GST-UNet: Spatiotemporal Causal Inference with Time-Varying Confounders

Miruna Oprescu [1] **David K. Park** [2] **Xihaier Luo** [2] **Shinjae Yoo** [2] **Nathan Kallus** [1]

Abstract

Estimating causal effects from spatiotemporal data is a key challenge in fields such as public health, social policy, and environmental science, where controlled experiments are often infeasible. However, existing causal inference methods relying on observational data face significant limitations: they depend on strong structural assumptions to address spatiotemporal challenges – such as interference, spatial confounding, and temporal carryover effects – or fail to account for *time-varying confounders*. These confounders, influenced by past treatments and outcomes, can themselves shape future treatments and outcomes, creating feedback loops that complicate traditional adjustment strategies. To address these challenges, we introduce the **GST-UNet** (**G**-computation **S**patio-**T**emporal **UNet**), a novel end-to-end neural network framework designed to estimate treatment effects in complex spatial and temporal settings. The GST-UNet leverages regression-based iterative G-computation to explicitly adjust for time-varying confounders, providing valid estimates of potential outcomes and treatment effects. To the best of our knowledge, the GST-UNet is the first neural model to account for complex, non-linear dynamics and time-varying confounders in spatiotemporal interventions. We demonstrate the effectiveness of the GST-UNet through extensive simulation studies and showcase its practical utility with a real-world analysis of the impact of wildfire smoke on respiratory hospitalizations during the 2018 California Camp Fire. Our results highlight the potential of GST-UNet to advance spatiotemporal causal inference across a wide range of policy-driven and scientific applications.

[1]Cornell University, Cornell Tech [2]Computing and Data Sciences, Brookhaven National Laboratory. Correspondence to: Miruna Oprescu <amo78@cornell.edu>.

Under Review.

1. Introduction

Environmental hazards, public health interventions, and large-scale socio-economic policies often require understanding complex cause-and-effect relationships across space and time (Reid et al., 2016b; Papadogeorgou et al., 2019; Song et al., 2020). For instance, when evaluating air quality interventions, policymakers need to assess how industrial regulations affect both immediate local health outcomes and broader regional impacts that evolve over time. These applications demand robust methods for estimating causal effects from observational spatiotemporal data.

However, spatiotemporal causal inference poses unique challenges. First, outcomes at each location are typically influenced by both local and neighboring covariates and interventions, leading to spatial confounding and interference. Second, observations exhibit strong temporal dependencies, with the effects of interventions "carrying over" in time. Finally, *time-varying confounders*–covariates that both influence and are influenced by past treatments and outcomes–create feedback loops that violate standard assumptions of independence over time. Consider, for example, how air quality policies are often implemented: governments may impose stricter regulations in response to poor air quality and high hospitalization rates, which in turn affect future air quality and health outcomes. Moreover, when estimating effects, we aim to understand the impact of different intervention sequences at specific locations given the observed history—a more challenging task than estimating average effects across space or time.

Current approaches to spatiotemporal causal inference either rely on restrictive modeling assumptions that may not capture real-world complexity, or employ flexible neural networks that are limited to single time points, independent time series, or settings without time-varying confounding (see Section 2 for a comprehensive overview). Further complicating matters, spatiotemporal settings often provide only a single realization of the process over time, rather than multiple independent time series.

To bridge this methodological gap, we propose GST-UNet (G-computation Spatiotemporal UNet), a novel end-to-end neural network framework designed to estimate treatment effects in complex spatiotemporal settings. GST-UNet addresses the key challenges of interference, spatial con-

founding, temporal carryover, and time-varying confounding by combining two powerful approaches: a U-Net-based encoder-decoder network for capturing spatial dependencies, and a regression-based, iterative G-computation procedure (Bang and Robins, 2005; Robins and Hernan, 2008) for adjusting for time-varying covariates. This integration enables valid estimation of potential outcomes while flexibly modeling complex dynamics across the spatial grid.

Our contributions are threefold: (1) We introduce the GST-UNet model architecture and provide theoretical foundations and implementation details for its iterative G-computation scheme—to the best of our knowledge, this is the first neural model to account for complex spatial dynamics and time-varying confounders in spatiotemporal interventions; (2) We demonstrate GST-UNet's effectiveness through extensive simulation studies designed to showcase key spatiotemporal complexities and time-varying confounders; and (3) We illustrate its practical utility through a real-world analysis of wildfire smoke impacts on respiratory hospitalizations during the 2018 Camp Fire in California.

2. Related Work

We summarize the most relevant prior work here, with a more detailed discussion in Appendix A.

Classical Spatiotemporal Causal Inference. Earlier spatiotemporal causal inference methods (e.g., spatial econometrics (Anselin, 2013), difference-in-differences (Keele and Titiunik, 2015), and synthetic controls (Ben-Michael et al., 2022)) rely on strong assumptions (e.g., parallel trends) and often fail to address interference or time-varying confounders. More recent classical approaches, on the other hand, typically estimate average effects at the regional level or rely on structural and modeling assumptions that may not hold in real-world spatiotemporal contexts (Wang, 2021; Christiansen et al., 2022; Papadogeorgou et al., 2022; Zhang and Ning, 2023; Zhou et al., 2024).

Machine Learning for Spatiotemporal Modeling. Machine learning models (e.g., convolutional and recurrent networks-based methods (Shi et al., 2015; Zhang et al., 2017), or graph-based approaches (Li et al., 2017; Wu et al., 2019)) capture spatiotemporal patterns for prediction but lack formal causal adjustments.

Time Series Causal Inference. Time-series causal inference often uses recurrent or transformer-based methods (Bica et al., 2020; Seedat et al., 2022; Melnychuk et al., 2022) but assumes independent time series, ignoring potential interference effects. Although iterative G-computation (Bang and Robins, 2005; Robins and Hernan, 2008) or marginal structural models (Robins et al., 2000) can handle time-varying confounders, most ML extensions (Lim, 2018; Li et al., 2021; Hess et al., 2024) exclude interference or cross-series confounding.

Neural-Based Spatiotemporal Causal Inference. In the context of neural spatiotemporal models, Tec et al. (2023) integrate spatial representations for causal inference, accounting for spatial confounding and leveraging temporal data to train a UNet model. However, they do not address feedback effects from lagged or time-varying confounders. Most similar to our work, Ali et al. (2024) present a climate-focused model that shares certain architectural similarities but emphasizes prediction rather than adjusting for time-varying confounders, leaving causal identification concerns largely unaddressed.

3. Problem Formulation

Spatiotemporal Data. We model the observed data as random variables on a discrete spatial domain represented by an $N_X \times N_Y$ lattice: $\mathcal{S} = \{(i,j) \mid i \in [N_X], j \in [N_Y]\}$, where $[N] = \{1, \dots, N\}$ denotes the index set. Time is indexed by $t \in [T]$. At each spatial location $s = (i,j)$ at time t, we observe a tuple $(\mathbf{X}_{s,t}, A_{s,t}, Y_{s,t})$, where $A_{s,t} \in \{0,1\}$ represents a binary treatment (or intervention), $Y_{s,t} \in \mathbb{R}$ is a continuous outcome of interest, and $\mathbf{X}_{s,t} \in \mathbb{R}^{d_X}$ is a vector of time-varying covariates (*e.g.* local weather conditions, pollution levels, or socioeconomic indicators). Additionally, each location s is associated with static features $V_s \in \mathbb{R}^{d_v}$ (*e.g.* geographical characteristics and socioeconomic indicators). While we focus on binary interventions for clarity, the methods generalize to more complex treatments. Conceptually, the data forms a 3D spatiotemporal tensor of size $T \times N_X \times N_Y$, though in practice, observations may be incomplete. Missing data can be accommodated using masking techniques during downstream modeling.

To streamline notation, we use boldface symbols for random variables defined over the entire spatial domain. For $U \in \{X, A, Y\}$, let $\mathbf{U}_t$ denote its value at time t, and let $\mathbf{U}_{t:t+\tau} = (\mathbf{U}_t, \dots, \mathbf{U}_{t+\tau})$ denote its value over a time interval. For a specific location s, we write $U_{s,t:t+\tau} = (U_{s,t}, \dots, U_{s,t+\tau})$. The history up to time t is denoted by $\mathbf{H}_{1:t} = (\mathbf{X}_{1:t}, \mathbf{A}_{1:t-1}, \mathbf{Y}_{1:t}, \mathbf{V})$ for the entire spatial domain and $H_{s,1:t} = (X_{s,1:t}, A_{s,1:t-1}, Y_{s,1:t})$ for a specific location s. Specific instantiations of these random variables are denoted using lowercase letters (e.g., $u \in \{x, a, y, h\}$).

Quantities of Interest. Our primary goal is to estimate location-specific Conditional Average Potential Outcomes (CAPOs) for a sequence of future spatiotemporal interventions, conditioned on observed history. Our approach builds on Rubin's potential outcomes framework (Rubin, 1978; Robins and Hernan, 2008; Robins et al., 2000), which we extend to accommodate spatiotemporal settings. More concretely, we consider a future time horizon of length $\tau \geq 1$ and a predetermined interventional sequence $\mathbf{a}_{t:t+\tau-1}$ ap-

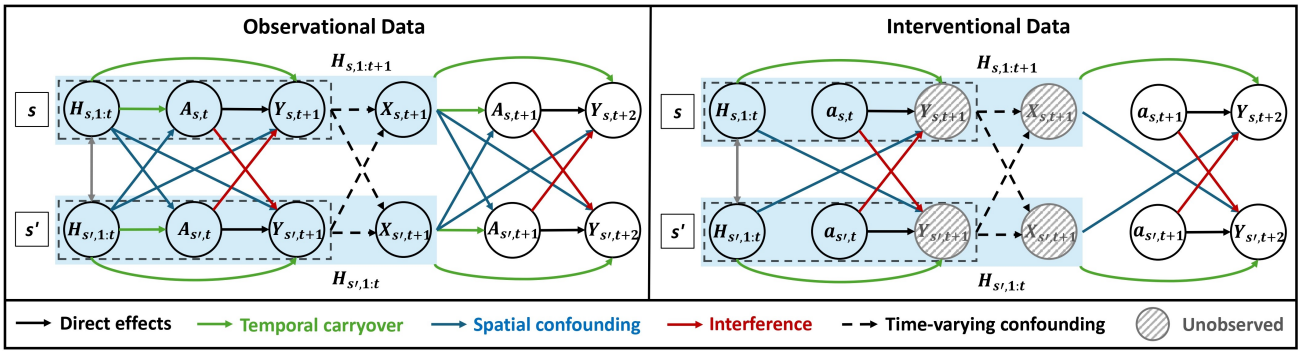

Figure 1. Observational data (left) versus interventional data (right) for a horizon $\tau = 2$ across multiple locations (s, s'). Green arrows indicate temporal carryover, blue arrows show spatial confounding, and red arrows depict interference; dashed arrows denote time-varying confounding, and dashed circles represent unobserved variables at inference time. Under the intervention (right), treatments are set independently of confounders, and the full history is not observed for the entire horizon.

plied across the spatial domain starting at time t. Our goal is to estimate the potential outcomes at time $t+\tau$, denoted as $\mathbf{Y}_{t+\tau}[\mathbf{a}_{t:t+\tau-1}]$. In particular, we aim to compute:

$$\mathbb{E}[\mathbf{Y}_{t+\tau}[\mathbf{a}_{t:t+\tau-1}] \mid \mathbf{H}_{1:t} = \mathbf{h}_{1:t}] \tag{1}$$

which represents the CAPOs at time $t+\tau+1$ under the given treatment sequence. Given two different interventional sequences $\mathbf{a}_{t:t+\tau}$ and $\mathbf{a}'_{t:t+\tau}$, a related secondary goal is to estimate the location specific Conditional Average Treatment Effect (CATE), given by:

$$\mathbb{E}[\mathbf{Y}_{t+\tau}[\mathbf{a}_{t:t+\tau-1}] - \mathbf{Y}_{t+\tau}[\mathbf{a}'_{t:t+\tau-1}] \mid \mathbf{H}_{1:t} = \mathbf{h}_{1:t}]$$

Although we focus primarily on CAPOs, CATEs and other effect measures can be derived similarly.

Assumptions. Identification of CAPOs from observational data relies on standard causal inference assumptions. Additionally, our setup relies on observing only *a single spatiotemporal outcome path*. This prevents direct estimation of the CAPOs in Eq. (1) – which relies on an expectation over multiple data samples – without additional assumptions. We therefore introduce the following assumptions:

Assumption 3.1 (Standard Causal Inference Assumptions)**.** We assume the following properties hold: *(Consistency)* $\mathbf{Y}_{t+\tau} = \mathbf{Y}_{t+\tau}[\mathbf{a}_{t:t+\tau-1}]$ whenever the observed sequence of treatments $\mathbf{A}_{t:t+\tau-1}$ satisfies $\mathbf{A}_{t:t+\tau-1} = \mathbf{a}_{t:t+\tau-1}$; (*Positivity*) $P(A_{s,t} = a_{s,t} \mid \mathbf{H}_{1:t} = \mathbf{h}_{1:t}) > 0$ for any $a_{s,t} \in \{0, 1\}$ and feasible realization of history $\mathbf{h}_{1:t}$; (*Sequential Unconfoundedness*) $\mathbf{Y}_{t+1:T}[\mathbf{a}_{t+1:T}] \perp \mathbf{A}_t \mid \mathbf{H}_{1:t}$, $\forall \mathbf{a}_{t+1:T} \in \{0, 1\}^{T-t}$, *i.e.* at each time step t, the treatment assignment is independent of future potential outcomes.

Assumption 3.2 (Representation-Based Time Invariance)**.** There exists a function (or embedding) $\phi : \mathcal{H} \times \mathcal{A} \to \mathcal{Z} \subseteq \mathbb{R}^h$ that maps $(\mathbf{H}_{1:t}, \mathbf{A}_t)$ to a finite-dimensional representation such that once we condition on $z = \phi(\mathbf{H}_{1:t}, \mathbf{A}_t)$, the distribution $(\mathbf{X}_{t+1}, \mathbf{Y}_{t+1})$ does not explicitly depend on t. Formally, for any $t, t' \in \{1, \ldots, T\}$ and $z \in \mathcal{Z}$, we have:

$$\begin{aligned} &p(\mathbf{X}_{t+1}, \mathbf{Y}_{t+1} \mid \phi(\mathbf{H}_{1:t}, \mathbf{A}_t) = z) \\ &\quad = p(\mathbf{X}_{t'+1}, \mathbf{Y}_{t'+1} \mid \phi(\mathbf{H}_{1:t'}, \mathbf{A}_{t'}) = z). \end{aligned}$$

Assumption 3.1 is standard in longitudinal causal inference settings (e.g., (Robins et al., 2000; Robins and Hernan, 2008; Bica et al., 2020; Li et al., 2021; Melnychuk et al., 2022; Hess et al., 2024)). Assumption 3.2 is specific to the single-time series setting, where pooling information across time is essential to enable estimation. We note that the single time-series setting frequently arises in causal inference, where assumptions such as stationarity or strict time homogeneity enable consistent estimation (Bojinov and Shephard, 2019; Papadogeorgou et al., 2022; Zhou et al., 2024). In contrast, our representation-based time invariance is *weaker*: rather than requiring $\mathbf{X}_t, \mathbf{Y}_t$ themselves to have a time-invariant distribution, we only assume that, once the history is summarized by $\phi(\mathbf{H}_{1:t}, \mathbf{A}_t)$, the transition to $(\mathbf{X}_{t+1}, \mathbf{Y}_{t+1})$ follows a single shared mechanism. This approach aligns with modern time-series causal inference that learn time-invariant latent embeddings to pool information across time steps (Lim, 2018; Li et al., 2021; Hess et al., 2024), thus leveraging more data for a single, stable representation rather than time-dependent parameters.

Identification and G-Computation. We splice our single time series into multiple "prefixes" of varying lengths such that for each $t \in \{1, \ldots, T - \tau\}$, we define the following

$$\mathbf{P}_t^\tau \;=\; \big(\mathbf{X}_{1:t+\tau}, \mathbf{A}_{1:t+\tau}, \mathbf{Y}_{1:t+\tau}, \mathbf{V}\big).$$

Under time-invariance (Assumption 3.2), conditioning on a suitable embedding of $\mathbf{P}_t^\tau$ renders the distribution of $\mathbf{Y}_{t+\tau}$ independent of t. We will thus write expectations over the prefixes given history embeddings as

$$\mathbb{E}_{\mathbf{P}}\big[\mathbf{Y}_{t+\tau} \;\big|\; \phi(\mathbf{H}_{1:t}, \mathbf{A}_t)\big],$$

where t here only identifies the location in the series (*i.e.* the segment's ending point) rather than implying a distinct distribution. Pooling across t then supplies $T-\tau$ conditionally independent segments from a *single* time series, enabling regression-based methods to estimate future outcomes from (embedded) histories.

Given $\mathbf{P}_t^\tau$, we show how to identify CAPOs from observational data. For $\tau \geq 2$, *future* covariates and outcomes (*i.e.* $\mathbf{X}_{t+1:t+\tau-1}, \mathbf{Y}_{t+1:t+\tau-1}$) may influence subsequent treatments, creating *time-varying confounding* (Coston et al., 2020). These confounders, shaped by past treatments and outcomes, can alter *future* assignments and outcomes, forming feedback loops that simple "condition-on-history" adjustments miss. Hence, adjustment—*e.g.* via iterative G-computation—is needed to avoid bias. Figure 1 illustrates these complexities by comparing observational data (left) and hypothetical interventional data (right) for $\tau = 2$. By contrast, when $\tau = 1$, conditioning on $\mathbf{H}_{1:t}$ is sufficient under standard assumptions, as no future confounders lie between $\mathbf{A}_t$ and $\mathbf{Y}_{t+1}$. In other words, the following naive identification does *not* generally hold for $\tau > 1$:

$$\mathbb{E}[\mathbf{Y}_{t+\tau}[\mathbf{a}_{t:t+\tau-1}] \mid \mathbf{H}_{1:t} = \mathbf{h}_{1:t}] \tag{2}$$
$$\neq \mathbb{E}[\mathbf{Y}_{t+\tau} \mid \mathbf{H}_{1:t} = \mathbf{h}_{1:t}, \mathbf{A}_{t:t+\tau-1} = \mathbf{a}_{t:t+\tau-1}] \tag{3}$$

To address time-varying confounding, we adapt *regression-based iterative G-computation* (Bang and Robins, 2005; Robins and Hernan, 2008) to the spatiotemporal setting, providing a systematic way to adjust for evolving confounders and achieve unbiased CAPO estimates. We integrate our single time-series spatiotemporal setting with the G-computation framework in the following result:

Theorem 3.3 (Identification with G-Computation)**.** *Assume that Assumption 3.1 and Assumption 3.2 hold. Further, let* $\mathbf{H}^{\mathbf{a}}_{1:t+k} := (\mathbf{X}_{1:t+k}, [\mathbf{A}_{1:t-1}, \mathbf{a}_{t:t+k-1}], \mathbf{Y}_{1:t+k})$ *denote the history where observed treatments from time* t *onward are replaced by* $\mathbf{a}_{t:t+k-1}$*. Define recursively:*

$$\begin{aligned}
&Q_\tau(\mathbf{H}_{1:t+\tau-1}, \mathbf{A}_{t+\tau-1}) \\
&\quad = \mathbb{E}_{\mathbf{P}}[\mathbf{Y}_{t+\tau} \mid \phi(\mathbf{H}_{1:t+\tau-1}, \mathbf{A}_{t+\tau-1})] \\
&Q_{\tau-1}(\mathbf{H}_{1:t+\tau-2}, \mathbf{A}_{t+\tau-2}) \\
&\quad = \mathbb{E}_{\mathbf{P}}[Q_\tau(\mathbf{H}^{\mathbf{a}}_{1:t+\tau-1}, \mathbf{a}_{t+\tau-1}) \mid \phi(\mathbf{H}_{1:t+\tau-2}, \mathbf{A}_{t+\tau-2})] \\
&\dots \\
&Q_1(\mathbf{H}_{1:t}, \mathbf{A}_t) \\
&\quad = \mathbb{E}_{\mathbf{P}}[Q_2(\mathbf{H}^{\mathbf{a}}_{1:t+1}, \mathbf{a}_{t+1}) \mid \phi(\mathbf{H}_{1:t}, \mathbf{A}_t)]
\end{aligned}$$

Then:

$$\begin{aligned}
&\mathbb{E}_{\mathbf{P}}[\mathbf{Y}_{t+\tau}[\mathbf{a}_{t:t+\tau-1}] \mid \phi(\mathbf{H}_{1:t}, \mathbf{a}_t) = \phi(\mathbf{h}_{1:t}, \mathbf{a}_t)] \\
&\quad = Q_1(\mathbf{h}_{1:t}, \mathbf{a}_t)
\end{aligned}$$

We provide a proof of Theorem 3.3 in Appendix B. This result naturally motivates a recursive regression approach for spatiotemporal CAPO estimation, fitting each $Q_k(\cdot)$ in reverse order and substituting interventional treatments where required. In the next section, we detail how our end-to-end GST-UNet framework implements these ideas, providing a principled neural solution for time-varying confounding in spatiotemporal settings.

4. Methodology

To address the challenges of time-varying confounders, spatial dependencies, and temporal carryover in observational data, we introduce **GST-UNet**, an end-to-end spatiotemporal neural model. Building on the identification guarantees in Theorem 3.3 and the single-series assumptions in Assumption 3.2, GST-UNet integrates iterative G-computation with a modular neural architecture to consistently estimate potential outcomes and treatment effects. We first convert the iterative G-computation result into a practical, recursive regression approach for CAPO estimation, before introducing the GST-UNet design in subsequent subsections.

4.1. Estimating CAPOs with Iterative G-Computation

While Theorem 3.3 motivates a recursive regression algorithm for each Q_k ($k = 1, \dots, \tau$), only Q_τ can be directly estimated from the prefix data. At the next step, $Q_{\tau-1}$ depends on $Q_\tau\big(\mathbf{H}^{\mathbf{a}}_{1:t+\tau-1}, \mathbf{a}_{t+\tau-1}\big)$—where the observed treatments $\mathbf{A}_{t:t+\tau-1}$ are replaced by $\mathbf{a}_{t:t+\tau-1}$—but such substituted outcomes are not observed in the prefix data. Therefore, for $k < \tau$, we propose a procedure where we generate *pseudo-outcomes* by predicting with the previously learned $\widehat{Q}_{k+1}$. Going forward, we use $\widehat{F}$ to denote any quantity F estimated from data. Formally, let $\phi \in \Phi$ be an embedding satisfying Assumption 3.2, and let $\mathcal{Q}$ be our function class for Q_k. We learn the sequence $\widehat{Q}_\tau, \dots, \widehat{Q}_1$ from prefix data $\{\mathbf{P}_t^\tau : t = 1, \dots, T-\tau\}$, via:

1. **Initialization.** With the prefix data, fit $\widehat{Q}_\tau$ to predict $\mathbf{Y}_{t+\tau}$ from the embedding $\phi(\mathbf{H}_{1:t+\tau-1}, \mathbf{A}_{t+\tau-1})$.
2. **Backward recursion.** For $k = \tau-1, \dots, 1$:
   (a) *Substitute interventions.* For each prefix $\mathbf{P}_t^\tau$, replace $\mathbf{A}_{t+k}$ by the interventional $\mathbf{a}_{t+k}$ to form the modified history $\mathbf{H}^{\mathbf{a}}_{1:t+k}$.
   (b) *Generate pseudo-outcomes.* Let $\widetilde{Y}_{t+k+1} = \widehat{Q}_{k+1}\big(\mathbf{H}^{\mathbf{a}}_{1:t+k}, \mathbf{a}_{t+k}\big)$, where $\widehat{Q}_{k+1}$ is the previously learned function. These $\widetilde{Y}_{t+k+1}$ act as surrogates for $\mathbf{Y}_{t+k+1}$ in the prefix data.
   (c) *Fit* $\widehat{Q}_k$. Regress $\widetilde{Y}_{t+k+1}$ on the current embedding $\phi\big(\mathbf{H}_{1:t+k-1}, \mathbf{A}_{t+k-1}\big)$ to learn $\widehat{Q}_k \in \mathcal{Q}$.
3. **Final step.** Given a new history $\mathbf{h}_{1:t}$ and an interventional path $\mathbf{a}_{t:t+\tau-1}$, we predict

$$\begin{aligned}
&\widehat{\mathbb{E}}_{\mathbf{P}}[\mathbf{Y}_{t+\tau}[\mathbf{a}_{t:t+\tau-1}] \mid \phi(\mathbf{H}_{1:t}, \mathbf{a}_t) = \phi(\mathbf{h}_{1:t}, \mathbf{a}_t)] \\
&\quad = \widehat{Q}_1\big(\mathbf{h}_{1:t}, \mathbf{a}_t\big).
\end{aligned}$$

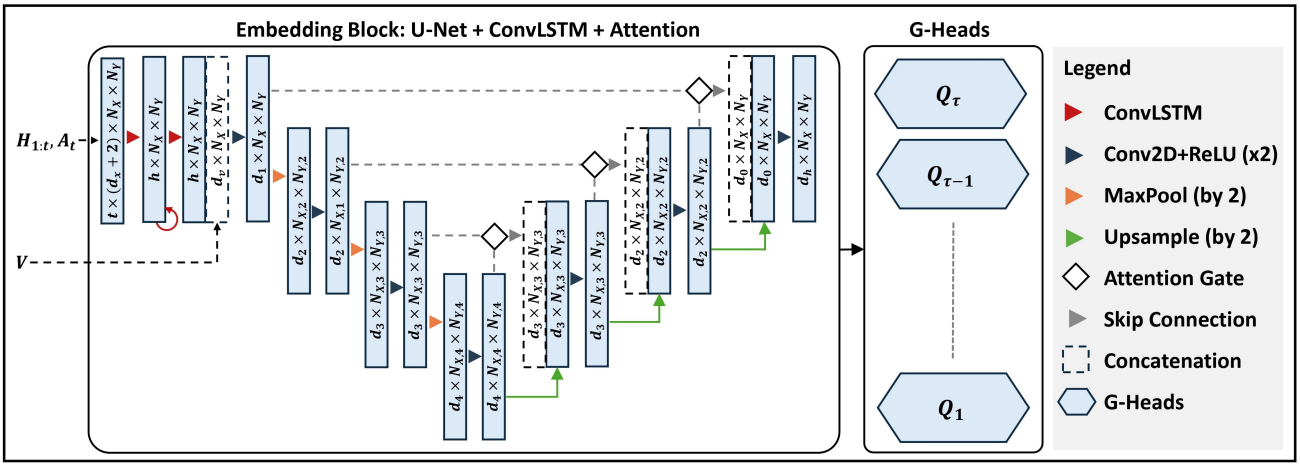

Figure 2. Overview of the GST-UNet architecture. The embedding block (left) is a U-Net augmented with a ConvLSTM layer and attention gates. Its final feature map is passed to a set of *G-heads* (right), where each G-head Q_k implements iterative G-computation (see Algorithm 1) to predict the potential outcomes or pseudo-outcomes at step k.

The iterative procedure yields a consistent estimator for the CAPOs if the Q_k's are estimated consistently from data (Laan and Robins, 2003). In the following subsections, we instantiate this procedure in our **GST-UNet** architecture, illustrating how to incorporate spatial dependencies and interference into ϕ and each Q_k, and implement a streamlined, end-to-end training strategy that unifies history embeddings and outcome predictions.

4.2. Model Architecture

The **GST-UNet** consists of two main components:

1. **Spatiotemporal Learning Module.** A U-Net-based network augmented with ConvLSTM and attention gates for spatiotemporal processing.
2. **Neural Causal Module.** τ G-computation heads, each mapping the spatiotemporal features to the final outcome predictions in the iterative procedure.

We illustrate the GST-UNet architecture in Figure 2 and describe its main components below.

Spatiotemporal Learning Module. *(1) Spatial Module.* To efficiently process high-dimensional spatial data, we employ U-Net (Ronneberger et al., 2015), a fully convolutional architecture originally developed for biomedical image segmentation. It employs an encoder-decoder design with skip connections: the encoder progressively downsamples the spatial grid through convolution and pooling, while the decoder upsamples it back to the original resolution, merging encoder features at each scale. *(2) Temporal Module.* U-Net has limitations in capturing temporal information. To address this, we integrate a Convolutional Long Short-Term Memory (ConvLSTM) layer (Shi et al., 2015) to the U-Net encoder. This module captures temporal dependencies by maintaining a hidden state across time steps while aggregating spatial information through convolutions. After computing the final ConvLSTM state, we append static (time-invariant) covariates $\mathbf{V}$ as additional feature channels, ensuring the subsequent U-Net encoder-decoder has direct access to both temporal dynamics and static location-specific information. In the decoder, we incorporate *attention gates* (Oktay et al., 2018) to selectively highlight relevant spatial regions, refining skip connections and emphasizing critical global or local patterns. The embedding module ultimately produces a d_h-dimensional feature map of size $N_X \times N_Y$, capturing essential spatiotemporal context—including interference, spatial confounding, and static covariates—for downstream G-computation.

Neural Causal Module. We attach τ *G-computation heads* to the U-Net's final feature maps, corresponding to the Q_k estimators in the iterative procedure (see Section 4.1). Each head can be a small convolutional module or a simple feed-forward network, depending on how much spatial structure remains to be captured. The information flow at the G-computation heads proceeds as follows: each head Q_k ($k = 1, \ldots, \tau$) receives the $d_h \times N_X \times N_Y$ U-Net embedding $\widehat{\phi}\big(\mathbf{H}_{1:t+k-1}, \mathbf{A}_{t+k-1}\big)$ (encompassing spatiotemporal and static context) and outputs an $N_X \times N_Y$ prediction for that time step. We refer to this as the *supervision step*, since Q_τ compares its predictions to the *real* observed outcomes $\mathbf{Y}_{t+\tau}$, anchoring the model in genuine data, while each $Q_{k<\tau}$ compares its predictions to pseudo-outcomes $\widetilde{\mathbf{Y}}_{t+k+1}$ provided by $\widehat{Q}_{k+1}$. These pseudo-outcomes arise in a subsequent *generation step*, wherein Q_{k+1} processes the intervened history $\widehat{\phi}\big(\mathbf{H}^{\mathbf{a}}_{1:t+k}, \mathbf{a}_{t+k}\big)$ in a *detached* forward pass (so $\widehat{Q}_{k+1}$ is not updated by Q_k's loss), thereby creating surrogate targets for Q_k. This procedure realizes the iterative G-computation logic from Section 4.1, enabling

Algorithm 1 GST-UNet Training and Inference

Input: Horizon τ, prefix data $\{\mathbf{P}_t^\tau\}_{t=1}^{T-\tau}$, interventions $\mathbf{a}_{t:t+\tau-1}$, curriculum schedule $\alpha_k^{(e)}$, total epochs E.
Initialize: parameters θ (U-Net embedding + G-heads).
for $e = 1 \dots E$ **do**
  for $k = \tau \dots 1$ **do**
    (Supervision) For each prefix i, predict outcomes:

$$\widehat{Y}_{t+k}^{(i)} = Q_k\big(\phi(\mathbf{H}_{1:t+k-1}^{(i)}, \mathbf{A}_{t+k-1}^{(i)}); \theta\big).$$

    (Generation (detached)) For each prefix i, generate pseudo-outcomes:

$$\widetilde{Y}_{t+k}^{(i)} = \begin{cases} Q_k\Big(\phi\big((\mathbf{H}^{\mathbf{a}}_{1:t+k-1})^{(i)}, \mathbf{a}_{t+k-1}^{(i)}\big); \theta\Big), & k < \tau, \\ Y_{t+\tau}^{(i)}, & k = \tau. \end{cases}$$

    where the observed $\mathbf{A}_{t:t+k-2}$'s were replaced with $\mathbf{a}_{t:t+k-2}$ in the history.
  end for
  (Loss aggregation) Compute the overall MSE loss

$$\mathcal{L}(\theta; e) = \frac{1}{\tau}\sum_{k=1}^{\tau} \alpha_k^{(e)} \sum_i \big(\widehat{Y}_{t+k}^{(i)} - \widetilde{Y}_{t+k+1}^{(i)}\big)^2.$$

  (Backward pass) Update θ by backpropagation.
end for
(Inference) Given a $\mathbf{h}_{1:t}$, return $Q_1(\phi(\mathbf{h}_{1:t}, \mathbf{a}_t); \widehat{\theta})$.

GST-UNet to estimate future outcomes under various counterfactual treatments. By separating the spatiotemporal embedding from the G-heads, we maintain a common representation for all prefix data (see Assumption 3.2) and flexibly capture interference and spatial confounding. Each G-head enforces the proper temporal adjustments to yield bias-free counterfactual inference.

4.3. Training and Inference

A key obstacle in learning from a single spatiotemporal series is that we must splice the data into many prefixes (Assumption 3.2), then estimate all Q_k in an iterative G-computation procedure (Section 4.1). In principle, we could train each G-head Q_k sequentially (from τ down to 1) by feeding its pseudo-outcomes to the next head. However, this creates complications when sharing the U-Net embedding ϕ across all heads: each Q_k might attempt to tailor ϕ to its own objective, leading to misaligned training signals.

Joint Loss and Multi-Task Training. To address this issue, we employ a *joint* (or *multi-task*) training approach (Caruana, 1997; Evgeniou and Pontil, 2004) by aggregating the loss terms from all G-heads into a single objective, then backpropagating once per batch. Concretely, for each head Q_k, let $\widetilde{Y}_{t+k+1}$ be the *real* outcomes if $k = \tau$ or *pseudo-outcomes* (generated by $\widehat{Q}_{k+1}$) if $k < \tau$. Our head-specific loss is a mean squared error (MSE) over all prefix samples:

$$\mathcal{L}_k(\theta) = \sum_{i=1}^{T-\tau} \Big[Q_k\big(\phi(\mathbf{H}_{1:t+k-1}^{(i)}, \mathbf{A}_{t+k-1}^{(i)}); \theta\big) - \widetilde{Y}_{t+k+1}^{(i)}\Big]^2,$$

where θ encompasses *all* model parameters (the shared U-Net embedding ϕ and the G-heads Q_k). Let $\alpha_k^{(e)}$ denote a *head-weight* for epoch e. We then form the overall training objective at epoch e by

$$\mathcal{L}(\theta; e) = \frac{1}{\tau}\sum_{k=1}^{\tau} \alpha_k^{(e)}\, \mathcal{L}_k(\theta). \tag{4}$$

By summing the losses and performing a single backward pass, we learn a common embedding $\widehat{\phi}$ that balances the needs of all G-heads, rather than fitting each head separately.

Curriculum Training. A naive implementation of (4) – where we give each G-head an equal weights – can be suboptimal: early in training, Q_τ (which sees real data) is inaccurate, and thus the pseudo-outcomes generated for $Q_{k<\tau}$ are effectively noise. Consequently, heads $Q_1, \dots, Q_{\tau-1}$ may overfit to poor targets before Q_τ has converged, leading to suboptimal solutions. To mitigate this, we employ a *curriculum* training approach (Bengio et al., 2009), gradually increasing the loss weight of earlier heads as Q_τ improves.

While many curricula are possible, we adopt a simple scheme controlled by a single hyperparameter e_c (the "curriculum period") so we can readily tune it. Let $p(e) = \min\{\tau, \lceil e/e_c \rceil\}$, which indexes a "phase" based on the current epoch e. We then define

$$\alpha_k^{(e)} = \begin{cases} 1/p(e), & \text{if } k \in \{\tau, \tau-1, \dots, \tau-p(e)+1\}, \\ 0, & \text{otherwise.} \end{cases}$$

Hence, during epochs $1 \le e \le e_c$ (phase $p(e) = 1$), only Q_τ is active with $\alpha_\tau^{(e)} = 1$; in the next interval $e_c < e \le 2e_c$ (phase $p(e) = 2$), Q_τ and $Q_{\tau-1}$ each have weight $1/2$, and so on, until phase τ sets all heads to uniform weight $1/\tau$. For $e > \tau\, e_c$, we continue training with $\alpha_k^{(e)} = 1/\tau$ across all G-heads for additional joint refinement. This progressive schedule ensures Q_τ becomes reasonably accurate on the observed data before earlier heads rely on its pseudo-outcomes. The single hyperparameter e_c succinctly controls how quickly we introduce each head, preventing excessive noise in early training.

We also adopt standard neural network practices, including mini-batch optimization and early stopping, to stabilize training and mitigate overfitting. At *inference* time, given a new history $\mathbf{h}_{1:t}$ and an interventional sequence $\mathbf{a}_{t:t+\tau-1}$, we compute $\widehat{Q}_1(\phi(\mathbf{h}_{1:t}, \mathbf{a}_t); \theta)$ as our target CAPO estimate. We sketch the overall training and inference procedure in Algorithm 1.

5. Experiments

We evaluate the proposed GST-UNet framework through two applications. First, we simulate synthetic data that incorporates key spatiotemporal causal inference challenges: interference, spatial confounding, temporal carryover, and time-varying confounding. Using this synthetic data generation process (DGP), we compare the GST-UNet algorithm against several baselines. Next, we demonstrate the utility of GST-UNet on a real-world dataset analyzing the impact of wildfire smoke on respiratory hospitalizations during the 2018 California Camp Fire.

Additional details–including exact simulation parameters, model architecture and execution setups, hyperparameter selection strategies, and validation procedures–can be found in Appendix C. Replication code is available at `https://github.com/moprescu/GSTUNet`.

5.1. Synthetic Data

We generate $T = 200$ time steps of a 64×64 ($N_X \times N_Y$) grid of observational data using the following system:

$$\mathbf{X}_t = \alpha_0 + \alpha_1 \mathbf{X}_{t-1} + \alpha_2 \mathbf{A}_{t-1} + \alpha_3 (K_X * \mathbf{X}_{t-1}) + \epsilon_X,$$

$$\mathbf{A}_t \sim \text{Bern}\Big(\sigma\big(\beta_1(\beta_0 + \tfrac{1}{L}\sum_{l=0}^{L-1} K_A * \mathbf{X}_{t-l})\big)\Big),$$

$$\mathbf{Y}_t = \gamma_0 + \gamma_1 \big(K_{YA} * \mathbf{A}_{t-1}\big) + \gamma_2 \tfrac{1}{L}\sum_{l=1}^{L} \big(K_{YX} * \mathbf{X}_{t-l}\big) + \gamma_3 \mathbf{Y}_{t-1} + \epsilon_Y,$$

where $d_X = 1$ (one feature), "$*$" denotes a 2D convolution over the $N_X \times N_Y$ grid, K_X, K_A, K_{YA}, K_{YX} are $d_k \times d_k$ convolution kernels, L is the number of temporal lags, and $\epsilon_X, \epsilon_Y \sim \mathcal{N}(0, 1)$ are i.i.d. noise terms. Each equation is evaluated at every location s in the grid, so $\mathbf{X}_t, \mathbf{A}_t, \mathbf{Y}_t$ represent $N_X \times N_Y$ matrices at time t. We choose parameter values such that the simulation remains stable (*i.e.*, the process does not diverge). This data-generating process (DGP) includes **interference** and **spatial confounding** from neighboring cells, as well as **temporal carryover**. Furthermore, $\mathbf{X}_t$ is a *time-varying* confounder, since its past values affect $\mathbf{A}$ and $\mathbf{Y}$, while current $\mathbf{A}$ influences future $\mathbf{X}$.

A concise example scenario is pollution control measures influencing health outcomes: $\mathbf{A}_t$ might represent binary interventions (e.g. traffic restrictions or industrial regulations), $\mathbf{X}_t$ could be spatially diffused air quality, and $\mathbf{Y}_t$ could track hospital visits or economic indicators. Government policies reacting to past pollution levels naturally create the feedback loops modeled here.

We vary β_1 to control the strength of time-varying confounding. When $\beta_1 = 0$, $\mathbf{X}_t$ does not directly affect $\mathbf{A}_t$, eliminating time-varying confounding; larger values of β_1 strengthen the time-varying confounding. For each β_1, we generate $n_{\text{test}} = 50$ test trajectories from random initial states, fix their histories $\mathbf{h}_{1:t}$, and simulate 100 potential τ-length futures to average the terminal outcomes and obtain each true CAPO. We evaluate horizon lengths $\tau \in \{5, 10\}$. We compare GST-UNet with: *(i)* **UNet+**, a naive variant using U-Net + ConvLSTM + Attention but no G-computation (the treatments are simply appended as static channels); and *(ii)* **STCINet** (Ali et al., 2024), a spatiotemporal causal forecasting model that similarly lacks an adjustment for time-varying confounders. Table 1 reports the RMSE $\pm$ standard deviation across the mean test trajectories for $\beta_1 \in \{0, 0.5, 1.0, 1.5, 2.0, 2.5, 3.0\}$. When $\beta_1 = 0$ (no time-varying confounding), all methods perform similarly. However, as β_1 increases, GST-UNet's error remains nearly constant, while the baselines degrade considerably, demonstrating increasing bias from time-varying confounders. Thus, these results highlight the value of GST-UNet's iterative G-computation in effectively adjusting for time-varying confounding in spatiotemporal settings.

5.2. Impact of Wildfires on Respiratory Health

Wildfire smoke has been associated with short-term adverse respiratory outcomes (Reid et al., 2016b;a; Cascio, 2018; Cleland et al., 2021; Letellier et al., 2025), with older adults especially vulnerable (DeFlorio-Barker et al., 2019). Recent events (e.g., ongoing Los Angeles wildfires) have amplified concerns about acute health impacts. In 2018, California experienced multiple severe wildfires (Wikipedia, 2025), including two major incidents: the *Carr Fire* (July–August) and the *Camp Fire* (November), which significantly worsened air quality across the state.

We analyze daily, county-level data from Letellier et al. (2025) that include $PM_{2.5}$ (particulate matter $< 2.5\,\mu m$), hospitalization counts for respiratory and cardiovascular conditions, and weather variables (temperature, precipitation, humidity, radiation, wind), plus population estimates from the California Department of Finance. Each of the weather variables can serve as a *time-varying confounder* because weather conditions affect future smoke levels and health outcomes, while also potentially being influenced by prior smoke levels.

We focus on the period from week 20 to week 48 (May 18–December 2, 2018), covering both the Carr and Camp fires. Following standard practice, we label a county as "treated" on any day it has mean $PM_{2.5}$ above $10\,\mu g/m^3$, and we use the raw hospitalization counts as the outcome (rather than per-10,000 incidence to avoid instability in small counties). To represent counties in a spatial grid, we interpolate each day's county-level data (treatment, outcome, five covariates) onto a 40×44 latitude–longitude lattice (discarding cells outside California), yielding a spatiotemporal tensor of size

Table 1. Comparison of models across horizons (τ). Values represent RMSE ± standard deviation across mean test trajectories. GST-UNet (ours) is highlighted in the last row. Bold values indicate the smallest number in each column for each horizon. Relative changes for GST-UNet compared to the best baseline are shown in **green** (improvement) or **red** (decrease).

τ	Model	$\beta_1 = 0.0$	$\beta_1 = 0.5$	$\beta_1 = 1.0$	$\beta_1 = 1.5$	$\beta_1 = 2.0$	$\beta_1 = 2.5$	$\beta_1 = 3.0$
5	UNet+	**0.51 ± 0.00**	0.62 ± 0.01	0.84 ± 0.01	1.03 ± 0.01	1.10 ± 0.01	1.16 ± 0.01	1.25 ± 0.01
	STCINet	0.52 ± 0.00	0.68 ± 0.01	0.93 ± 0.01	1.11 ± 0.01	1.20 ± 0.01	1.33 ± 0.01	1.32 ± 0.01
	GST-UNet	0.55 ± 0.01	**0.61 ± 0.01**	**0.60 ± 0.01**	**0.61 ± 0.01**	**0.64 ± 0.01**	**0.58 ± 0.01**	**0.64 ± 0.01**
		(**+7.8%**)	(**-1.6%**)	(**-28.6%**)	(**-40.8%**)	(**-41.8%**)	(**-50.0%**)	(**-48.8%**)
10	UNet+	0.56 ± 0.00	0.53 ± 0.00	0.68 ± 0.00	0.85 ± 0.00	1.00 ± 0.01	1.02 ± 0.01	1.01 ± 0.01
	STCINet	0.57 ± 0.00	0.59 ± 0.00	0.74 ± 0.00	0.86 ± 0.01	1.11 ± 0.01	1.15 ± 0.01	1.26 ± 0.01
	GST-UNet	**0.50 ± 0.00**	**0.44 ± 0.01**	**0.45 ± 0.01**	**0.57 ± 0.01**	**0.48 ± 0.01**	**0.53 ± 0.01**	**0.49 ± 0.01**
		(**-10.7%**)	(**-16.7%**)	(**-33.8%**)	(**-32.9%**)	(**-52.1%**)	(**-48.0%**)	(**-51.5%**)

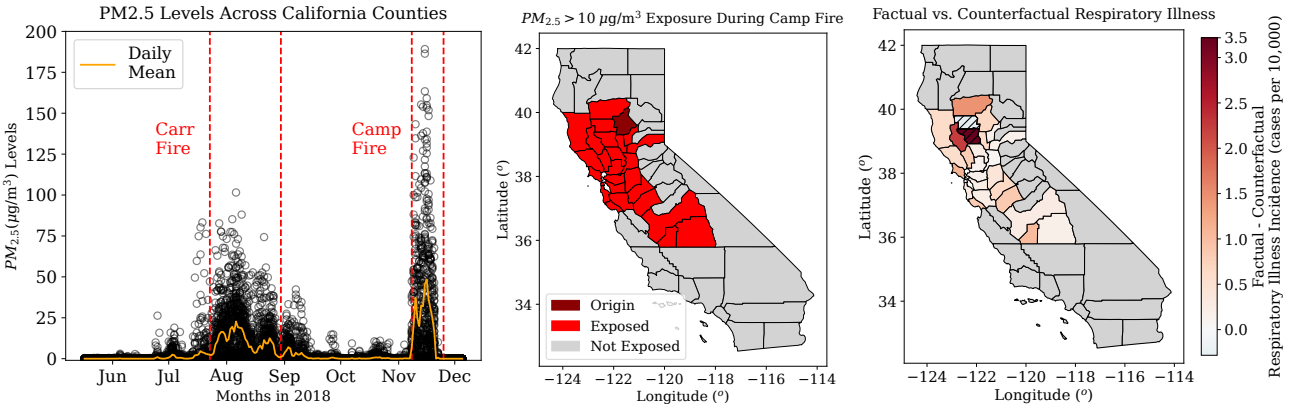

Figure 3. **(Left)** Daily $PM_{2.5}$ levels across California from May to December 2018, with red lines marking major wildfires. **(Center)** Counties exposed to average $PM_{2.5} > 10\,\mu g/m^3$ during the Camp Fire (red), origin county in dark red. **(Right)** Factual minus CAPO-predicted daily respiratory admissions during peak Camp Fire. Hashed areas indicate small-population counties ($< 30{,}000$).

$203 \times 7 \times 40 \times 44$. Interpolation ensures that each grid cell approximates the region it overlaps (weighted by intersection area), so the downstream model captures spatial gradients in $PM_{2.5}$, weather, and hospitalizations. We train GST-UNet with a horizon $\tau = 10$. The earlier wildfire period (Jun–July) is used for validation and hyperparameter tuning. After model selection, we generate counterfactual predictions for the peak phase of the Camp Fire, November 8–17, 2018. See Appendix C for details on data preprocessing, interpolation, and masking.

Figure 3 (left) shows the rise in $PM_{2.5}$ levels during the mid-late 2018 wildfire season, and (center) highlights counties with daily $PM_{2.5} > 10\,\mu g/m^3$. Using GST-UNet, we estimate the daily CAPOs had the Camp Fire never occurred (i.e., treating all counties as if $PM_{2.5} \leq 10\,\mu g/m^3$). Figure 3 (right) compares these CAPOs to the factual daily mean incidence (hospitalization cases per 10,000 residents). Hatching marks low-population counties ($< 30{,}000$ compared to $> 70{,}000$ for other exposed counties) with higher uncertainty; we exclude these from the analysis. Over November 8–17, GST-UNet predicts **approximately 4,650 excess respiratory hospitalizations** (465 per day) attributable to the Camp Fire, with the highest incidence near the fire source. This result is qualitatively consistent with Letellier et al. (2025), who report around 259 excess daily cases *averaged* over November 8–December 5—a longer window including lower-intensity days, thus yielding a smaller daily estimate. Overall, GST-UNet captures the spatiotemporal variation in smoke exposure and health outcomes, illustrating its promise for real-world causal inference in environmental health and policy.

6. Conclusion

We introduced GST-UNet, a neural framework for estimating causal effects in spatiotemporal settings by combining a U-Net-based architecture for spatial modeling with iterative G-computation to adjust for time-varying confounders. GST-UNet addresses key challenges such as interference, spatial confounding, temporal carryover, and time-varying confounders. Through simulations, the GST-UNet outperformed existing baselines, and in a real-world case study of wildfire smoke exposure during the 2018 Camp Fire, it provided fine-grained, location-specific, and credible effect estimates. These results highlight GST-UNet's potential to improve causal inference in fields such as public health, environmental science, and social policy.

Acknowledgements

This material is based upon work supported by the National Science Foundation under Grant No. 1846210 and by the U.S. Department of Energy, Office of Science, Office of Advanced Scientific Computing Research, under Award Number DE-SC0023112. Part of this work was conducted while Miruna Oprescu was a research intern at Brookhaven National Laboratory.

Impact Statement

This work advances the field of machine learning by developing a spatiotemporal causal inference framework that enables more accurate estimation of treatment effects in complex real-world settings. GST-UNet has broad applications in public health, environmental science, and social policy, where understanding intervention effects can inform evidence-based decision-making. While our method is designed to improve causal inference from observational data, care must be taken when applying it to high-stakes policy decisions, ensuring robustness against biases in data collection and model assumptions. We encourage responsible use, particularly in applications affecting vulnerable populations.

References

S. Ali, O. Faruque, and J. Wang. Estimating direct and indirect causal effects of spatiotemporal interventions in presence of spatial interference. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 213–230. Springer, 2024.

L. Anselin. *Spatial econometrics: methods and models*, volume 4. Springer Science & Business Media, 2013.

H. Bang and J. M. Robins. Doubly robust estimation in missing data and causal inference models. *Biometrics*, 61 (4):962–973, 2005.

E. Ben-Michael, A. Feller, and J. Rothstein. Synthetic controls with staggered adoption. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 84 (2):351–381, 2022.

Y. Bengio, J. Louradour, R. Collobert, and J. Weston. Curriculum learning. In *Proceedings of the 26th annual international conference on machine learning*, pages 41–48, 2009.

I. Bica, A. M. Alaa, J. Jordon, and M. van der Schaar. Estimating counterfactual treatment outcomes over time through adversarially balanced representations. *arXiv preprint arXiv:2002.04083*, 2020.

I. Bojinov and N. Shephard. Time series experiments and causal estimands: exact randomization tests and trading. *Journal of the American Statistical Association*, 2019.

R. Caruana. Multitask learning. *Machine learning*, 28: 41–75, 1997.

W. E. Cascio. Wildland fire smoke and human health. *Science of the total environment*, 624:586–595, 2018.

R. Christiansen, M. Baumann, T. Kuemmerle, M. D. Mahecha, and J. Peters. Toward causal inference for spatiotemporal data: Conflict and forest loss in colombia. *Journal of the American Statistical Association*, 117(538): 591–601, 2022.

S. E. Cleland, M. L. Serre, A. G. Rappold, and J. J. West. Estimating the acute health impacts of fire-originated pm2. 5 exposure during the 2017 california wildfires: Sensitivity to choices of inputs. *Geohealth*, 5(7):e2021GH000414, 2021.

A. Coston, E. Kennedy, and A. Chouldechova. Counterfactual predictions under runtime confounding. *Advances in neural information processing systems*, 33:4150–4162, 2020.

S. DeFlorio-Barker, J. Crooks, J. Reyes, and A. G. Rappold. Cardiopulmonary effects of fine particulate matter exposure among older adults, during wildfire and non-wildfire periods, in the united states 2008–2010. *Environmental health perspectives*, 127(3):037006, 2019.

T. Evgeniou and M. Pontil. Regularized multi–task learning. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 109–117, 2004.

X. Glorot and Y. Bengio. Understanding the difficulty of training deep feedforward neural networks. In *Proceedings of the 13th International Conference on Artificial Intelligence and Statistics (AISTATS)*, pages 249–256. JMLR Workshop and Conference Proceedings, 2010.

K. Hess, D. Frauen, V. Melnychuk, and S. Feuerriegel. G-transformer for conditional average potential outcome estimation over time. *arXiv preprint arXiv:2405.21012*, 2024.

K. Jordahl, J. V. den Bossche, M. Fleischmann, J. Wasserman, J. McBride, J. Gerard, J. Tratner, M. Perry, A. G. Badaracco, C. Farmer, G. A. Hjelle, A. D. Snow, M. Cochran, S. Gillies, L. Culbertson, M. Bartos, N. Eubank, maxalbert, A. Bilogur, S. Rey, C. Ren, D. Arribas-Bel, L. Wasser, L. J. Wolf, M. Journois, J. Wilson, A. Greenhall, C. Holdgraf, Filipe, and F. Leblanc. geopandas/geopandas: v0.8.1, July 2020. URL https://doi.org/10.5281/zenodo.3946761.

L. J. Keele and R. Titiunik. Geographic boundaries as regression discontinuities. *Political Analysis*, 23(1):127–155, 2015.

D. P. Kingma. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

M. J. Laan and J. M. Robins. *Unified methods for censored longitudinal data and causality*. Springer, 2003.

N. Letellier, M. Hale, K. U. Salim, Y. Ma, F. Rerolle, L. Schwarz, and T. Benmarhnia. Applying a two-stage generalized synthetic control approach to quantify the heterogeneous health effects of extreme weather events: A 2018 large wildfire in california event as a case study. *Environmental Epidemiology*, 9(1):e362, 2025.

R. Li, S. Hu, M. Lu, Y. Utsumi, P. Chakraborty, D. M. Sow, P. Madan, J. Li, M. Ghalwash, Z. Shahn, et al. G-net: a recurrent network approach to g-computation for counterfactual prediction under a dynamic treatment regime. In *Machine Learning for Health*, pages 282–299. PMLR, 2021.

Y. Li, R. Yu, C. Shahabi, and Y. Liu. Diffusion convolutional recurrent neural network: Data-driven traffic forecasting. *arXiv preprint arXiv:1707.01926*, 2017.

B. Lim. Forecasting treatment responses over time using recurrent marginal structural networks. *Advances in neural information processing systems*, 31, 2018.

V. Melnychuk, D. Frauen, and S. Feuerriegel. Causal transformer for estimating counterfactual outcomes. In *International Conference on Machine Learning*, pages 15293–15329. PMLR, 2022.

O. Oktay, J. Schlemper, L. L. Folgoc, M. Lee, M. Heinrich, K. Misawa, K. Mori, S. McDonagh, N. Y. Hammerla, B. Kainz, et al. Attention u-net: Learning where to look for the pancreas. *arXiv preprint arXiv:1804.03999*, 2018.

G. Papadogeorgou, F. Mealli, and C. M. Zigler. Causal inference with interfering units for cluster and population level treatment allocation programs. *Biometrics*, 75(3): 778–787, 2019.

G. Papadogeorgou, K. Imai, J. Lyall, and F. Li. Causal inference with spatio-temporal data: estimating the effects of airstrikes on insurgent violence in iraq. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 84(5):1969–1999, 2022.

A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Kopf, E. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai, and S. Chintala. Pytorch: An imperative style, high-performance deep learning library. NeurIPS, 2019.

C. E. Reid, M. Brauer, F. H. Johnston, M. Jerrett, J. R. Balmes, and C. T. Elliott. Critical review of health impacts of wildfire smoke exposure. *Environmental health perspectives*, 124(9):1334–1343, 2016a.

C. E. Reid, M. Jerrett, I. B. Tager, M. L. Petersen, J. K. Mann, and J. R. Balmes. Differential respiratory health effects from the 2008 northern california wildfires: A spatiotemporal approach. *Environmental research*, 150: 227–235, 2016b.

J. Robins and M. Hernan. Estimation of the causal effects of time-varying exposures. *Chapman & Hall/CRC Handbooks of Modern Statistical Methods*, pages 553–599, 2008.

J. M. Robins, M. A. Hernan, and B. Brumback. Marginal structural models and causal inference in epidemiology, 2000.

O. Ronneberger, P. Fischer, and T. Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18*, pages 234–241. Springer, 2015.

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

N. Seedat, F. Imrie, A. Bellot, Z. Qian, and M. van der Schaar. Continuous-time modeling of counterfactual outcomes using neural controlled differential equations. *arXiv preprint arXiv:2206.08311*, 2022.

X. Shi, Z. Chen, H. Wang, D.-Y. Yeung, W.-K. Wong, and W.-c. Woo. Convolutional lstm network: A machine learning approach for precipitation nowcasting. *Advances in neural information processing systems*, 28, 2015.

C. Song, Y. Wang, X. Yang, Y. Yang, Z. Tang, X. Wang, and J. Pan. Spatial and temporal impacts of socioeconomic and environmental factors on healthcare resources: a county-level bayesian local spatiotemporal regression modeling study of hospital beds in southwest china. *International Journal of Environmental Research and Public Health*, 17(16):5890, 2020.

M. Tec, J. G. Scott, and C. M. Zigler. Weather2vec: Representation learning for causal inference with non-local confounding in air pollution and climate studies. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 14504–14513, 2023.

Y. Wang. Causal inference under temporal and spatial interference. *arXiv e-prints*, pages arXiv–2106, 2021.

Wikipedia. Camp Fire (2018) — Wikipedia, the free encyclopedia. http://en.wikipedia.org/w/index.php?title=Camp%20Fire%20(2018)&oldid=1271689743, 2025. [Online; accessed 29-January-2025].

Z. Wu, S. Pan, G. Long, J. Jiang, and C. Zhang. Graph wavenet for deep spatial-temporal graph modeling. *arXiv preprint arXiv:1906.00121*, 2019.

J. Zhang, Y. Zheng, and D. Qi. Deep spatio-temporal residual networks for citywide crowd flows prediction. In *Proceedings of the AAAI conference on artificial intelligence*, volume 31, 2017.

W. Zhang and K. Ning. Spatiotemporal heterogeneities in the causal effects of mobility intervention policies during the covid-19 outbreak: A spatially interrupted time-series (sits) analysis. *Annals of the American Association of Geographers*, 113(5):1112–1134, 2023.

L. Zhou, K. Imai, J. Lyall, and G. Papadogeorgou. Estimating heterogeneous treatment effects for spatio-temporal causal inference: How economic assistance moderates the effects of airstrikes on insurgent violence. *arXiv preprint arXiv:2412.15128*, 2024.

A. Extended Literature Review

Classical Spatiotemporal Causal Inference. Early spatiotemporal causal inference methods—including spatial econometrics (Anselin, 2013), difference-in-differences (Keele and Titiunik, 2015), and synthetic controls (Ben-Michael et al., 2022)—provide useful frameworks for estimating treatment effects across regions but rely on strong assumptions such as parallel trends or stable treatment assignment. These approaches struggle with interference, nonlinear dependencies, and time-varying confounders, limiting their applicability in complex settings. More recent classical approaches estimate average treatment effects across regions or impose structural and modeling assumptions that may not generalize well to real-world spatiotemporal contexts (Wang, 2021; Christiansen et al., 2022; Papadogeorgou et al., 2022; Zhang and Ning, 2023; Zhou et al., 2024).

Machine Learning for Spatiotemporal Modeling. The rise of large-scale spatiotemporal datasets has led to the adoption of machine learning models for predictive tasks. Convolutional and recurrent neural networks (Shi et al., 2015; Zhang et al., 2017) effectively capture spatial and temporal dependencies but lack explicit causal adjustment mechanisms. Graph-based deep learning methods (Li et al., 2017; Wu et al., 2019) model spatial interactions but typically ignore feedback effects from time-varying confounders. Some recent work integrates spatial representations for causal inference—e.g., Tec et al. (2023) incorporate geographic confounders using a UNet-based model—but these methods do not explicitly model iterative dependencies over time or adjust for time-varying confounders.

Time-Series Causal Inference. In the longitudinal domain, time-series causal inference typically employs recurrent networks, Transformers, or propensity-based models (Bica et al., 2020; Seedat et al., 2022; Melnychuk et al., 2022), but these approaches often assume independent time series, overlooking spatial interference and cross-series confounding. Handling time-varying confounders has relied on marginal structural models (Robins et al., 2000) or iterative G-computation (Bang and Robins, 2005; Robins and Hernan, 2008), but machine learning adaptations (Lim, 2018; Li et al., 2021; Hess et al., 2024) continue to assume independent observations, making them ill-suited for fully spatiotemporal settings where interference is prevalent.

Neural-Based Spatiotemporal Causal Inference. Recent efforts have explored neural models for spatiotemporal causal inference. Tec et al. (2023) integrate spatial confounding adjustments using a UNet-based framework but do not model feedback effects from time-varying confounders. The most similar work to ours, Ali et al. (2024), introduces a climate-focused neural model that shares certain architectural components but is designed primarily for predictive tasks rather than causal identification, leaving time-varying confounders unaddressed.

Our Contribution. GST-UNet bridges these gaps by combining a U-Net-based architecture for spatiotemporal grids with iterative G-computation to adjust for time-varying confounders. Unlike prior methods, GST-UNet explicitly accounts for interference, nonlinear spatial-temporal dynamics, and feedback loops, ensuring valid causal identification under standard assumptions. This enables fine-grained, location-specific estimates of potential outcomes, improving spatiotemporal causal inference in domains where observational data is abundant but controlled experiments are infeasible.

B. Proof of Theorem 3.3

We aim to show that under Assumption 3.1 and Assumption 3.2, the CAPOs in Equation (1) can be identified recursively from a single time series via a sequence of conditional expectations.

Step 1: Recursive decomposition for the intractable expectation We first demonstrate the recursive decomposition of the intractable expectation in the CAPO definition (Equation (1)). While this expectation is theoretically well-defined, it cannot be directly estimated in practice due to the limited availability of data. Specifically, we only observe a single time series, meaning we have just one sample of the history at time $t+\tau$ for each t. Nevertheless, as we will show, we can convert these expectations into expectations over prefix-based segments that allow us to estimate these quantities from the data.

Starting from $\mathbb{E}[\mathbf{Y}_{t+\tau}[\mathbf{a}_{t:t+\tau-1}] \mid \mathbf{H}_{1:t} = \mathbf{h}_{1:t}]$, we have:

$$
\begin{aligned}
&\mathbb{E}[\mathbf{Y}_{t+\tau}[\mathbf{a}_{t:t+\tau-1}] \mid \mathbf{H}_{1:t} = \mathbf{h}_{1:t}] \\
&= \mathbb{E}[\mathbf{Y}_{t+\tau}[\mathbf{a}_{t:t+\tau-1}] \mid \mathbf{H}_{1:t} = \mathbf{h}_{1:t}, \mathbf{A}_t = \mathbf{a}_t] && \text{(Sequential ignorability and positivity (Assumption 3.1))} \\
&= \mathbb{E}\big[\mathbb{E}[\mathbf{Y}_{t+\tau}[\mathbf{a}_{t:t+\tau-1}] \mid \mathbf{H}^{\mathbf{a}}_{1:t+1}] \mid \mathbf{H}_{1:t} = \mathbf{h}_{1:t}, \mathbf{A}_t = \mathbf{a}_t\big] && \text{(Law of total probability)}
\end{aligned}
$$

$$
\begin{aligned}
&= \mathbb{E}\big[\mathbb{E}[\mathbf{Y}_{t+\tau}[\mathbf{a}_{t:t+\tau-1}] \mid \mathbf{H}^{\mathbf{a}}_{1:t+1}, \mathbf{A}_{t+1} = \mathbf{a}_{t+1}] \mid \mathbf{H}_{1:t} = \mathbf{h}_{1:t}, \mathbf{A}_t = \mathbf{a}_t\big] && \text{(Sequential ignorability and positivity)}\\
&= \mathbb{E}\Big[\mathbb{E}\big[\mathbb{E}[\mathbf{Y}_{t+\tau}[\mathbf{a}_{t:t+\tau-1}] \mid \mathbf{H}^{\mathbf{a}}_{1:t+2}] \;\big|\; \mathbf{H}^{\mathbf{a}}_{1:t+1}, \mathbf{A}_{t+1} = \mathbf{a}_{t+1}\big] \;\Big|\; \mathbf{H}_{1:t} = \mathbf{h}_{1:t}, \mathbf{A}_t = \mathbf{a}_t\Big] && \text{(Law of total probability)}\\
&= \mathbb{E}\Big[\mathbb{E}\big[\mathbb{E}[\mathbf{Y}_{t+\tau}[\mathbf{a}_{t:t+\tau-1}] \mid \mathbf{H}^{\mathbf{a}}_{1:t+2}, \mathbf{A}_{t+2} = \mathbf{a}_{t+2}] \;\big|\; \mathbf{H}^{\mathbf{a}}_{1:t+1}, \mathbf{A}_{t+1} = \mathbf{a}_{t+1}\big] \;\Big|\; \mathbf{H}_{1:t} = \mathbf{h}_{1:t}, \mathbf{A}_t = \mathbf{a}_t\Big] \\
& && \text{(Sequential ignorability and positivity)}\\
&\dots\\
&= \mathbb{E}\Big[\dots\mathbb{E}\big[\mathbb{E}[\mathbf{Y}_{t+\tau}[\mathbf{a}_{t:t+\tau-1}] \mid \mathbf{H}^{\mathbf{a}}_{1:t+\tau-1}, \mathbf{A}_{t+\tau-1} = \mathbf{a}_{t+\tau-1}] \\
&\qquad\qquad \big| \; \mathbf{H}^{\mathbf{a}}_{1:t+\tau-2}, \mathbf{A}_{t+\tau-2} = \mathbf{a}_{t+\tau-2}\big] \\
&\qquad\qquad \big| \; \dots \\
&\qquad\qquad \Big| \; \mathbf{H}_{1:t} = \mathbf{h}_{1:t}, \mathbf{A}_t = \mathbf{a}_t\Big] && \text{(Sequential ignorability and positivity)}\\
&= \mathbb{E}\Big[\dots\mathbb{E}\big[\mathbb{E}[\mathbf{Y}_{t+\tau} \mid \mathbf{H}^{\mathbf{a}}_{1:t+\tau-1}, \mathbf{A}_{t+\tau-1} = \mathbf{a}_{t+\tau-1}] \\
&\qquad\qquad \big| \; \mathbf{H}^{\mathbf{a}}_{1:t+\tau-2}, \mathbf{A}_{t+\tau-2} = \mathbf{a}_{t+\tau-2}\big] \\
&\qquad\qquad \big| \; \dots \\
&\qquad\qquad \Big| \; \mathbf{H}_{1:t} = \mathbf{h}_{1:t}, \mathbf{A}_t = \mathbf{a}_t\Big] && \text{(Consistency)}
\end{aligned}
$$

Thus, if we had multiple spatiotemporal time-series samples, we could directly estimate this nested expression from data, since the right-hand side depends solely on observed quantities, ensuring identifiability.

Step 2: From intractable to prefix-based expectations We now show how to estimate the nested expectations using the prefix data. First, by Assumption 3.2, we can rewrite the inner-most expectation as

$$
\begin{aligned}
\mathbb{E}[\mathbf{Y}_{t+\tau} \mid \mathbf{H}^{\mathbf{a}}_{1:t+\tau-1}, \mathbf{A}_{t+\tau-1} = \mathbf{a}_{t+\tau-1}] &= \mathbb{E}_{\mathbf{P}}[\mathbf{Y}_{t+\tau} \mid \phi(\mathbf{H}^{\mathbf{a}}_{1:t+\tau-1}, \mathbf{a}_{t+\tau-1})]\\
&= Q_\tau(\mathbf{H}^{\mathbf{a}}_{1:t+\tau-1}, \mathbf{a}_{t+\tau-1}). && \text{(Definition of } Q_\tau\text{)}
\end{aligned}
$$

Thus, by using Assumption 3.1, we can write this expectation over the prefix data which we have many samples of. Now consider the next nested expectation:

$$
\begin{aligned}
&\mathbb{E}[Q_\tau(\mathbf{H}^{\mathbf{a}}_{1:t+\tau-1}, \mathbf{a}_{t+\tau-1}) \mid \mathbf{H}^{\mathbf{a}}_{1:t+\tau-2} = \mathbf{h}^{\mathbf{a}}_{1:t+\tau-2}, \mathbf{A}_{t+\tau-2} = \mathbf{a}_{t+\tau-2}]\\
&= \int Q_\tau(\mathbf{h}^{\mathbf{a}}_{1:t+\tau-1}, \mathbf{a}_{t+\tau-1}) p(x_{t+\tau-1}, y_{t+\tau-1} \mid \mathbf{h}^{\mathbf{a}}_{t+\tau-2}, \mathbf{a}_{t+\tau-2}) d(x_{t+\tau-1}, y_{t+\tau-1})\\
&= \int_P Q_\tau(\mathbf{h}^{\mathbf{a}}_{1:t+\tau-1}, \mathbf{a}_{t+\tau-1}) p(x_{t+\tau-1}, y_{t+\tau-1} \mid \phi(\mathbf{h}^{\mathbf{a}}_{t+\tau-2}, \mathbf{a}_{t+\tau-2})) d(x_{t+\tau-1}, y_{t+\tau-1}) && \text{(Assumption 3.2)}\\
&= \mathbb{E}_P[Q_\tau(\mathbf{H}^{\mathbf{a}}_{1:t+\tau-1}, \mathbf{a}_{t+\tau-1}) \mid \phi(\mathbf{H}^{\mathbf{a}}_{1:t+\tau-2}, \mathbf{A}_{t+\tau-2}) = \phi(\mathbf{h}^{\mathbf{a}}_{1:t+\tau-2}, \mathbf{a}_{t+\tau-2})]\\
&= Q_{\tau-1}(\mathbf{h}^{\mathbf{a}}_{1:t+\tau-2}, \mathbf{a}_{t+\tau-2})
\end{aligned}
$$

Tracing this argument recursively through the nested expectation in Step 1, we obtain:

$$\mathbb{E}[\mathbf{Y}_{t+\tau}[\mathbf{a}_{t:t+\tau-1}] \mid \mathbf{H}_{1:t} = \mathbf{h}_{1:t}] = Q_1(\mathbf{h}_{1:t}, \mathbf{a}_t),$$

as desired. Thus, Q_1 – which can be estimated from the prefix data – recovers the CAPOs, under our assumptions, even from a single chain.

C. Experimental Details

In this appendix, we provide further information on the simulation experiments (Section 5.1) and the real-world wildfire application (Section 5.2), including exact parameter settings, model architecture and execution details, hyperparameter selection strategies, and validation procedures. All code for generating, preprocessing, and analyzing both the synthetic and real-world datasets—and for training and evaluating GST-UNet—is available at `https://github.com/moprescu/GSTUNet`, with step-by-step replication instructions in the repository's `README.md`.

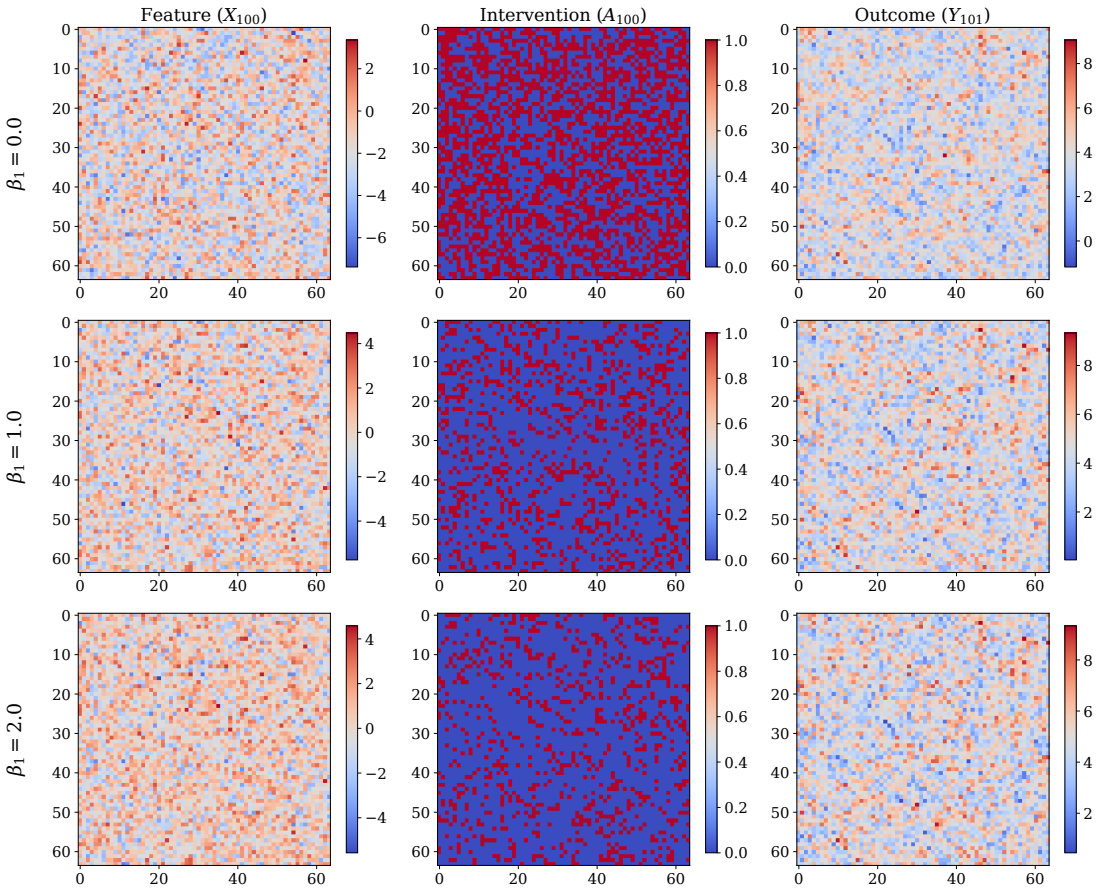

Figure 4. Samples from the DGP at $t = 100$, comparing feature X_{100} (left), intervention A_{100} (center), and outcome Y_{101} (right) for varying $\beta_1 \in \{0.0, 1.0, 2.0\}$.

For both applications, GST-UNet employs a U-Net backbone with a single ConvLSTM layer (hidden dimension 32) and a contracting-expanding path of channel sizes $16 \to 32 \to 64 \to 128 \to 256$. The G-computation heads are implemented as shallow feed-forward neural networks that operate on the U-Net feature maps at each grid cell for G-computation. In practice, to ensure stable ConvLSTM training and reduce computational overhead, we truncate the input history to a fixed length. All neural networks are implemented via the `nn` module in `PyTorch` (Paszke et al., 2019). Experiments were conducted on an NVIDIA A100 (Ampere) GPU using the Perlmutter system at the National Energy Research Scientific Computing Center (NERSC). The synthetic experiments required roughly 55 minutes per hyperparameter set, while the wildfire experiment completed in about 5 minutes.

C.1. Synthetic Experiments

Data Simulation Process. For our primary simulation experiments, we generate $T = 200$ time steps on a 64×64 grid. The simulation parameters in the generating equations

$$\mathbf{X}_t = \alpha_0 + \alpha_1 \mathbf{X}_{t-1} + \alpha_2 \mathbf{A}_{t-1} + \alpha_3 (K_X * \mathbf{X}_{t-1}) + \epsilon_X,$$

$$\mathbf{A}_t \sim \text{Bern}\Big(\sigma\big(\beta_1\big(\beta_0 + \tfrac{1}{L}\sum_{l=0}^{L-1} K_A * \mathbf{X}_{t-l}\big)\big)\Big),$$

$$\mathbf{Y}_t = \gamma_0 + \gamma_1\big(K_{YA} * \mathbf{A}_{t-1}\big) + \gamma_2 \tfrac{1}{L}\sum_{l=1}^{L}\big(K_{YX} * \mathbf{X}_{t-l}\big) + \gamma_3 \mathbf{Y}_{t-1} + \epsilon_Y,$$

Table 2. Hyperparameters and their ranges. We boldface the values that provided the best validation performance.

Hyperparameter	**Model(s)**	**Value Range**
Batch size	All models	$\{2, \mathbf{4}, 8\}$
Learning rate	All models	$\{10^{-4}, \mathbf{5 \times 10^{-4}}, 10^{-3}\}$
Scheduler patience	All models	$\{3, \mathbf{5}, 10\}$
Early stopping patience	All models	$\{5, \mathbf{10}\}$
Curriculum period	GST-UNet	$\{1, 3, \mathbf{5}, 7\}$
Curriculum learning rate	GST-UNet	$\{10^{-4}, \mathbf{5 \times 10^{-4}}, 10^{-3}\}$
UNet output dim d_h	GST-UNet	$\{8, \mathbf{16}, 32\}$
G-head hidden size	GST-UNet	$\{\mathbf{8}, 16\}$
G-head layers	GST-UNet	$\{\mathbf{1}, 2, 3\}$

are given by:

- $\mathbf{X}_t$:

$$\alpha_0 = 0.5,\ \alpha_1 = 0.5,\ \alpha_2 = -2.0,\ \alpha_3 = 0.2,\ K_X = \begin{pmatrix} 0 & 0.45 & 0 \\ 0.15 & 0 & 0.35 \\ 0 & 0.05 & 0 \end{pmatrix}.$$

  where K_X influences how $\mathbf{X}$ diffuses across neighboring cells, with an asymmetry due to advection.
- $\mathbf{A}_t$:

$$\beta_0 = -1.0,\ \beta_1 \in \{0.0, 0.5, 1.0, 1.5, 2.0, 2.5, 3.0\},\ K_A = \frac{1}{16}\begin{pmatrix} 1.0 & 1.0 & 1.0 \\ 1.0 & 8.0 & 1.0 \\ 1.0 & 1.0 & 1.0 \end{pmatrix}.$$

- $\mathbf{Y}_t$:

$$\gamma_0 = 2.0,\ \gamma_1 = 1.5,\ \gamma_2 = 0.5,\ \gamma_3 = 0.5,\ K_{YX} = \frac{1}{16}\begin{pmatrix} 1.0 & 1.0 & 1.0 \\ 1.0 & 8.0 & 1.0 \\ 1.0 & 1.0 & 1.0 \end{pmatrix},\ K_{YA} = \frac{1}{16}\begin{pmatrix} 1.0 & 1.0 & 1.0 \\ 1.0 & 8.0 & 1.0 \\ 1.0 & 1.0 & 1.0 \end{pmatrix}.$$

We use $L = 5$ temporal lags for $\mathbf{X}$ and $\mathbf{Y}$, a seed of 42 for reproducibility. See Figure 4 for representative $t = 100$ snapshots of X_{100}, A_{100}, and Y_{101} under varying β_1.

For each β_1, we first generate a factual dataset of length $T = 200$ (*i.e.*, $\{(\mathbf{X}_t, \mathbf{A}_t, \mathbf{Y}_t)\}_{t=1}^{200}$). We then create $n_{\text{test}} = 50$ test histories of length $l_H = 10$. For each test history, we simulate 100 trajectories under a randomly chosen (yet fixed over the test data) counterfactual intervention of length $\tau = 10$, and average the outcomes at each step to approximate the true CAPOs. This procedure yields a final test set of shape $n_{\text{test}} \times (l_H + \tau + 1) \times 64 \times 64$, *i.e.*, $50 \times 21 \times 64 \times 64$.

Neural Architectures. The **GST-UNet** comprises a single ConvLSTM layer (hidden dimension 32), followed by a U-Net with channel sizes $16 \to 32 \to 64 \to 128 \to 256$. Its G-computation heads are shallow feed-forward networks operating on the final U-Net feature maps at each grid cell; both the U-Net's output dimension (d_h) and the G-head architecture (number of layers, hidden size) are treated as hyperparameters. The **UNet+** baseline uses the same ConvLSTM+U-Net backbone as GST-UNet but outputs a single channel ($d_h = 1$), omitting any G-computation. For direct comparison, we also implement **STCINet** (Ali et al., 2024) with an identical ConvLSTM+U-Net backbone, and retaining their original Latent Factor Model (LFM) details.

Training Details. We randomly initialize all model parameters (GST-UNet and baselines) with Xavier uniform weights (Glorot and Bengio, 2010). We use the Adam optimizer (Kingma, 2014) with an initial learning rate, halving it whenever the validation loss plateaus for a specified scheduler patience. To mitigate overfitting, we adopt early stopping when the validation loss fails to improve for a specified early stopping patience epochs. Validation uses 40 of the 190 training prefixes, and the total training is capped at 100 epochs. We tune the following hyperparameters: *(i)* batch size, learning rate, scheduler patience, and early stopping patience (common to all models); *(ii)* for GST-UNet, the curriculum period and learning rate for curriculum phases, the U-Net output dimension d_h, and the number and width of hidden layers in the feed-forward G-heads. Table 2 lists the hyperparameter ranges considered, with the values yielding the best validation performance in **bold**.

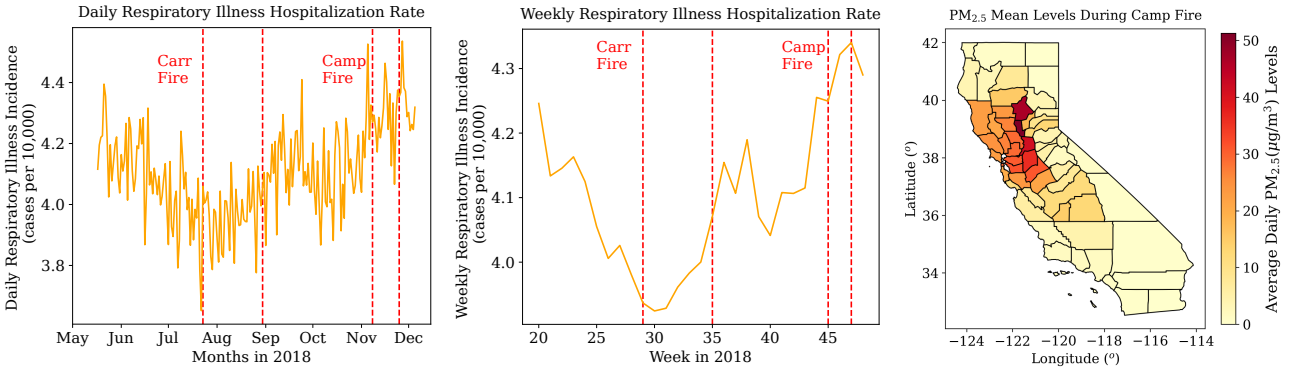

Figure 5. **(Left)** Daily respiratory illness incidence (cases per 10,000). **(Center)** Weekly aggregated incidence. **(Right)** Average daily $PM_{2.5}$ during the Camp Fire.

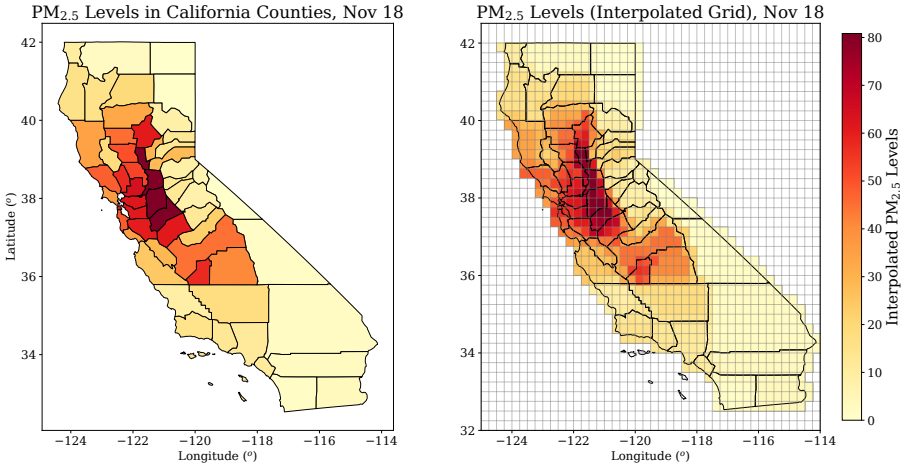

Figure 6. An example of county-level (*left*) vs. grid-interpolated (*right*) $PM_{2.5}$ levels on November 18 (during the Camp Fire). The grid interpolation produces a 40×44 lattice of area-weighted estimates aligned with our spatiotemporal framework.

Evaluation Procedure. We evaluate each model by averaging the root mean square error (RMSE) of the estimated CAPOs against ground truth across 50 test trajectories. Table 1 in the main text reports RMSE $\pm$ standard deviation for horizon lengths $\tau \in \{5,\ 10\}$ and $\beta_1 \in \{0, 0.5, 1.0, 1.5, 2.0, 2.5, 3.0\}$.

C.2. Wildfire Application

Data Preprocessing and Interpolation We utilize 2018 daily county-level $PM_{2.5}$, respiratory/cardiovascular hospitalizations, weather variables (temperature, precipitation, humidity, radiation, wind) from Letellier et al. (2025), along with population data from the California Department of Finance. Our study period spans weeks 20–48 (May 18–December 2, 2018), covering both the Carr and Camp fires. As illustrated in Figure 5, daily and weekly aggregated respiratory illness rates rise around these events, while $PM_{2.5}$ levels also surge during the Camp Fire.

To align with our spatiotemporal framework, we use `geopandas` (Jordahl et al., 2020) to interpolate county-level covariates, $PM_{2.5}$, and hospitalizations onto a latitude–longitude grid from 32°N to 42°N latitude and -125° to -114° longitude, at a resolution of $0.25°$. Each grid cell's values are an area-weighted average of the counties it intersects, yielding a 40×44 spatial lattice. We mask out non-California cells by setting them to zero, thus obtaining a consistent dataset for further analysis. As an example, Figure 6 illustrates how the raw county-level data compare to the interpolated grid for $PM_{2.5}$ on November 18.

Model Training and Validation We train GST-UNet with prediction horizon $\tau = 10$ days. The loss function is MSE with two key modifications: (1) we mask grid cells outside California's boundaries to exclude them from loss computation, and (2) we apply cell-specific weights proportional to the number of cells per county to prevent bias towards geographically larger counties. For validation and hyperparameter tuning, we use data from the first 50 days of the wildfire season. The

GST-UNet hyperparameters are: batch size $= 40$, learning rate $= 5 \times 10^{-4}$, scheduler patience $= 5$, early stopping patience $= 10$, curriculum period $= 5$, curriculum learning rate $= 5 \times 10^{-4}$, UNet output dimension $d_h = 16$, G-head hidden layer size $= 8$, and G-head layers $= 1$. Using this configuration, we generate counterfactual predictions for the Camp Fire peak period (November 8–17) by iteratively applying the trained model with increasing history lengths. We note that counties with populations below $20{,}000$–$30{,}000$ can yield unreliable incidence rate estimates, given baseline daily rates of approximately 4 cases per 10,000 individuals. In Figure 3, we denote these high-uncertainty counties with hatched markings.