Reply to Reviewer #2:

This manuscript presents a thoughtful re-evaluation of machine learning-based weather normalization (ML-WN) methods in the context of short-term air quality interventions. It reveals a critical shortcoming in traditional ML-WN, which underestimates emission reductions following abrupt measures. The authors propose a refined method (MacLeWN), supported by synthetic experiments and real-world application during the COVID-19 lockdown in London, which improves estimation accuracy. The work is timely and policy-relevant, offering an improved framework for assessing short-term regulatory impacts. I find the methodological innovation and policy implications and recommend publication after minor clarification.

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.

L101, the dataset is split into 80% for model training and 20% for evaluation at each site. Is this split performed randomly? Given the temporal continuity in meteorological conditions and emissions, data correlation may affect the validity of this approach.

Thank you for your comments. For machine learning research, data solutions can be divided into 90/10, 80/20, and 70/30 approaches. The choice of 80/20 split was followed the "80/20 hold-out rule" (often traced back to the Pareto principle and formalised for model evaluation in standard ML references such as (Goodfellow et al. 2016; Nguyen et al. 2021). This split is also widely adopted in air-quality ML studies, including (Grange et al. 2018) and (Vu et al. 2019), because it leaves a sufficiently large, unseen subset for honest skill assessment while retaining enough data for stable training. We added the related literatures for that sentence in the revised manuscript.

L223, this underestimation occurs because the ML-WN method may not fully capture abrupt changes in emission patterns "due to the smoothing effects inherent in machine learning models". What are the smoothing effects mentioned? Since the proposed MacLeWN method is also based on machine learning, doesn't it also exhibit similar smoothing effects?

Thank you for highlighting the need for precision. In the revised manuscript we now amend the sentence to make it clear. "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 will be discussed further in the discussion.". One main source of ML-WN's bias is that it resamples weather conditions from the entire historical record and splices them into the intervention window, combining genuinely reduced emissions with weather conditions that never occurred and thus smoothing the step change. MacLeWN avoids this issue by first quantifying the meteorological contribution for each hour from the emission normalised condition and then subtracting that influence from observations. We have inserted additional sentences in the Discussion to make this reasoning explicit in the revised manuscript "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."

Figure 3 in Sec 3.2, in the three-month lockdown scenario, both methods appear to perform similarly, and neither seems to compare with the observed trends particularly well. Is this interpretation correct?

Thank you for your comments. In Figure 3, the blue bears (observed NO_x concentration changes) capture both the emission reductions from the COV-19 lockdown and the concurrent meteorological effects. Because these bars are consistently higher than the corresponding weather-normalised estimates from ML-WN and MacLeWN, it indicates that the lockdown period benefited from meteorological conditions that promoted pollutant dispersion.

When we compare the two weather-normalised method directly, their estimates diverge by about 17% for a one-week lockdown effects, decrease to 10% when the lockdown is extended to one month, and fall to 6% over three months. This progressive convergence aligns with the idealised scenario analysis in Section 3.1, and we have expanded Discussion section to emphasise this point.

"… 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 NOx reductions from raw observations consistently exceed 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. 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 signals could lead to systematic underestimation and MacLeWN shows clear larger policy intervention effects under this real-world policy implementations…."

Reference

Goodfellow, Ian, Yoshua Bengio, Aaron Courville, and Yoshua Bengio. 2016. *Deep learning* (MIT press Cambridge).

Grange, Stuart K, David C Carslaw, Alastair C Lewis, Eirini Boleti, and Christoph Hueglin. 2018. 'Random forest meteorological normalisation models for Swiss PM 10 trend analysis', *Atmospheric Chemistry and Physics*, 18: 6223-39.

Nguyen, Quang Hung, Hai-Bang Ly, Lanh Si Ho, Nadhir Al-Ansari, Hiep Van Le, Van Quan Tran, Indra Prakash, and Binh Thai Pham. 2021. 'Influence of data splitting on performance of machine learning models in prediction of shear strength of soil', *Mathematical Problems in Engineering*, 2021: 4832864.

Vu, Tuan V, Zongbo Shi, Jing Cheng, Qiang Zhang, Kebin He, Shuxiao Wang, and Roy M Harrison. 2019. 'Assessing the impact of clean air action on air quality trends in Beijing using a machine learning technique', *Atmospheric Chemistry and Physics*, 19: 11303-14.