission regimes.

The root of these uncertainties lies in the interplay between model architecture, variable interdependencies, and real-world complexity. The accuracy of ML-WN depends on the predictive robustness of its underlying algorithms, which must generalize beyond training data to capture sudden shifts in emissions. However, temporal variables (e.g., hour, day, season) and meteorological parameters (e.g., temperature, wind speed) are deeply intertwined in environmental systems. Temporal indicators generally act as proxies for human activity-driven emissions, yet they correlate systematically with meteorological cycles, such as solar radiation peaks at midday, atmospheric stability varies diurnally, and seasonal weather patterns drive recurring emission scenarios (e.g., heating demand in winter). These collinearities create a “proxy trap”, where models may conflate emission-driven trends with weather-driven fluctuations. Tree-based ensembles, while resilient to multicollinearity in prediction tasks, face interpretability challenges: variable importance metrics become unstable when predictors are correlated, splitting attribution across redundant features. Consequently, the “brute force” resampling central to ML-WN

360 (i.e., simulating counterfactual meteorological conditions across time), may inadvertently dilute the signal of abrupt policy impacts, particularly when interventions disrupt established correlations between time and emissions.

Further limitations arise from the fidelity of input data and the physical plausibility of resampled scenarios (Kilkenny and Robinson, 2018). ML-WN assumes meteorological variables can be independently perturbed, yet real-world weather systems exhibit tightly coupled dynamics (e.g., temperature-humidity relationships, land-sea breeze cycles). Resampling risks
365 generating unphysical combinations, for instance, applying wintertime temperature inversions to summer datasets could distort ozone chemistry or particulate dispersion pathways (Vu et al., 2019). Moreover, meteorological conditions absent during the model training phase can compromise predictive accuracy, especially in urban areas subject to complex and routine meteorological events. In coastal urban areas, for instance, diurnal breeze patterns regulate pollution advection (Geddes et al., 2021; Di Bernardino et al., 2021), but models trained on sparse temporal data may fail to resolve these
370 mesoscale processes, leading to biased normalisations.

The application of ML-WN and MacLeWN to COVID-19 lockdown data highlights the practical relevance of our findings. Contrary to the uniform emission cuts assumed in the idealised scenarios, the lockdown produced reductions that were highly variable in both space and time. The observed concentration changes represent a convolution of emission abatement and concurrent meteorological influences. Because percentile NO_x reductions from raw observations consistently exceed
375 those of the weather-normalised estimates generated by ML-WN and MacLeWN, it indicates that the lockdown period coincided with meteorological conditions conducive to pollutant dispersion (Acosta-Ramirez and Higham, 2022; Shi et al., 2021; Lee et al., 2020a). A direct comparison of the two weather-normalised methods shows that their estimates differ by roughly 17 % for a one-week lockdown, narrowing to 10 % for a one-month lockdown and 6 % for a three-month lockdown. These results are consistent with our simulations under idealised conditions (Fig. 2), where ML-WN’s smoothing of transient
380 signals could lead to systematic underestimation and MacLeWN shows clear larger policy intervention effects under this real-world policy implementations. Instead of resampling historical weather conditions while keeping the original emission proxies, MacLeWN estimates the influence from weather for each hour by comparing observations relative to pollutant neutral, “normalised emission” baseline, and then it subtracts weather impacts from observations.

From a regulatory aspect, the foregoing analysis indicates that for brief measures (less than 4–6 weeks), MacLeWN scheme
385 should be the preferred approach; for longer programmes (more than 3 months), ML-WN bias falls below 5 %, well within normal error bounds. Policies of intermediate length merit dual reporting with both approaches, giving policymakers a clear span of likely outcomes and sharper grounds for action. It is also important to acknowledge that even the MacLeWN approach may not entirely capture all high-frequency, weather-like variability of air quality. The validity of any weather-normalised scheme ultimately depends on the reliability of the underlying learning model. Reliance on temporal variables as
390 proxies for emissions, rather than direct emission factors, means some meteorological effects correlated with time (e.g., temperature variations throughout the day) may still confound the model; when addressing secondary pollutants such as PM_{2.5} or O₃, the predictor set must include proxies for precursor abundance so that the algorithm can disentangle chemistry–meteorology coupling rather than mis-assign chemical production to “weather” effects. Model performance also remains

Formatted: Not Highlight

Formatted: Not Highlight

Formatted: Subscript

context-dependent. In tropical or arid areas, the weak seasonality, deep convection, and episodic dust plumes can shorten meteorological autocorrelation and undermine resampling stability, while mountainous terrain introduces local circulations that are seldom captured by single-station inputs. Nonetheless, MacLeWN represents an improvement in assessing the immediate impacts of short-term policy interventions.

5 Conclusions

In this work, our logical analysis shows that the widely used machine-learning weather normalisation (ML-WN) approach could markedly underestimate the immediate benefits of short-term air quality interventions. Across eight idealised but plausible NO_x-reduction scenarios at London's Marylebone Road, the ML-WN framework missed up to 42 % of the one-week reduction signal and still understated one-month effects by ~14 %. We proposed a refined weather normalisation method MacLeWN that significantly reduced such biases, bringing re-normalised concentrations into near-identity with the known synthetic truth. When applied to the real-world COVID-19 lockdown on Marylebone Road, London, the ML-WN tended to yield more conservative estimates compared to MacLeWN, particularly for shorter intervention periods at one week (~52% vs. 69%). This further highlight that the ML-WN smooths away a substantial fraction of the abrupt emission signal.

While the proposed MacLeWN refinement is self-consistent, and potentially reduces underestimation biases in weather normalization, its reliability remains contingent on the explainability of the underlying machine learning models, necessitating cautious interpretation and continuous evaluation. Importantly, improving model transparency can increase confidence in the assessments provided by machine learning. The performance of machine learning models is directly influenced by the quality and relevance of their input variables (Geiger et al., 2020). Incorporating specific, causally relevant predictive variables—such as traffic counts, fleet compositions, and industrial emission data—can improve both model performance and explainability in air quality simulations. Strategies such as assessing multicollinearity using statistical measures like the Variance Inflation Factor (Thompson et al., 2017), applying dimensionality reduction techniques like Principal Component Analysis (Abdi and Williams, 2010), employing alternative importance measures less sensitive to correlation (e.g., permutation-based methods) (Mi et al., 2021), and using model interpretation tools like partial dependence plots (Greenwell, 2017) and SHapley Additive exPlanations (SHAP) values (Gebreyesus et al., 2023) can be employed. These approaches not only improve predictive accuracy but also enhance model robustness and interpretation, making them more reliable tools for evaluating the effectiveness of environmental policies.

Our findings have significant implications for the evaluation of air quality interventions and formulation of environmental policies. Although the underestimation of pollutant reductions by the ML-WN decreases for interventions sustained over three months (< 5%), the potential to overlook immediate benefits remains a concern, especially for short-term and emergency measures that necessitate precise evaluation for timely public health responses. Examples include policies implemented during events like sports gatherings and festivals (Yao et al., 2019; Singh et al., 2010; Andrews, 2008),

emergency responses to air pollution episode (Tian et al., 2019), and abrupt air pollution incidents such as fireworks displays, industrial accidents, volcanic eruptions, warfare, and wildfires. Additionally, significant social changes, such as those observed during the COVID-19 pandemic, have demonstrated how rapid shifts in human activity can affect air quality (Zangari et al., 2020; Gualtieri et al., 2020). Underestimation of benefits may lead to underappreciation of policy measures and improper resource allocation, affecting public confidence and future support for environmental initiatives. Moreover, weather normalisation errors propagate into downstream methodologies like the Synthetic Control Method (SCM), which constructs a synthetic treated unit from a combination of control units that were not subjected to the intervention, aiming to estimate how pollutant concentrations would have evolved in the absence of the policy (Mork et al., 2024). SCM Controlling for meteorological factors is important to isolate the effects of the intervention from natural weather variations that influence pollutant behaviour and dispersion (Dai et al., 2024). However, if the weather normalisation method introduces inherent underestimation bias in the treated unit, it would lead to a smaller apparent difference between the treated and synthetic control units, thereby skewing the assessment of the true effect of policy interventions (Ben-Michael et al., 2021; Xu, 2017).

Code Availability

Code for Machine learning-based Weather Normalisation is accessible at <https://github.com/clnair-ascm/aqpet>.

440 Data Availability

Air quality data at Marylebone Road and North Kensington can be retrieved from the Automatic Urban and Rural Network (AURN): <https://uk-air.defra.gov.uk/networks/network-info?view=aurn>. Meteorological data can be accessed at NOAA integrated surface database (ISD): <https://www.ncei.noaa.gov/products/land-based-station/integrated-surface-database>.

Competing interests

445 At least one of the (co-)authors is a member of the editorial board of Atmospheric Chemistry and Physics.

Author contribution

YD designed the experiments, performed the simulations, and prepared the manuscript with contributions from all co-authors. ARMK and ZS supervised the work.

Acknowledgements

450 This work is funded by Wellcome Trust (Ref: 227150/Z/23/Z); We also acknowledge the support from the University of Birmingham Institute of Global Innovation Clean Air Theme.

References

Abdi, H. and Williams, L. J.: Principal component analysis, Wiley Interdiscip. Rev. Comput. Stat., 2, 433-459, <https://doi.org/10.1002/wics.101>, 2010.

455 [Acosta-Ramírez, C. and Higham, J.E.: Effects of meteorology and human-mobility on UK's air quality during COVID-19. Meteorological Applications, 29\(3\), 2061, https://doi.org/10.1002/met.2061, 2022.](#)

Andrews, S. Q.: Inconsistencies in air quality metrics: 'Blue Sky' days and PM10 concentrations in Beijing, Environ. Res. Lett., 3, 034009, <https://doi.org/10.1088/1748-9326/3/3/034009>, 2008.

Bannister, E. J., Cai, X., Zhong, J., and MacKenzie, A. R.: Neighbourhood-scale flow regimes and pollution transport in cities, Bound.-Lay. Meteorol., 179, 259-289, <https://doi.org/10.1007/s10546-020-00593-y>, 2021.

460 Ben-Michael, E., Feller, A., and Rothstein, J.: The augmented synthetic control method, J. Am. Stat. Assoc., 116, 1789-1803, <https://doi.org/10.1080/01621459.2021.1929245>, 2021.

[Betancourt, C., Li, C. W., Kleinert, F., and Schultz, M. G.: Graph machine learning for improved imputation of missing tropospheric ozone data, Environmental science & technology, 57, 18246-18258, https://doi: 10.1021/acs.est.3c05104, 2023.](#)

465 [Bigi, A. and Harrison, R. M.: Analysis of the air pollution climate at a central urban background site, Atmos. Environ., 44, 2004-2012, https://doi.org/10.1016/j.atmosenv.2010.02.028, 2010.](#)

Cole, M. A., Elliott, R. J., and Liu, B.: The impact of the Wuhan Covid-19 lockdown on air pollution and health: a machine learning and augmented synthetic control approach, Environ. and Resour. Econ., 76, 553-580, <https://doi.org/10.1007/s10640-020-00483-4>, 2020.

470 [Dai, Q., Hou, L., Liu, B., Zhang, Y., Song, C., Shi, Z., Hopke, P. K., and Feng, Y.: Spring Festival and COVID-19 lockdown: disentangling PM sources in major Chinese cities, Geophys. Res. Lett., 48, e2021GL093403, https://doi.org/10.1029/2021GL093403, 2021.](#)

Dai, Y., Cai, X., Zhong, J., and MacKenzie, A. R.: Chemistry, street canyon geometry, and emissions effects on NO₂ “hotspots” and regulatory “wiggle room”, *npj Clim. and Atmos. Sci.*, 5, 102, <https://doi.org/10.1038/s41612-022-00323-w>, 2022.

Dai, Y., Liu, B., Tong, C., and Shi, Z.: aqpet—An R package for air quality policy evaluation, *Enviro. Model. Softw.*, 106052, <https://doi.org/10.1016/j.envsoft.2024.106052>, 2024.

Davies, N. G., Barnard, R. C., Jarvis, C. I., Russell, T. W., Semple, M. G., Jit, M., and Edmunds, W. J.: Association of tiered restrictions and a second lockdown with COVID-19 deaths and hospital admissions in England: a modelling study, *Lancet Infect. Dis.*, 21, 482-492, [https://doi.org/10.1016/S1473-3099\(20\)30984-1](https://doi.org/10.1016/S1473-3099(20)30984-1), 2021.

Di Bernardino, A., Iannarelli, A. M., Casadio, S., Mevi, G., Campanelli, M., Casasanta, G., Cede, A., Tiefengraber, M., Siani, A. M., and Spinei, E.: On the effect of sea breeze regime on aerosols and gases properties in the urban area of Rome, Italy, *Urban Clim.*, 37, 100842, <https://doi.org/10.1016/j.uclim.2021.100842>, 2021.

Ding, J., Dai, Q., Li, Y., Han, S., Zhang, Y., and Feng, Y.: Impact of meteorological condition changes on air quality and particulate chemical composition during the COVID-19 lockdown, *J. Environ. Sci.*, 109, 45-56, <https://doi.org/10.1016/j.jes.2021.02.022>, 2021.

Ding, J., Dai, Q., Fan, W., Lu, M., Zhang, Y., Han, S., and Feng, Y.: Impacts of meteorology and precursor emission change on O₃ variation in Tianjin, China from 2015 to 2021, *J. Environ. Sci.*, 126, 506-516, <https://doi.org/10.1016/j.jes.2022.03.010>, 2023.

Faustini, A., Rapp, R., and Forastiere, F.: Nitrogen dioxide and mortality: review and meta-analysis of long-term studies, *European Respiratory Journal*, 44, 744-753, <https://doi.org/10.1183/09031936.00114713>, 2014.

Fuller, R., Landrigan, P. J., Balakrishnan, K., Bathan, G., Bose-O'Reilly, S., Brauer, M., Caravanos, J., Chiles, T., Cohen, A., and Corra, L.: Pollution and health: a progress update, *Lancet Planet. Health.*, 6, e535-e547, [https://doi.org/10.1016/S2542-5196\(22\)00090-0](https://doi.org/10.1016/S2542-5196(22)00090-0), 2022.

Gebreyesus, Y., Dalton, D., Nixon, S., De Chiara, D., and Chinnici, M.: Machine learning for data center optimizations: feature selection using Shapley additive exPlanation (SHAP), *Future Internet*, 15, 88, <https://doi.org/10.3390/fi15030088>, 2023.

Geddes, J. A., Wang, B., and Li, D.: Ozone and nitrogen dioxide pollution in a coastal urban environment: The role of sea breezes, and implications of their representation for remote sensing of local air quality, *J. Geophys. Res.: Atmos.*, 126, e2021JD035314, <https://doi.org/10.1029/2021JD035314>, 2021.

- Geiger, R. S., Yu, K., Yang, Y., Dai, M., Qiu, J., Tang, R., and Huang, J.: Garbage in, garbage out? Do machine learning application papers in social computing report where human-labeled training data comes from?, Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency, 325-336, <https://doi.org/10.1145/3351095.3372862>, 2020.
- 505 [Goodfellow, I., Bengio, Y., Courville, A., and Bengio, Y.: Deep learning, 2, http://www.deeplearningbook.org, MIT press Cambridge2016.](http://www.deeplearningbook.org)
- Grange, S. K., Carslaw, D. C., Lewis, A. C., Boleti, E., and Hueglin, C.: Random forest meteorological normalisation models for Swiss PM 10 trend analysis, Atmos. Chem. Phys., 18, 6223-6239, <https://doi.org/10.5194/acp-18-6223-2018>, 2018.
- Greenwell, B. M.: pdp: An R package for constructing partial dependence plots, R J., 9, 421, <https://doi.org/10.32614/RJ-2017-016>, 2017.
- 510 Gualtieri, G., Brilli, L., Carotenuto, F., Vagnoli, C., Zaldei, A., and Gioli, B.: Quantifying road traffic impact on air quality in urban areas: A Covid19-induced lockdown analysis in Italy, Environ. Pollut., 267, 115682, <https://doi.org/10.1016/j.envpol.2020.115682>, 2020.
- Harrison, R. M., Van Vu, T., Jafar, H., and Shi, Z.: More mileage in reducing urban air pollution from road traffic, Environ. Int., 149, 106329, <https://doi.org/10.1016/j.envint.2020.106329>, 2021.
- 515 Jeanjean, A. P., Buccolieri, R., Eddy, J., Monks, P. S., and Leigh, R. J.: Air quality affected by trees in real street canyons: The case of Marylebone neighbourhood in central London, Urban Forestry & Urban Greening, 22, 41-53, <https://doi.org/10.1016/j.ufug.2017.01.009>, 2017.
- Kilkenny, M. F. and Robinson, K. M.: Data quality:“Garbage in–garbage out”, Health Inf. Manag. J., <https://doi.org/10.1177/1833358318774357>, 2018.
- 520 LeDell, E. and Poirier, S.: H2o automl: Scalable automatic machine learning, Proceedings of the AutoML Workshop at ICML, available at: https://www.automl.org/wp-content/uploads/2020/07/AutoML_2020_paper_61.pdf (last access: 19 May 2025) 2020.
- Lee, J. D., Drysdale, W. S., Finch, D. P., Wilde, S. E., and Palmer, P. I.: UK surface NO2 levels dropped by 42% during the COVID-19 lockdown: impact on surface O3, Atmos. Chem. Phys., 20, 15743-15759, <https://doi.org/10.5194/acp-20-15743-2020>, 2020a.
- 525 Lee, K. K., Bing, R., Kiang, J., Bashir, S., Spath, N., Stelzle, D., Mortimer, K., Bularga, A., Doudesis, D., and Joshi, S. S.: Adverse health effects associated with household air pollution: a systematic review, meta-analysis, and burden estimation study, Lancet Glob. Health, 8, e1427-e1434, [https://doi.org/10.1016/S2214-109X\(20\)30343-0](https://doi.org/10.1016/S2214-109X(20)30343-0), 2020b.

- 530 Masson, V., Lemonsu, A., Hidalgo, J., and Voogt, J.: Urban climates and climate change, *Annu. Rev. Environ. and Resourc.*, 45, 411-444, <https://doi.org/10.1146/annurev-environ-012320-083623>, 2020.
- Meng, X., Liu, C., Chen, R., Sera, F., Vicedo-Cabrera, A. M., Milojevic, A., Guo, Y., Tong, S., Coelho, M. d. S. Z. S., and Saldiva, P. H. N.: Short term associations of ambient nitrogen dioxide with daily total, cardiovascular, and respiratory mortality: multilocation analysis in 398 cities, *bmj*, 372, <https://doi.org/10.1136/bmj.n534>, 2021.
- 535 Mi, X., Zou, B., Zou, F., and Hu, J.: Permutation-based identification of important biomarkers for complex diseases via machine learning models, *Nat. Commun.*, 12, 3008, <https://doi.org/10.1038/s41467-021-22756-2>, 2021.
- Mork, D., Delaney, S., and Dominici, F.: Policy-induced air pollution health disparities: Statistical and data science considerations, *Science*, 385, 391-396, <https://doi.org/10.1126/science.adp1870>, 2024.
- Seo, J., Park, D.-S. R., Kim, J. Y., Youn, D., Lim, Y. B., and Kim, Y.: Effects of meteorology and emissions on urban air quality: a quantitative statistical approach to long-term records (1999–2016) in Seoul, South Korea, *Atmos. Chem. Phys.*, 18, 16121-16137, <https://doi.org/10.5194/acp-18-16121-2018>, 2018.
- Shi, Z., Song, C., Liu, B., Lu, G., Xu, J., Van Vu, T., Elliott, R. J., Li, W., Bloss, W. J., and Harrison, R. M.: Abrupt but smaller than expected changes in surface air quality attributable to COVID-19 lockdowns, *Sci. Adv.*, 7, eabd6696, <https://doi.org/10.1126/sciadv.abd6696>, 2021.
- 545 Singh, D., Gadi, R., Mandal, T., Dixit, C., Singh, K., Saud, T., Singh, N., and Gupta, P. K.: Study of temporal variation in ambient air quality during Diwali festival in India, *Environ. Monit. Assess.*, 169, 1-13, <https://doi.org/10.1007/s10661-009-1145-9>, 2010.
- Thompson, C. G., Kim, R. S., Aloe, A. M., and Becker, B. J.: Extracting the variance inflation factor and other multicollinearity diagnostics from typical regression results, *Basic Appl. Soc. Psych.*, 39, 81-90, <https://doi.org/10.1080/01973533.2016.1277529>, 2017.
- 550 Tian, J., Cai, T., Shang, J., Schauer, J. J., Yang, S., Zhang, L., and Zhang, Y.: Effects of the emergency control measures in Beijing on air quality improvement, *Atmos. Pollut. Res.*, 10, 580-586, <https://doi.org/10.1016/j.apr.2018.10.005>, 2019.
- Vu, T. V., Shi, Z., Cheng, J., Zhang, Q., He, K., Wang, S., and Harrison, R. M.: Assessing the impact of clean air action on air quality trends in Beijing using a machine learning technique, *Atmos. Chem. Phys.*, 19, 11303-11314, <https://doi.org/10.5194/acp-19-11303-2019>, 2019.
- 555 <https://doi.org/10.5194/egusphere-2024-1237>, 2024.
- Woolley, G. J., Rutter, N., Wake, L., Vionnet, V., Derksen, C., Essery, R., Marsh, P., Tutton, R., Walker, B., and Lafaysse, M.: Multi-physics ensemble modelling of Arctic tundra snowpack properties, *The Cryosphere*, 18, 5685-5711, <https://doi.org/10.5194/egusphere-2024-1237>, 2024.

- Xu, Y.: Generalized synthetic control method: Causal inference with interactive fixed effects models, *Polit. Anal.*, 25, 57-76,
560 <https://doi.org/10.1017/pan.2016.2>, 2017.
- Yao, L., Wang, D., Fu, Q., Qiao, L., Wang, H., Li, L., Sun, W., Li, Q., Wang, L., and Yang, X.: The effects of firework
regulation on air quality and public health during the Chinese Spring Festival from 2013 to 2017 in a Chinese megacity,
Environ. Int., 126, 96-106, <https://doi.org/10.1016/j.envint.2019.01.037>, 2019.
- Zangari, S., Hill, D. T., Charette, A. T., and Mirowsky, J. E.: Air quality changes in New York City during the COVID-19
565 pandemic, *Sci. Total Environ.*, 742, 140496, <https://doi.org/10.1016/j.scitotenv.2020.140496>, 2020.
- Zhong, J., Cai, X.-M., and Bloss, W. J.: Coupling dynamics and chemistry in the air pollution modelling of street canyons: A
review, *Environ. Pollut.*, 214, 690-704, <https://doi.org/10.1016/j.envpol.2016.04.052>, 2016.