Responses to Reviewers Feedback

We thank the reviewers for the useful feedback. We have significantly improved upon the two main criticisms raised across the reviews by including many new comparison benchmarks (VAE, PCA, CRAE, WX-reg, ResNets) and a more pointed comparison with other methods in air pollution and climate science.

Reviewer #1

* Spatial stationarity:
Thanks for your clarifications! We have made sure to mention this more thoroughly in the revision. We already took some actions to reduce the effect of padding/normalization during our initial experiments (we already did some of this in our experiments). For instance, our NARR dataset covers more than just the mainland US. There is a large area of sea and we use a mask to only predict outcomes inland. Therefore, even though we are using padding, it is less affected. We also normalized each input variable so that zero-paddings amounts to mean imputation. Next, we replaced batch norm with FRN layers (Singh & Krishnan, 2020), which have been preferred over batch norm in tasks like Bayesian inference (Izmailov et al. 2020) for the same reasons mentioned by the reviewer.

* U-net receptive field: we will more clearly state the limitations imposed by the architecture. A new appendix I visualizes receptive fields for varying depths using data from App. 2.

* What does Figure 3 show?: We replaced it with a table showing the bias and MSE of causal estimates for a larger number of benchmark models.

* Is the proposed approach specific to U-Nets?: No. Section 3 already mentioned that ResNets could be an alternative when using a large enough number of residual layers. And we have emphasized this more in the revision. Appendix J in the revision contains a brief discussion and comparison of the simulation task.

* Computation of R^2: Appendix G explains the computation. The R^2 is the squared correlation between the prediction Xhat_{s + delta}(Zs) and Xs' averaged by the magnitude of the displacement delta.

Reviewer #3

* Code not available: Perhaps an error? We did submit the code. See also the anonymized link.

* Experiments are just conducted on the dataset constructed in this paper/Comparison with previous research on air pollution and climate: We did not find a benchmark task in the literature that we could fairly compare with. We believe this work is the first to formalize the problem of NLC, and the literature does not yet provide tools specific to NLC. We precisely chose our set of

experiments because of their quasi-experimental nature, which offers a point of comparison for evaluating causal effects (since our goal is not prediction/modeling). We aimed strongly to evaluate the impact of these case studies attending to the goals of this AAAI AISI Special Track. Both experiments do build on and compare with existing literature. They use different outcomes (O3 and SO4), spatial domains, and control datasets, except for the atmospheric covariates in common. Our construction of the NARR dataset is not arbitrary either. For instance, the study of PM2.5 and climate by Shen, Mickley, and Murray (2017) uses the same set of covariates and year ranges. The revised paper will clarify the connections with existing work more clearly.

* Not reported when the method fails/stress tests: We agree this is important. Our experiments highlighted the sensitivity of the U-net propensity score method to overfitting (sparse setup). The new benchmarks in the linear setting in the simulation now also show when alternative methods are as good or better (such as WX regression). In addition, the purpose of the formal theory of Sec. 2 is to establish the limits to the validity of the method. Like many confounding adjustment methods, it is likely to fail if there is unobserved confounding. The revision emphasizes this point more strongly and also discusses the risks of underspecifying the radius in the self-supervised features.

* Section "Learning NLC representations via supervision are too simple: This section explicitly links the proposal to the causal inference literature adapting seminal work by Rubin. There is a vast literature on propensity score regression, and the purpose of this section is simply to highlight that under NLC, one should use an NLC-ready propensity score statistical model. We will make sure to clarify the context and give examples in this section.

Reviewer #6

* Lack of baselines: The revision added several new benchmarks in Table 1. PCA, CRAE, and VAE require to operate on patches. PCA is surprisingly stronger than convolutional autoencoders. WX-regression comes from the econometric literature (mentioned in related work) and is the oracle model for the linear simulation task. We also added a small comparison with ResNets as an alternative in the appendix.

* The paper lacks information about the total training data, split into training and test: Thank you for the remark. The revision adds that information throughout. We put our attention to the error in estimating the average causal effect (which is an in-sample quantity). Yet we agree this information is important. We apologize for the imprecision in our statement about overfitting not being relevant in dimensionality reduction.

Reviewer #7

* Definitions and assumptions could be illustrated and commented. Thank you for the suggestions. While our presentation of these assumptions closely follows the standards of the causal inference literature of potential outcomes, we have made our best effort in the revision to improve it and connect the assumptions and definitions to examples in our applications.