Reply to Reviewer #1:

This manuscript presents a significant methodological advancement in weather normalization techniques by rigorously identifying and quantifying the underestimation bias inherent in traditional ML-WN approaches when assessing short-term air quality interventions. This work is well-motivated, with clear relevance to air quality policy assessment, and the methodological framework (e.g. synthetic intervention scenarios and COVID-19 lockdown case study) is innovative. Overall, I find the work valuable and recommend it for publication after the following concerns are addressed:

Thank you for taking the time to review our work and for your encouraging feedback. We appreciate your positive assessment and will address the minor clarifications promptly.

1. The resampling methodology (Eq. 4-6) needs more detailed explanation. What is the statistical justification for the number of resamples (n=300)? Was convergence tested?

   Thank you for your suggestions. We have expanded the explanation of the resampling method (Equations 3–6). For additional clarification, please refer to Question 2.

   In practice our choice of n = 300 resamples is based on literature precedent and to make our results comparable to other works. For example, in rmweather R package, the de-facto implementation of Grange et al.'s meteorological-normalisation method uses n_samples = 300 as its default and recommended setting (https://cran.r-project.org/web/packages/rmweather/index.html); early applications of the technique likewise adopt 250–350 draws (e.g., Vu et al., 2019 for Beijing air-quality trends https://doi.org/10.5194/acp-19-11303-2019), establishing this range as a community benchmark, as they found good convergency after 300. We added related references into the revised manuscript to support the choice of resampling number.

2. The description of how MacLeWN removes temporal variation correlated with emissions could be expanded. Currently, equations (3)-(6) are concise but lack an intuitive explanation of how MacLeWN differentiates policy-driven from meteorology-driven variability.

   Thank you for very helpful suggestions. We broadened the description of the MacLeWN approach in the Methods section of the revised manuscript.

   "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." and section 2.2 for MacLeWN approach.

3. The claim that MacLeWN "explicitly accounts for intervention timing" needs elaboration. How does the model distinguish abrupt policy signals from stochastic noise?

   We thank the reviewer for this request for clarification. We found the above statement potentially confusing, and we amended the sentence in the abstract.

   MacLeWN separates policy signals from random noise through a two-stage filter. First, it averages out all diurnal- and weekly-emission proxies (hour, weekday, season), producing a "neutral-emission" baseline that retains only long-term trends and cuts the high-frequency variability normally linked to anthropogenic cycles (e.g. traffic). Second, it computes an hour-specific meteorological factor by contrasting observed concentrations with this baseline. When that factor is removed from the raw observations, any remaining step-like deviation is the part that cannot be explained by the stochastic spread of meteorology and should be attributed to emissions. For additional clarification, please refer to Question 2.

4. The synthetic intervention approach is creative but raises questions about ecological validity. How well do these idealized scenarios represent real-world policy implementations where emission changes may be more gradual or heterogeneous?

   We appreciate this important point and have added clarifying text. In brief,

   1) Range of temporal profiles. Besides the one-week "step" scenario, our test matrix contains phased-out (S6, S7) and cyclic (S8) patterns that mimic staggered or variable real-world controls. MacLeWN shows the same advantage over ML-WN across all three profiles, indicating that its benefit is not limited to an instantaneous step change.
   2) We added more sentences accordingly to make this clear. Specifical, In Section 2.1, "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." In Section 3.1, "Importantly, 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."

5. The policy implications of these findings should be expanded. For instance, how should air quality managers choose between methods when evaluating interventions of different durations?

   Thank you for your suggestions. We added relative content in the discussion section. "From a regulatory aspect, the foregoing analysis indicates that for brief measures (less than 4–6 weeks), MacLeWN scheme 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."

6. The manuscript indicates that ML-WN has "black-box" challenge, whether MacLeWN has this kind of challenge. I recommend the authors include SHAP (SHapley Additive exPlanations) or partial dependence plots for key variables to quantify variable contributions or interactions in MacLeWN.

   Thank you for very helpful suggestions. Because both ML-WN and the MacLeWN employ machine learning models to normalise meteorological influences from pollutant trends, they face the same "black box" interpretability challenge. In fact, the underlying ML model for both approaches is the same. To enhance transparency for the model, we have now computed partial dependence plots (PDPs) for meteorological predictors to the revised Supplementary Information (Figures S13-S14).

7. This study only focuses on NOx at two London sites, please discuss whether MacLeWN's improvements would hold for other pollutants (e.g., PM2.5, O3) where meteorological influences and emission sources differ. Discuss potential limitations when applying MacLeWN in regions with different climatology (e.g, tropical or arid zones) and complex terrain.

   Thank you for your suggestions. We limited our proof-of-concept to NOx concentrations at London sites because (1) NOx can be considered as a passive scalar without involving chemistry; and (2) London Marylebone Road site has significant traffic volume and could see the biggest impact from COVID-19 lockdown. We discussed the above limitations in the discussion section. "It is also important to acknowledge that even the MacLeWN approach may not entirely capture all high-frequency, weatherlike 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 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 PM2.5 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 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."

8. Line 81-84: The 3-hour threshold for linear interpolation of missing data seems arbitrary. Please justify or reference established practices in similar studies.

   Thank you for your suggestion. There are several recent peer reviewed papers that explicitly apply the same missing-data strategy (i.e., linear interpolation for very short gaps <3h), including:

   1) *Betancourt, C., Li, C.W., Kleinert, F. and Schultz, M.G., 2023. Graph machine learning for improved imputation of missing tropospheric ozone data. Environmental science & technology, 57(46), pp.18246-18258.*
   2) *Woolley, G.J., Rutter, N., Wake, L., Vionnet, V., Derksen, C., Essery, R., Marsh, P., Tutton, R., Walker, B., Lafaysse, M. and Pritchard, D., 2024. Multi-physics ensemble modelling of Arctic tundra snowpack properties. The Cryosphere, 18(12), pp.5685-5711.*

   We added the related references at the end of original sentence accordingly.

9. Line 370: "unproper" → "improper"

   Thank you for pointing this out. This typo has been corrected.

10. In Figure 2, the image resolution is insufficient, making axis labels and annotations difficult to read. There is visible overlap between panel labels (a-d) the actual figure content, requiring layout adjustment. Consider adding error bars to quantify variability in the "actual" vs estimated effects.

    Thank you for drawing our attention to the presentation quality of Figure 2. The figure has been regenerated and the (a–d) identifiers moved outside the plotting area to eliminate overlap.

    In the synthetic-scenario experiment the "actual" series is analytically prescribed and the MacLeWN/ML-WN estimates are deterministic point outputs of the trained models; hence conventional sampling error bars are not applicable. The only stochastic component is the Monte-Carlo resampling error, with a sufficiently large number of resamples (as noted in Q1), this error becomes negligible.

11. In Figure 3, could the authors please clarify why the observed NOx concentrations show a larger percentage reduction (-53.7%) than ML-WN (-51.3%), despite exhibiting smaller absolute decreases (58.1 vs. 71.9 μg/m^3)? This apparent contradiction warrants explanation, particularly regarding how the different baseline concentrations influence these percentage comparisons. Additionally, could you comment on whether this phenomenon affects the interpretation of model performance differences, especially for longer intervention periods?

    Thank you for your comments. The apparent inconsistency is mainly because a baseline effect driven by weather. During the three-month policy window, the observed pre-lockdown NOx was about 108 μg/m^3, wheras the ML-WN "deweathered" value was about 140 μg/m^3. This gap shows that, over the entire intervention period, meteorological conditions improved pollution dispersion processes in London, which has also been reported in previous studies (https://doi.org/10.5194/acp-20-15743-2020; https://doi.org/10.1002/met.2061; DOI: 10.1126/sciadv.abd6696).

The baseline discrepancy does not alter our appraisal of model skill. Because the weather-normalised series starts from a higher pre-lockdown level, a given absolute fall is divided by a larger denominator, yielding a smaller percentage change; once that scaling is recognised, the absolute–percentage divergence disappears. Importantly, for longer interventions the ML-WN bias we quantify (<5 % beyond three months) is already so small that the choice of absolute versus relative metrics leaves the ranking of the two methods unchanged. Hence the baseline effect is a matter of presentation, not of substantive model-performance difference.

12. In Table 3, footnote should specify if uncertainties represent 1σ or 95% CI.

Thank you for pointing this out. We have updated Table 3 footnote to specify that the reported uncertainties correspond to one standard error. "Note: In each case, the data are detrended following the method in Sect. 2.3.2; the uncertainties are expressed as the standard error."