Weather2vec: Representation Learning for Causal Inference with Non-local Confounding in Air Pollution and Climate Studies

Mauricio Tec[1*], **James G. Scott**[2,3], **Corwin M. Zigler**[2]

[1]Department of Biostatistics, Harvard University
[2]Department of Statistics and Data Sciences, The University of Texas at Austin
[3]Department of Information, Risk, and Operations Management, The University of Texas at Austin
mauriciogtec@hsph.harvard.edu, james.scott@mccombs.utexas.edu, cory.zigler@austin.utexas.edu

Abstract

Estimating the causal effects of a spatially-varying intervention on a spatially-varying outcome may be subject to non-local confounding (NLC), a phenomenon that can bias estimates when the treatments and outcomes of a given unit are dictated in part by the covariates of other nearby units. In particular, NLC is a challenge for evaluating the effects of environmental policies and climate events on health-related outcomes such as air pollution exposure. This paper first formalizes NLC using the potential outcomes framework, providing a comparison with the related phenomenon of causal interference. Then, it proposes a broadly applicable framework, termed *weather2vec*, that uses the theory of balancing scores to learn representations of non-local information into a scalar or vector defined for each observational unit, which is subsequently used to adjust for confounding in conjunction with causal inference methods. The framework is evaluated in a simulation study and two case studies on air pollution where the weather is an (inherently regional) known confounder.

Introduction

Causal effects of spatially-varying exposures on spatially-varying outcomes may be subject to *non-local confounding* (NLC), which occurs when the treatments and outcomes for a given unit are affected by *covariates* of other nearby units (Cohen-Cole and Fletcher 2008; Florax and Folmer 1992; Chaix, Leal, and Evans 2010; Elhorst 2010). In simple cases, NLC can be resolved using simple summaries of non-local data, such as the averages of the covariates over pre-specified neighborhoods. But in many realistic settings, NLC is caused by the complex interaction of spatial factors, and thus it cannot be resolved using simple *ad hoc* summaries of neighboring covariates. For such scenarios, we propose *weather2vec*, a framework that uses a U-net (Ronneberger, Fischer, and Brox 2015) to learn representations that encode NLC information and can be used in conjunction with standard causal inference tools. The method is broadly applicable to settings where the covariates are available over a grid of spatial units, and where the outcome and treatment are observed in some subset of the grid.

The name *weather2vec* stems from its motivation to address limitations in current methods for estimating causal effects in environmental studies where meteorological processes are known confounders, aiming to contribute to the development of new flexible machine learning tools to assess the *causal effect* of policies and climate-related events on health-relevant outcomes: a task which has been recently identified by Rolnick et al. (2022) as a pressing outstanding challenge for tackling the effects of climate change.

Two applications will be discussed in detail. The first application follows an earlier analysis by Papadogeorgou, Choirat, and Zigler (2019), who estimated the air quality impact of power plant emissions controls. This case study evaluates the method's ability to reduce NLC under sparsely observed treatments (in combination with with propensity matching methods (Rubin 2005)). The second application is in *meteorological detrending* (Wells et al. 2021), and uses *weather2vec* to deconvolve climate variability from policy changes when characterizing long-term air quality trends. These two examples are accompanied by a simulation study comparing alternative adjustments to account for NLC.

In summary, this article has three aims:

1. Provide a rigorous characterization of NLC using the potential outcomes framework, clarifying some connections with causal interference.
2. Expand the library of NN methods in causal inference by proposing a U-net as a viable model to account for NLC in conjunction with standard causal inference tools.
3. Establish a promising research direction for addressing NLC in scientific studies of air pollution exposure – in which NLC is a common problem (driven by meteorology) for which widely applicable tools are lacking.

We investigate two mechanisms to obtain the representations: one supervised, and one self-supervised. The supervised one formally links the representation of NLC to the balancing property of propensity (and prognostic) scores in the causal inference literature (Rubin 2008; Hansen 2008). This approach requires that the outcome and treatment are densely available throughout the covariates' grid. By contrast, the self-supervised approach first learns representations encoding neighboring covariate information into a low-dimensional vector, which can subsequently be included as confounders in downstream causal analyses when the outcomes and treatments are sparsely observed on the grid.

Mathematical proofs of all the propositions in the paper, details of the experiments, and additional explanations

*Corresponding author.

can be found in the web technical appendix of the paper (Tec, Scott, and Zigler 2022). The code is available at https://github.com/mauriciogtec/weather2vec-reproduce.

Related work Previous research has investigated NNs for the (non-spatial) estimation of balancing scores (Keller, Kim, and Steiner 2015; Westreich, Lessler, and Funk 2010; Setoguchi et al. 2008) and counterfactual estimation (Shalit, Johansson, and Sontag 2017; Johansson, Shalit, and Sontag 2016; Shi, Blei, and Veitch 2019). But none of these works specifically consider NLC.

Relevant applications of U-nets in environmental studies include forecasting (Larraondo et al. 2019; Sadeghi et al. 2020), estimating spatial data distributions from satellite images (Hanna et al. 2021; Fan et al. 2021), indicating that U-nets are powerful tools to manipulate rasterized weather data. Also relevant, Lu and Chang (2005) give a specific application of NNs for meteorological detrending, although without considering adjusting for neighboring covariates. Shen, Mickley, and Murray (2017) do consider regional dependencies by applying patch-wise PCA to extract meteorological features to improve the prediction of air pollution. Approaches to learning summaries of neighboring covariates for regression-based causal inference have been investigated in the econometrics literature. For example, WX-regression models (Elhorst 2010) formulate the outcome as a linear function of the treatment and the covariates of some pre-specified neighborhood. CRAE (Blier-Wong et al. 2020) uses an autoencoder over pre-extracted patches of regional census data that is fed into an econometric regression. In contrast to predictive regression-based approaches, *weather2vec* aims at learning balancing scores, which have known benefits that include the ability to empirically assess the threat of residual confounding and offer protection against model misspecification that arises when modeling outcomes directly (Rubin 2008).

There is also a maturing literature on adjusting for unobserved spatially-varying confounding (Reich et al. 2021). Spatial random effect methods are popular in practice, although Khan and Calder (2020) have highlighted their sensitivity to misspecification for the purposes of confounding adjustment. The distance-adjusted propensity score matching (DAPSm) (Papadogeorgou, Choirat, and Zigler 2019) matches units based jointly on estimated propensity scores and spatial proximity under the rationale that spatial proximity can serve as a proxy for similarity in spatially-varying covariates. In the same spirit, Veitch, Wang, and Blei (2019) use graph embeddings to account for proximity within a network as a proxy for confounding. In general, the primary target of spatial confounding methods are settings where confounding is local conditional on the unobserved spatially-varying confounders–in contrast to NLC.

Finally, NLC is distinct from *causal interference* (Tchetgen and VanderWeele 2012; Forastiere, Airoldi, and Mealli 2021; Sobel 2006; Zigler and Papadogeorgou 2021; Ogburn and VanderWeele 2014; Bhattacharya, Malinsky, and Shpitser 2020), although both phenomena arise from spatial (or network) interaction, and they both impose limitations on standard causal inference methods. While forms of NLC have often been acknowledge in the literature of interference, to the best of our knowledge, flexible statistical methods specifically addressing NLC by learning the dependencies with respect to neighboring covariates do not exist.

Potential Outcomes and NLC

We now recall the potential outcomes framework, also known as the Rubin Causal Model (RCM) (Rubin 2008), and we later adapt it to the case of NLC confounding. The RCM distinguishes between the observed outcome Y_s at unit s and those that would be observed under counterfactual (potential) treatments $Y_s(a)$ (formally defined below). We start with some notation. The assigned treatment is denoted A_s. It is assumed to be binary for ease of presentation, although the ideas generalize to more general treatments. For instance, in our Application 1, the treatment is whether or not a catalytic device is installed on a power plant to reduce pollutant emissions. $\mathbb{S}$ is the set where the outcome and treatment are measured (e.g., the location of the power plants); $\mathbb{G} \supset \mathbb{S}$ is a grid containing the rasterized covariates $\{\boldsymbol{X}_s \in \mathbb{R}^d \colon s \in \mathbb{G}\}$; for any $B \subset \mathbb{G}$, $\mathbf{X}_B = \{\boldsymbol{X}_s \mid s \in B\}$; $X \perp\!\!\!\perp Y \mid Z$ denotes conditional independence of X and Y given Z; lastly, $p(\cdot)$ denotes a generic probability or density function. We will assume throughout that $\boldsymbol{X}_s$ only contains pre-treatment covariates, meaning they are not affected by the treatment or outcome.

Definition 1 (Potential outcomes)**.** *The potential outcome $Y_s(\boldsymbol{a})$ is the outcome value that would be observed at location s under the global treatment assignment $\boldsymbol{a} = (a_1, \dots, a_{|\mathbb{S}|})$.*

For $Y_s(\boldsymbol{a})$ to depend only on a_s, the RCM needs an additional condition called the *stable unit treatment value assumption*, widely known as SUTVA, and encompassing notions of *consistency* and ruling out *interference*.

Assumption 1 (SUTVA)**.** *(1) Consistency: there is only one version of the treatment. (2) No interference: the potential outcomes for one location do not depend on treatments of other locations. Together, these conditions imply that $Y_s(\boldsymbol{a}) = Y_s(a_s)$ for any assignment vector $\boldsymbol{a} \in \{0,1\}^{|\mathbb{S}|}$, and that the observed outcome is the potential outcome for the observed treatment, i.e., $Y_s = Y_s(A_s)$.*

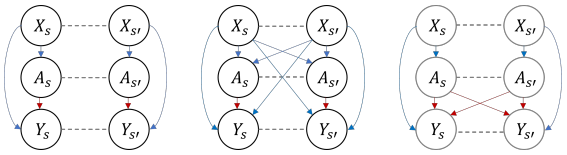
To contextualize SUTVA in our power plant example, observe that it would be violated if the pollution measured at s depends not only on whether or not the catalytic device was installed at that power plant (that is, on the assignment A_s) but also on whether or not the device was installed on other power plants ($A_{s'}$ for $s \neq s$). We assume SUTVA throughout as it is common in many causal inference studies. Then, the potential outcomes allow defining an important estimand of interest: the average treatment effect.

Definition 2 (ATE)**.** *The average treatment effect (ATE) is the quantity $\tau_{ATE} = |\mathbb{S}|^{-1} \sum_{s \in \mathbb{S}} \{Y_s(1) - Y_s(0)\}$.*

One cannot estimate the ATE directly since one never simultaneously observes $Y_s(0)$ and $Y_s(1)$. The next assumption in the RCM formalizes conditions for estimating the

ATE, (or other causal estimands) with observed data by stating that any observed association between A_s and Y_s is not due to an unobserved factor.

Assumption 2 (Treatment Ignorability)**.** *The treatment A_s is ignorable with respect to some vector of controls $\boldsymbol{L}_s$ if and only if $Y_s(1), Y_s(0) \perp\!\!\!\perp A_s \mid \boldsymbol{L}_s$.*



(a) Local Confounding. (b) NLC, no interference. (c) Interference, no NLC.

Figure 1: Confounding types.

This ignorability assumption would fail where there exist unobserved confounders. For the sake of brevity, we will say that $\boldsymbol{L}_s$ is *sufficient* to mean that the treatment is ignorable conditional on $\boldsymbol{L}_s$. We now introduce NLC, which occurs when non-local covariates are among the confounders. It is formally stated as follows:

Definition 3 (Non-local confounding)**.** *We say there is non-local confounding (NLC) when there exist neighborhoods $\{\mathcal{N}_s \subset \mathbb{G} \mid s \in \mathbb{S}\}$ such that $\boldsymbol{L}_s = \mathbf{X}_{\mathcal{N}_s}$ is sufficient and the neighborhoods are necessarily non-trivial ($\mathcal{N}_s \neq \{s\}$).*

In our power plant example, atmospheric vectors $X_{s'}$ are associated with the air pollution outcomes at other locations Y_s (Shen, Mickley, and Murray 2017), as well as the probability of installing a catalytic device on a power plant, A_s. Figs. 1a and 1b show a graphical representation of local confounding versus NLC. Horizontal dotted lines emphasize spatial correlations in the covariate, treatment, and outcome processes that do not result in confounding. For contrast, Fig. 1c shows the distinct phenomenon of (direct) interference, in which $A_{s'}$ affects A_s (Ogburn and VanderWeele 2014). (This depiction of is only one of the forms that interference can take. For instance, it may also happen through contagion (Ogburn and VanderWeele 2014).)

Subsequent discussion of the size of the NLC neighborhood, $\mathcal{N}_s$, will make use of the following proposition stating that a neighborhood containing sufficient confounders can be enlarged without sacrificing the sufficiency.

Proposition 1. *Let $\boldsymbol{L}_s$ be a sufficient set of controls including only pre-treatment covariates. and let $\boldsymbol{L}'_s$ be another set of controls satisfying $\boldsymbol{L}'_s \supset \boldsymbol{L}_s$. Then, $\boldsymbol{L}'_s$ is also sufficient.*

We can now state a classic result regarding the identifiability of causal effects from observed data under the above assumptions.

Proposition 2. *Assume SUTVA holds and that $\boldsymbol{L}_s$ is sufficient. Then*

$$\mathbb{E}\left[\mathbb{E}[Y_s \mid \boldsymbol{L}_s, A_s = 1] - \mathbb{E}[Y_s \mid \boldsymbol{L}_s, A_s = 0]\right], \qquad (1)$$

is an unbiased estimator of τ_{ATE} (where s is taken uniformly at random from $\mathbb{S}$).

Eq. (1) already offers a way to estimate causal effects from observed data (by estimating the two inner conditional expectations). However, it can be highly sensitive to the specification of the expected outcome model. There are many alternatives, one of which is to use *inverse probability of treatment weighting* (IPTW) (Cole and Hernán 2008) (described in the technical appendix for completeness).

Adjustment for NLC with Weather2vec

Accounting for NLC would be fairly straightforward provided infinite data and the right set of confounders. By virtue of Proposition 1, one could, in principle, specify a non-linear regression $Y_s \approx f(A_s, \mathbf{X}_{\mathbb{G}}, s)$ that includes every non-local covariate $\boldsymbol{X}_{s'} \in \mathbf{X}_{\mathbb{G}}$ as part of the regressors. With large model capacity, and infinite repeated samples per location, this regression would perfectly estimate $\mathbb{E}[Y_s \mid \boldsymbol{L}_s, A_s = a]$, and thus be able to estimate the ATE using Proposition 2. But this scenario is far from realistic. Most commonly, there will be only one observation for each s, requiring additional structure to enable statistical estimation. Thus, we consider the question: what kind of statistical and functional model (e.g., to predict the probability of treatment) reflects the causal structure of NLC and allows for flexible statistical models under such restrictions?

One desirable statistical property to consider is *spatial stationarity*. Intuitively, it entails that the distributions of Y_s and A_s with respect to a neighboring covariate $\boldsymbol{X}_{s'}$ should only depend on $\delta = s - s'$ (their relative position). Formally, it requires that for any set $B \subset \mathbb{G}$, displacement vector δ, and $s \in \mathbb{G}$, the following identity holds $p(A_s, Y_s \mid \mathbf{X}_B = x) = p(A_{s+\delta}, Y_{s+\delta} \mid \mathbf{X}_{B+\delta} = x)$. For *weather2vec*, we focus on the U-net (Ronneberger, Fischer, and Brox 2015), which are neural network models composed of convolution operators that are approximately spatially stationary and allow learning predictions from neighboring inputs at every point of a grid. An overview of U-nets is provided in the next section for completeness. A key property is that a U-net f_θ can transform the input covariates $\mathbf{X}_{\mathbb{G}}$ onto an output grid $\mathbf{Z}_{\theta,\mathbb{G}} := f_\theta(\mathbf{X}_{\mathbb{G}})$ of same spatial dimensions in which each scalar or vector $\boldsymbol{Z}_{\theta,s} \in \mathbf{Z}_{\theta,\mathbb{G}}$ localizes contextual spatial information from the input grid.

U-nets are not the only neural architecture with these properties. For instance, one could adapt residual networks (He et al. 2016) as a shallow alternative to a U-net. The essence of *weather2vec* is to define appropriate learning tasks to obtain the NN weights θ. Two such tasks are considered, summarized below and described in detail in subsequent sections.

1. (**Supervised**) Assuming the treatment is densely available over $\mathbb{G}$, estimate $\mathbf{Z}_{\theta,\mathbb{G}}$ as the probability of treatment conditional on non-local covariates.
2. (**Self-supervised**) If the treatment is not densely available over $\mathbb{G}$, then learn $\boldsymbol{Z}_{\theta,s}$ so that it is highly predictive of $\boldsymbol{X}_{s'}$ for any s' within a specified radius of s. Then use $\boldsymbol{Z}_{\theta,s}$ as an input in a second-stage model to learn the treatment probability.

These strategies allow learning a *propensity score*, $p(A_s = 1 \mid \boldsymbol{Z}_{\theta,s})$, which can be used within a well-

established causal inference technique such as IPTW to produce robust causal estimates of τ_{ATE}. Later we also consider a variant based on *prognostic scores* (Hansen 2008), which are predictive functions of the untreated outcomes.

An Overview of the U-net for Summarizing NLC

The U-net transformation involves two parts: a *contractive* stage and a symmetric *expansive stage*. These steps use convolutions with learnable parameters and non-linear functions to aggregate information from the input grid spatially and create rich high-level features. The convolutions in the contractive path duplicate the number of latent features at each layer. Then, these intermediate outputs go through *pooling* layers which halve the spatial dimensions. Together, these operations augment the dimensionality of each point of the grid, combining information at many spatial points with richer information contained at fewer points. Convolutions propagate information spatially, and the deeper they are in the contractive path, the larger their propagation reach (in the original scale of the input grid). The expansive path, on the other hand, uses *up-sampling* to progressively interpolate the deep higher-level features back to a finer spatial lattice, and then uses convolutions to reduce back the latent dimensionality at each grid point; with the characteristic that, in contrast to the input grid, every point now localizes spatial information. The output vector can have any arbitrary dimension after possibly applying an additional layer after the expansive path (or before the contractive path, or both). The technical appendix provides a visual example of the U-net architecture. See also the original work by Ronneberger, Fischer, and Brox (2015).

The unknown weights θ learn what non-local information is summarized by $\boldsymbol{Z}_{\theta,s}$. The depth of the U-net (number of down/up layers) dictates the maximum radius of spatial aggregation. Shallow U-nets operating on fine-grained grids may have limited spatial aggregation capabilities. Convolutions, pointwise activations, pooling, and upsampling layers are all spatially stationary operations. However, some commonly used operations, such as padding and batch normalization layers, may affect stationarity. Some strategies that can be implemented to reduce their impact is removing padding, masking outputs, and replacing batch normalization with FRN layers (Singh and Krishnan 2020) or other valid normalization. We implement these strategies further in the details of our applications.

Learning NLC Representations Via Supervision

The supervised approach links the proposed representation learning to the seminal work of Rubin (1978) on propensity scores for causal inference. We briefly summarize balancing scores following the standard presentation.

Definition 4 (Propensity score)**.** $b(\boldsymbol{L}_s)$ *is a* balancing score *iff* $A_s \perp\!\!\!\perp \boldsymbol{L}_s \mid b(\boldsymbol{L}_s)$*. The coarsest balancing score is* $b(\boldsymbol{L}_s) := p(A_s = 1 \mid \boldsymbol{L}_s)$*, widely known as the* propensity score*.*

Definition 5 (Prognostic score)**.** $b(\boldsymbol{L}_s)$ *is a* prognostic score *iff* $Y_s(0) \perp\!\!\!\perp \boldsymbol{L}_s \mid b(\boldsymbol{L}_s)$*. The coarsest prognostic score is* $b(\boldsymbol{L}_s) := \mathbb{E}[Y_s(0) \mid \boldsymbol{L}_s]$*.*

The propensity score blocks confounding through the treatment (Rubin 2005); prognostic scores do so through the outcome (Hansen 2008). The importance of these definitions is summarized by the next well-known result.

Proposition 3. *If $b(\boldsymbol{L}_s)$ is a balancing score, then $\boldsymbol{L}_s$ suffices to control for confounding iff $b(\boldsymbol{L}_s)$ does. The same result holds for the prognostic score under the additional assumption of no effect modification.*

This result suggests considering $\boldsymbol{L}_s$ to be implicitly defined by the full covariates grid "centered" at s, letting the network weights learn the effective radius of dependence. We can equate $\boldsymbol{Z}_{\theta,s}$ to either the propensity score or the prognostic score via direct regression, which amounts to minimizing the binary classification and regression loss:

$$\mathcal{L}^{\text{prop}}_{\text{sup}}(\theta) = \textstyle\sum_{s\in\mathbb{S}} \text{CrossEnt}(A_s, \boldsymbol{Z}_{\theta,s}) \tag{2}$$

$$\mathcal{L}^{\text{prog}}_{\text{sup}}(\theta) = \textstyle\sum_{s\in\mathbb{S}:A_s=0} (Y_s - \boldsymbol{Z}_{\theta,s})^2. \tag{3}$$

Notice that Eq. (3) applies only to untreated units. The learned propensity score can be directly plugged into a robust estimator such as IPTW or it can be used as a covariate in the case of the prognostic score. Learning θ through supervision results in an efficient scalar $\boldsymbol{Z}_{\theta,s}$ compressing NLC information, allowing for θ to just attend to relevant neighboring covariate information that pertains to confounding. Yet supervision may not be possible with small-data studies where Y_s and A_s are only measured sparsely. In such cases, the supervised model will likely overfit the data. For example, in application 1 of the paper, $\mathbb{S}$ consists only of measurements at 473 power plants, while the size of $\mathbb{G}$ is 128×256. Overfitting would invalidate common causal inference methods like IPTW that rely on unbiased estimates of the propensity score.

Representations Via Self-supervised Dimensionality Reduction

Self-supervision frames the representation learning problem as dimension reduction without reference to the treatment or outcome. The representations are then used to learn a balancing score for causal effect estimation in a second analysis stage. This approach requires specification of a fixed neighborhood $\mathcal{N}_s$ (parameterized by a radius R) and latent dimension k, resulting on different representations for different hyper-parameter choices, which can be selected using standard model selection techniques (such as AIC) in the second stage. The dimension reduction's objective is that $\boldsymbol{Z}_{\theta,s}$ encodes predictive information of any $\boldsymbol{X}_{s+\delta}$ for $(s+\delta) \in \mathcal{N}_s$. A simple predictive model $\boldsymbol{X}_{s+\delta} \approx g_\phi(\boldsymbol{Z}_{\theta,s}, \delta)$ is proposed. First, let $\boldsymbol{\Gamma}_\phi(\cdot)$ be a function taking an offset δ as input and yielding a $k \times k$ matrix, and let $h_\psi(\cdot)\colon \mathbb{R}^k \to \mathbb{R}^d$ be a decoder with output values in the covariate space. The idea is to consider $\boldsymbol{\Gamma}_\phi(\delta)$ as a selection operator acting on $\boldsymbol{Z}_{\theta,s}$. The task loss function can be written succinctly as

$$\begin{aligned} &\mathcal{L}_{\text{self}}(\theta, \phi, \psi \mid R) = \\ &\textstyle\sum_{s\in\mathbb{G}} \sum_{\{\delta\colon \|\delta\|\le R\}} \left(\boldsymbol{X}_{s+\delta} - h_\psi(\boldsymbol{\Gamma}_\phi(\delta)\boldsymbol{Z}_{\theta,s})\right)^2. \end{aligned} \tag{4}$$

The technical appendix provides additional intuition about Eq. (4) and a connection with PCA. While Eq. (4) is formulated for spatial dimensionality reduction only, an advantage

	patch-based		*self-supervised*	*supervised*				*spatial + supervised*	
Task	VAE	CRAE	W2V-SELF	WX	LOCAL	AVS	W2V-SUP	CAR	W2V-CAR
Linear	0.58	0.04	0.02	0.58	0.58	0.53	**0.01**	0.58	0.04
Linear-sparse	0.59	**0.06**	**0.06**	0.58	0.58	0.57	0.1	0.59	0.1
Non-linear	0.58	0.13	0.11	0.58	0.58	0.58	**0.07**	0.58	0.18
Non-linear-sparse	0.59	0.15	**0.14**	0.57	0.57	0.57	0.15	0.58	0.19

Table 1: Comparisons in average causal effect error (*Bias* $= \sum_{i=1}^{n} n^{-1}(\hat{\tau}^{(i)}_{\text{IPTW}} - \tau_{\text{ATE}})|$) for different propensity score models in simulated datasets across $n = 10$ random seeds. *Dense task*: A_s and Y_s are observed on the full 128×256 grid. *Sparse task*: A_s and Y_s are observed in 1000 points scattered throughout the grid.

of this expression is that it can be easily extended to multi-task settings and dimensionality reduction in the temporal axis for spatiotemporal data. We plan to explore these possibilities for future work.

NLC and Interference

The introduction briefly contrasted NLC with the related problem of interference, a topic that we expand on here. We first formalize the concept of interference, following closely the form of interference considered in Forastiere, Airoldi, and Mealli (2021), which replaces SUTVA with the following neighborhood-level assumption, termed the *stable unit neighborhood treatment value assignment* (SUTNVA).

Assumption 3 (SUTNVA)**.** *(1) Consistency: there is only one version the treatment. (2) Neighborhood-level interference: for each location s, there is a neighborhood $\mathcal{N}_s$ such that the potential outcomes depend only on the treatments at $\mathcal{N}_s$. Together, these conditions imply that $Y_s(\boldsymbol{a}) = Y_s(\boldsymbol{a}_{\mathcal{N}_s})$ for any assignment vector $\boldsymbol{a} \in \{0,1\}^{|\mathbb{S}|}$, and that the observed outcome is the potential outcome for the observed treatment, i.e., $Y_s = Y_s(\mathbf{A}_{\mathcal{N}_s})$.*

This definition of interference only considers *direct* interference, leaving aside indirect mechanisms such as contagion (Ogburn 2018; Shalizi and Thomas 2011). Investigating the role of NLC in such scenarios is left for future work. We now describe one generalization of the ATE for this type of direct interference. The statement uses potential outcomes of the form $Y_s(a_s = a, \mathbf{A}_{\mathcal{N}_s \setminus \{s\}})$ – a short-hand notation for the potential outcome that assigns the treatments of all the neighbors of s to their observed treatments in the data.

Definition 6 (DATE)**.** *The direct average treatment effect (DATE) is the quantity $\tau_{DATE} = |\mathbb{S}|^{-1} \sum_{s \in \mathbb{S}} \{Y_s(a_s = 1, \mathbf{A}_{\mathcal{N}_s \setminus \{s\}}) - Y_s(a_s = 0, \mathbf{A}_{\mathcal{N}_s \setminus \{s\}})\}$.*

The following proposition by Forastiere, Airoldi, and Mealli (2021) states two conditions under which one can "ignore" interference.

Proposition 4. *Assume SUTNVA. Conditions (1) and (2) correspond to the notions of neighborhood-level ignorability and conditional independence of the neighboring treatments. If (1) $\mathbf{A}_{\mathcal{N}_s} \perp\!\!\!\perp Y_s(\boldsymbol{a}) \mid \boldsymbol{L}_s$ for all $\boldsymbol{a} \in \{0,1\}^{|\mathcal{N}_s|}$ and (2) $A_s \perp\!\!\!\perp A_{s'} \mid \boldsymbol{L}_s$ for all $s \in \mathbb{S}, s' \in \mathcal{N}_s$. Then Eq.* (1) *is an unbiased estimator of τ_{DATE}.*

When NLC is present (the arrows from $X_{s'}$ in Fig. 1b), conditions (1) and (2) can be violated. To see this, consider

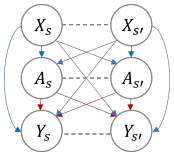


Figure 2: Interference + NLC.

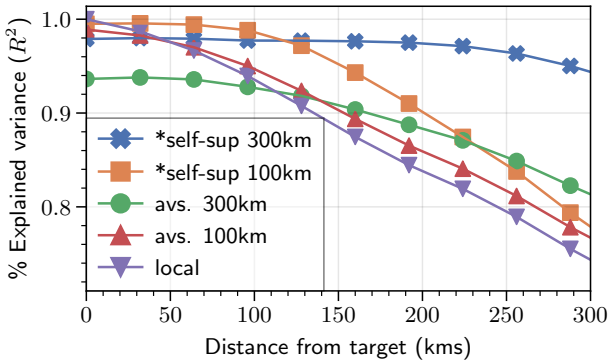
Fig. 2 representing the co-occurrence of interference and NLC. Adjusting only for local covariates would violate condition (1) with a spurious correlation between Y_s and $A_{s'}$ (through the backdoor path $Y_s \leftarrow \boldsymbol{X}_{s'} \rightarrow A_{s'}$). Similarly, a spurious correlation between A_s and $A_{s'}$ would persist via the path $A_s \leftarrow \boldsymbol{X}_{s'} \rightarrow A_{s'}$. For such cases, *weather2vec* can play an important role in satisfying (1) and (2) since, after controlling for NLC (consisting in Fig. 2 of adjusting for both $\boldsymbol{X}_s$ and $\boldsymbol{X}_{s'}$ and blocking the incoming arrows from neighboring covariates into one's treatments and outcomes), the residual dependencies would more closely resemble those of Fig. 1c. In summary, adjusting for NLC with *weather2vec* can aid satisfaction of the conditional independencies required to estimate causal effects with the same estimator used to estimate the ATE absent interference. Notice that τ_{DATE} is not the only estimand of interest, for instance, in future work we wish to explore the role of non-local covariates when estimating *spill-over* effects (Ogburn 2018).

Simulation Study

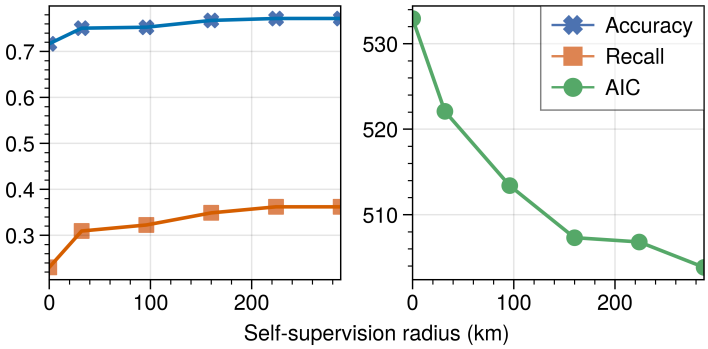
We conduct a simulation study that roughly mimics a dataset where pollution dispersion is influenced by non-local meteorological covariates as in our applications. We briefly describe the setup and results here. The technical appendix contains additional details and visualizations.

Data generation summary. Two-dimensional covariates simulating wind vectors are generated from the gradient field of a random spatial process. The treatment probability and the outcome (simulating air pollution) are non-local functions of the covariates such that areas with lower outcomes have a higher probability of treatment, with a fixed treatment effect of $\tau_{\text{ate}} = 0.1$. Two varying factors are considered: whether $\mathbb{S}$ is dense or sparse; and whether the simulated data is linear or non-linear on the covariates. The implicit radius of NLC is determined by using 13×13 convolution kernels to simulate the treatment and outcomes with non-linear operations.

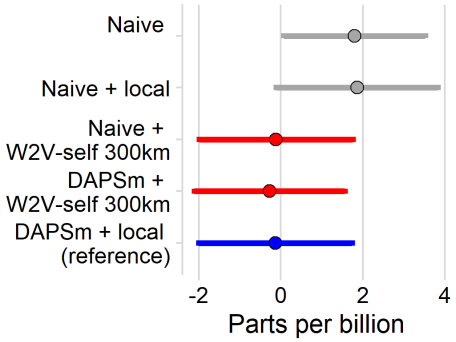
Baselines. We implement the supervised (W2V-SUP) and



(a) Self-supervision, NARR data



(b) Fit metrics, propensity score model



(c) Estimated causal effects

Figure 3: Application 1: The effectiveness of catalytic devices to reduce power plant ozone emissions.

self-supervised methods (W2V-SELF) using depth-2 U-nets. We then compare them with several baselines: first, no adjustment (UN), computed as the difference in means of treated and non-treated; LOCAL, which uses local covariates only; AVG, which appends averages of neighboring covariates, assuming the neighborhood size is known. Next, we use two convolutional autoencoders baselines of dimension reduction that operate on pre-extracted patches of the oracle size; CRAE (Blier-Wong et al. 2020) and VAE (Kingma and Welling 2013). Notice that although we include these baselines for reference, patch-based estimates do not scale to large datasets. Next, we consider WX linear logistic classification (Elhorst 2010) using a larger kernel of the oracle size. Finally, an approach based solely on spatial modeling CAR (Besag 1974), and a hybrid method combining the spatial term with the supervised U-net (W2V-CAR). 10 random seeds are run for each configuration.

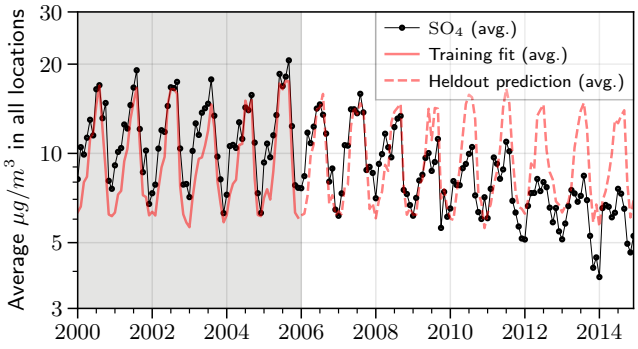
Causal estimation. All the estimates are based on IPTW from a learned propensity score. For LOCAL, avg, and methods based on dimension reduction, we fit the learned vectors through a two-layer feed-forward network (FFN) for the propensity score. We consider four latent dimensions for the self-supervised method and all dimension reduction baselines.

Results summary. The results are summarized in Table 1. When $\mathbb{S}$ is dense, the supervised *weather2vec* outperforms all others, exhibiting near-zero bias in the linear case and a small amount of finite-sample bias in the nonlinear case. The self-supervised version is competitive in all scenarios, performing better than the alternatives in the nonlinear sparse case.

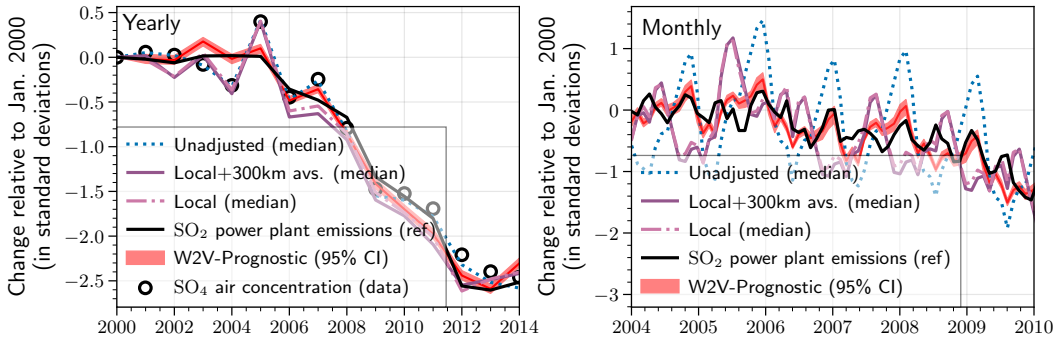
Applications in Air Pollution and Climate

Application 1: Quantifying the impact of power plant emission reduction technologies The study aims to quantify the impact of SCR/SNCR catalytic devices (Muzio, Quartucy, and Cichanowiczy 2002) to reduce emissions among coal-fired power plants in the U.S (Papadogeorgou 2016). See the technical appendix for a detailed description of the dataset. Since air quality regulations are inherently regional and power plants are concentrated in regions with similar weather and economic demand factors, regional weather correlates with the assignment of the intervention. Further, weather patterns (such as wind vectors, precipitation, and humidity) dictate regional differences in the formation and dispersion of ambient air pollution. Thus, the weather is a potential confounding factor that cannot be entirely characterized by local measurements.

Self-supervised features from NARR. We construct a dataset of atmospheric covariates following Shen, Mickley,



(a) Prognostic score fit averaged over the entire grid $\mathbb{G}$.



(b) Detrended series at $\mathbb{S}^*$ resembles power plant emissions. (*Left*) Yearly trend $\delta_{\text{year}(t)}$. (*Right*) Monthly trend $\delta_{\text{year}(t)} + \gamma_{\text{month}(t)}$

Figure 4: Application 2: Meteorological detrending of SO_4.

and Murray (2017). We downloaded monthly NARR data (Mesinger et al. 2006) containing averages of gridded atmospheric covariates across the mainland U.S. for the period 2000-2014. We considered 5 covariates: temperature at 2m, relative humidity, total precipitation, and north-sound and east-west wind vector components. For each variable, we also include its year-to-year average. Our dataset is identical to Shen, Mickley, and Murray (2017), except that they project it to a lower resolution, while we keep it so that each grid cell covers roughly a 32×32 km area, forming a 128×256 grid. We implemented the self-supervised *weather2vec* with a lightweight U-net of depth 2, 32 hidden units, and only one convolution per level; see the appendix for additional details. To measure the quality of the encoding, Fig. 3a shows the percentage of variance explained (R^2), comparing with neighbor averaging and local values. This metric is computed as the coefficient determination, which is essentially the average squared correlation between the prediction and the actual data, aggregated by distance to the center. The results show that the 32-dimensional self-supervised features provide a better reconstruction than averaging and using the local values. For instance, the 300km averages only capture 82% of the variance, while the self-supervised *weather2vec* features capture 95%. See the appendix for details on the calculation of the R^2 and neural network architecture.

Estimated pollution reduction. We evaluate different propensity score models for different neighborhood sizes of the June 2004 NARR *weather2vec*-learned features with the same logistic model and other covariates as in DAPSm, augmented with the self-supervised features. We selected the representation using features within a 300km radius based on its accuracy, recall, and AIC in the propensity score model relative to other considered neighborhood sizes (Figure 3b). The causal effects are then obtained by performing 1:1 nearest neighbor matching on the estimated propensity score as in DAPSm. Figure 3c compares treatment effect estimates for different estimation procedures. Overall, standard (naive) matching using the self-supervised features is comparable to DAPSm, but without requiring the additional spatial adjustments introduced by DAPSm. The same conclusion does not hold when using local weather only , which (as in the most naive adjustment) provides the scientifically uncredible result that emissions reduction systems significantly *increase* ozone pollution. Do notice the wide confidence intervals which are constructed using conditional linear models fitting the matched data sets (Ho et al. 2007). Thus, while the mean estimate shows a clear improvement, the intervals show substantial overlap, warranting caution.

Application 2: Meteorological detrending of sulfate We investigate meteorological detrending of the U.S. sulfate (SO_4) time series with the goal (common to the regulatory policy and atmospheric science literature) of adjusting long-

term pollution trends by factoring out meteorologically-induced changes and isolating impacts of emission reduction policies (Wells et al. 2021). We focus on SO_4 because it is known that its predominant source in the U.S. is SO_2 emissions from coal-fired power plants, on which observed data are available for comparison. Thus, we hypothesize that an effectively detrended SO_4 time series will closely resemble that of the power plant emissions.

Prognostic score. We obtained gridded SO_4 concentration data publicly available from van Donkelaar et al. (2021), consisting of average monthly values in the mainland U.S. in 2000–2014. The data is aggregated into 32km-by-32km cells to match the resolution of atmospheric covariates. The model uses a U-net with quadratic loss for the (log) concentrations of SO_4. Since the prognostic score is defined based on outcome data in the absence of treatment, we leverage the fact that the power plant emissions were relatively constant for the period 2000-2005 and using 2006 as test data – regarding this period as absent of treatment. The model predictions, aggregated by all points in the grid is shown in Figure 4a. The difference between the red line (the prognostic score fit) and the black dotted line (the SO_4) observations during 2000 - 2006 is a proxy for the meteorology-induced changes in the absence of treatment.

Trend estimation. For comparability we adhere to the recommended detrending model by (Wells et al. 2021). Accordingly, we specify a regression with a year and seasonal fixed-effect term. Rather than pursue an entirely new methodology for detrending, we intentionally adhere to standard best practices and merely aim to evaluate whether augmenting this approach with the *weather2vec* representation of the prognostic score offers improvement. The outcome $\log(Y_{s,t})$ for untreated units is regressed using the predictive model

$$\mu_{s,t} = \alpha + \delta_{\text{year}(t)} + \gamma_{\text{month}(t)} + \textstyle\sum_{j=1}^{p} \beta_p X_{st}^p \qquad (5)$$

for all $s \in \mathbb{S}^*$ and $t = 1, \ldots, T$; and where δ_ℓ is the year effect for $\ell = 2000, \ldots, 2014$; γ_κ is the seasonal (monthly) effect for $\kappa = 1, \ldots, 12$; $\mathbb{S}^* \subset \mathbb{S}$ are the locations of the power plants; and X_{st}^p are the controls with linear coefficients $\beta_{s,p}$. These controls are obtained from a B-spline basis of degree 3 using: 1) local weather only, and 2) local weather plus the *weather2vec* prognostic score. The model is fitted using Bayesian inference with a Gibbs sampler. Figure 4b shows the fitted (posterior median) yearly and monthly trends, which resemble the power plant emissions trends much more closely than the predicted trends from models that include local or neighborhood average weather. Notice the "double peak" per year in the monthly power plant emissions (owing to seasonal power demand), which is only captured by the detrended *weather2vec* series.

Discussion and Future Work

While notions of NLC have been acknowledged in causal inference, a potential-outcomes formalization of NLC and flexible tools to address it are lacking. We offer such a formalization, along with a flexible representation learning approach to account for NLC with gridded covariates and treatments and outcomes measured (possibly sparsely) on the same grid. Our proposal is most closely tailored to problems in air pollution and climate science, where key relationships may be confounded by meteorological features, and promising results from two case studies evidence the potential of *weather2vec* to improve causal analyses over those with more typical accounts of local weather. A limitation of the approach is that the learned *weather2vec* representations are not as interpretable as direct weather covariates and using them could impede transparency when incorporated in policy decisions. Future work could explore new methods for interpretability. Other extensions could include additional data domains, such as graphs and longitudinal data with high temporal resolution. The links to causal interference explored in Section *NLC and Interference* also offer clear directions for future work to formally account for NLC in the context of estimating causal effects with interference and spill-over.

Acknowledgments

Part of this work was done while Mauricio Tec was a doctoral student at the Department of Statistics and Data Sciences at the University of Texas at Austin. The work was supported by the National Institutes of Health (NIH-R01ES26217, NIH-R01ES030616) and the US Environmental Protection Agency (EPA-RD835872). Its contents are solely the responsibility of the grantee and do not necessarily represent the official views of the US EPA. Further, the US EPA does not endorse the purchase of any commercial products or services mentioned in the publication. Some of the computations in this paper were run on the FASRC Cannon cluster supported by the FAS Division of Science Research Computing Group at Harvard University.

References

Besag, J. 1974. Spatial interaction and the statistical analysis of lattice systems. *Journal of the Royal Statistical Society: Series B (Methodological)*, 36(2): 192–225.

Bhattacharya, R.; Malinsky, D.; and Shpitser, I. 2020. Causal inference under interference and network uncertainty. In *Uncertainty in Artificial Intelligence*, 1028–1038.

Blier-Wong, C.; Baillargeon, J.-T.; Cossette, H.; Lamontagne, L.; and Marceau, E. 2020. Encoding neighbor information into geographical embeddings using convolutional neural networks. In *The Thirty-Third International Flairs Conference*.

Chaix, B.; Leal, C.; and Evans, D. 2010. Neighborhood-level confounding in epidemiologic studies: unavoidable challenges, uncertain solutions. *Epidemiology*, 21(1): 124–127.

Cohen-Cole, E.; and Fletcher, J. M. 2008. Is obesity contagious? social networks vs. environmental factors in the obesity epidemic. *Journal of health economics*, 27(5): 1382–1387.

Cole, S. R.; and Hernán, M. A. 2008. Constructing inverse probability weights for marginal structural models. *American journal of epidemiology*, 168(6): 656–664.

Elhorst, J. P. 2010. Applied spatial econometrics: raising the bar. *Spatial economic analysis*, 5(1): 9–28.

Fan, J.; Chen, D.; Wen, J.; Sun, Y.; and Gomes, C. P. 2021. Resolving Super Fine-Resolution SIF via Coarsely-Supervised U-Net Regression. In *NeurIPS 2021 Workshop on Tackling Climate Change with Machine Learning*.

Florax, R.; and Folmer, H. 1992. Specification and estimation of spatial linear regression models: Monte Carlo evaluation of pre-test estimators. *Regional science and urban economics*, 22(3): 405–432.

Forastiere, L.; Airoldi, E. M.; and Mealli, F. 2021. Identification and estimation of treatment and interference effects in observational studies on networks. *Journal of the American Statistical Association*, 116(534): 901–918.

Hanna, J.; Mommert, M.; Scheibenreif, L. M.; and Borth, D. 2021. Multitask Learning for Estimating Power Plant Greenhouse Gas Emissions from Satellite Imagery. In *NeurIPS 2021 Workshop on Tackling Climate Change with Machine Learning*.

Hansen, B. B. 2008. The prognostic analogue of the propensity score. *Biometrika*, 95(2): 481–488.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 770–778.

Ho, D. E.; Imai, K.; King, G.; and Stuart, E. A. 2007. Matching as nonparametric preprocessing for reducing model dependence in parametric causal inference. *Political analysis*, 15(3): 199–236.

Johansson, F.; Shalit, U.; and Sontag, D. 2016. Learning representations for counterfactual inference. In *International conference on machine learning*, 3020–3029. PMLR.

Keller, B.; Kim, J.-S.; and Steiner, P. M. 2015. Neural networks for propensity score estimation: Simulation results and recommendations. In *Quantitative psychology research*, 279–291. Springer.

Khan, K.; and Calder, C. A. 2020. Restricted spatial regression methods: Implications for inference. *Journal of the American Statistical Association*, 1–13.

Kingma, D. P.; and Welling, M. 2013. Auto-encoding variational bayes. *arXiv:1312.6114*.

Larraondo, P. R.; Renzullo, L. J.; Inza, I.; and Lozano, J. A. 2019. A data-driven approach to precipitation parameterizations using convolutional encoder-decoder neural networks. *arXiv preprint 1903.10274*.

Lu, H.-C.; and Chang, T.-S. 2005. Meteorologically adjusted trends of daily maximum ozone concentrations in Taipei, Taiwan. *Atmospheric Environment*, 39(35): 6491–6501.

Mesinger, F.; DiMego, G.; Kalnay, E.; Mitchell, K.; Shafran, P. C.; Ebisuzaki, W.; Jović, D.; Woollen, J.; Rogers, E.; Berbery, E. H.; et al. 2006. North American regional reanalysis. *Bulletin of the American Meteorological Society*, 87(3).

Muzio, L.; Quartucy, G.; and Cichanowiczy, J. 2002. Overview and status of post-combustion NOx control: SNCR, SCR and hybrid technologies. *International Journal of Environment and Pollution*, 17(1-2): 4–30.

Ogburn, E. L. 2018. Challenges to estimating contagion effects from observational data. In *Complex Spreading Phenomena in Social Systems*, 47–64. Springer.

Ogburn, E. L.; and VanderWeele, T. J. 2014. Causal diagrams for interference. *Statistical science*, 29(4): 559–578.

Papadogeorgou, G. 2016. Power Plant Emissions Data V2. *Harvard Dataverse*. https://doi.org/10.7910/DVN/M3D2NR. Accessed 2022-07-01.

Papadogeorgou, G.; Choirat, C.; and Zigler, C. M. 2019. Adjusting for unmeasured spatial confounding with distance adjusted propensity score matching. *Biostatistics*, 20(2).

Reich, B. J.; Yang, S.; Guan, Y.; Giffin, A. B.; Miller, M. J.; and Rappold, A. 2021. A review of spatial causal inference methods for environmental and epidemiological applications. *International Statistical Review*, 89(3): 605–634.

Rolnick, D.; Donti, P. L.; Kaack, L. H.; Kochanski, K.; Lacoste, A.; Sankaran, K.; Ross, A. S.; Milojevic-Dupont, N.; Jaques, N.; Waldman-Brown, A.; et al. 2022. Tackling climate change with machine learning. *ACM Computing Surveys (CSUR)*, 55(2): 1–96.

Ronneberger, O.; Fischer, P.; and Brox, T. 2015. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*.

Rubin, D. B. 1978. Bayesian inference for causal effects: The role of randomization. *The Annals of statistics*, 34–58.

Rubin, D. B. 2005. Causal inference using potential outcomes: Design, modeling, decisions. *Journal of the American Statistical Association*, 100(469).

Rubin, D. B. 2008. For objective causal inference, design trumps analysis. *The Annals of Applied Statistics*, 2(3): 808–840.

Sadeghi, M.; Nguyen, P.; Hsu, K.; and Sorooshian, S. 2020. Improving near real-time precipitation estimation using a U-Net convolutional neural network and geographical information. *Environmental Modelling & Software*, 134.

Setoguchi, S.; Schneeweiss, S.; Brookhart, M. A.; Glynn, R. J.; and Cook, E. F. 2008. Evaluating uses of data mining techniques in propensity score estimation: a simulation study. *Pharmacoepidemiology and drug safety*, 17(6): 546–555.

Shalit, U.; Johansson, F. D.; and Sontag, D. 2017. Estimating individual treatment effect: generalization bounds and algorithms. In *International Conference on Machine Learning*, 3076–3085. PMLR.

Shalizi, C. R.; and Thomas, A. C. 2011. Homophily and contagion are generically confounded in observational social network studies. *Sociological methods & research*, 40(2): 211–239.

Shen, L.; Mickley, L. J.; and Murray, L. T. 2017. Influence of 2000–2050 climate change on particulate matter in the United States: results from a new statistical model. *Atmospheric Chemistry and Physics*, 17(6): 4355–4367.

Shi, C.; Blei, D. M.; and Veitch, V. 2019. Adapting neural networks for the estimation of treatment effects. In *Proceedings of the 33rd International Conference on Neural Information Processing Systems*, 2507–2517.

Singh, S.; and Krishnan, S. 2020. Filter response normalization layer: Eliminating batch dependence in the training of deep neural networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 11237–11246.

Sobel, M. E. 2006. What do randomized studies of housing mobility demonstrate? Causal inference in the face of interference. *Journal of the American Statistical Association*, 101(476): 1398–1407.

Tchetgen, E. J. T.; and VanderWeele, T. J. 2012. On causal inference in the presence of interference. *Statistical methods in medical research*, 21(1).

Tec, M.; Scott, J.; and Zigler, C. 2022. Weather2vec: Representation Learning for Causal Inference with Non-Local Confounding in Air Pollution and Climate Studies. *arXiv preprint arXiv:2209.12316*.

van Donkelaar, A.; Hammer, M. S.; Bindle, L.; Brauer, M.; Brook, J. R.; Garay, M. J.; Hsu, N. C.; Kalashnikova, O. V.; Kahn, R. A.; Lee, C.; et al. 2021. Monthly global estimates of fine particulate matter and their uncertainty. *Environmental Science & Technology*, 55(22): 15287–15300.

Veitch, V.; Wang, Y.; and Blei, D. 2019. Using embeddings to correct for unobserved confounding in networks. *Advances in Neural Information Processing Systems*, 32.

Wells, B.; Dolwick, P.; Eder, B.; Evangelista, M.; Foley, K.; Mannshardt, E.; Misenis, C.; and Weishampel, A. 2021. Improved estimation of trends in US ozone concentrations adjusted for interannual variability in meteorological conditions. *Atmospheric Environment*, 248.

Westreich, D.; Lessler, J.; and Funk, M. J. 2010. Propensity score estimation: machine learning and classification methods as alternatives to logistic regression. *Journal of clinical epidemiology*, 63(8).

Zigler, C. M.; and Papadogeorgou, G. 2021. Bipartite causal inference with interference. *Statistical science: a review journal of the Institute of Mathematical Statistics*, 36(1): 109.