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

Anonymous submission

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.

1 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 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. Importantly, here we do not consider the alternate objective of predicting or forecasting modeling outcomes (such as climate and air pollution modeling) but to infer the *causal* effect that an intervention had on these outcomes. For this purpose, we leverage representation learning methods computer vision and ground these methods in the formal theory of causal inference using potential outcomes (Rubin 2008).

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 example is an application to the problem of 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 (Rubin 2005), clarifying some connections with interference and related methods (Tchetgen and VanderWeele 2012; Forastiere, Airoldi, and Mealli 2021; Sobel 2006).
2. Expand the library of NN methods in causal inference by proposing a U-net (Ronneberger, Fischer, and Brox 2015) 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.

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). None of these works, however, 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.

Approaches to learn 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. Similarly, CRAE (Blier-Wong et al. 2020) uses an autoencoder to encode pre-extracted patches of regional census data into a lower-dimensional vector that is fed into an econometric regression, somewhat analogous to the dimensionality reduction step in the self-supervised formulation of *weather2vec*, but requiring to pre-process the data into patches. In a similar spirit and highly relevant to this paper, Shen, Mickley, and Murray (2017) apply patch-wise PCA to meteorological data to improve predictive power of air pollution. However, their method is framed within a regression framework, and not for the purpose of estimating average causal effects. 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; Veitch, Wang, and Blei 2019; Papadogeorgou, Choirat, and Zigler 2019). 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. As an alternative, 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. We use this model as a baseline in one of our applications. Veitch, Wang, and Blei (2019) take a related approach in a context where network proximity is viewed analogously to spatial proximity, and show that, under certain regularity conditions, network proximity can be used as a proxy for a network-level unobserved confounder. They propose a mechanism to learn embeddings that capture confounding information and used them together with augmented inverse probability weighting to obtain unbiased causal estimates. Importantly, they only consider the "pure homophily" case (Shalizi and Thomas 2011), where the entirety of the confounding is assumed to be encoded by relative position in the network. While some of these methods could be useful for NLC, they all primarily target settings where confounding is local.

Finally, NLC is distinct from, but notionally similar to, *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). Both interference and NLC arise from spatial (or network) interaction, and they both impose limitations on standard causal inference methods. Various works in this literature have discussed the role of conditional ignorability given neighboring covariates; for instance, in Vansteelandt (2007) and Forastiere, Airoldi, and Mealli (2021). However, to the best of our knowledge, flexible statistical methods specifically addressing NLC by learning the dependencies with respect to neighboring covariates have been ignored.

2 Potential outcomes and NLC

This section follows the standard presentation of the potential outcomes framework, also known as the Rubin Causal Model (RCM) (Rubin 2008), and adapts it to the case of NLC confounding by introducing its formal definition. 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 notaton. 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 the emissions of some pollutant. Next, $\mathbb{S}$ 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 that contains the rasterized covariates $\{\boldsymbol{X}_s \in \mathbb{R}^d \colon s \in \mathbb{G}\}$; $\mathbf{X}_B$ where $B \subset \mathbb{G}$ is a set means $\mathbf{X}_B = \{\boldsymbol{X}_s \mid s \in B\}$; and $X \perp\!\!\!\perp Y \mid Z$ means that X and Y are conditionally independent given Z. Throughout $\boldsymbol{X}_s$ is assumed to consist of pre-treatment covariates only, meaning they are not affected by the treatment or outcome. Finally, we use the generic notation $p(\cdot)$ to denote a density or probability function.

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 $\boldsymbol{a} = (a_1, \ldots, a_{|\mathbb{S}|})$.*

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 to define 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.

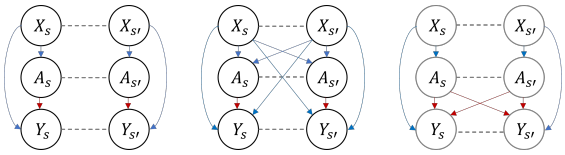
Assumption 2 (Treatment Ignorability)**.** *The treatment* A_s *is ignorable with respect to some vector of controls* $\boldsymbol{L}_s$ *if and only if* $Y_s(1), Y_s(0) \perp\!\!\!\perp A_s \mid \boldsymbol{L}_s$*.*

This ignorability assumption would fail to hold when there is a confounding variable U not in $\boldsymbol{L}_s$ that is associate with both the outcome and the treatment. This variable U induces a non-causal correlation between the treatment and outcome which leads to wrong conclusion about causal effects when not accounted for. 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 (*$\mathcal{N}_s \neq \{s\}$*).*

To put the definition into context, in our power plat 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 their respective treatments A_s, such as the installation of a catalytic device. This phenomenon is not this setting. For instance, it can occur in other spatial scenarios such as with socioeconomic and demographic data (Blier-Wong et al. 2020). In summary, in this paper we relax the typical assumption of ignorability and replace it with a version in which it only holds after controlling for an appropriate neighborhood of covariates. We then propose computer vision methods that allow to do this in a statistically efficient and flexible way.

Figs. 1a and 1b show a graphical representation of local confounding and NLC respectively. Horizontal dotted lines



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

Figure 1: Confounding types.

emphasize that there may be 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).) The key point is that Fig. 1b shows a scenario where the neighboring covariates suffice for ignorability and SUTVA still holds, whereas this is generally not the case with interference. Interference is further discussed in Section 4.

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.*

All the proofs are in Appendix B. We conclude this section with a classic result stating that any sufficient $\boldsymbol{L}_s$ can be used to estimate the ATE from quantities and relations in the observed data.

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* $\mathbb{S}$*).*

While Eq. (1) already offers a way to estimate causal effects from observed data (by estimating the two inner conditional expectations), it is often highly sensitive to the specification of the expected outcome model. Thus, alternative causal estimators are often preferred, such as inverse probability of treatment weighting (IPTW) (Cole and Hernán 2008); described in Appendix A for completeness, and which makes use of estimates of the *propensity score* introduced in the next section.

3 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

$\mathbb{E}[\mathbb{E}[Y_s \mid \boldsymbol{L}_s, A_s = a]$ 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, and $\mathbb{S}$ can also be small, 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 $p(A_s, Y_s \mid \mathbf{X}_B = x) = p(A_{s+\delta}, Y_{s+\delta} \mid \mathbf{X}_{B+\delta} = x)$.

Here we propose to use convolutional structures as base computational models enabling learning from stationarity. We focus specifically on the case of the U-net (Ronneberger, Fischer, and Brox 2015) for representation learning, since they are approximately spatially stationary and adhere to the causal structure of NLC shown in Fig. 1b. 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 stack residual layers (He et al. 2016) as a shallow alternative to a U-net. Residual layers have also been considered in combination with U-nets (Liu et al. 2020). We shall focus on the case of the U-net here for simplicity, since our aim here is to enable causal estimation under NLC. The essence of *weather2vec* is to define appropriate learning tasks to obtain the NN weights θ. Two such tasks are considered, summarized below and in Appendix K, 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 to learn a *propensity score* $p(A_s = 1 \mid \boldsymbol{Z}_{\theta,s})$ (formally introduced later in this section) which can be used within a well-established causal inference technique such as IPTW (Appendix A) to produce robust causal estimates of τ_{ATE}. The key innovation with respect to traditional causal propensity score methods is the inclusion of NLC information. An alternative to these propensity score methods is to use *prognostic scores* (also defined formally below), which learn instead a predictive model of the untreated outcomes and are used to estimate causal effects mainly through Eq. (1), losing some of the robustness properties of propensity score estimation but being more naturally applicable in some scenarios. See Hansen (2008) for a review on the prognostic score and Application 2.

An overview of the U-net for summarizing NLC

In this section, we provide a brief description of the U-net's functional form, referring the reader to Ronneberger, Fischer, and Brox (2015) for additional details. The U-net transformation involves two parts: a *contractive* stage and an symmetric *expansive stage*. Both of 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 linear or convolutional layer followings the expansive path (or before the contractive path, or both). Fig. 6 in the Appendix provides a visual example of the U-net architecture.

The unknown weights θ can 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 lares are all spatially stationary operations. However, attention must be paid to padding and batch normalization layers sometimes used in U-nets, which may affect stationarity. Some strategies can be implemented to reduce their impact is removing padding, masking outputs and replacing batch normalization with alternatives such as FRN layers (Singh and Krishnan 2020). We implement some of 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. Our presentation follows closely the standard material in this literature. As mentioned earlier, the key innovation with existing approaches is the use of convolutional NN structures to flexibly accommodate NLC in the learned score. We begin with the definition of balancing score and the related prognostic score.

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 the following strategy. We will consider $\boldsymbol{L}_s$ to be implicitly defined by the full grid $\mathbf{X}$ "centered" at s, where the effective size will be determined by the learned NN weights. Then, we can equate $\boldsymbol{Z}_{\theta,s}$, to either the propensity score or the prognostic score via direct regression, which amounts to minimizing the loss functions:

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

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

where CrossEnt is the binary cross-entropy loss. Observe that Eq. (3) applies only to untreated units, and it is thus not equivalent to standard regression. The propensity score can be directly plugged into a robust estimator such as with IPTW (Appendix A) and the prognostic score as a covariate for causal estimation via Eq. (1). 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.

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 to the data. For example, in application 1 in section 6, $\mathbb{S}$ consists only of measurements at 473 power plants, while the size of $\mathbb{G}$ is 128×256. An over-fitted propensity score would result in insufficient "overlap" (Stuart 2010) by assigning zero probability to the unobserved treatment, resulting in causal inferences that would rely on model extrapolation to areas where covariate information is not represented in both treatment groups. To avoid this, the self-supervised approach targets scenarios with sparse $\mathbb{S}$.

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 an 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

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

Appendix C provides additional intuition about Eq. (4). A connection with PCA is also described in Appendix D. While Eq. (4) is formulated for spatial dimensionality reduction only, an advantage 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.

4 NLC and Interference

Section 1 briefly contrasted NLC with the related problem of interference, a topic that we expand here. We first formalize the concept of interference. To ground the discussion, we follow 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 (Ogburn and VanderWeele 2014), 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* $\tau_{DATE} = |\mathbb{S}|^{-1}\sum_{s\in\mathbb{S}}\{Y_s(a_s = 1, \mathbf{A}_{\mathcal{N}_s\setminus\{s\}}) - Y_s(a_s = 0, \mathbf{A}_{\mathcal{N}_s\setminus\{s\}})\}$.

Notice that τ_{DATE} is not the only estimand of interest. For instance, the literature often considers *spill-over* effects (Ogburn 2018). But for this work, we only focus on τ_{DATE}, leaving other estimands for future work. Now consider the question: how does NLC affect the estimation of τ_{DATE} based on equation 1? We provide a partial answer based on the following result by Forastiere, Airoldi, and Mealli (2021), which

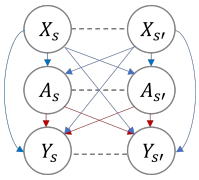


Figure 2: Interference + NLC.

states two conditions under which equation 1 is an unbiased estimator of τ_{DATE}.

Proposition 4. *Assume SUTNVA. If (1)* $\mathbf{A}_{\mathcal{N}_s} \perp\!\!\!\perp (Y_s(\boldsymbol{a}))_{\boldsymbol{a}\in\{0,1\}^{|\mathcal{N}_s|}} \mid \boldsymbol{L}_s$ *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}*.*

Conditions (1) and (2) correspond, respectively, to the notions of neighborhood-level ignorability and conditional independence of the neighboring treatments. It turns out that when NLC is present (the arrows from $X_{s'}$ in Fig. 1b), conditions (1) and (2) in the proposition can be violated. To see this, consider Fig. 2 representing the co-occurrence of interference and NLC. Adjusting only for local covariates would violate neighborhood ignorability condition (1) with a spurious correlation between Y_s and $A_{s'}$ (through the back-door path $Y_s \leftarrow \boldsymbol{X}_{s'} \rightarrow A_{s'}$). Similarly, a spurious correlation between A_s and $A_{s'}$ would persist (via the 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.

5 Simulation study

We conduct a simulation study that roughly mimics a dataset where pollution is dispersed in accordance with non-local meteorological covariates as in our applications. We briefly describe the setup here and the results. Appendix F contains a detailed explanation on data generation, baselines, and additional visualizations.

Data generation summary. The 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.

Causal estimation with *weather2vec* summary We use the self-supervised (W2V-SELF) and supervised (W2V-SUP) variants in a propensity score model for causal estimation using IPTW (Appendix A). The dimension-reduced vectors will then be passed through a two-layer feed-forward network (FFN) for propensity score estimation. We consider 4 latent dimensions for the self-supervised method and all dimension reduction baselines.

Baselines summary. We consider the following baselines for comparison: no adjustment (UN), which is simply the difference in expectations of treated and non-treated; LOCAL, which uses the same FFN; AVG, which is similar but appends averages to neighboring covariates, assuming the confounding neighborhood size is known. Next we include three baselines of dimensionality reduction that operate on pre-extracted patches of the oracle size: PCA; convolutional regional auto-encoder CRAE (Blier-Wong et al. 2020) and variational auto-encoders VAE (Kingma and Welling 2013), and then use the FFN for propensity score estimation. Notice that although we include these baselines for reference, patch-based estimates do not scale to large datasets. We also consider WX regression (Elhorst 2010) as a purely regression-based method that does not use IPTW and instead uses a single-layer convolution and takes the treatment as indicator variable input in linear regression. Finally, we also include an approach from spatial modeling using conditional auto-regression CAR for propensity score estimation (Besag 1974), and a hybrid method combining the spatial term with the supervised U-net (W2V-CAR). A total of 10 experiments are conducted for each configuration. The results are shown in Table 1.

Results summary. When $\mathbb{S}$ is dense, the supervised *weather2vec* outperforms all other methods, exhibiting near zero bias in the linear case and a small amount of finite-sample bias for the non-linear task. The self-supervised version is competitive in all scenarios, performing better than the alternatives in the non-linear sparse case. PCA was a surprisingly strong baseline in call cases, performing better than more sophisticated non-linear alternatives. While perhaps additional tuning would lead to increased performance of non-linear dimension reduction methods, it also suggests that when patch extraction is affordable, PCA may be a safe choice, particularly since one can always increase the number of latent features.

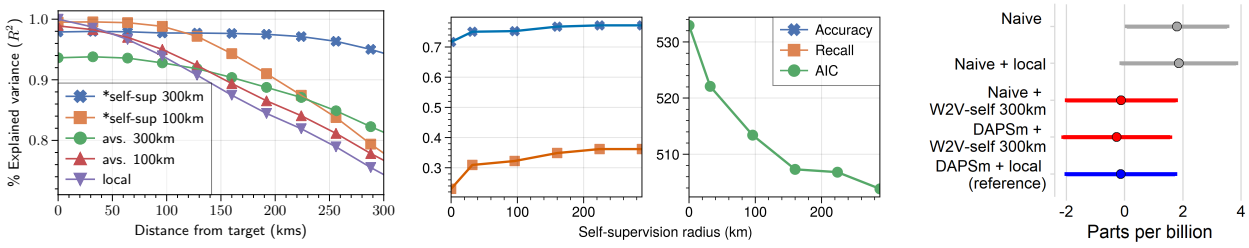
6 Applications in Air Pollution and Climate

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). Appendix G provides a 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 which cannot be entirely characterized by local measurements.

Self-supervised features from NARR. We construct a dataset of atmospheric covariates following Shen, Mickley, and Murray (2017). We downloaded monthly NARR data

		patch-based			*self-supervised*	*outcome reg.*	*sup. prop. score*			*spatial sup. prop. score*	
Task	*Metric*	PCA	VAE	CRAE	W2V-SELF	WX	LOCAL	AVS	W2V-SUP	CAR	W2V-SUP+CAR
Linear dense	↓ *Bias*	0.02	0.57	0.07	0.03	**0.00** (oracle)	0.59	0.59	**0.00**	0.59	**0.00**
	↓ *MSE*	0.01	3.46	0.05	0.01	**0.00** (oracle)	3.39	3.35	**0.00**	3.42	**0.00**
Linear sparse	↓ *Bias*	0.10	0.59	0.12	0.09	**0.02** (oracle)	0.60	0.60	0.16	0.60	0.34
	↓ *MSE*	0.15	3.54	0.17	0.12	**0.03** (oracle)	3.57	3.6	1.29	3.66	1.50
Non-linear dense	↓ *Bias*	0.06	0.58	0.10	0.06	0.20	0.58	0.58	0.05	0.58	**0.03**
	↓ *MSE*	0.04	3.42	0.11	0.04	0.38	3.39	3.35	0.03	3.42	**0.01**
Non-linear sparse	↓ *Bias*	0.17	0.60	0.21	**0.15**	0.22	0.59	0.59	0.34	0.60	0.36
	↓ *MSE*	0.31	3.66	0.57	**0.26**	0.48	3.53	3.55	1.35	3.58	2.02

Table 1: Comparisons in average causal effect error for different propensity score models in simulated datasets across $n = 10$. *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. *Bias* $:= \sum_{i=1}^{n} n^{-1}(\hat{\tau}^{(i)}_{\text{IPTW}} - \tau_{\text{ATE}})|$ and *MSE* $:= \sum_{i=1}^{n} n^{-1}(\hat{\tau}^{(i)}_{\text{IPTW}} - \tau_{\text{ATE}})^2$. In all simulations $\tau_{\text{ATE}} = 0.1$. Self-supervised methods use the oracle neighborhood size, which in practice needs to be determined via model selection.



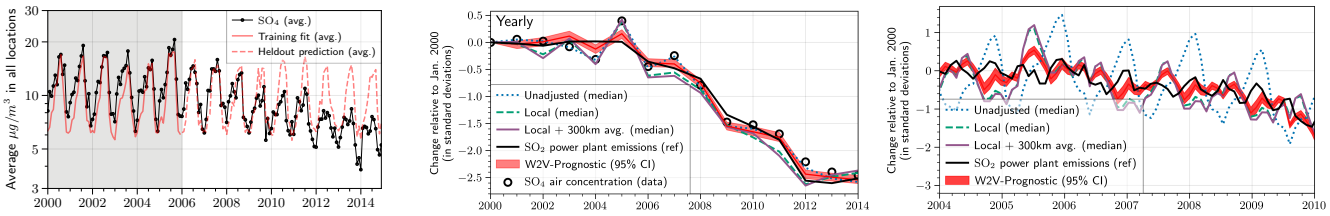
(a) Self-supervision, NARR data (b) Fit metrics, propensity score model (c) Estimated causal effects

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

(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 Appendix G for more details and a schematic of the U-net architecture. 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 Appendix H 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 on the basis of 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 un-credible result that emissions reduction systems significantly *increase* ozone pollution. Do notice the wide confidence intervals which are constructed using conditional linear models fit to the matched data sets (Ho et al. 2007). Thus, while the mean estimate shows a clear improvement, the intervals shows 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 emis-



(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.

sions 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 the Atmospheric Composition Analysis Group (Group 2001; van Donkelaar et al. 2021), consisting of average monthly value for each raster cell in the mainland U.S. for the period of study 2000–2014. The data is aggregated into 32km-by-32km grids 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

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

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^p_{st} 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.

7 Discussion and Future Work

While notions of NLC have been acknowledged in causal inference (most explicitly in spatial econometrics but also alluded to in literature on spatial confounding and interference), 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 4 also offer clear directions for future work to formally account for NLC in the context of estimating causal effects with interference and spill-over.

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.

Group, A. C. A. 2001. Surface PM2.5. Accessed September 2021. URL: https://sites.wustl.edu/acag.

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 Ba, J. 2014a. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Kingma, D. P.; and Ba, J. 2014b. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

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*.

Liu, N.; He, T.; Tian, Y.; Wu, B.; Gao, J.; and Xu, Z. 2020. Common-azimuth seismic data fault analysis using residual UNet. *Interpretation*, 8(3).

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.

Luo, W.; Li, Y.; Urtasun, R.; and Zemel, R. 2016. Understanding the effective receptive field in deep convolutional neural networks. *Advances in neural information processing systems*, 29.

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. 2001. DAPSm-Analysis. Accessed September 2021. URL: https://github.com/gpapadog/DAPSm-Analysis.

Papadogeorgou, G. 2016. Power Plant Emissions Data.

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

Rasmussen, C. E. 2003. Gaussian processes in machine learning. In *Summer school on machine learning*, 63–71. Springer.

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.

Stuart, E. A. 2010. Matching methods for causal inference: A review and a look forward. *Statistical science: a review journal of the Institute of Mathematical Statistics*, 25(1).

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

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.

Vansteelandt, S. 2007. On confounding, prediction and efficiency in the analysis of longitudinal and cross-sectional clustered data. *Scandinavian journal of statistics*, 34(3): 478–498.

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.

Appendix

A Inverse Probability of Treatment Weighting (IPTW)

IPTW is one of the most standard causal inference techniques (Rubin 2005; Cole and Hernán 2008). This estimator emulates a pseudo-population in which confounders are equally distributed between the treated and untreated units. Instead of actually constructing this pseudo-population, it estimates the ATE using weights that are inversely proportional to some propensity score estimate $\hat{\mu}_s = \hat{p}(A_s = 1 \mid \boldsymbol{L}_s)$ using the formula

$$\begin{aligned}\hat{\tau}_{\text{IPTW}} =& |\mathbb{S}|^{-1} \sum_{s \in \mathbb{S}} \{(Y_s/\hat{\mu}_s)\mathbb{I}(A_s = 1) \\ & - (Y_s/(1-\hat{\mu}_s))\mathbb{I}(A_s = 0)\}.\end{aligned}$$

When $\mathcal{L}_s$ is sufficient and the propensity scores are known (instead of estimated), this formula yields unbiased causal estimates of the ATE.

B Proofs

Proof of Proposition 1. For convenience, drop the subscript s and boldface notations, and denote $L^c = L' \setminus L$. We will use a graphical argument based on the backdoor criterion (Pearl 1988, ch. 4.3). Suppose that $L^c \to A$ (here $\to$ means causation) and observe the two following facts: first, a path $Y \to L^c \to A$ would violate the assumption of pre-treatment covariates; second, a path $Y \leftarrow L^c \to A$ would need to be absent or be blocked by L due to sufficiency. If blocked, it must be of the form $Y \leftarrow L \leftarrow L^c \leftarrow A$ since a reversed first arrow would violate the pre-treatment assumption, implying $L^c \perp\!\!\!\perp Y \mid L$ (and as a consequence, conditionally independent of $Y(0), Y(1)$). An analogous argument shows that assuming $L^c \to Y$ would imply L^c is conditionally independent from A given L. In summary, conditioning on L^c does not open any new (backdoor) paths from A to Y. And the result follows from the backdoor criterion. □

Proof of Proposition 2. This is a standard result in introductory expositions of potential outcomes. For each $a \in \{0, 1\}$ we have that

$$\begin{aligned}\mathbb{E}[\mathbb{E}[Y_s \mid \boldsymbol{L}_s, A_s = a]] &= \mathbb{E}[\mathbb{E}[Y_s(a) \mid \boldsymbol{L}_s, A_s = a]] \\ &= \mathbb{E}[\mathbb{E}[Y_s(a) \mid \boldsymbol{L}_s]] \\ &= \mathbb{E}[Y_s(a)].\end{aligned}$$

The first equality follows from SUTVA; the second from sufficiency; the third from the law of iterated expectation. Finally, $\mathbb{E}(Y_s(a)) = |\mathbb{S}|^{-1} \sum_s Y_s(a)$ by definition, implying the proposition's statement. □

Proof of Proposition 3. We follow (Hansen 2008)'s formulation of the prognostic score, and prove the results along the lines of (Rosenbaum and Rubin 1983, theorems 1-3). We'll proceed in three steps. All which are somewhat informative of the role of the prognostic score. Again, we drop the subscript s and boldface from the notation for clarity.

Step 1. Conditional expectation of the outcome is a prognostic score. Denote $\psi(L) = \mathbb{E}[Y(0) \mid L]$. We want to show the balancing property: $Y(0) \perp\!\!\!\perp L \mid \psi(L)$.

Recall the definition of conditional expectation (see (Williams 1991, ch. 9.2)): $Z = \mathbb{E}[Y \mid L]$ iff $\mathbb{E}[Y\mathbb{I}(L \in A)] = \mathbb{E}[Z\mathbb{I}(L \in D)]$ for any L-measurable set D. We will use this definition and show that $p(Y(0) \in C \mid L) = p(Y(0) \in C \mid \psi(L))$, implying the required independence. (Conditioning on $(L, \psi(L))$ is equivalent to only condition on L.)

Now, since $\psi(L)$ is a function of L, the event $\psi(L) \in D$ can be re-written as $L \in \psi^{-1}(D)$ using the pre-image notation. Then,

$$\begin{aligned}&\mathbb{E}[\mathbb{I}(Y(0) \in C)\mathbb{I}(\psi(L) \in D)] \\ &\quad = \mathbb{E}[\mathbb{I}(Y(0) \in C)\mathbb{I}(L \in \psi^{-1}(D))] \\ &\quad = \mathbb{E}[\mathbb{E}[(Y(0) \in C) \mid L]\mathbb{I}(L \in \psi^{-1}(D))],\end{aligned}$$

implying that $\mathbb{E}[\mathbb{I}(Y(0) \in C) \mid L] = \mathbb{E}[\mathbb{I}(Y(0) \in C) \mid \psi(L)]$. The result then follows from noting that probabilities are expectations of indicator functions.

Step 2. Any other prognostic score $b(L)$ is finer than $\psi(L)$. Suppose it is not the case, then there are ℓ_1, ℓ_2 such that $\psi(\ell_1) \neq \psi(\ell_2)$ but $b(\ell_1) = b(\ell_2)$. But by the balancing property we have that

$$\begin{aligned}&\mathbb{E}[Y(0) \mid b(L) = b(\ell_1)] \\ &\quad = \mathbb{E}[Y(0) \mid b(L) = b(\ell_1), L = \ell_1] \\ &\quad = \psi(\ell_1),\end{aligned}$$

which would imply that $\mathbb{E}[Y(0) \mid b(L) = b(\ell_1)] \neq \mathbb{E}[Y(0) \mid b(L) = b(\ell_2)]$, violating the assumption that $b(\ell_1) = b(\ell_2)$ and leading to a contradiction. Thus $\psi(\ell_1) = \psi(\ell_2)$ implies that $b(\ell_1) = b(\ell_2)$, which in turn implies the existence of some function $\psi(L) = f(b(L))$ and thus $\psi(L)$ is coarser.

Step 3. If b is a prognostic score, then L is sufficient iff $b(L)$ is also sufficient. First, if $b(L)$ is sufficient, then the proof is trivial. So let's consider the opposite case. First we show that $p(Y(0) \in C \mid A, b(L)) = p(Y(0) \in C \mid b(L))$.

The proof follows from the following identities

$$\begin{aligned}&p(Y(0) \in C \mid A, b(L)) \\ &\quad = \mathbb{E}[\mathbb{I}(Y(0) \in C) \mid b(L)] \\ &\quad = \mathbb{E}[\mathbb{E}[\mathbb{I}(Y(0) \in C) \mid L] \mid A, b(L)] \\ &\quad = \mathbb{E}[\mathbb{E}[\mathbb{I}(Y(0) \in C) \mid \psi(L)] \mid A, b(L)] \\ &\quad = \mathbb{E}[\mathbb{I}(Y(0) \in C) \mid \psi(L)] \\ &\quad = p(Y(0) \in C \mid b(L)).\end{aligned}$$

The first equality is by definition, the second by iterated expectation, the third one by the sufficiency of L; the fourth one is because $\psi(L)$ is balancing (Step 1); the fifth one is because $\psi(L)$ is a function of $b(L)$ by Step 2; the last one is by definition.

Finally, for the treated outcome $Y(1)$, the assumption of no effect modification means that the same argument carries on for $Y(1)$ (since $Y(1) - Y(0)$ is independent of A). □

Proof of Proposition 4. Let $a \in \{0, 1\}$. By the assumption of conditional independence of the treatments given $\boldsymbol{L}_s$ (assumption (2) in the proposition), we have that

$$\mathbb{E}[Y_s \mid \boldsymbol{L}_s, A_s = a]] = \mathbb{E}[Y_s \mid \boldsymbol{L}_s, A_s = a, \mathbf{A}_{\mathcal{N}_s \setminus \{s\}}]$$

Having noted this, the proof is identical to that of Proposition 2

$$\begin{aligned}
&\mathbb{E}[\mathbb{E}[Y_s \mid \boldsymbol{L}_s, A_s = a, \mathbf{A}_{\mathcal{N}_s \setminus \{s\}}]] \\
&\quad = \mathbb{E}[\mathbb{E}[Y_s(a_s = a, \mathbf{A}_{\mathcal{N}_s \setminus \{s\}}) \mid \boldsymbol{L}_s, A_s = a, \mathbf{A}_{\mathcal{N}_s \setminus \{s\}}]] \\
&\quad = \mathbb{E}[\mathbb{E}[Y_s(a_s = a, \mathbf{A}_{\mathcal{N}_s \setminus \{s\}}) \mid \boldsymbol{L}_s]] \\
&\quad = \mathbb{E}[Y_s(a_s = a, \mathbf{A}_{\mathcal{N}_s \setminus \{s\}})]
\end{aligned}$$

where the first identity is due to SUTNVA; the second one is by neighborhood-level sufficiency (assumption (1) in the proposition); and the third one is by the law of iterated expectation. Finally, $E[Y_s(a_s = a, \mathbf{A}_{\mathcal{N}_s \setminus \{s\}})] = (1/|\mathbb{S}|) \sum_s Y_s(a_s = a, \mathbf{A}_{\mathcal{N}_s \setminus \{s\}})$ since the randomness in the expectation is due to s uniformly from $\mathbb{S}$. □

C Motivating example for the self-supervised model: perfect encoding

Assume that the covariates $\boldsymbol{X}_s$ have dimension $d = 1$ and that the self-supervision task is to learn the adjacent values in the grid (north, west, south, east) and the central point $s = (i, j)$ using the representation $\boldsymbol{Z}_{\theta,s}$. If we set the representation dimension to $k = 5$, then the obvious candidate for the representation is

$$\boldsymbol{Z}_{(i,j)} = (\boldsymbol{X}_{(i-1,j)}, \boldsymbol{X}_{(i,j-1)}, \boldsymbol{X}_{(i+1,j)}, \boldsymbol{X}_{(i,j+1)}, \boldsymbol{X}_{(i,j)})^\top$$

Now let $\boldsymbol{\gamma}(\ell) = (\boldsymbol{\gamma}(\ell)_1, \ldots, \boldsymbol{\gamma}(\ell)_5)$ be the ℓ-th indicator vector with $\boldsymbol{\gamma}(\ell)_j = \mathbb{I}(j = \ell)$. Then

$$\boldsymbol{\gamma}(1)^\top \boldsymbol{Z}_{(i,j)} = \boldsymbol{X}_{(i-1,j)}, \cdots \qquad \boldsymbol{\gamma}(5)^\top \boldsymbol{Z}_{(i,j)} = \boldsymbol{X}_{(i,j)}$$

Hence $\boldsymbol{Z}$ is a perfect encoding and the $\boldsymbol{\gamma}(\ell)$'s are perfect classifiers for each offset ℓ. To generalize this idea to higher dimensions $d > 1$, we can take $\boldsymbol{\gamma}(\ell)$ to be a $d \times k$ matrix for each ℓ. Then $\boldsymbol{\gamma}(\ell)^\top \boldsymbol{Z}_{\theta,s}$ is a d-dimensional vector for each offset ℓ. The same idea is behind the self-supervised model, which takes $\boldsymbol{\Gamma} = \gamma^\top$ as a $k \times k$ matrix and adds a decoder neural network. Rather than using indicator functions for Γ, the method formulates it as a neural network that is a function of the offset.

D A connection between the self-supervised model and PCA

Principal components analysis (PCA) is closely related to a special case of the self-supervised *weather2vec* when using a single $(2R+1)\times(2R+1)$-convolution instead of the U-net, leaving h_ψ as the identity function, and defining the offset embedding $\boldsymbol{\Gamma}(\delta)$ as independent $d \times k$ vectors for each offset δ (rather than a neural network with δ as a continuous input). The equivalence is in the sense of reconstruction since both methods can be seen as minimizing the reconstruction error. However, in the self-supervised case there is no guarantee that the latent dimensions of $\boldsymbol{Z}_{\theta,s}$ will be orthogonal as in PCA applied to each patch of size $(2R + 1) \times (2R + 1)$.

E Software and Hardware

We use open-source software PyTorch 1.10 (Paszke et al. 2019) on Python 3.9 (Van Rossum and Drake Jr 1995) for training all the models on a single laptop with an Nvidia GPU 980M (8GB) and a CPU Intel i7-4720HQ at 2.60GHz. The code uses fairly standard functions for NN training and we did not attempt to optimize it for speed. We also use R 3.6 (R Core Team 2021) for downloading and pre-processing atmospheric data from NARR, as well as for comparison with DAPSm in application 1 (see Appendix G).

The code for Bayesian inference in Application 2 is implemented in pure Python as a straightforward Gibbs sampler since the model is Gaussian.

F Details of the simulation study

The simulated data mimics the meteorological data in our applications and the matches setup of Section 2 with SUTVA and NLC.

Data simulation and basic linear task. The covariates $\boldsymbol{X}_s$ are the gradient field (the first differences along rows and columns) of an unobserved Gaussian Process (Rasmussen 2003) defined over a 128×256 grid. To fix ideas, the simulation is carried out to roughly mimic a study of pollution sources where pollution is dispersed in accordance with non-local weather covariates, so, $\boldsymbol{X}_s = (\boldsymbol{X}_s^1, \boldsymbol{X}_s^2)$ can be roughly interpreted as "wind vectors". The treatment assignment probability (the propensity score) and the outcome are computed as a "non-local" function of $\boldsymbol{X}_s$, simulated to correspond to higher probability of treatment in areas that tend to disperse more pollution. Such an assignment can be performed using a convolution operation. More precisely, let $\mu = \sum_{j \in \{1,2\}} K_j \star \mathbf{X}^j$ be the result of convolving $\mathbf{X}$ with a specially designed convolution kernel K. Then, the treatment assignment probability is $A_s \sim \text{Bernoulli}(\mu_s)$ and the outcome is $Y_s = -\mu_s + \epsilon_s + \tau A_s$, where τ is the treatment effect and ϵ_s is a mixture of spatial and random noise of unit variance. $\tau = 0.1$ in all experiments. K has dimensions 13×13 (its size determines the radius of NLC). K_1 contains -1's in the upper half, +1's in the lower half, and 0's in the middle row. $K_2 = K_1^\top$. Convolving a a gradient field with K is an approximate form of identifying valleys and hills in the potential of the gradient field. In this basic formulation, $\mathbb{S} = \mathbb{G}$, meaning that A_s and Y_s are densely available over the grid $\mathbb{G}$.

Additional task variants. In one variant, we consider a sparse configuration in which the outcome and treatment are only sparsely available in a subset $\mathbb{S}$ of 500 randomly selected points in $\mathbb{G}$. In another variant, we evaluate the results on a *non-linear* version of the treatment assignment logits, computed as $\mu = \sum_{j \in \{1,2\}} K_j \star \text{sign}(\mathbf{X}^j)$. This small amount of non-linearity strongly increases the complexity of the problem. Both tasks variants are also combined, resulting in 4 total tasks.

Figure 5 illustrates the data used in the simulation study. (a) shows the simulated covariates $\mathbf{X} = (\mathbf{X}^1, \mathbf{X}^2)$ as the gradient vector field of an unobserved potential function (sampled from a Gaussian Process) $\mathbf{F}$, whose level curves overlay the covariates (vector field), represented by arrows. (b) and (c) jointly compose the kernel used to generate the confounding factor. (d) is the resulting treatment assignment probability for the linear task, and (e) is the corresponding

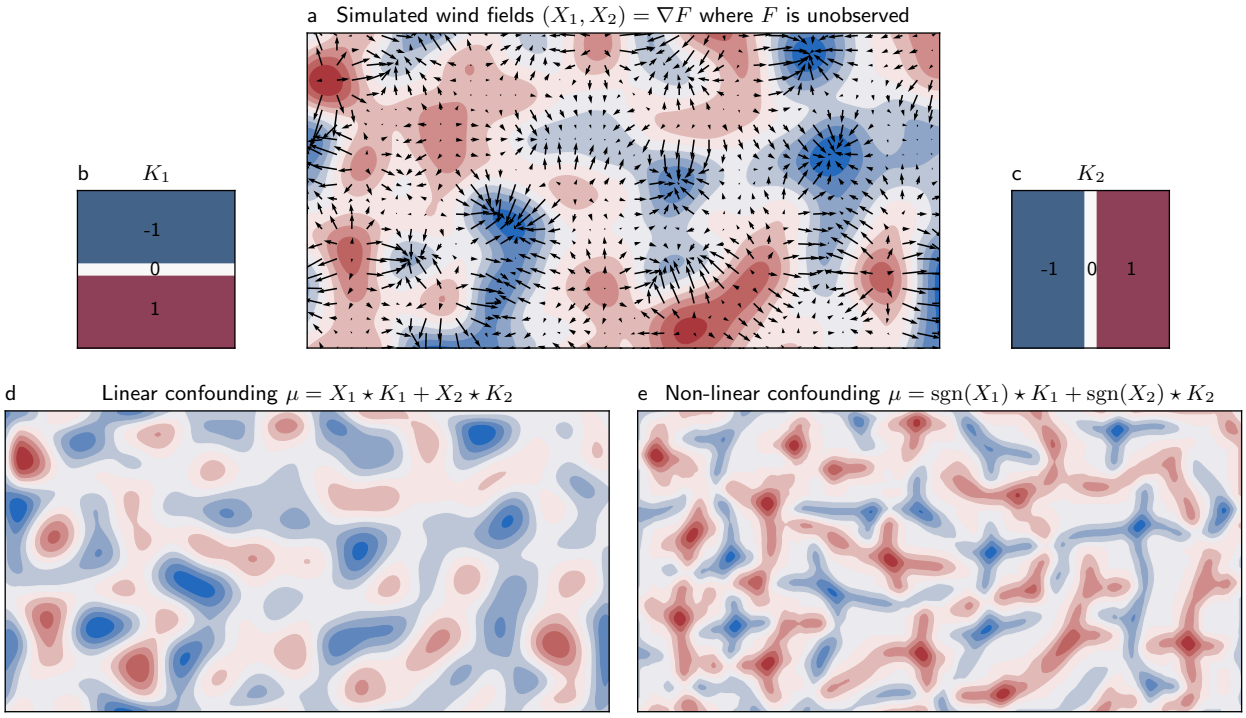


Figure 5: Simulations, components and variants in the simulation study

non-linear variant. In both, convolutions approximately correspond to valleys and hills of the unobserved potential.

Causal estimation procedure. We first estimate a propensity score model $\hat{\mu}_s$ using the learned $\boldsymbol{Z}_{\theta,s}$. For the supervised variant, $\hat{\mu}_s = \text{sigmoid}(\boldsymbol{Z}_{\theta,s})$. But for the self-supervised, it is constructed from an feed-forward network with one or two hidden layers. In all cases, the final estimate of τ is produced using IPTW. Although more sophisticated causal estimators could be used, the exercise only intends to measure the degree to which a propensity score anchored to $\boldsymbol{Z}_{\theta,s}$ encodes the necessary NLC information.

Details on the neural network architectures. The NNs used in the study are very lightweight, since the data consists of only one image of $128x256$. Typical NN sizes with millions of parameters would easily overfit to this task. The basic U-net architecture used for *weather2vec* is in Figure 6, but using the simulated gradient fields instead of atmospheric covariates. All convolutions and linear layers are followed by batch normalization and SiLU activations (Elfwing, Uchibe, and Doya 2018), except in the last layer.

- **Supervised *weather2vec***. The propensity score model uses two hidden units and depth 2. The model has 1.2k (trainable) parameters.
- **Self-supervised *weather2vec***. The auto-encoder uses 16 hidden units and depth two. The offset model $\boldsymbol{\Gamma}_\phi$ is a two-layer feed-forward network with 16 hidden units. The decoder h_ψ is feed-forward network with one hidden layer of also 16 units. In total, the auto-encoder has 77k parameters. In addition, the propensity score model uses a feed-forward network with two hidden layers of 16 units, resulting in 600 parameters.
- **Local and Local+Averages.** These baselines use the same propensity score model as the self-supervised one. Due to their smaller input size, they have around 400 parameters.
- **Spatial RE**. Rather than a neural network, we used a conditionally auto-regressive (CAR) (Besag 1974) model such that $A_s \sim \text{Bernoulli}(\text{sigmoid}(\boldsymbol{Z}_{\theta,s}))$, $\boldsymbol{Z}_{\theta,s} \sim \text{CAR}(\lambda)$ and $\lambda \sim \text{Gamma}(1, 1)$. The CAR portion of the negative loglikelihood penalizes the (squared) differences of adjacent values of $\boldsymbol{Z}_{\theta,s}$ in the grid by a factor of λ. Notice that λ here is learned along with the model. We remark that CAR models are more scalable alternatives to Gaussian process for applications requiring only smoothing and interpolation.
- **Supervised *weather2vec* + spatial RE**. This variant formulates the representation as $\boldsymbol{Z}_{\theta,s} = \tilde{Z}_{\theta,s} + \xi_s$, where $\tilde{\boldsymbol{Z}}_{\theta,s}$ is the output of the U-net and ξ_s has a CAR prior and is restricted to $\sum_s \xi_s = 0$ for identifiability. Intuitively, the term ξ_s captures the errors in the propensity score model that have a strong spatial distribution.

Details on the training procedures and hyper-parameters. In all cases, we use a fixed learning rate of 10^{-4}, a weight decay of 10^{-4}, and 20,000 gradient steps with the ADAM optimizer (Kingma and Ba 2014a). The full simulation study takes about 8 hours to finish running two baselines in parallel. The values of weight decay, training epochs and learning

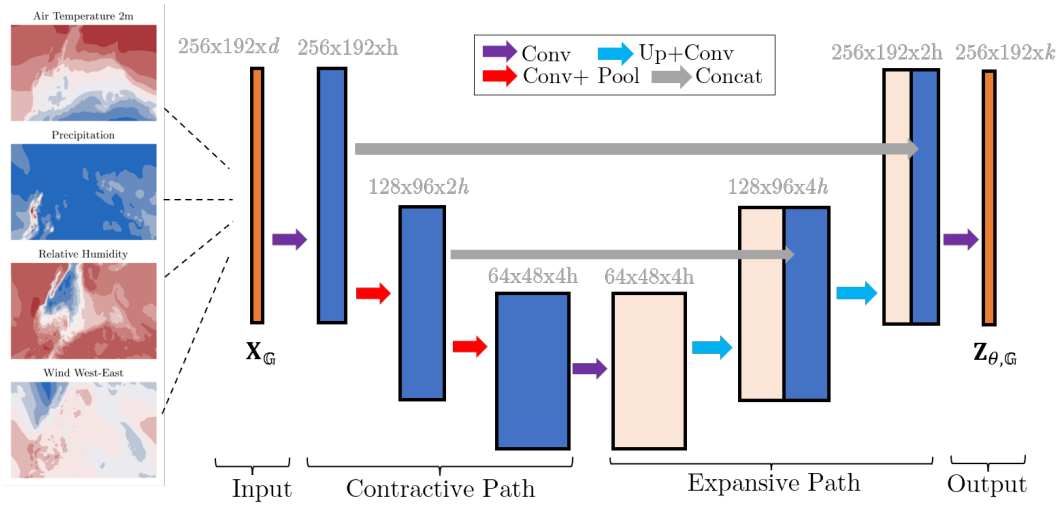


Figure 6: Basic U-net architecture used in the two applications and the simulation study.

rate were chosen as reasonable values without much additional optimization. The number of layers and architectures were chosen by inspection after a few runs, aiming to find a model small enough as to avoid over-fitting without requiring tuning the regularization hyper-parameters or early stopping.

G Additional details of Application 1

Atmospheric data download. The NARR (Mesinger et al. 2006) dataset associated with the application can be downloaded via FTP with the R script provided in the code accompanying the paper. The data is also publicly available for download from the website of the National Oceanic and Atmospheric Administration (NOOA) (National Oceanic and Atmospheric Administration). We could not find any license information for the dataset. We could not find any license attached to the dataset.

Power plants data. Information for the largest 473 coal-fired power plants emitting SO_2 during 2000–2014 was obtained from (Papadogeorgou 2016) (publicly available under creative commons license CC0 1.0).

Neural network architecture for self-supervised features. The auto-encoder uses 32 hidden units and depth 3. The offset model $\mathbf{\Gamma}_\phi$ is a two-layer feed-forward network with 32 hidden units. The decoder h_ψ is feed-forward network with one hidden layer of also 32 units. In total, the auto-encoder has 1.2M parameters. The architecture is shown in Figure 6. Convolutions are followed by FRN normalization layers (Singh and Krishnan 2020) and SiLU activations. Pooling uses *MaxPool2d* and upsampling use *Bilinear Upsampling2d* as implemented in PyTorch. The model architecture was not tuned since the model with 32 hidden units seemed to work well.

Details on the training procedures and hyper-parameters. The model is trained for 300 epochs using batch size 4, a linear decay learning rate from 10^{-2} to 10^{-4} using the ADAM (Kingma and Ba 2014b) optimizer (no weight decay). The number of epochs and learning rate were tuned by inspection after a few runs simply to ensure the model was learning at a reasonable speed, but not tuned otherwise. The atmospheric covariates were standardized before training.

We do not split in training and validating datasets since the model is a compression/dimensionality reduction technique, and thus it cannot over-fit. (In fact, an "over-fitting" here would be a desirable property, since it would mean a perfect dimensionality reduction.)

Computation of the explained variance (R^2). The traditional R^2 is defined as one minus the ratio of sum-of-squares between the prediction errors and the centered targets. Since the covariates are standardized, the latter quantity is simply N. In each training epoch we collect the sum of squared prediction errors for all time periods. Denote this quantity as SSE_j where j indicates the covariate dimension for $j = 1, ..., d$. Then $R_2 = 1 - (Nd)^{-1} \sum_j SSE_j$ is the proposed estimator of the fraction of the variance explained.

Comparison with DAPSm. We modified the DAPSm authors implementation from Github (Papadogeorgou 2001) (no license provided) to include the *weather2vec* self-supervised features as another predictor in their otherwise unchanged propensity score model. The modified R script is in the code accompanying this paper.

H Additional details of Application 2

Neural network architecture for supervised features. The U-net architecture for the prognostic score model follows the same general architecture as application 1 (figure 6), but with dimension one in the last layer (k). In addition we use depth 2 and $d = 8$ (hidden features multiple) to reduce the number of parameters since we empirically observed overfitting with the same parameters as the first application. However, we do not conduct an explicit hyper-parameter sweep. We used the period 2000-2005 as training data (because emissions are mostly flat, see figure 4b) and choose the model weights that minimize the test mean squared error

using data from 2006.

SO_4 data download. We downloaded the dataset the SO_4 grid for inland US from the website of the Atmospheric Composition Analysis Group's (van Donkelaar et al. 2021) website (Group 2001). We could not find any license information for the dataset. Instructions for replications are provided in the code.

Missing data. Data for some observations in the SO_4 grid are missing and a few have clearly erroneous (near infinite) values. In addition, some locations have data but present zeros through the entire period. We excluded these values using a binary mask in the likelihood by removing non-finite values and keeping only locations with positive observations throughout. Doing so greatly improved the quality of the fitted model. The final locations cover most of the inland U.S., with missing areas mostly outside the U.S, oceans, or the Rocky West and Great Basin.

Details on the training procedures and hyper-parameters. The hyper-parameters are also the same as in the self-supervised model except that we use a weight decay of 10^{-4} to reduce over-fitting. We did not tune this parameter, however, we did not notice a significant difference by increasing or decreasing its value by a factor of 10.

I Visualization of the U-net receptive field

To visualize the ability of the U-net to capture non-local dependencies we conducted an experiment to visualize the receptive fields of the U-net using data from Application 2 where the outcome $\boldsymbol{Y}_t$ is SO_4 for the years 2000–2014 across the mainland U.S. and the predictive variables $\mathbf{X}_t$ are the NARR atmospheric covariates. We fitted a regression model using the U-net for varying depths in $\{1, \ldots, 5\}$. To avoid an explosion of the latent parametters and overfitting, we do not duplicate the number of latent features in the contractive path of the U-net, and instead keep it constant at 32. Finally, we estimate the receptive field $RF(\cdot)$ using standard gradient techniques (Luo et al. 2016) by choosing a target output location s_* over the grid, and compute the average gradient (across time points) of the output Y_{t,s_*} with respect to inputs $\boldsymbol{X}_{t,s}$ from all locations over the grid as

$$RF(s) = \frac{1}{180}\sum_{t=1}^{180}\left\|\frac{\partial Y_{t,s_*}}{\partial \boldsymbol{X}_{t,s}}\right\|_2.$$

Fig. 7 show the results. The orange color indicates a higher gradient. The target outcome location s^* is shown as a blue dot. We can see that as the depth increases, so does the receptive field around the point. However, after depth 3, increasing to depth 4 maintains a higher focus on the same areas west to the target location. This observation suggests that the U-net learned the effective sources of non-local correlation with the outcome.

The case when the depth is 5 is also interesting, since it shows a potential problem when setting the depth too high. Some areas of spurious correlation appear (for example in the lower border of the image). This phenomenon could be due to padding effects since, after 5 contractive layers, the dimension 128×256 is reduced to 4×8. The previous observations are confirmed in Fig. 8 by looking at the out-of-sample performance by masking 10% of the training data when fitting the model. The performance metrics suggest that in this case, the optimal depth would be 4; although as we previously observed, increasing from depth 3 to depth 4 has a smaller effect.

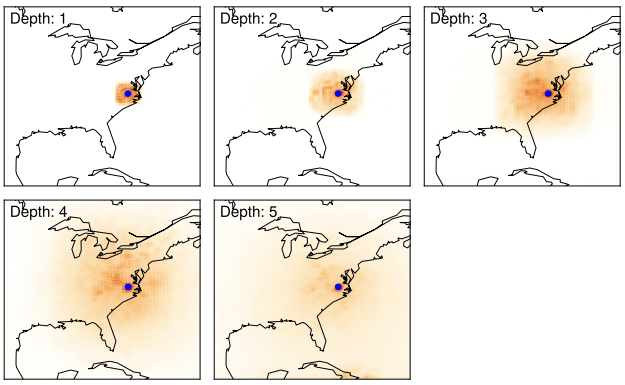


Figure 7: Receptive field by varying depths of the U-net.

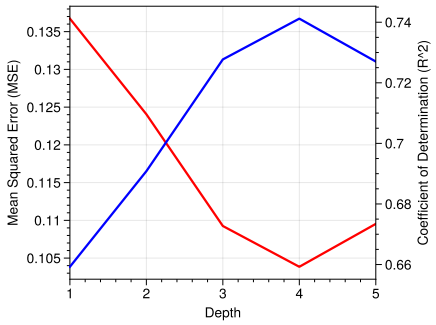


Figure 8: Performance in validation set by varying depths.

J Comparison with ResNets

K Algorithms

Algorithms 1 and 2 summarize the two main approaches to learning representation of NLC using the U-net.

Algorithm 1: Supervised representation learning and causal estimation

Input: Covariates $\boldsymbol{X}_{\mathbb{G}}$; outcome and treatment Y_s and A_s at a dense subset $\mathbb{S} \subset \mathbb{G}$; and U-net model f_θ with 1-dimensional outputs.

1: Obtain neural network weights $\hat{\theta}$ minimizing the supervised loss (equations 2 or 3) and compute output balancing score grid from optimal weights $\hat{\mathbf{Z}}_{\mathbb{G}} = f_{\hat{\theta}}(\boldsymbol{X}_{\mathbb{G}})$.
2: Use standard causal inference methods (e.g., IPTW in Appendix A) to obtain $\hat{\tau}_{\text{ATE}}$ adjusting for $\hat{\boldsymbol{Z}}_s$ and other relevant local confounders at each unit s.

Algorithm 2: Self-supervised non-local representation learning for causal estimation

Input: Covariates $\boldsymbol{X}_{\mathbb{G}}$; outcome and treatment Y_s and A_s at any subset $\mathbb{S} \subset \mathbb{G}$; U-net model f_θ with k-dimensional outputs; and grid of candidate radii $\mathcal{R} = \{R_1, \ldots, R_{\text{max}}\}$.

1: For each $R \in \mathcal{R}$, obtain neural network weights $\hat{\theta}^R$ minimizing the self-supervised loss (equations 4) and compute $\hat{\boldsymbol{Z}}_{R,\mathbb{G}} = f_{\hat{\theta}^R}(\boldsymbol{X}_{\mathbb{G}})$.
2: Use a propensity/prognostic score model to choose the optimal r. For example, use a logistic regression taking inputs $\hat{\boldsymbol{Z}}_{R,s}$ (and other relevant local confounders) and compute the Akaike information criterion (AIC). Then choose $\hat{r}$ that minimizes the AIC.
3: Estimate the τ_{ATE} using the learned propensity score model, e.g., using IPTW (Appendix A).

Appendix 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.

Elfwing, S.; Uchibe, E.; and Doya, K. 2018. Sigmoid-weighted linear units for neural network function approximation in reinforcement learning. *Neural Networks*, 107: 3–11.

Group, A. C. A. 2001. Surface PM2.5. Accessed September 2021. URL: https://sites.wustl.edu/acag.

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

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).

National Oceanic and Atmospheric Administration. ???? North American regional reanalysis. Accessed March 2021. URL: https://psl.noaa.gov/data/gridded/data.narr.html.

Paszke, A.; Gross, S.; Massa, F.; Lerer, A.; Bradbury, J.; Chanan, G.; Killeen, T.; Lin, Z.; Gimelshein, N.; Antiga, L.; Desmaison, A.; Kopf, A.; Yang, E.; DeVito, Z.; Raison, M.; Tejani, A.; Chilamkurthy, S.; Steiner, B.; Fang, L.; Bai, J.; and Chintala, S. 2019. PyTorch: An Imperative Style, High-Performance Deep Learning Library. In Wallach, H.; Larochelle, H.; Beygelzimer, A.; d'Alché-Buc, F.; Fox, E.; and Garnett, R., eds., *Advances in Neural Information Processing Systems 32*, 8024–8035.

Pearl, J. 1988. *Probabilistic reasoning in intelligent systems: networks of plausible inference*. Morgan kaufmann.

R Core Team. 2021. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Rosenbaum, P. R.; and Rubin, D. B. 1983. The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1): 41–55.

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.

Van Rossum, G.; and Drake Jr, F. L. 1995. *Python reference manual*. Centrum voor Wiskunde en Informatica Amsterdam.

Williams, D. 1991. *Probability with martingales*. Cambridge university press.