

Meteorological Normalization or Deweathering for Predicting Air Pollutant Concentration: Pitfalls and Limitations

Valentino Petrić¹, Mario Lovrić^{2,3}, and Bernhard C. Geiger^{4,5}

¹Ascalia d.o.o., 40000 Čakovec, Croatia

²Centre for bioanthropology, Institute for Anthropological Research, 10000 Zagreb, Croatia

³Lisbon Council, BE-1040 Brussels, Belgium

⁴Know-Center, 8010 Graz, Austria

⁵Signal Processing and Speech Communication Laboratory, Graz University of Technology, 8010 Graz, Austria

Abstract. Meteorological normalization is a key concept in studying anthropogenic effects on air pollutant concentrations and its temporal trends. While apparently successful in revealing anthropogenic effects and often used, there are downsides to the methods and limitations which should be taken into account when using it. This study examines the concept of meteorological normalization in the context of predicting air pollutant concentrations, particularly focusing on the method's theoretical background and implicit assumptions. We provide a rigorous analysis that outlines the conditions under which meteorological normalization can effectively estimate expected values of response variables, which are commonly time series of air pollutant concentrations such as PM₁₀. Through our analysis, we identify critical assumptions, such as the independence of meteorological and non-meteorological variables, which can significantly impact the validity of the method's outcomes. We also explore potential failure modes of meteorological normalization when these assumptions are violated, offering practical examples and visualizations that highlight the limitations of this technique. Our findings suggest that while meteorological normalization can be a useful tool, careful consideration of its applicability is necessary to avoid misleading conclusions in air quality assessments.

1 Introduction

While it is known that air pollutants are mostly driven by emission from a diversity of sources, such as traffic (Colville et al., 2001), residential (Bari et al., 2011) and industrial combustion (Bing Chen and Zhang, 2011) there is also a significant contribution driven by local and global meteorological factors such as temperature (Hassan et al., 2020), wind (Munir et al., 2017), pressure and humidity (Speranza and Caggiano, 2023) for, e.g., PM₁₀ and PM_{2.5} (particulate matter with aerodynamic diameters <10 µm and <2.5 µm, respectively) but also for pollutants such as NO₂ and carbon fractions (Šimić et al., 2020).

The EU directive on air quality (Directive 2008/50/EC) sets limit values for particulate matter in Europe, i.e., 40 µg m⁻³ for PM₁₀ and 25 µg m⁻³ for PM_{2.5}, while there are also limits for daily values, e.g., 50 µg m⁻³ for PM₁₀. Currently, the effect induced by the adoption of restricted traffic areas on PM is a debated topic due to the difficulties in disentangling the effect from of other effects such as seasonality and the influence of weather conditions (Munir et al., 2017). Meteorological conditions can hence influence the exacerbation, dispersion, and reactivity of PM in urban areas (Speranza and Caggiano, 2023).

A well-adopted strategy for disentangling the effects of meteorological factors and emission is meteorological normalization (Grange and Carslaw, 2019; Grange et al., 2018) or de-weathering. This is especially useful when studying imission, i.e., concentrations of air pollutants measured at monitoring stations or sensors which can be influence by local meteorology. While the concept was known, it gained a lot of momentum due to its successful synergy with machine learning and the availability of an easy-to-use R-scripting language library¹ (Grange et al., 2018). Indeed, according to Google Scholar and at the time of writing this paper, the two establishing works (Grange et al., 2018) and (Grange and Carslaw, 2019) were cited over 300 and 200 times respectively. This points to the increasing popularity of the method, which we also used in our previous works in a Python adaption (Lovrić et al., 2021, 2022) to assess the effect of the COVID-19 lockdowns in 2020.

Meteorological normalization uses data-driven regression models, most often Random Forests (Breiman, 2001) and gradient boosted models (Gagliardi and Andenna, 2021) as they have shown superior performance over generalized linear and additive models (Wong et al., 2023), to predict daily or hourly pollutant concentrations given seasonality and meteorological variables. Once the predictive model becomes trained and maps the relationships to the predictive meteorological (X) and non-meteorological, often seasonal (T) features, one permutes or shuffles the meteorological variables while the seasonal ones remain fixed. This results in a “corrected model” which is excluding the meteorological contribution.

While the approach often shows reasonable results, there is also criticism towards the method. Qiu et al. (2022) argue that, despite the extensive use of various meteorology correction methods in numerous studies, there remains significant uncertainty about their effectiveness in accounting for meteorological variability to accurately estimate counterfactual air quality and identify the true effects of changes in anthropogenic emissions. They also point out that while many studies lean heavily on the predictive accuracy of their statistical models, their predictive performance does not necessarily equate to accurate estimations of counterfactual scenarios and causal impacts. They also argue that emission-driven trends without influence from meteorological variability cannot be derived from data, especially for geographical locations where meteorologically influenced environmental processes are the main drivers of, e.g., dust emissions. Further criticism is oriented towards the inability to use the “corrected” model for extrapolation to scenarios that are not covered in the data (Nicely et al., 2020; Zabrocki et al., 2022), such as a new wind direction at the measurement station.

In this work, we provide a mathematical and at the same time accessible analysis of meteorological normalization and derive conditions under which it converges to an expectation conditioned on seasonal variables only (Section 2). Selectively violating these conditions, we illustrate the failure modes of meteorological normalization, i.e., we provide easy-to-understand examples where this technique does not behave as expected, but which are still of immense practical relevance (Section 3). With these insights, in Section 4 we discuss situations in which meteorological normalization can be, or should *not* be applied, and outline ideas to extend the technique to be applicable in these latter situations. To keep the exposition accessible, all proofs and derivations, as well as details for the numerical experiments, are deferred to the Supplementary Material.

¹<https://github.com/skgrange/rmweather>

2.1 Problem Setup and Notation

We consider the setting where a certain response variable should be estimated from a set of descriptors, of which a subset is referred to as “meteorological” variables. The descriptors are hence independent variables such as temperature (temp), relative humidity (rh), pressure (press), wind speed (ws), etc. and are used as predictive variables when estimating the response variable, such as pollutant concentration PM_{10} (Grange et al., 2018). Assuming a probabilistic setting, the descriptors and the response variable are random variables (RVs). Specifically, if X denotes a vector of real-valued meteorological RVs (e.g. temp, rh, press, ws), T denotes real-valued descriptor RVs that are not meteorological (e.g. Julian day), and Y denotes a real-valued response variable that should be estimated (such as PM_{10}), then the behavior of (X, T, Y) is determined by the joint distribution $p_{X,T,Y}$.² We do not make any assumptions on the causal direction of the learning problem, i.e., we do not assume that the target Y is the output of a causal mechanism to inputs (X, T) (but see also Section 4.2). We admit, however, that the causal setting, i.e., where the response variable Y is a causal effect of both meteorological and non-meteorological variables, is the more common and scientifically often more interesting setting.

To learn the relationship between (X, T) and Y , we assume to have access to a dataset \mathcal{D} of N independent realizations (e.g., hourly or daily meteorological measurements and immission data) of (X, T, Y) drawn from $p_{X,T,Y}$,

$$\mathcal{D} := \{(x_i, t_i, y_i)\}_{i=1, \dots, N}. \quad (1)$$

Using this dataset and assuming an additive loss function $l: \mathbb{R}^2 \rightarrow \mathbb{R}_0^+$ (such as the squared Euclidean loss or the 0-1 loss) we can learn a (machine learning) model f

$$\hat{f} \in \arg \min_{f \in \mathcal{F}} \frac{1}{N} \sum_{i=1}^N l(y_i, f(x_i, t_i)) \quad (2)$$

where \mathcal{F} is the hypothesis class from which we select \hat{f} . For example, in the original paper (Grange et al., 2018), \mathcal{F} was chosen as the set of random forest models.

2.2 Meteorological Normalization using Mathematical Notation

Meteorological normalization proceeds with the learned model \hat{f} as follows: To compensate the effect of meteorological variables X (e.g. temp, press, rh, ws) on the target Y for a fixed set of (non-meteorological) descriptors $T = t$ (e.g., Julian day), the learned model \hat{f} is evaluated for realizations r_i of X drawn from the dataset \mathcal{D} with replacement.³ We denote the random draws from $\mathcal{D}_x = \{x_i\}_{i=1, \dots, N}$ as

$$\mathcal{R} = \{r_i: r_i \in \mathcal{D}_x\}_{i=1, \dots, R} \quad (3)$$

²While we believe that our results apply to more general settings, we here assume that this distribution is absolutely continuous w.r.t. the Lebesgue or the counting measure, which allows us to use densities.

³The R package **rmweather** also allows to sample without replacement, cf. (? , p. 13).

and note that R can be larger, smaller, or equal to N (if $R > N$, then some realizations of the meteorological variable X are sampled multiple times). With this, meteorological normalization M is defined as an average of the model predictions obtained by feeding the model with the correct non-meteorological descriptors, but with randomly resampled meteorological
 85 descriptors. Mathematically, the meteorological normalization is thus defined as the function

$$M(t) = \frac{1}{R} \sum_{i=1}^R \hat{f}(r_i, t). \quad (4)$$

The following proposition characterizes the properties of $M(t)$ in settings where the number of random samples and the dataset size grow to infinity.

Proposition 1. *For R approaching infinity (i.e., for many resampling operations), meteorological normalization converges to
 90 the average of the model (2) w.r.t. the meteorological variables, i.e.,*

$$M(t) \xrightarrow{R \rightarrow \infty} \frac{1}{N} \sum_{i=1}^N \hat{f}(x_i, t). \quad (5a)$$

If further N approaches to infinity (i.e., for large datasets), meteorological normalization converges to the expectation of the model w.r.t. the meteorological variables, i.e.,

$$M(t) \xrightarrow{R, N \rightarrow \infty} \mathbb{E}_{X \sim p_X} [\hat{f}(X, t)] \quad (5b)$$

95 where p_X is the marginal distribution of X induced by $p_{X,T,Y}$.

Hence, meteorological normalization does not inherently make a statement about the response variable Y , but rather about the average/expected outcome of the learned model. The following result, which is positive, shows that insights about the response variable Y are possible under certain conditions:

Proposition 2. *Suppose that i) learning pollutant concentration from seasonal meteorological variables was successful in the
 100 sense that*

$$\hat{f}(x, t) = \mathbb{E}_{Y \sim p_{Y|x,t}} [Y] \quad (6a)$$

where $p_{Y|x,t}$ is the shorthand notation for the conditional distribution of Y given that $X = x$ and $T = t$, and that ii) the marginal distribution $p_{X,T}$ factorizes as

$$p_{X,T} = p_X p_T. \quad (6b)$$

105 Then, as R and N tend to infinity,

$$M(t) \xrightarrow{R, N \rightarrow \infty} \mathbb{E}_{Y \sim p_{Y|t}} [Y] \quad (6c)$$

where $p_{Y|t}$ is the shorthand notation for the conditional distribution of Y given that $T = t$.

In essence, the statement of Proposition 2 is that, under certain conditions, meteorological normalization estimates the expectation of the response variable (such as PM_{10}) conditioned only on the non-meteorological variables (such as the Julian date in (Grange et al., 2018)). It is these certain conditions that (may) justify the usage of meteorological normalization, and the violation of which may yield a meteorological normalization $M(t)$ that behaves unexpectedly, and even undesirably.

Let us now briefly discuss whether the conditions of Proposition 2 are fulfilled in realistic scenarios. The factorization in (6b) is equivalent to the requirement that X and T are marginally independent, which is a severe restriction. For example, the common setting in which T contains an indicator of time (such as Julian day) and in which X contains variables such as temperature, precipitation, etc., is ruled out by this assumption, as temperature shows a seasonal dependence on Julian day. Instead of the factorization in (6b), Proposition 2 can also be shown to hold if X and Y are conditionally independent given T , i.e., if $Y - T - X$ is a Markov tuple. In this case, $\mathbb{E}_{Y \sim p_{Y|X,t}}[Y] = \mathbb{E}_{Y \sim p_{Y|t}}[Y]$. If learning is successful in the sense of (6a), it follows that $\hat{f}(x, t) = \hat{f}'(t)$, i.e., the learned model does not depend on the meteorological descriptors. In such a setting, meteorological normalization becomes obsolete. Also this setting is quite unlikely, as descriptors that are not directly connected to the response (as X in this case) will rarely be the subject of study.

The requirement that learning was successful in the sense of (6a) is not less restrictive. On the one hand, to obtain (6a) one must choose the loss function l in (2) such that the learned function \hat{f} coincides with the conditional expectation; for example, l could be chosen to be the squared Euclidean distance for a real-valued response variable Y . On the other hand, learning must be successful in the sense that \hat{f} *interpolates and extrapolates* well, i.e., (6a) should hold not only for (x, t) tuples in the dataset \mathcal{D} , but also for (x, t) combinations not seen during training (i.e., model generalization). This is important not only in the standard setting of meteorological normalization (e.g., because due to random resampling there will be (r, t) combinations not seen during training), but especially also if meteorological normalization is used in forecasting settings, i.e., in settings where T includes time indicators and $M(t)$ is evaluated for t outside the training range.

3 Experiments

The following examples will investigate meteorological normalization in settings violating the assumptions of Proposition 2 and illustrate its limitations. In the experiments, we compare the results of meteorological normalization, $M(t)$, with the value of the response only due to direct (i.e., unmediated) influence of T , $y_{\text{direct}}(t)$, and the expected response when conditioned only on T , $\mathbb{E}_{Y \sim p_{Y|t}}[Y]$.⁴ All our numerical experiments are performed using Python 3.11 with the libraries numpy (version 2.0.1) and scikit-learn (version 1.5.1) for calculation and model development, alongside matplotlib (version 3.9.0) and scipy (version 1.14.0) for displaying results. Details on the individual experiments are given in the Supplementary Materials.

⁴In the literature on causality, $y_{\text{direct}}(t)$ and $\mathbb{E}_{Y \sim p_{Y|t}}[Y]$ are related to the *direct* and *total effect* of one variable on another, respectively.

3.1 Experiment 1: Meteorological Normalization with Correlated Inputs

The first experiment is based on a setting in which (6b) does not hold, i.e., where meteorological variables may depend on non-meteorological variables. Specifically, we consider a linear model for the response variable:

$$Y = \alpha T - \beta X + \gamma, \quad (7)$$

140 where T is, e.g., a seasonal non-meteorological variable like the Julian date and X are deemed as meteorological contributors to the prediction models, while γ is some offset. For example, Y could indicate the concentration of a pollutant that is increasing linearly in time T ($\alpha > 0$) and that is affected by precipitation X , which is a binary variable (e.g., rain/no rain). Specifically, we assume that precipitation ($X = 1$) reduces the pollutant concentration by a value of $\beta > 0$ due to “washing out”. We further assume that the probability of precipitation depends on the time T . Specifically, we assume that

$$145 \quad \mathbb{P}(X = 1|T = t) = 0.5 + 0.5 \cos(2\pi t) \quad (8)$$

i.e., in this hypothetical scenario, the probability of precipitation changes seasonally with time. An illustration of this behavior is shown in Figure 1.

The response due to direct influence of T can be computed, under a linear model, via the coefficient that connects T to the Y , from which we obtain

$$150 \quad y_{direct}(t) = \alpha t + \gamma. \quad (9a)$$

The expected response conditioned only on T also includes effects mediated via X :

$$\mathbb{E}_{Y \sim p_{Y|t}}[Y] = \alpha t + \gamma - \beta(0.5 + 0.5 \cos(2\pi t)). \quad (9b)$$

If we assume that \mathcal{D} contains values of T equidistantly sampled from some interval of integer length, then meteorological normalization will yield, for N sufficiently large and $R \rightarrow \infty$,

$$155 \quad M(t) = \alpha t + \gamma - \frac{\beta}{N} \sum_{i=1}^N x_i \approx \alpha t + \gamma - \beta \mathbb{P}(X = 1) = \alpha t + \gamma - 0.5\beta. \quad (9c)$$

Hence, as anticipated by the fact that the conditions of Proposition 2 are violated, $M(t) \neq \mathbb{E}_{Y \sim p_{Y|t}}[Y] \neq y_{direct}(t)$.

The three solutions of (9) are displayed, as a function of t , in Figure 1. In the numerical experiments, we selected $T \in [0, 1]$, $\alpha = 5$, $\beta = 2$, and $\gamma = 3$. Indeed, $y_{direct}(t)$ only considers the direct path from T to Y , while $\mathbb{E}_{Y \sim p_{Y|t}}[Y]$ also takes into account the expected effect of precipitation on pollution *at the considered time* t . Specifically, we see that in seasons
160 with little precipitation, the expected pollutant concentration is elevated w.r.t. seasons with heavy precipitation. In contrast, meteorological normalization $M(t)$ – based on a linear regression model to ensure successful learning in this simple scenario – ignores the dependence between T and X and responds with an estimate of pollutant concentration that assumes a precipitation level “averaged” over the entire observation period (e.g., over a year). Thus, while pollutant concentration can be expected to be lower during the wet season (simply because rain is more likely according to (8)), meteorological normalization predicts
165 pollutant concentration for a given Julian date t via the all-year average precipitation probability.

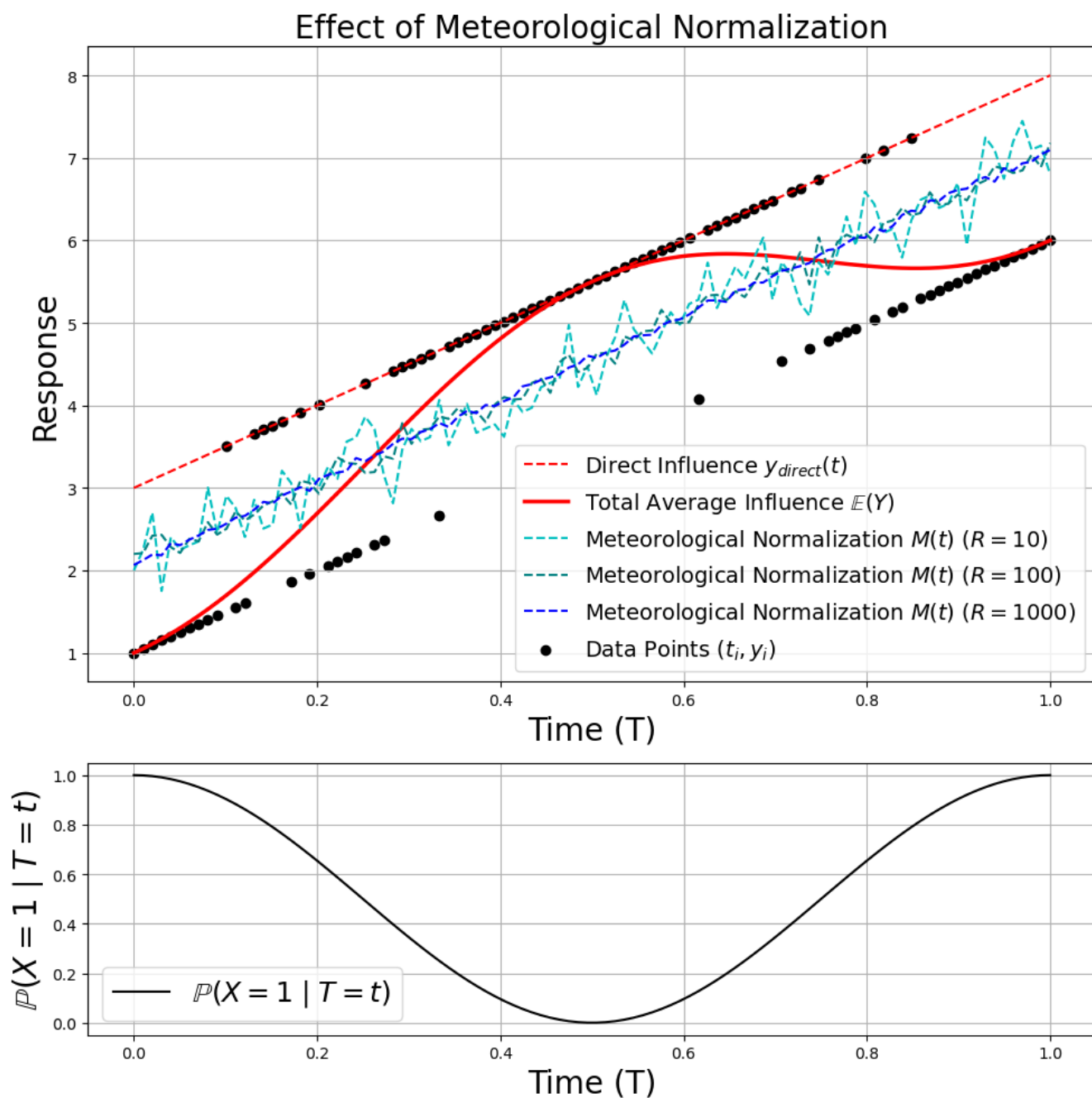


Figure 1. Illustration of Experiment 1. The top figure shows the dataset \mathcal{D} , $y_{direct}(t)$, the result $M(t)$ from meteorological normalization using a linear model, and the expectation $E_{Y \sim p_{Y|t}}[Y]$ based only on t , indicating that they are all different (see text for an explanation). The bottom figure illustrates the statistical dependency between X and T according to (8).

3.2 Experiment 2: Meteorological Normalization in Forecasting Settings

We next investigate a setting where meteorological and non-meteorological variables are independent, i.e., where (6b) holds, and where we can even assume that learning is successful in the sense of (2) on the domain described by the training dataset \mathcal{D} . However, we assume here that meteorological normalization is used in a forecasting setting, i.e., we use it to anticipate some
170 pollutant concentration for (future) values of T .

To this end, we re-use the model (7) from the previous experiment, but assume that precipitation has a constant probability $\mathbb{P}(X = 1|T = t) = 0.5$ for all Julian dates. We still obtain

$$y_{direct}(t) = \alpha t + \gamma \tag{10a}$$

from the original model (7). Assuming that learning is successful and that thus both conditions of Proposition 2 are satisfied,
175 we get, at least for t covered by the training dataset \mathcal{D} ,

$$M(t) \approx \mathbb{E}_{Y \sim p_{Y|t}}[Y] \approx \alpha t + \gamma - 0.5\beta \tag{10b}$$

for N and R sufficiently large. Meteorological normalization thus agrees with the expectation of the pollutant concentration for a given Julian date.

We performed experiments with the same settings as in the previous section, with the main difference that training data was
180 selected for $T \in [0, 1]$, but meteorological normalization was performed for the larger set of Julian dates $[0, 2]$. We furthermore trained a random forest regressor for the learned model \hat{f} , instead of a linear regressor as in Section 3.1. As it can be seen in Figure 2, meteorological normalization largely agrees with the expectation $\mathbb{E}_{Y \sim p_{Y|t}}[Y]$ on $[0, 1]$, but fails to forecast the linear trend for $[1, 2]$. The reason is that random forests do not extrapolate well beyond the training data but are predict constant values for all datapoints outside of the range seen during training. Since random forests are common model classes for meteorological
185 normalization, cf. (Grange et al., 2018), this technique can be used in forecasting scenarios only with considerable caution (or by including prior knowledge in the selection of the model class \mathcal{F} , see Section 4.2).

3.3 Experiment 3: Meteorological Normalization in Extrapolation Settings

In this last experiment, we again assume that learning is successful on the support of the training data, but show another setting in which meteorological normalization can fail if X and T are not independent, i.e., if (6b) is violated. The experiment
190 in this section is a mixture of those in Sections 3.1 and 3.2: The dependence between X and T will cause meteorological normalization to sample tuples (r_i, t_i) that are, as in forecasting, *not* covered by the training dataset \mathcal{D} , and that hence require good extrapolation properties of \hat{f} .⁵

For this experiment we assumed that (X, T, Y) are jointly Gaussian, with X and T being zero mean, unit variance, and (strongly) correlated with correlation coefficient ρ . Furthermore,

$$195 \quad Y = \alpha X + \beta T + \eta \tag{11}$$

⁵In Section 3.1 this was not a problem, as the linear model \hat{f} extrapolates well to unseen inputs due to its simplicity.

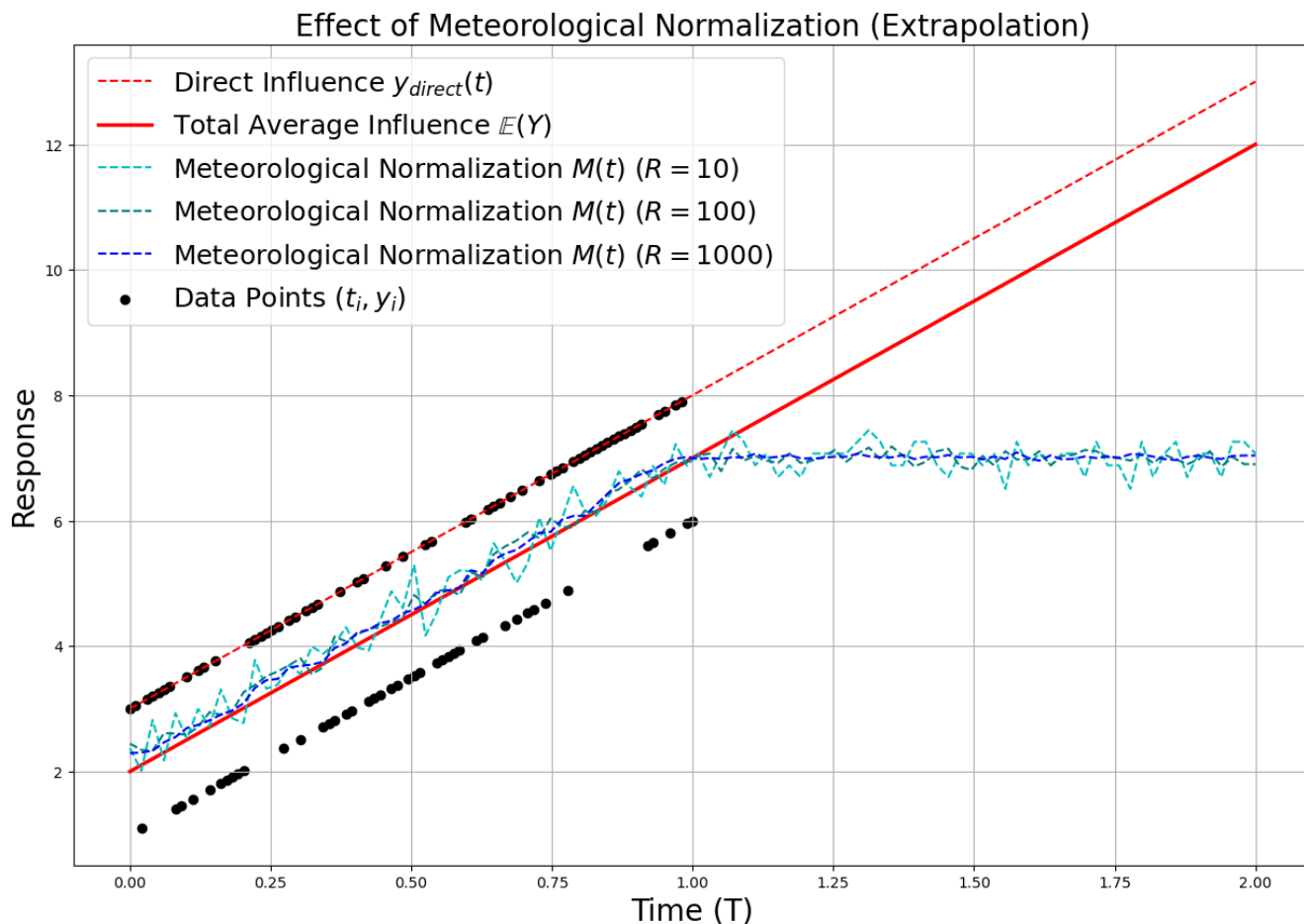


Figure 2. Illustration of Experiment 2. The figure shows the dataset \mathcal{D} , $y_{direct}(t)$, the result $M(t)$ from meteorological normalization using a linear model, and the expectation $\mathbb{E}_{Y \sim p_{Y|t}}[Y]$ based only on t . While the latter two are similar for $T \in [0, 1]$, the result from meteorological normalization cannot reproduce the linear trend for $T \in [1, 2]$.

where η is a Gaussian noise with zero mean and a small variance of σ^2 . It can be shown that the expectation $\mathbb{E}_{Y \sim p_{Y|t}}[Y]$ can be computed as

$$\mathbb{E}_{Y \sim p_{Y|t}}[Y] = (\alpha\rho + \beta)t. \quad (12a)$$

The response due to direct influence is a RV due to η , and averaging over this RV yields

$$\mathbb{E}_\eta[Y_{direct}(t)] = \beta t. \quad (12b)$$

Setting $\alpha = 1.5$, $\beta = 1.2$, $\sigma^2 = 0.01$, and a correlation coefficient of 0.8 for the correlation between X and T yields the situation depicted in Figure 3. A sample of $N = 5000$ datapoints exhibits an ellipsoid shape (confirming the strong correlation between X and T). Furthermore, the expected relationship $\mathbb{E}_{Y \sim p_{Y|x,t}}[Y]$ and the decision tree model \hat{f} trained with default hyperparameters are also depicted, showing a good agreement between – i.e., successful learning in the sense of (2) – within the support of the training data. It can also be seen, however, that the trained model \hat{f} does not agree well with $\mathbb{E}_{Y \sim p_{Y|x,t}}[Y]$ outside of the training data domain, e.g., for small values of X and large values of T (or vice-versa).

Performing meteorological normalization with the trained decision tree \hat{f} shows that $M(t)$ and $\mathbb{E}_{Y \sim p_{Y|t}}[Y]$ disagree substantially, see Figure 4. This disagreement is caused by the fact that meteorological normalization, via resampling, breaks any statistical dependence between X and T , which under model (11) reinforces the response (see also (12a)). Indeed, both X and T have monotonic effects on Y thanks to their positive coefficients α and β , and their strong positive correlation ensures that both variables assume values in similar ranges. There is, however, also a noticeable disagreement between meteorological normalization $M(t)$ and the expected direct effect $\mathbb{E}_\eta[Y_{direct}(t)]$, especially for extreme values of T . The reason for this is that meteorological normalization resamples values of X according to (4), which corresponds to evaluating \hat{f} on a circular, instead of an elliptic domain (see Figure 5). Hence, even though the model \hat{f} was successfully learned, it does not extrapolate well to combinations of (X, T) values unseen during training. In this particular case, small values of T would require sampling only small values of X ; but according to Proposition 1, meteorological normalization will sample X from a zero-mean Gaussian distribution that is independent of T . As a consequence, and with reference to Figure 3, for small values of T , a large sample for X will lead to a situation where \hat{f} overestimates the response, explaining the discrepancy in Figure 4.

We believe that the situation depicted in this experiment is a common situation for meteorological normalization, since i) badly extrapolating tree-based models are quite common hypothesis classes \mathcal{F} and ii) meteorological variables are indeed often correlated with Julian date. A model that has not been trained on snow, or low temperatures (X), for summer days (T) will behave unpredictably during meteorological normalization, where such combinations of X and T will be sampled with high probability (if R is large). We believe that, hence, this example is particularly relevant for understanding the limitations of meteorological normalization.

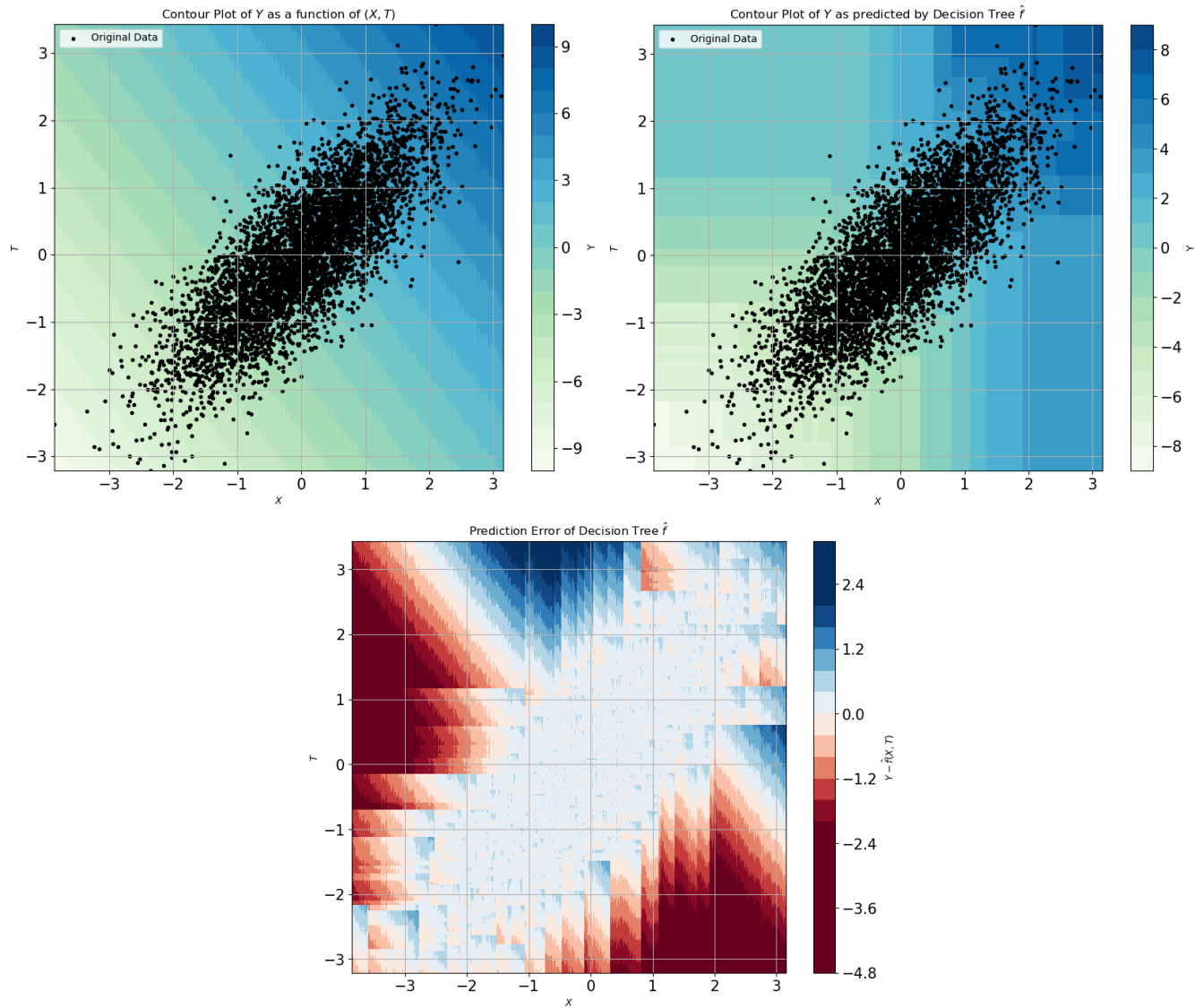


Figure 3. Illustration of Experiment 3. The left figure shows the dataset and a contour plot of the model (11), while the right figure shows a contour plot of the learned function \hat{f} . The bottom figure shows the model error. As it can be seen, the trained decision tree agrees well with the model in regions where training data was available.

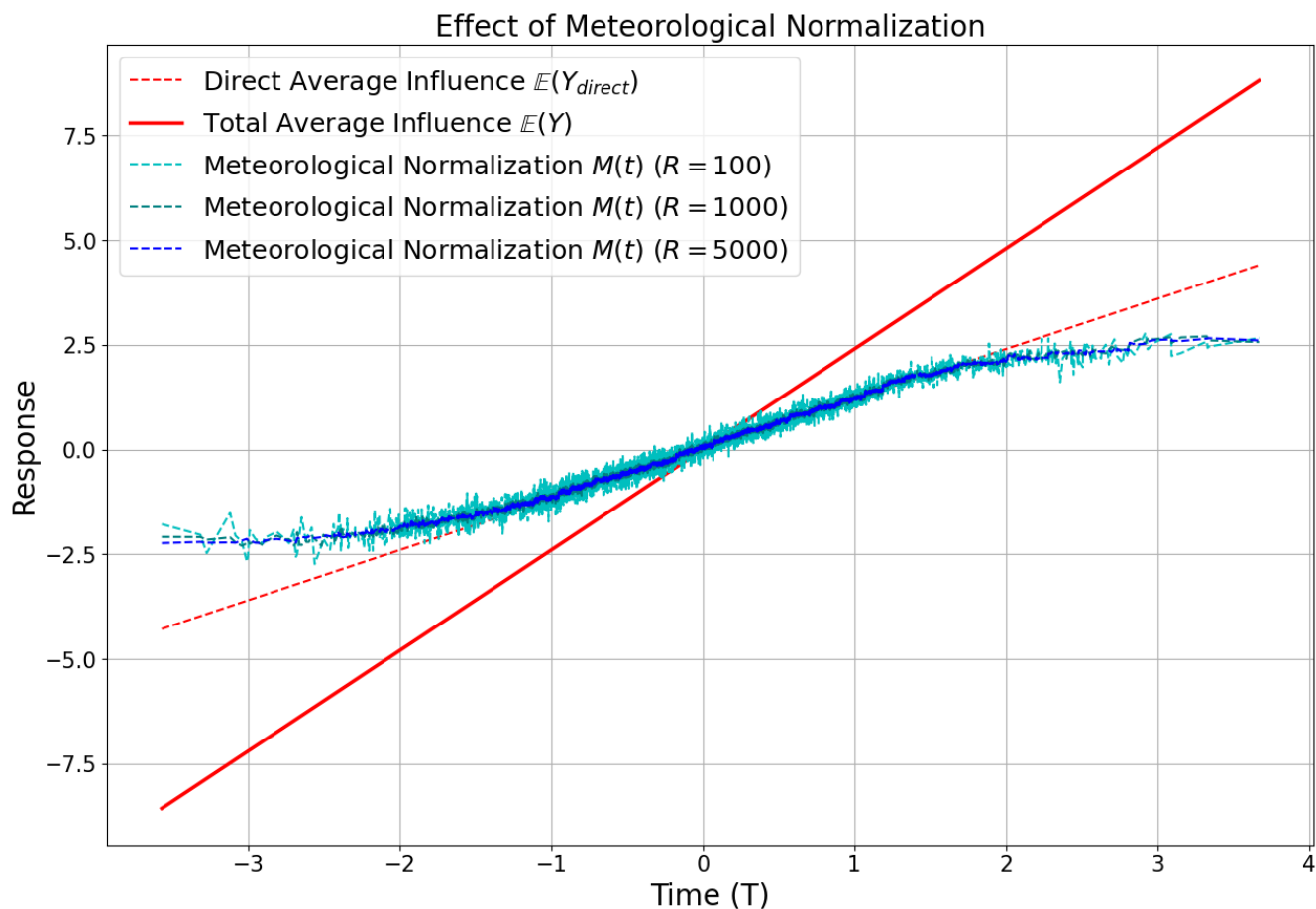


Figure 4. Illustration of Experiment 3. The figure shows the expected direct influence $\mathbb{E}_{\eta}[Y_{direct}(t)]$, the result $M(t)$ from meteorological normalization using a decision tree model, and the expectation $\mathbb{E}_{Y \sim p_{Y|t}}[Y]$ based only on t . It can be seen that neither of these results agree due to the fact that meteorological normalization assumes independence between X and T and required evaluating \hat{f} outside of the training data domain.

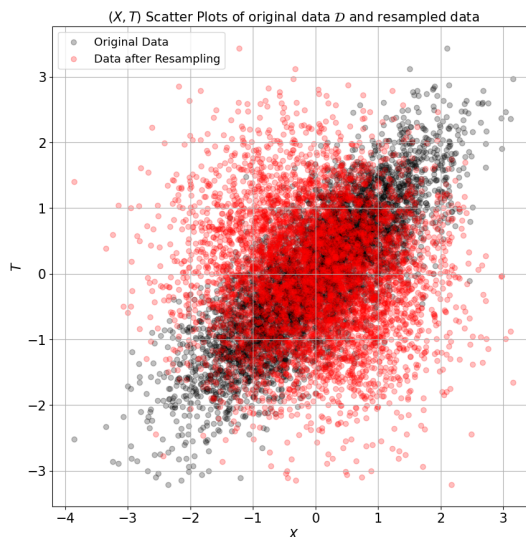


Figure 5. When performing meteorological normalization, the resampling operation eliminates the statistical dependence between meteorological and other variables and thus requires evaluating the learned model \hat{f} outside of the training data domain.

225 4 Discussion

4.1 The Pitfalls of Meteorological Normalization

In this work we have investigated meteorological normalization from a mathematical perspective and showed that, given certain conditions, the result of meteorological normalization coincides with the expected value of the response variable conditioned on the non-meteorological features. These conditions include i) successful learning in the sense that the model used for meteorological normalization is itself (close to) an expected value, and ii) that meteorological and non-meteorological variables are marginally independent. While the first condition is unproblematic in settings for which sufficient data is available, the second condition appears to be critical in many situations of practical interest. Indeed, in many such settings there is substantial statistical dependence between these variables, especially if the non-meteorological variables contain a notion of absolute time (such as Julian day): Meteorological factors such as temperature, humidity, precipitation levels, etc. change seasonally and are, thus, strongly dependent on time.

Such a dependence leads to the situation that the training dataset used for developing the machine learning model does not cover the product of the supports of the respective marginal distributions, but only a subset of it. Meteorological normalization, however, evaluates this model on the product of supports of these marginal distributions – and hence for inputs for which the machine learning model was not trained. This extrapolation setting, which we investigated in Section 3.3, was already discussed as being critical in (Nicely et al., 2020; Zabrocki et al., 2022). A similar extrapolation setting appears when meteorological normalization is used for forecasting (cf. Section 3.2). We thus agree with Nicely et al. (2020); Zabrocki et al. (2022) and call for caution when using meteorological normalization (if it must be used at all).

4.2 Addressing Pitfalls and Alternatives to Meteorological Normalization

Since the most critical problem appears to be the fact that meteorological normalization requires good extrapolation properties for the predominant situation when meteorological and non-meteorological variables are statistically dependent, one option would be to ensure respective extrapolation properties by selecting the hypothesis class \mathcal{F} accordingly. Indeed, in our experiments in Sections 3.2 and 3.3, we chose tree-based models; had we chosen linear models, extrapolation would have been less critical, and the agreement between the expectation and the result from meteorological normalization would have been better. (Situations such as the one illustrated in Section 3.1 would still appear, however.) Choosing an appropriate inductive bias requires a deep understanding of the relationship one intends to learn, which is often not possible in a purely data-driven setting (but which may be achievable in theory-inspired settings when prior knowledge is available, cf. (Hoffer et al., 2022)).

Another option is to adapt the algorithmic approach to meteorological normalization in (4) such that the independence assumption (6b) is replaced by a less restrictive condition. Indeed, if the dataset \mathcal{D} can be clustered into (or covered by) subsets $\mathcal{D}^{(\ell)}$, where for each ℓ the contained variables x_i and t_i appear to be drawn from a factorized distribution, the following adaptation to meteorological normalization could be successful:

$$M'(t) = \frac{1}{R} \sum_{i=1}^R \hat{f}(r'_i, t) \quad (13)$$

where the set $\{r'_i\}$ is obtained by sampling x_i values from only the cluster $\mathcal{D}^{(\ell)}$ that corresponds to the selected value of t , and not from \mathcal{D} as a whole. This would correspond to the situation in which for a day in July, meteorological variables are randomly sampled from other days in July only. Future work shall investigate whether such an approach (e.g., based on k-means clustering of \mathcal{D} or Gaussian mixture models for (X, T)) yields more stable results, by both theoretical analyses and numerical, real-world experiments.

Finally, there is a rich branch of the literature concerned with effect estimation, and which is strongly rooted in the field of causality. Especially if some of the non-meteorological variables can be intervened upon (e.g., via policies), direct and total effects can be estimated (see, e.g., (Pearl, 2001)). Indeed, intervening on the non-meteorological variable by setting it to $T = t$ (e.g., via a policy) but keeping the meteorological variables unaffected leads to the definition of the *average natural direct effect* (Pearl, 2001, Def. 5). Note that the direct effect is related to the response due to direct influence of T , which we denoted as $y_{direct}(t)$ in our work; the quantity $\mathbb{E}_{Y \sim p_{Y|t}}[Y]$, and hence the results $M(t)$ of meteorological normalization, is thus related to the total effect. Future work shall investigate these connections in deeper details and explore approaches how meteorological normalization could benefit from this perspective of effect estimation rooted in causal theory.

5 Conclusion and the future of meteorological normalization

In this work we present theoretical analyses and experimental evidence on properties and limitations of meteorological normalization. Recognizing that meteorological normalization is heavily influenced by the dependency of meteorological descriptors on seasonality, our discussion emphasizes the need for future work to explore how to cope with these limitations. Following the

theoretical analysis presented in this work, one avenue to solving this is de-seasoning meteorological data to guarantee their independence. While in linear models this de-seasoning is related to effect estimation, such an endeavour becomes rather complicated with non-linear models where effect estimation is non-trivial. Excluding the temporal component of meteorological variables at a higher level in nonlinear regression models may help mitigating issues and improve the overall validity of using meteorological normalization. An alternative approach to making the method more reliable is considering the distributions of the data and hence avoiding impossible combinations of meteorological and non-meteorological variables (such as simultaneous snow, rain, and sun in summer) during resampling. Future work shall be devoted to determine the success and applicability of these approaches for the prediction of air pollutant concentrations.

Acknowledgments

This was supported by the affiliated institutions, the Horizon Europe EDIAQI (Grant agreement ID: 101057497) and Horizon Europe URBREATH (Grant agreement ID: 101139711). Know-Center is a COMET center within COMET – Competence Centers for Excellent Technologies. This program is funded by the Austrian Federal Ministries for Climate Policy, Environment, Energy, Mobility, Innovation and Technology (BMK) and for Labor and Economy (BMAW), represented by Österreichische Forschungsförderungsgesellschaft mbH (FFG), Steirische Wirtschaftsförderungsgesellschaft mbH (SFG) and the Province of Styria, Vienna Business Agency and Standortagentur Tirol.

Appendix A: Proof of Proposition 1

Proof. Since \mathcal{R} is obtained by uniformly resampling from \mathcal{D}_x with replacement, we can consider \mathcal{R} as R independent and identically distributed (i.i.d.) draws from a uniform distribution $u_{\mathcal{D}_x}$ over \mathcal{D}_x , i.e., $\mathbb{P}(R = x_i) = 1/N$ for every $i \in [N] = \{1, \dots, N\}$. As $R \rightarrow \infty$, the average in (4) converges to the expectation

$$M(t) \xrightarrow{R \rightarrow \infty} \mathbb{E}_{R \sim u_{\mathcal{D}_x}} [\hat{f}(R, t)] = \frac{1}{N} \sum_{i=1}^N \hat{f}(x_i, t). \quad (\text{A1})$$

The second part follows immediately from the law of large numbers. □

Appendix B: Proof of Proposition 2

Proof. We insert (6a) in (5b) to obtain

$$M(t) \xrightarrow{R, N \rightarrow \infty} \mathbb{E}_{X \sim p_X} [\mathbb{E}_{Y \sim p_{Y|x,t}} [Y]] \quad (\text{B1})$$

and noticing that with (6b) we have that $p_X = p_{X|t}$, we get

$$M(t) \xrightarrow{R, N \rightarrow \infty} \mathbb{E}_{X \sim p_{X|t}} [\mathbb{E}_{Y \sim p_{Y|x,t}} [Y]] = \mathbb{E}_{X, Y \sim p_{Y, X|t}} [Y] = \mathbb{E}_{Y \sim p_{Y|t}} [Y]. \quad (\text{B2})$$

300 □

Appendix C: Derivation and Experimental Details for Section 3.1

While the derivation of $y_{direct}(t)$ is straightforward, we expand the derivation of $\mathbb{E}_{Y \sim p_{Y|t}}[Y]$ in more detail here. Specifically,

$$\begin{aligned}\mathbb{E}_{Y \sim p_{Y|t}}[Y] &= \mathbb{E}_{X \sim p_{X|t}}[\alpha t - \beta X + \gamma] \\ &= \alpha t - \mathbb{E}_{X \sim p_{X|t}}[\beta X] + \gamma \\ &= \alpha t + \gamma - \beta \mathbb{P}(X = 1|T = t) \\ &= \alpha t + \gamma - \beta(0.5 + 0.5 \cos(2\pi t))\end{aligned}$$

where the first equality follows from the model (7), the second from the linearity of expectation and the fact that αt and γ are constants, the third from the fact that the expectation of a $\{0, 1\}$ -valued RV is equal to the probability that this RV assumes 1, and the last from (8).

We performed numerical experiments with $T \in [0, 1]$, $\alpha = 5$, $\beta = 2$, $\gamma = 3$. We sampled $N = 100$ values $\{t_i\}$ equidistantly from $[0, 1]$ (using the `linspace` function from `numpy`), and for each t_i , one x_i was sampled from the Bernoulli distribution in (8). The values y_i were calculated using (7). The resulting samples (x_i, t_i, y_i) comprise the dataset \mathcal{D} . With this dataset, we trained a simple linear regression model \hat{f} to ensure successful learning. This model was used for meteorological normalization with $R = \{10, 100, 1000\}$. Specifically, in all experiments we selected R random permutations of the meteorological variables x_i of the dataset to perform meteorological normalization simultaneously for all instances of t_i in the dataset \mathcal{D} . For the resulting tuples (t_i, r_i) , \hat{f} was re-evaluated and results were averaged according to (4). The curves in Figure 1 for $y_{direct}(t)$ and $\mathbb{E}_{Y \sim p_{Y|t}}[Y]$ were obtained from evaluating (9) directly.

Appendix D: Experimental Details for Section 3.2

We performed numerical experiments with the same settings as in Section 3.1, with the difference that, for every t_i , x_i was sampled from a Bernoulli distribution with constant success probability $\rho = 0.5$. We further did not train a linear regression model, but chose \hat{f} to be a random forest regressor. The model was trained using the `scikit-learn` library and consisted of 100 trees with a depth of at most 5 each. We again performed meteorological normalization with $R = \{10, 100, 1000\}$, where, to simulate forecasting, the time variable T was equidistantly sampled from $[0, 2]$ in this case, and the $N = 100$ instances of x_i were permuted over these samples of t_i . The curves in Figure 2 for $y_{direct}(t)$ and $\mathbb{E}_{Y \sim p_{Y|t}}[Y]$ were obtained from evaluating (10) directly.

Appendix E: Derivation and Experimental Details for Section 3.3

To derive the expected response conditioned only on non-meteorological variables, we made use of the fact that (X, T) follow a jointly Gaussian distribution with zero mean vector and covariance matrix

$$\Sigma = \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix}. \quad (\text{E1})$$

330 Then, the distribution of X given $T = t$ can again shown to be Gaussian with mean ρt and variance $1 - \rho^2$. As a consequence, X can be written as

$$X = \rho t + \epsilon \quad (\text{E2})$$

where ϵ is Gaussian with zero mean and variance $1 - \rho^2$. Inserting this in the model (11) yields, for a given $T = t$,

$$Y = \alpha(\rho t + \epsilon) + \beta t + \eta \quad (\text{E3})$$

335 and taking the expectation results in (12a) (ϵ and η are zero mean and independent).

Using Equation (E3), we created the dataset and then conducted numerical experiments using it to predict Y from X and T with a decision tree. The decision tree model was trained without hyperparameter optimization or selection. The results of predicting Y from the equation X are shown in Figure 3, where Y is presented as a function of (X, T) . Additionally, black points indicate a strong correlation between X and T . This graph also illustrates the predictions of the trained model. Figure 3

340 also displays the model's error. The curves in Figure 4 were obtained when meteorological normalization was applied with $R = \{100, 1000, 5000\}$. $y_{direct}(t)$ and $\mathbb{E}_{Y \sim p_{Y|t}[Y]}$ were obtained using (12a) direct observations.

References

- Bari, M. A., Baumbach, G., Kuch, B., and Scheffknecht, G.: Air Pollution in Residential Areas from Wood-fired Heating, *Aerosol and Air Quality Research*, 11, 749–757, <https://doi.org/10.4209/aaqr.2010.09.0079>, 2011.
- 345 Bing Chen, Jinsheng Chen, J. Z. and Zhang, F.: Particulate Air Pollution from Combustion and Construction in Coastal and Urban Areas of China, *Journal of the Air & Waste Management Association*, 61, 1160–1165, <https://doi.org/10.1080/10473289.2011.603995>, 2011.
- Breiman, L.: Random Forests, *Machine Learning*, 45, 5–32, <https://doi.org/10.1023/A:1010933404324>, 2001.
- Colville, R., Hutchinson, E., Mindell, J., and Warren, R.: The transport sector as a source of air pollution, *Atmospheric Environment*, 35, 1537–1565, [https://doi.org/https://doi.org/10.1016/S1352-2310\(00\)00551-3](https://doi.org/https://doi.org/10.1016/S1352-2310(00)00551-3), 2001.
- 350 Gagliardi, R. and Andenna, C.: Machine learning meteorological normalization models for trend analysis of air quality time series, *International Journal of Environmental Impacts*, 4, 375–387, <https://doi.org/10.2495/EI-V4-N4-375-387>, 2021.
- Grange, S. K. and Carslaw, D. C.: Using meteorological normalisation to detect interventions in air quality time series, *Science of The Total Environment*, 653, 578–588, <https://doi.org/10.1016/j.scitotenv.2018.10.344>, 2019.
- Grange, S. K., Carslaw, D. C., Lewis, A. C., Boleti, E., and Hueglin, C.: Random forest meteorological normalisation models for Swiss PM₁₀ trend analysis, *Atmospheric Chemistry and Physics*, 18, 6223–6239, <https://doi.org/10.5194/acp-18-6223-2018>, 2018.
- 355 Hassan, H., Latif, M. T., Juneng, L., Amil, N., Khan, M. F., Yik, D. J., and Abdullah, N. A.: Interaction of PM₁₀ concentrations with local and synoptic meteorological conditions at different temporal scales, *Atmospheric Research*, 241, 104975, <https://doi.org/https://doi.org/10.1016/j.atmosres.2020.104975>, 2020.
- Hoffer, J. G., Ofner, A. B., Rohrhofer, F. M., Lovrić, M., Kern, R., Lindstaedt, S., and Geiger, B. C.: Theory-Inspired Machine Learning – Towards a Synergy between Knowledge and Data, *Welding in the World*, <https://doi.org/10.1007/s40194-022-01270-z>, open access, 2022.
- 360 Lovrić, M., Pavlović, K., Vuković, M., Grange, S. K., Haberl, M., and Kern, R.: Understanding the true effects of the COVID-19 lockdown on air pollution by means of machine learning, *Environmental Pollution*, 274, 115900, <https://doi.org/10.1016/j.envpol.2020.115900>, 2021.
- Lovrić, M., Antunović, M., Šunić, I., Vuković, M., Kecorius, S., Kröll, M., Bešlić, I., Godec, R., Pehnec, G., Geiger, B. C., Grange, S. K., and Šimić, I.: Machine Learning and Meteorological Normalization for Assessment of Particulate Matter Changes during the COVID-19 Lockdown in Zagreb, Croatia, *International Journal of Environmental Research and Public Health*, 19, 6937, <https://doi.org/10.3390/ijerph19116937>, 2022.
- 365 Munir, S., Habeebullah, T. M., Mohammed, A. M., Morsy, E. A., Rehan, M., and Ali, K.: Analysing PM_{2.5} and its Association with PM₁₀ and Meteorology in the Arid Climate of Makkah, Saudi Arabia, *Aerosol and Air Quality Research*, 17, 453–464, <https://doi.org/10.4209/aaqr.2016.03.0117>, 2017.
- 370 Nicely, J. M., Duncan, B. N., Hanisco, T. F., Wolfe, G. M., Salawitch, R. J., Deushi, M., Haslerud, A. S., Jöckel, P., Josse, B., Kinnison, D. E., Klekociuk, A., Manyin, M. E., Marécal, V., Morgenstern, O., Murray, L. T., Myhre, G., Oman, L. D., Pitari, G., Pozzer, A., Quaglia, I., Revell, L. E., Rozanov, E., Stenke, A., Stone, K., Strahan, S., Tilmes, S., Tost, H., Westervelt, D. M., and Zeng, G.: A machine learning examination of hydroxyl radical differences among model simulations for CCMI-1, *Atmospheric Chemistry and Physics*, 20, 1341–1361, <https://doi.org/10.5194/acp-20-1341-2020>, 2020.
- 375 Pearl, J.: Direct and indirect effects, in: *Proceedings of the Seventeenth Conference on Uncertainty in Artificial Intelligence*, UAI’01, p. 411–420, Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, ISBN 1558608001, 2001.
- Qiu, M., Zigler, C., and Selin, N. E.: Statistical and machine learning methods for evaluating trends in air quality under changing meteorological conditions, *Atmospheric Chemistry and Physics*, 22, 10551–10566, <https://doi.org/10.5194/acp-22-10551-2022>, 2022.

Speranza, A. and Caggiano, R.: Meteorological variables and PM10 exceedances effect on aerosol particles in a low emission zone using com-
positional data analysis, *Journal of Geochemical Exploration*, 255, 107 322, <https://doi.org/https://doi.org/10.1016/j.gexplo.2023.107322>,
2023.

Wong, Y. J., Yeganeh, A., Chia, M. Y., Shiu, H. Y., Ooi, M. C. G., Chang, J. H. W., Shimizu, Y., Ryosuke, H.,
Try, S., and Elbeltagi, A.: Quantification of COVID-19 impacts on NO2 and O3: Systematic model selection and
hyperparameter optimization on AI-based meteorological-normalization methods, *Atmospheric Environment*, 301, 119 677,
<https://doi.org/https://doi.org/10.1016/j.atmosenv.2023.119677>, 2023.

Zabrocki, L., Alari, A., and Benmarhnia, T.: Improving the design stage of air pollution studies based on wind patterns, *Scientific Reports*,
12, <https://doi.org/10.1038/s41598-022-11939-6>, 2022.

Šimić, I., Lovrić, M., Godec, R., Kröll, M., and Bešlić, I.: Applying machine learning methods to better understand, model
and estimate mass concentrations of traffic-related pollutants at a typical street canyon, *Environmental Pollution*, 263, 114 587,
<https://doi.org/https://doi.org/10.1016/j.envpol.2020.114587>, 2020.