Deep Poisson Case Time Series Model for Complex Environmental Exposure Modeling on Acute Health Outcomes

Keyu Li KEYU.LI@DUKE.EDU
Department of Electrical and Computer Engineering
Duke University
Durham, North Carolina, USA
Elaona Lemoto ELAONA.LEMOTO@DUKE.EDU
Department of Biostatistics and Bioinformatics
Duke University
Durham, North Carolina, USA
Zachary D. Calhoun ZACHARY.CALHOUN@DUKE.EDU
Department of Civil and Environmental Engineering
Duke University
Durham, North Carolina, USA
Charles T. Wood CHARLES.WOOD@DUKE.EDU
Department of Pediatrics
Duke University
Durham, North Carolina, USA
Nrupen A. Bhavsar NRUPEN.BHAVSAR@DUKE.EDU
Department of Surgery
Duke University
Durham, North Carolina, USA
David Carlson DAVID.CARLSON@DUKE.EDU
Department of Biostatistics and Bioinformatics
Duke University
Durham, North Carolina, USA

Abstract

Modern environmental health studies increasingly collect rich spatio-temporal exposure data, yet existing methods for assessing acute health effects still rely on low-dimensional summaries that discard informative structure. Self-matched designs, such as the case time series, effectively control for stable between-area confounding but do not readily extend to complex, high-dimensional exposure histories. To address this limitation, we introduce the Deep Poisson Case Time Series (DPCTS), a self-matched deep learning framework for area-level health outcomes that integrates within-unit stratification with end-to-end representation learning from complex exposures. DPCTS employs a neural exposure encoder trained with a Poisson objective, incorporates temporal context and multi-scale supervision to capture lagged effects across multiple resolutions, and supports tabular, image-based, and multimodal inputs. In addition, we propose novel evaluation metrics that isolate exposure-driven performance from between-area baseline differences by measuring within-

stratum fit, predictive accuracy, and discrimination. Across synthetic, semi-synthetic, and real-world experiments, DPCTS recovers nonlinear exposure-response relationships and fine-scale risk patterns more accurately than classical baselines. These findings establish that representation learning can be embedded within self-matched epidemiologic study designs to analyze complex exposure data while preserving the within-unit comparisons essential for studying short-term health risk.

1. Introduction

Acute environmental exposures can trigger immediate health impacts, often observed as short-term increases in health event counts (Rothman et al., 2008; Liu et al., 2019; Sun et al., 2021; Gariazzo et al., 2023). A common modeling setup is a (panel) time series problem: we observe sequences of outcomes and exposures over time, often across multiple units (e.g., individuals, neighborhoods), and aim to learn short-term exposure–outcome relationships from temporally indexed data (Bhaskaran et al., 2013). However, classical time-series models do not, by default, control for time-invariant unit effects when multiple units are modeled jointly (Bhaskaran et al., 2013). Without explicit unit fixed effects, persistent or slowly varying unit-specific risk factors, such as population composition, underlying health burden, or healthcare access, can confound exposure–outcome associations when multiple units are analyzed jointly (Wooldridge, 2010; Bartolucci et al., 2015).

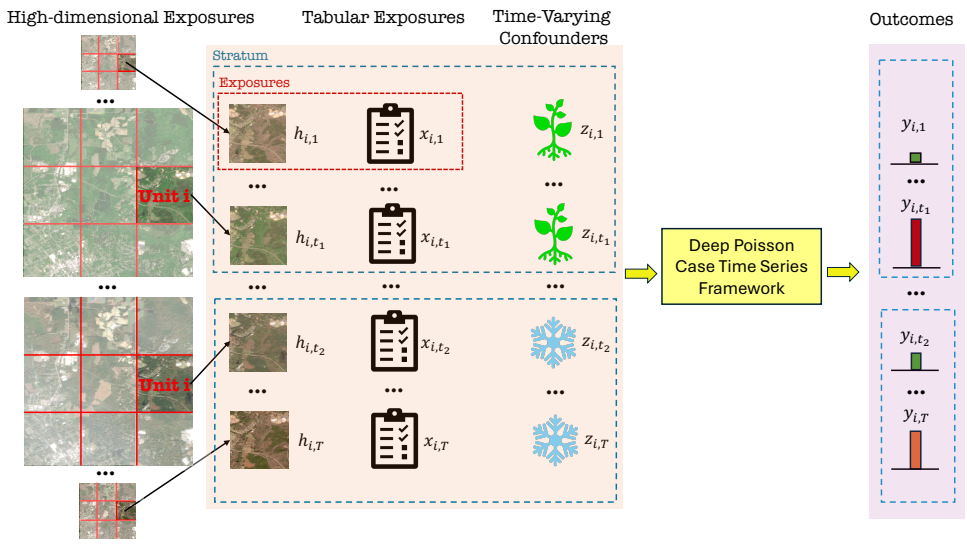

Figure 1: Schematic of the spatio-temporal data structure in DPCTS. For each unit $i = 1, \ldots, I$ and day $t = 1, \ldots, T$, we observe high-dimensional exposures h_{it} (e.g., images), tabular exposures x_{it}, and time-varying confounders z_{it}. Units are shown as regular grid cells for clarity, but in practice may be irregular geographic areas such as census block groups. For such units, h_{it} can be created by cropping or masking satellite imagery using boundary shapefiles. Time-varying confounders are shown as seasons for simplicity and can be used to create strata; see Section 3.1.

Self-matched designs such as time-stratified case-crossover address this via "control-by-design" (Maclure, 1991; Carracedo-Martínez et al., 2010), comparing exposures within

the same unit over time and conditioning out unit-specific intercepts. However, classical case-crossover approaches are limited in flexibly adjusting for time-varying confounders and scaling to complex, temporally structured, or high-dimensional exposures (Janes et al., 2005; Tobias et al., 2024; Gasparrini, 2021). The case time series (CTS) design bridges the first gap by combining within-unit self-matching with the longitudinal structure of time-series models (Gasparrini, 2021). CTS offers a flexible, robust and highly adaptable framework for transient risk estimation and can be adapted to areally aggregated data. However, existing CTS workflows are primarily designed for low-dimensional exposure time series (Gasparrini, 2011, 2021). Since human-engineered summaries of complex spatio-temporal objects (e.g., satellite image sequences or dense sensor networks) may discard informative structure (Himeur et al., 2022), a gap remains for analyzing high-dimensional spatio-temporal exposures while accounting for between-unit confounding.

Closing this gap would allow the use of complexf data sources (e.g., dense sensor networks for heat, pollution plumes, or satellite-derived wildfire smoke) and temporal history. We propose the Deep Poisson Case Time Series (DPCTS) model to fill this gap. DPCTS is a self-matched deep learning framework for acute-effect estimation with areal count outcomes under complex exposures. It retains self-matched identification via conditional Poisson modeling (Peng et al., 2006; Armstrong et al., 2014) while incorporating a neural exposure encoder trained end-to-end with the Poisson log-likelihood, enabling nonlinear exposure–outcome learning from complex spatio-temporal inputs. Figure 1 illustrates the data structure. DPCTS also uses both fine- and aggregated-scale supervision within one coherent objective, leveraging Poisson additivity to improve robustness under sparse health counts. Figure A1 in Appendix A compares DPCTS with the aforementioned methods.

Besides modeling challenges, conventional evaluation metrics can miss key model behavior in self-matched analyses, because apparent performance may be driven by between-unit baseline differences instead of the within-unit temporal variation that defines the estimand. Standard metrics rather than RMSE and AUC are inadequate in this setting, as they obscure exposure-driven effects and yield misleading model comparisons. To address this, we introduce three complementary evaluation metrics, each computed within strata and aggregated across units, that assess complementary aspects of model performance: Stratified Deviance R^2 for overall goodness-of-fit, Stratum-Normalized RMSE (snRMSE) for within-stratum predictive accuracy, and Multi-Rank AUC (MR-AUC) for within-stratum discrimination. Together, they provide assessments aligned with the self-matched study design.

Our main contributions are summarized as follows:

- A deep CTS model that integrates within-unit stratification with neural representation learning, enabling acute-effect analysis under multimodal, high-dimensional exposures.
- Multi-scale learning for transient risk modeling across resolutions.
- Evaluation tailored to self-matched count modeling, enabling comparisons of how well models capture exposure-driven variation rather than baseline fixed effects.

Generalizable Insights about Machine Learning in the Context of Healthcare

Our study highlights three lessons that extend beyond this application:

- First, in healthcare settings with substantial baseline heterogeneity across units—such as persistent differences in population characteristics, healthcare access, and socioeconomic conditions—self-matched ML design can preserve clinically meaningful identification while still enabling flexible representation learning.
- Second, modern exposure sources relevant to health, such as multi-source spatio-temporal fields, may contain predictive signal that is lost when reduced to hand-crafted summaries, motivating end-to-end models that operate directly on rich inputs.
- Third, evaluation for healthcare machine learning should align with the scientific estimand: in self-matched acute-effect studies, design-aware metrics are necessary to distinguish true recovery of within-unit exposure effects from spurious performance driven by between-unit baseline differences.

2. Related Work

2.1. High-dimensional Exposure Modeling

Work on high-dimensional exposure modeling in health has largely followed two lines. The first studies structured but still tabular exposure mixtures, using methods such as weighted quantile sum regression (Carrico et al., 2015; Czarnota et al., 2015), Bayesian kernel machine regression (Bobb et al., 2015), and their lagged extensions (Wilson et al., 2022) to capture correlation, nonlinearity, and time-varying mixture effects; these methods are valuable for chemical mixtures, but they typically assume a moderate number of hand-defined exposure variables rather than raw images or other richly structured inputs.

The second line has turned to machine learning and deep learning for the exposome, with reviews highlighting growing use of random forests, boosting, neural networks, and multimodal AI to handle complex, correlated environmental data (Oskar and Stingone, 2020); however, these studies also note persistent challenges around temporal structure, interpretability, heterogeneous data sources, and integration with epidemiologic study design (Isola et al., 2024). In environmental health specifically, deep learning from images has been proposed as a way to estimate multiple exposures jointly from satellite, street-view, and related geospatial data (Weichenthal et al., 2019), but the main emphasis has been exposure assessment or cross-sectional risk prediction rather than self-matched inference for acute count outcomes (Clark et al., 2025; Himeur et al., 2022). Our work builds on this literature by learning representations directly from high-dimensional exposure fields while preserving a self-matched framework for short-term health risk estimation.

2.2. Environmental Exposures on Health Outcomes

A large environmental-health literature has linked short-term exposures such as heat, air pollution (Holloway et al., 2021), and pollen to acute outcomes including mortality, hospital admissions, and emergency visits, traditionally using time-series or case-crossover designs (Bhaskaran et al., 2013). In parallel, high-dimensional exposure studies increasingly use remote sensing, street-view imagery, and other geospatial data to characterize neighborhoods and relate them to health outcomes (Weichenthal et al., 2019; Levy et al., 2021; Nguyen et al., 2019). Representative applications include deep learning from satellite imagery to estimate

county mortality (Levy et al., 2021), street-view based characterization of neighborhood features associated with chronic disease and premature mortality (Nguyen et al., 2019, 2021), satellite-image analysis linked to cardiometabolic disease prevalence (Chen et al., 2024), and combined satellite-plus-street-view models for neighborhood obesity Chen et al. (2025). Exposome studies have also begun combining environmental, molecular, and clinical data with machine learning to build risk scores (Guimbaud et al., 2024). However, most work remains cross-sectional, focused on chronic outcomes or exposure estimation, and rarely addresses *acute* short-term health effects from transient exposures. Our paper targets that gap by studying acute count outcomes from high-dimensional environmental exposures in a longitudinal, small-area, self-matched framework.

3. Methods

3.1. Problem Setup

In our analysis, a *unit* refers to a geographic area, such as a census block group or grid cell, as shown in Figure 1. Let units be indexed by $i = 1, \ldots, I$ and days by $t = 1, \ldots, T$. Following Gasparrini (2021), we partition each unit's time series into temporal strata indexed by $k = 1, \ldots, K$. Each stratum groups observations sharing a discrete temporal characteristic, such as day of week, calendar month, or their intersection. This stratification controls slower-varying or structured temporal confounding shared within strata. Defined separately for each unit, these strata eliminate all time-invariant confounding within unit i and stratum-invariant temporal confounding within stratum k by design.

Given this stratified structure, we define the observed data and outcome model. We observe outcome counts Y_{it} with tabular exposures $x_{it} \in \mathbb{R}^p$ (where p is the feature dimension) and high-dimensional exposures h_{it}. When h_{it} is an image, such as a satellite scene, we have $h_{it} \in \mathbb{R}^{C \times H \times W}$ with C spectral channels over a spatial grid of height H and width W. We denote the lag window by L, giving temporal tensors $\mathbf{X}_{it} = [x_{it}, x_{i,t-1}, \ldots, x_{i,t-L}]$ and $\mathbf{H}_{it} = [h_{it}, h_{i,t-1}, \ldots, h_{i,t-L}]$. Measurable time-varying confounders such as time or age are denoted as z_{it} for each (i, t). As in previous work (Gasparrini, 2021), for count-type outcomes Y_{it}, we assume a Poisson model,

$$Y_{it} \,\big|\, \mathbf{X}_{it}, \mathbf{H}_{it}, z_{it} \sim \text{Poisson}\big(\mu_{it}\big),$$

and define the canonical parameter via

$$\log \mu_{it} = \eta_{it}.$$

We target acute (short-lag) exposure effects while controlling for time-invariant factors.

3.2. Case Time Series (CTS)

CTS (Gasparrini, 2021) combines within-unit matching and control of temporal variation in risks. A standard CTS framework for counts can be written as

$$Y_{it} \sim \text{Poisson}(\mu_{it}), \tag{1}$$

$$\log \mu_{it} = \xi_{i(k)} + f\big(x_{it}, L\big) + s(t) + v(z_{it}), \tag{2}$$

where:

- $\xi_{i(k)}$ are unit-specific intercepts that contain baseline risks for stratum k;
- $f(\cdot)$ models the exposure–outcome relationship over lag window L. For lag $\ell \in \{0, \ldots, L\}$, it can represent the contribution of exposure at time $t - \ell$ to the outcome at time t, allowing a distributed lag structure;;
- $s(t)$ is a smooth calendar-time function shared across units (long-term/seasonal control) at different timescales;
- $v(z_{it})$ models other measurable time-varying confounders z_{it} (e.g., age, holidays, time since a specific intervention).

In the CTS framework, the shared calendar control $s(t)$ removes population-wide temporal structure (e.g., long-term trends, seasonality); $\xi_{i(k)}$ removes time-invariant confounding within unit i, and absorbs stratum-specific level shifts (e.g., "Mondays in area i"); $v(z_{it})$ adjusts for observed within-unit, time-varying confounders; and $f(x_{it}, L)$ represents the object of interest, the log-relative risk of the exposure history.

For Poisson count outcomes, i represents an area, and conditional Poisson regression with a log link defines the estimators. To preserve the self-matched identification logic of the CTS design, several assumptions are required; these are summarized in Appendix B.

4. Deep Poisson Case Time Series

We propose a general framework for capturing relationships between complex exposures and count-type outcomes, called the *Deep Poisson Case Time Series* (DPCTS). DPCTS also enables multi-scale learning by supervising the model at multiple granularities.

For areal count outcomes, let $\mu_{it}^{(0)} > 0$ be a fixed baseline mean. We learn in log space,

$$\log \mu_{it} - \log \mu_{it}^{(0)} = r_{it}, \tag{3}$$

where r_{it} captures deviations from the baseline. This residual formulation is well-motivated for areal count models. First, it corresponds to the standard multiplicative decomposition of a Poisson log-linear model (Coxe et al., 2009), where the baseline term $\log \mu_{it}^{(0)}$ captures stable spatial heterogeneity and r_{it} represents short-term, exposure-driven relative risk changes (Wakefield, 2007). This aligns with the conditional likelihood structure of existing self-matched design methods (Lumley and Levy, 2000; Gasparrini, 2021). Second, separating the baseline ensures $\mu_{it} > 0$ and helps prevent the model from confounding signals in high-dimensional exposures with inherent spatial baseline trends, thereby improving identifiability and stability of exposure-effect estimation, which is particularly important when large spatial differences in baseline rates dominate outcome risk (Wilson and Wakefield, 2018).

Let $e_{it} = (x_{it}; h_{it})$ denote the exposure features (possibly including tabular and high-dimensional components). We define the exposure history

$$E_{i,t-L:t} = \big(e_{it}, \ldots, e_{i,t-L}\big),$$

and let $\mathcal{F}_\theta(\cdot)$ be a sequence summary produced by any chosen instantiation, such as a transformer. Then

$$r_{it} = \mathcal{F}_\theta\big(E_{i,t-L:t}, c_{it}, z_{it}\big), \tag{4}$$

where c_{it} denotes optional context (e.g., encodings, dates, missingness markers). Depending on the network, this mechanism can handle high-dimensional or multimodal exposures across diverse data structures. For example, deep architectures can learn DLNM-like lag shapes (Gasparrini et al., 2010).

4.1. Transformer Network Structure

While $\mathcal{F}_\theta(\cdot)$ can be chosen from a variety of structures, for concreteness, we show the realization of a transformer where the network is composed of three different parts. For the first part, we form fused embeddings u_{it} from the exposures and optional context,

$$u_{it} = \begin{bmatrix} \phi_e(e_{it}) \\ \phi_c(c_{it}) \end{bmatrix}, \tag{5}$$

where $\phi_e(\cdot)$ is typically a deep neural network and $\phi_c(\cdot)$ embeds context.

The second part focuses on handling the sequence of information. Let $\mathcal{T} \subseteq \{t-L, \ldots, t\}$ be the set of valid time steps used in a given step. An attention encoder then produces

$$d_{it} = \left[\mathrm{Enc}_\theta \left(\left([u_{i\tau}; z_{i\tau}] \right)_{\tau \in \mathcal{T}_{it}} \right) \right]_t, \tag{6}$$

where a transformer block (Vaswani et al., 2017) takes the ordered sequence of exposure, context, and covariate embeddings and returns temporal representations, from which we use the output corresponding to the current time point t as d_{it}. We can impose causal attention masks so that only the current and past time points affect the prediction at time t.

Finally, the value used in our predictions is obtained via a learned MLP head ψ,

$$r_{it} = \psi(d_{it}). \tag{7}$$

These components together give the structure of $\mathcal{F}_\theta(\cdot)$. Similar networks can be constructed from GRUs, LSTMs, or simple neural networks.

4.2. Training Objectives and Inference

Poisson Negative Log-likelihood Loss. Let $Y_{it} \sim \mathrm{Poisson}(\mu_{it})$ with $\log \mu_{it} = \log \mu_{it}^{(0)} + r_{it}$. First, we define a Poisson negative log-likelihood (NLL) loss to encourage the model head to predict counts that match observed y_{it},

$$\mathcal{L}_{\mathrm{nll}} = \tfrac{1}{N} \textstyle\sum_{i,t} \left[\mu_{it} - y_{it} \log(\mu_{it}) \right], \tag{8}$$

where N is the total number of observed (i,t) pairs.

Multi-scale Poisson Aggregation Loss. To improve calibration both at the observation level and after temporal aggregation, we introduce a multi-scale learning objective. This is motivated by two practical considerations. First, daily healthcare counts can be noisy at fine spatial and temporal resolution, making single-scale supervision unstable. Second, predictive accuracy after aggregation over clinically meaningful windows (e.g., days or weeks) is often important. We therefore add an aggregation loss after summing outcomes and predictions within predefined groups g, such as calendar-based windows at different temporal resolutions.

This construction is principled under the additivity of conditionally independent Poisson variables (Serfozo, 1972). Specifically, for each group g, we define

$$Y_g = \sum_{(i,t)\in g} y_{it}, \qquad \hat{Y}_g = \sum_{(i,t)\in g} \hat{\mu}_{it}. \tag{9}$$

We then apply a Poisson deviance loss at the group level:

$$\mathcal{L}_{\text{agg}} = 2 \sum_g \left[Y_g \log\left(\frac{Y_g}{\hat{Y}_g}\right) - \left(Y_g - \hat{Y}_g\right) \right]. \tag{10}$$

By optimizing the row-level Poisson objective with this aggregation loss, the model is encouraged to make predictions accurate both locally and after temporal aggregation.

Areal Classification Loss. For areal cases, we add an optional cross-entropy classification loss for unit ID prediction (e.g., census block group ID). This auxiliary area-ID loss acts as a lightweight regularizer, encouraging embeddings to preserve stable, area-specific exposure structure while complementing the Poisson objective for short-term variation. Such multi-task and pretext losses often improve representation quality and generalization (Caruana, 1997; Doersch et al., 2015). Let there be B areas, with $p_i \in \Delta^{B-1}$ denoting predicted class probabilities for area i. Letting b_i be the true area index, the auxiliary loss is

$$\mathcal{L}_{\text{area}} = -\frac{1}{I} \sum_{i=1}^{I} \log p_{i,b_i}. \tag{11}$$

We also add regularization terms into our training objectives. A residual amplitude penalty

$$\mathcal{R}_{\text{res}} = \lambda_{\text{res}} \sum_{i,t} r_{it}^2 \tag{12}$$

is used to discourage the residual head from absorbing baseline effects, which stabilizes training and improves identifiability of the baseline vs. residual split. The total loss is calculated using the Poisson NLL and a weighted sum of the remaining loss terms to balance the trade-off between observation-level fit, daily calibration, and regularization:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{nll}} + \alpha_{\text{agg}} \mathcal{L}_{\text{agg}} + \alpha_{\text{area}} \mathcal{L}_{\text{area}} + \mathcal{R}_{\text{res}}. \tag{13}$$

DPCTS Assumptions. While our DPCTS framework employs a novel deep learning architecture, it builds on the causal identification assumptions of the classical CTS model and additionally invokes the causal and modeling assumptions detailed in Appendix B.

Post-hoc Affine Calibration. Deep neural networks can be poorly calibrated (Guo et al., 2017; Dheur and Taieb, 2024). We thus conduct post-hoc calibration (Niculescu-Mizil and Caruana, 2005; Kuleshov et al., 2018) by learning to rescale predictions with an affine transformation on validation data,

$$\begin{aligned} \mu_{it}^{\text{cal}}(\alpha, \beta) &= \mu_{it}^{(0)} \exp(\alpha + \beta r_{it}), \\ (\hat{\alpha}, \hat{\beta}) &= \arg\min_{\alpha,\, \beta \geq 0} \sum_{(i,t)\in \text{VAL}} \left[\mu_{it}^{\text{cal}}(\alpha, \beta) - y_{it} \log \mu_{it}^{\text{cal}}(\alpha, \beta) \right] + \lambda_{\text{cal}} (\beta - 1)^2. \end{aligned} \tag{14}$$

with a mild ridge prior on β and bounded α with L-BFGS (Liu and Nocedal, 1989; Byrd et al., 1995). At test time we freeze these fitted parameters $(\hat{\alpha}, \hat{\beta})$ and compute

$$\log \tilde{\mu}_{it} = \log \mu_{i,t}^{(0)} + \hat{\alpha} + \hat{\beta}\, r_{i,t}. \tag{15}$$

We then use $\tilde{\mu}_{it}$ for evaluation.

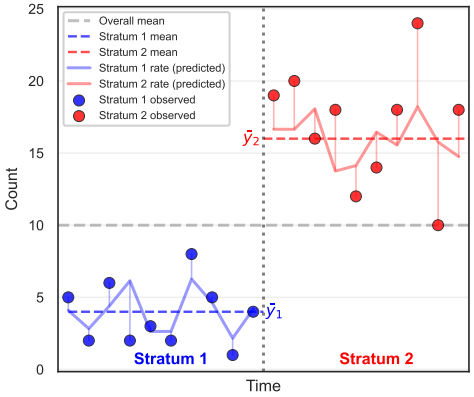

Figure 2: Illustration of Stratum-Normalized RMSE (snRMSE). Instead of computing global RMSE, we measure prediction error within each stratum and normalize it by the stratum mean. The same stratification principle is used for snRMSE, Stratified Deviance R^2, and Multi-Rank AUC, so evaluation focuses on exposure-driven temporal variation rather than between-unit baseline differences.

5. Evaluation

We evaluate performance using three complementary metrics aligned with the self-matched data structure. Each is computed within strata and aggregated across units, so evaluation reflects exposure-driven temporal variation rather than between-unit baseline differences. Together, they assess goodness-of-fit, within-stratum predictive accuracy, and within-stratum discrimination for fair model comparison.

Stratum-Normalized RMSE (snRMSE): Within-Stratum Predictive Accuracy. This metric assesses the magnitude of prediction error within each stratum, normalized to that stratum's own outcome scale to ensure equitable comparison. We illustrate the idea in Figure 2, and the same concept is applied in all our proposed metrics.

For each stratum k, define the stratum-specific RMSE as

$$\mathrm{RMSE}_k = \sqrt{\frac{1}{T_k}\sum_{t=1}^{T_k}\left(y_{kt} - \hat{\mu}_{kt}\right)^2}, \tag{16}$$

where T_k denotes the number of observation days in stratum k. Let $\bar{y}_k$ be the mean count in stratum k. The normalized RMSE for stratum k is

$$\mathrm{nRMSE}_k = \frac{\mathrm{RMSE}_k}{\bar{y}_k}, \tag{17}$$

which is defined only for strata with $\bar{y}_k > 0$; strata with $\bar{y}_k = 0$ are degenerate for this normalized metric and are excluded from the snRMSE calculation. The overall snRMSE is then obtained by averaging across the remaining strata:

$$\mathrm{snRMSE} = \frac{1}{K}\sum_{k=1}^{K}\mathrm{nRMSE}_k. \tag{18}$$

This averaging ensures each stratum contributes equally, and that error is interpreted relative to local outcome levels, providing a direct measure of relative predictive error within strata for practical prediction tasks within identifiable units.

Stratified Deviance R^2: Overall Goodness-of-Fit. This metric quantifies the proportion of explainable within-stratum variation captured by the model's exposure terms. While deviance is a likelihood-based measure for Poisson models (McCullagh, 2019), its raw

value is not standardized. Standard Pseudo-R^2 measures (Cameron and Windmeijer, 1996) often use an *unstratified* null model, which is suboptimal when the logical null includes stratum fixed effects and can misrepresent explainable variation. Our contribution is not deviance-based R^2 itself, which connects to prior local deviance measures (Di Mari et al., 2023), but its adaptation to fixed, pre-specified self-matched strata, where the stratum mean serves as the design-aligned null. To adapt pseudo-R^2 to the stratified case, let y_{kt} denote the observed count in stratum $k = 1, \ldots, K$ at time $t = 1, \ldots, T_k$, and let $\hat{\mu}_{kt}$ be the corresponding fitted mean. The scaled Poisson deviance is

$$D\big(y, \hat{\mu}\big) = 2 \sum_{k=1}^{K} \sum_{t=1}^{T_k} \left[y_{kt} \log\left(\frac{y_{kt}}{\hat{\mu}_{kt}}\right) - (y_{kt} - \hat{\mu}_{kt}) \right], \tag{19}$$

with the convention that $y_{kt} \log(y_{kt}/\hat{\mu}_{kt}) = 0$ when $y_{kt} = 0$.

The null model contains only the stratum-specific intercepts (i.e., the self-matched baseline), so that

$$\tilde{\mu}_{kt} = \bar{y}_k = \tfrac{1}{T_k} \textstyle\sum_{t=1}^{T_k} y_{kt}, \tag{20}$$

and its deviance is $D\big(y, \tilde{\mu}\big)$. Stratified Deviance R^2 is then defined as

$$R^2_{\text{sdev}} = 1 - \frac{D\big(y,\hat{\mu}\big)}{D\big(y,\tilde{\mu}\big)}. \tag{21}$$

R^2_{sdev} aligns with the model's Poisson likelihood. It is bounded above by 1 for perfect fit; 0 indicates no improvement over stratum means, and negative values indicate worse fit. Due to inherent Poisson uncertainty, even the ground truth model will not achieve 1, as highlighted in the results.

Multi-Rank AUC (MR-AUC): Within-Stratum Discrimination. This metric evaluates the model's ability to correctly rank the relative risk of different time points within the same stratum. For each stratum k, consider all unordered time pairs (t, t') with $t \neq t'$. For each pair, define a label

$$z_{ktt'} = \begin{cases} 1, & \text{if } y_{kt} > y_{kt'}, \\ 0, & \text{if } y_{kt} < y_{kt'}, \end{cases} \tag{22}$$

and a corresponding score based on the model's predictions:

$$s_{ktt'} = \hat{\mu}_{kt} - \hat{\mu}_{kt'}. \tag{23}$$

Pairs with $y_{kt} = y_{kt'}$ are treated as ties.

Pooling all such pairs across strata yields a collection $\{(z_q, s_q)\}_{q=1}^{Q}$, where each q indexes a within-stratum comparison. The MR-AUC is then defined as the usual area under the ROC curve computed from these labels and scores:

$$\text{MR-AUC} = \text{AUC}\big(\{(z_q, s_q)\}_{q=1}^{Q}\big). \tag{24}$$

An MR-AUC of 0.5 indicates no discriminative ability beyond random ordering within a stratum, while a score of 1.0 signifies perfect within-stratum ranking.

Collectively, Stratified Deviance R^2, snRMSE, and MR-AUC provide a comprehensive assessment of a model's fit, calibration, and discriminatory power specifically for the within-stratum variation central to self-matched designs.

6. Experiment Results

We evaluate models across four scenarios: (i) nonlinear exposure–outcome relationships, (ii) high-dimensional longitudinal inputs, (iii) multimodal exposures, and (iv) a real-world case study. All scenarios use geographically aggregated count outcomes (e.g., counts summed per census block group), reflecting standard epidemiologic practice where outcomes are often reported as spatial aggregates rather than individual records. Results for scenario (i) are reported in Appendix E.1, with full details for all scenarios in Appendix D.

6.1. Modeling High-Dimensional and Temporal Exposures

6.1.1. Experiment Setup and Model Performance

We evaluate DPCTS using two image-based semi-synthetic datasets. The first is a semi-synthetic dataset based on the MNIST digit dataset (LeCun et al., 2010). The second is based on a real-world Sentinel-2 satellite-image dataset (Agency, 2021), which is known to capture environmental quantities, and uses simulated events designed to mimic counts that you could get from electronic medical records at a feasible scale (e.g., census block groups).

For each dataset, we evaluate whether DPCTS can learn exposure-outcome relationships for high-dimensional exposures such as images, under (i) *image-only* and (ii) *image time series* settings, to illustrate our method's versatility. Details and results on the image-only setting are included in Appendix E.2. For the image-time-series setting, we include five comparison models: Linear-CTS; Spline-CTS; DPCTS-w/o, which uses only current exposures; Neural-Poisson, which removes the stratum-specific baseline while using the same encoder as DPCTS; and DPCTS-engineered, which uses engineered/PCA inputs while retaining the DPCTS residual objective. As the baseline methods cannot directly use imagery, all CTS methods employ engineered image statistics and PCA features as their exposure representations. We report the same self-matched metrics as elsewhere, with details deferred to Appendix D.2. Results are summarized in Table 1. Additional robustness analyses, including loss-weight sensitivity and learned affine calibration parameters, are reported in Appendix E.5 and Appendix E.6.

We further conduct an ablation study to assess the contribution of different loss components and the post-hoc affine calibration step to overall performance; see Appendix E.3.

MNIST. We first create a semi-synthetic MNIST digit dataset by simulating block-by-day outcomes in which the latent risk depends on image-derived signals while MNIST digits are used as daily exposures. For each block $i \in \{1, \dots, 200\}$ and day t over three years, we sample an MNIST image $I_{i,t}$ and generate counts

$$Y_{i,t} \sim \text{Poisson}(\lambda_{i,t}) \tag{25}$$

$$\log \lambda_{i,t} = \alpha_i + \gamma_{s(t)} + f(h_{i,t}) + \phi \log \lambda_{i,t-1} + \sum_{\ell \in \mathcal{L}} \eta_\ell \, f(h_{i,t-\ell}) \tag{26}$$

where α_i is a block fixed effect and $\gamma_{s(t)}$ is a shared calendar effect (day-of-week). We consider two MNIST settings. (i) Image-only: $\phi = 0$ and $\mathcal{L} = \emptyset$, isolating cross-sectional learning from pixels. (ii) Image time series: $\phi > 0$ and $\mathcal{L} = \{1, \dots, 7\}$, inducing temporal dependence through a lagged image signal.

Table 1: Image time-series exposure results on MNIST and Sentinel-2 datasets. All values are averages over 5 independent runs. (GT = ground truth model)

Dataset	Model	↑ Stratified Dev R^2	↓ snRMSE(×1e-3)	↑ MR-AUC
MNIST	DPCTS	**0.239** ± 0.004	**1.663** ± 0.007	**0.732** ± 0.002
	DPCTS-w/o	0.211 ± 0.003	1.689 ± 0.006	0.719 ± 0.002
	DPCTS-eng.	0.229 ± 0.004	1.670 ± 0.006	0.728 ± 0.002
	Neural-Poisson	0.228 ± 0.025	1.669 ± 0.021	0.726 ± 0.007
	Linear-CTS	0.208 ± 0.008	1.692 ± 0.009	0.714 ± 0.003
	Spline-CTS	0.224 ± 0.010	1.672 ± 0.012	0.725 ± 0.004
	GT	0.247	1.655	0.737
Sentinel-2	DPCTS	**0.842** ± 0.004	**0.721** ± 0.006	**0.926** ± 0.002
	DPCTS-w/o	0.824 ± 0.005	0.733 ± 0.007	0.906 ± 0.003
	DPCTS-eng.	0.835 ± 0.004	0.727 ± 0.006	0.923 ± 0.002
	Neural-Poisson	0.833 ± 0.018	0.728 ± 0.016	0.921 ± 0.006
	Linear-CTS	0.822 ± 0.006	0.735 ± 0.008	0.904 ± 0.003
	Spline-CTS	0.827 ± 0.005	0.730 ± 0.007	0.919 ± 0.003
	GT	0.853	0.712	0.931

Table 1 shows that DPCTS outperforms comparison models. In the image time-series setting, both DPCTS and Spline-CTS outperform DPCTS-w/o, highlighting the importance of modeling full exposure history when temporal dependence, such as autocorrelation and lagged effects, is present. The Neural-Poisson and DPCTS-engineered ablations further separate these improvements, indicating that both representation learning and the CTS-aligned residual framework contribute to DPCTS performance.

Sentinel-2. To demonstrate effectiveness for high-dimensional fine-scale exposures, we use Sentinel-2 Level-2A surface reflectance data, which provides high-resolution (10–60 m), multi-spectral imagery with frequent revisit. We downloaded available scenes over Durham County, NC, for 2023–2024, represented as six channels: four surface-reflectance bands (B2, B3, B4, B8) and two auxiliary maps (aerosol optical thickness and water vapour). After tile creation and missing-value imputation, each channel is standardized using training-set statistics. An example tile is shown in Figure D2 in Appendix D.2.3. The same setup and preprocessing are used in Sections 6.2 and 6.3. We then simulate count outcomes as in the MNIST semi-synthetic experiment and train DPCTS and comparison models. As shown in Table 1, DPCTS outperforms all comparison models. Neural-Poisson and DPCTS-engineered show the same pattern on satellite imagery, supporting that gains are not solely due to the neural encoder or residual formulation alone. The gap between DPCTS-w/o and Spline-CTS further underscores the importance of longitudinal exposure modeling. We also evaluate different missingness imputation settings in Appendix E.4.

6.2. Multimodal Exposures

We evaluate multimodal exposures using a semi-synthetic Sentinel-2 dataset with synthetic longitudinal tabular exposures (temperature history) and real-world longitudinal high-dimensional exposures (satellite images). Each observation is a (block, t) pair with (i) current temperature and lags; (ii) a six-channel Sentinel-2 tile for the same block and date; and (iii) cyclical day-of-year features as time-varying confounders.

We compare against Linear-, Spline-, and DLNM-CTS baselines using engineered image statistics and PCA features, and DPCTS-w/o, which removes the temporal component.

Details are in Appendix D.3. As shown in Figure 3, DPCTS shows a clear advantage over comparison models and remains stable across independent runs. Among baselines, DLNM-CTS performs best, reflecting its ability to capture cumulative exposure effects. The lower performance of DPCTS-w/o further underscores the importance of temporal modeling in our framework.

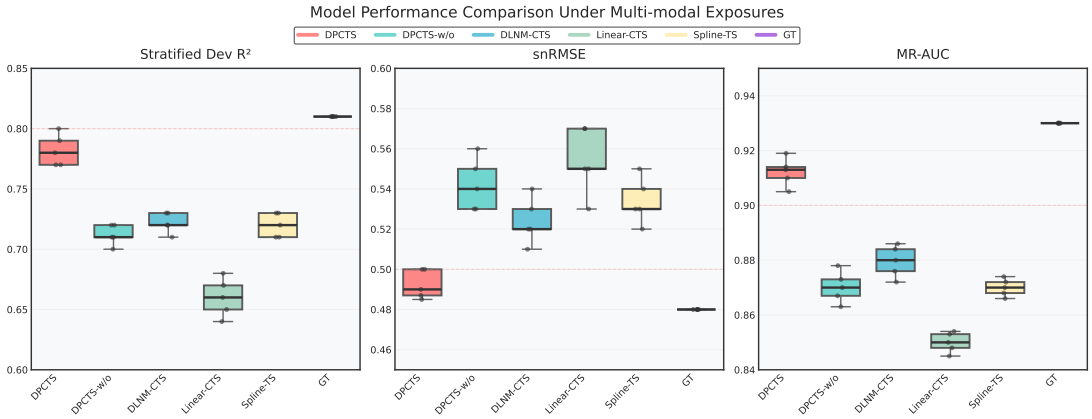

Figure 3: Performance of different models on multimodal exposures. The figure compares three metrics across model architectures over five independent runs, demonstrating our proposed model's effectiveness in learning from multimodal exposures (GT = ground truth model that uses true μ for predicting outcomes). The red dashed line in each panel is a fixed horizontal reference line added for visual comparison (0.8 for Stratified Dev R^2, 0.5 for snRMSE, and 0.9 for MR-AUC.

6.3. A Case Study on Real-World Data

6.3.1. Data Description

Exposures. We collected multimodal environmental exposure data for Durham County, North Carolina, from 2021–2023, including satellite imagery and temperature measurements:

- **Sentinel-2 satellite imagery:** Following the preprocessing pipeline in Section 6.1, we downloaded and processed Sentinel-2 imagery and partitioned each scene into spatial tiles. To construct block-group-level image sequences, we obtained year-matched TIGER/Line shapefiles for Durham County from the U.S. Census Bureau (United States Census Bureau, 2025). For each block group, we computed its geographic centroid and assigned the Sentinel-2 tile with the closest center. The assigned tiles were organized into temporal sequences, yielding a consistent block-group-to-tile mapping over time.

- **Temperature tabular data:** We collected block-group-level hourly air temperature data in Durham County from May–September 2021–2023 using Weather Underground observations. Observations were interpolated to 500m resolution using Gaussian process regression, with block-group temperature calculated as the mean of gridded

points within each area and hour. These data were organized into time-series tabular format aligned with the spatial units.

Together, these multimodal data provide rich spatio-temporal environmental exposures for modeling associations with the health outcomes described below.

Health Outcome. The study population and outcomes were obtained from Duke University Health System (DUHS) EHR, the primary healthcare provider in Durham County, NC. We considered cardiovascular disease emergency department visits from May–September 2021–2023 as the outcome of interest. Relevant encounters included visit types ED, EI, or OS with at least one ICD-10 diagnosis code between I10 and I79, yielding 19,140 encounters. Patient addresses were geocoded and linked to block-group-level FIPS codes, then aggregated to daily block-group counts as our areal health outcome.

6.3.2. Experiment Details and Results

We defined temporal strata by block group, calendar month, and day-of-week. We did not split strata by year; instead, corresponding weekdays within the same calendar month were pooled across years to increase observations per stratum and improve stability in this sparse count setting. For internal validation, we repeated a month-level hold-out procedure three times, each time holding out one randomly selected month and training on the remaining months.

We compared DPCTS against two baselines: (i) CTS models with linear and spline exposure functions using engineered image statistics and PCA features; and (ii) DPCTS-w/o, which removes the temporal component. All models used the same design-aware metrics, with results in Table 2. DPCTS achieved the best performance across all metrics, suggesting better capture of nonlinear relationships between complex environmental exposures and acute cardiovascular outcomes. We also provide a counterfactual temperature-response analysis in Appendix E.7.

Metric values should be interpreted in light of prediction granularity. We predict cardiovascular encounter counts at the block-group-by-day level, where sparse and variable daily counts make prediction difficult and lead to modest absolute performance. snRMSE is further affected by normalization: because it divides error by the stratum-specific mean and averages equally across strata, it can be large when many strata have low baseline counts. Thus, here snRMSE is most useful for relative model comparison rather than raw-magnitude interpretation.

To evaluate performance at a more stable temporal resolution, we aggregated daily predictions and observed counts to the block-group-by-week level without retraining, using Poisson additivity. This weekly evaluation complements the daily self-matched metrics, since aggregation smooths sparse stochastic variation but does not preserve original day-of-week strata. As shown in Table 3, DPCTS again performs best, suggesting it captures exposure-related signal that is clearer at a less sparse temporal scale.

7. Discussion

Technical, Clinical, and Social Implications. This work shows that flexible deep models for high-dimensional spatio-temporal exposures can be integrated with self-matched

Table 2: Predictive performance on real-world cardiovascular ED visits using satellite imagery and daily temperature. Results are averaged over three repeated month-level hold-out runs.

Model	Stratified Dev R^2 ↑	snRMSE ↓	MR-AUC ↑
DPCTS	**0.072** ± 0.011	**18.13** ± 1.423	**0.591** ± 0.012
DPCTS-w/o	0.056 ± 0.010	20.79 ± 1.652	0.567 ± 0.013
Linear-CTS	0.033 ± 0.009	24.15 ± 1.888	0.535 ± 0.014
Spline-CTS	0.047 ± 0.010	21.92 ± 1.715	0.549 ± 0.013

Model	Weekly Dev R^2 ↑	Weekly nRMSE ↓
DPCTS	**0.318** ± 0.041	**2.85** ± 0.247
DPCTS-w/o	0.264 ± 0.044	3.21 ± 0.273
Linear-CTS	0.147 ± 0.043	3.92 ± 0.319
Spline-CTS	0.196 ± 0.043	3.56 ± 0.350

Table 3: Weekly aggregated performance on the real-world cardiovascular ED dataset. Daily predictions and counts were aggregated to block-group-by-week level without retraining.

epidemiologic designs, improving expressiveness while remaining aligned with the causal and clinical structure of the problem.

Beyond predictive performance, our study shows how ML methods can be adapted to healthcare settings with complex spatio-temporal exposures and structured confounding. Combining case time series with deep exposure encoders improves modeling flexibility while preserving an epidemiologically meaningful design principle: comparisons are made within matched strata, reducing the influence of stable between-unit differences. This matters because apparent predictive gains may otherwise reflect persistent background differences across neighborhoods or populations rather than clinically meaningful short-term exposure effects. Our findings suggest that health ML models should be evaluated not only by predictive accuracy, but also by whether their structure aligns with the scientific question and data-generating process.

These properties are relevant for environmental epidemiology and public health applications involving complex, multimodal exposures and acute areal count outcomes. By learning directly from raw spatio-temporal inputs such as satellite imagery, DPCTS offers a flexible alternative to pipelines that depend on separately estimated exposure products, often calibrated using sparse monitoring networks. In principle, this could support earlier, localized identification of health risks during heat waves or pollution episodes, enabling more targeted warnings, surveillance, and resource allocation. More broadly, fine-scale risk estimation while accounting for stable area-level heterogeneity may support more precise and equitable public health responses.

Relationship to Causal Methods. Many deep causal estimators target low-dimensional treatments, typically binary or scalar-valued, and rely on covariate adjustment in cross-sectional or i.i.d. data (Johansson et al., 2016; Shalit et al., 2017; Nie et al., 2021). Recent methods can incorporate high-dimensional covariates such as images (Jiang et al., 2023), but generally do not model the longitudinal, self-matched structure needed to identify acute effects from within-unit temporal variation. In contrast, DPCTS is built around a self-matched design for high-dimensional, temporally structured exposures.

This design-based perspective places our work within causal methods that identify effects through study design rather than rich confounder modeling. Synthetic control targets sustained interventions in few treated units using untreated donors (Abadie and Gardeazabal, 2003; Abadie et al., 2010), but is not suited to the acute, continuously varying exposures and many-unit settings considered here. DPCTS instead builds on self-matched longitudinal designs such as case-crossover and case time series (Maclure, 1991; Gasparrini, 2021; Gasparrini et al., 2022). Combined with flexible representation learning, this enables exposure–response estimation from complex, high-dimensional modalities while preserving self-matching.

Limitations. Our work has several limitations. First, DPCTS assumes within-unit exposure variation is exogenous conditional on observed time-varying confounders; unmeasured time-varying factors could still bias estimates. Second, although the model scales to moderately high-dimensional inputs, computational demands grow with encoder complexity and the number of strata, challenging applications with millions of fine spatial units. Finally, the framework prefers complete exposure series; missing exposure data, especially systematic missingness (e.g., cloud cover), requires careful imputation or modeling adjustments. Future work could address these through improved missing-data methods, distributed training, and instrumental variable designs.

8. Conclusion

We introduce DPCTS, a self-matched deep learning framework for modeling acute count outcomes under complex spatio-temporal and multimodal exposures. DPCTS combines Case Time Series structure with neural exposure encoding and multi-scale supervision, enabling flexible exposure-response modeling while preserving within-unit comparisons. We also propose design-aware evaluation metrics for self-matched settings. Across synthetic, semi-synthetic, and real-world experiments, DPCTS improves recovery of exposure-driven risk patterns over existing approaches, supporting its use for modern environmental health data.

Acknowledgments. This project was supported by a grant through The Duke Endowment.

References

Alberto Abadie and Javier Gardeazabal. The economic costs of conflict: A case study of the basque country. *American Economic Review*, 93(1):113–132, 2003.

Alberto Abadie, Alexis Diamond, and Jens Hainmueller. Synthetic control methods for comparative case studies: Estimating the effect of california's tobacco control program. *Journal of the American Statistical Association*, 105(490):493–505, 2010.

European Space Agency. Copernicus Sentinel-2 msi level-2a boa reflectance product. collection 1, 2021. URL https://doi.org/10.5270/S2_-znk9xsj.

Ben G Armstrong, Antonio Gasparrini, and Aurelio Tobias. Conditional poisson models: a flexible alternative to conditional logistic case cross-over analysis. *BMC medical research methodology*, 14(1):122, 2014.

Francesco Bartolucci, Federico Belotti, and Franco Peracchi. Testing for time-invariant unobserved heterogeneity in generalized linear models for panel data. *Journal of Econometrics*, 184(1):111–123, 2015.

Krishnan Bhaskaran, Antonio Gasparrini, Shakoor Hajat, Liam Smeeth, and Ben Armstrong. Time series regression studies in environmental epidemiology. *International journal of epidemiology*, 42(4):1187–1195, 2013.

Jennifer F Bobb, Linda Valeri, Birgit Claus Henn, David C Christiani, Robert O Wright, Maitreyi Mazumdar, John J Godleski, and Brent A Coull. Bayesian kernel machine regression for estimating the health effects of multi-pollutant mixtures. *Biostatistics*, 16 (3):493–508, 2015.

Richard H Byrd, Peihuang Lu, Jorge Nocedal, and Ciyou Zhu. A limited memory algorithm for bound constrained optimization. *SIAM Journal on scientific computing*, 16(5):1190–1208, 1995.

A. Colin Cameron and Frank A. G. Windmeijer. R-squared measures for count data regression models with applications to health-care utilization. *Journal of Business & Economic Statistics*, 14(2):209–220, 1996.

Eduardo Carracedo-Martínez, Margarita Taracido, Aurelio Tobias, Marc Saez, and Adolfo Figueiras. Case-crossover analysis of air pollution health effects: a systematic review of methodology and application. *Environmental health perspectives*, 118(8):1173, 2010.

Caroline Carrico, Chris Gennings, David C Wheeler, and Pam Factor-Litvak. Characterization of weighted quantile sum regression for highly correlated data in a risk analysis setting. *Journal of agricultural, biological, and environmental statistics*, 20(1):100–120, 2015.

Rich Caruana. Multitask learning. *Machine learning*, 28(1):41–75, 1997.

Zhuo Chen, Jean-Eudes Dazard, Yassin Khalifa, Issam Motairek, Catherine Kreatsoulas, Sanjay Rajagopalan, and Sadeer Al-Kindi. Deep learning–based assessment of built environment from satellite images and cardiometabolic disease prevalence. *JAMA cardiology*, 9(6):556–564, 2024.

Zhuo Chen, Tong Zhang, Jean-Eudes Dazard, Sai Rahul Ponnana, Weichuan Dong, Skanda Moorthy, Santosh Kumar Sirasapalli, Haitham Khraishah, Salil Deo, Sanjay Rajagopalan, et al. Ai-enhanced analysis of built environment imagery and neighborhood obesity in us cities. *JAMA Network Open*, 8(9):e2534612, 2025.

Lara P Clark, Daniel Zilber, Charles Schmitt, David C Fargo, David M Reif, Alison A Motsinger-Reif, and Kyle P Messier. A review of geospatial exposure models and approaches for health data integration. *Journal of Exposure Science & Environmental Epidemiology*, 35(2):131–148, 2025.

Stephanie E Cleland, William Steinhardt, Lucas M Neas, J Jason West, and Ana G Rappold. Urban heat island impacts on heat-related cardiovascular morbidity: A time series analysis of older adults in us metropolitan areas. *Environment international*, 178:108005, 2023.

Stefany Coxe, Stephen G West, and Leona S Aiken. The analysis of count data: A gentle introduction to poisson regression and its alternatives. *Journal of personality assessment*, 91(2):121–136, 2009.

Jenna Czarnota, Chris Gennings, and David C Wheeler. Assessment of weighted quantile sum regression for modeling chemical mixtures and cancer risk. *Cancer informatics*, 14: CIN–S17295, 2015.

Victor Dheur and Souhaib Ben Taieb. Probabilistic calibration by design for neural network regression. In *International Conference on Artificial Intelligence and Statistics*, pages 3133–3141. PMLR, 2024.

Roberto Di Mari, Salvatore Ingrassia, and Antonio Punzo. Local and overall deviance r-squared measures for mixtures of generalized linear models. *Journal of classification*, 40 (2):233–266, 2023.

Carl Doersch, Abhinav Gupta, and Alexei A Efros. Unsupervised visual representation learning by context prediction. In *Proceedings of the IEEE international conference on computer vision*, pages 1422–1430, 2015.

Claudio Gariazzo, Matteo Renzi, Alessandro Marinaccio, Paola Michelozzi, Stefania Massari, Camillo Silibello, Giuseppe Carlino, Paolo Giorgi Rossi, Sara Maio, Giovanni Viegi, et al. Association between short-term exposure to air pollutants and cause-specific daily mortality in italy. a nationwide analysis. *Environmental Research*, 216:114676, 2023.

Antonio Gasparrini. Distributed lag linear and non-linear models in r: the package dlnm. *Journal of statistical software*, 43:1–20, 2011.

Antonio Gasparrini. The case time series design. *Epidemiology*, 32(6):829–837, 2021.

Antonio Gasparrini, Ben Armstrong, and Mike G Kenward. Distributed lag non-linear models. *Statistics in medicine*, 29(21):2224–2234, 2010.

Antonio Gasparrini et al. A tutorial on the case time series design for small-area analysis. *BMC Medical Research Methodology*, 2022.

Jean-Baptiste Guimbaud, Alexandros P Siskos, Amrit Kaur Sakhi, Barbara Heude, Eduard Sabidó, Eva Borràs, Hector Keun, John Wright, Jordi Julvez, Jose Urquiza, et al. Machine learning-based health environmental-clinical risk scores in european children. *Communications Medicine*, 4(1):98, 2024.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. On calibration of modern neural networks. In *International conference on machine learning*, pages 1321–1330. PMLR, 2017.

Yassine Himeur, Bhagawat Rimal, Abhishek Tiwary, and Abbes Amira. Using artificial intelligence and data fusion for environmental monitoring: A review and future perspectives. *Information Fusion*, 86:44–75, 2022.

Tracey Holloway, Daegan Miller, Susan Anenberg, Minghui Diao, Bryan Duncan, Arlene M Fiore, Daven K Henze, Jeremy Hess, Patrick L Kinney, Yang Liu, et al. Satellite monitoring for air quality and health. *Annual review of biomedical data science*, 4(1):417–447, 2021.

Stefania Isola, Giuseppe Murdaca, Silvia Brunetto, Emanuela Zumbo, Alessandro Tonacci, and Sebastiano Gangemi. The use of artificial intelligence to analyze the exposome in the development of chronic diseases: a review of the current literature. *Informatics*, 11(4):86, 2024. doi: 10.3390/informatics11040086.

Holly Janes, Lianne Sheppard, and Thomas Lumley. Case–crossover analyses of air pollution exposure data: referent selection strategies and their implications for bias. *Epidemiology*, 16(6):717–726, 2005.

Ziyang Jiang, Zhuoran Hou, Yiling Liu, Yiman Ren, Keyu Li, and David Carlson. Estimating causal effects using a multi-task deep ensemble. In *Proceedings of the 40th International Conference on Machine Learning (ICML)*, 2023.

Fredrik D. Johansson, Uri Shalit, and David Sontag. Learning representations for counterfactual inference. In *Proceedings of the 33rd International Conference on Machine Learning (ICML)*, 2016.

Volodymyr Kuleshov, Nathan Fenner, and Stefano Ermon. Accurate uncertainties for deep learning using calibrated regression. In *International conference on machine learning*, pages 2796–2804. PMLR, 2018.

Yann LeCun, Corinna Cortes, and CJ Burges. Mnist handwritten digit database. *ATT Labs [Online]. Available: http://yann.lecun.com/exdb/mnist*, 2, 2010.

Joshua J Levy, Rebecca M Lebeaux, Anne G Hoen, Brock C Christensen, Louis J Vaickus, and Todd A MacKenzie. Using satellite images and deep learning to identify associations between county-level mortality and residential neighborhood features proximal to schools: A cross-sectional study. *Frontiers in public health*, 9:766707, 2021.

Cong Liu, Renjie Chen, Francesco Sera, Ana M Vicedo-Cabrera, Yuming Guo, Shilu Tong, Micheline SZS Coelho, Paulo HN Saldiva, Eric Lavigne, Patricia Matus, et al. Ambient particulate air pollution and daily mortality in 652 cities. *New England Journal of Medicine*, 381(8):705–715, 2019.

Dong C Liu and Jorge Nocedal. On the limited memory bfgs method for large scale optimization. *Mathematical programming*, 45(1):503–528, 1989.

Thomas Lumley and David Levy. Bias in the case-crossover design: implications for studies of air pollution. *Environmetrics*, 11(6):689–704, 2000.

Malcolm Maclure. The case-crossover design: a method for studying transient effects on the risk of acute events. *American journal of epidemiology*, 133(2):144–153, 1991.

Peter McCullagh. *Generalized linear models*. Routledge, 2019.

Quynh C Nguyen, Sahil Khanna, Pallavi Dwivedi, Dina Huang, Yuru Huang, Tolga Tasdizen, Kimberly D Brunisholz, Feifei Li, Wyatt Gorman, Thu T Nguyen, et al. Using google street view to examine associations between built environment characteristics and us health outcomes. *Preventive medicine reports*, 14:100859, 2019.

Quynh C Nguyen, Jessica M Keralis, Pallavi Dwivedi, Amanda E Ng, Mehran Javanmardi, Sahil Khanna, Yuru Huang, Kimberly D Brunisholz, Abhinav Kumar, and Tolga Tasdizen. Leveraging 31 million google street view images to characterize built environments and examine county health outcomes. *Public Health Reports*, 136(2):201–211, 2021.

Alexandru Niculescu-Mizil and Rich Caruana. Predicting good probabilities with supervised learning. In *Proceedings of the 22nd international conference on Machine learning*, pages 625–632, 2005.

Lizhen Nie, Mao Ye, Qiang Liu, and Dan Nicolae. Vcnet and functional targeted regularization for learning causal effects of continuous treatments. *International Conference on Learning Representations (ICLR)*, 2021.

Sabine Oskar and Jeanette A Stingone. Machine learning within studies of early-life environmental exposures and child health: Review of the current literature and discussion of next steps. *Current environmental health reports*, 7(3):170–184, 2020.

Roger D Peng, Francesca Dominici, and Thomas A Louis. Model choice in time series studies of air pollution and mortality. *Journal of the Royal Statistical Society Series A: Statistics in Society*, 169(2):179–203, 2006.

Kenneth J Rothman, Sander Greenland, Timothy L Lash, et al. *Modern epidemiology*, volume 3. Wolters Kluwer Health/Lippincott Williams & Wilkins Philadelphia, 2008.

Richard F Serfozo. Processes with conditional stationary independent increments. *Journal of applied probability*, 9(2):303–315, 1972.

Uri Shalit, Fredrik D. Johansson, and David Sontag. Estimating individual treatment effect: generalization bounds and algorithms. In *Proceedings of the 34th International Conference on Machine Learning (ICML)*, 2017.

Shengzhi Sun, Kate R Weinberger, Amruta Nori-Sarma, Keith R Spangler, Yuantong Sun, Francesca Dominici, and Gregory A Wellenius. Ambient heat and risks of emergency department visits among adults in the united states: time stratified case crossover study. *Bmj*, 375, 2021.

Aurelio Tobias, Yoonhee Kim, and Lina Madaniyazi. Time-stratified case-crossover studies for aggregated data in environmental epidemiology: a tutorial. *International journal of epidemiology*, 53(2):dyae020, 2024.

United States Census Bureau. Census mapping files: TIGER/Line shapefiles, 2025. URL https://www.census.gov/. Accessed February 8, 2025.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

Jon Wakefield. Disease mapping and spatial regression with count data. *Biostatistics*, 8(2): 158–183, 2007.

Scott Weichenthal, Marianne Hatzopoulou, and Michael Brauer. A picture tells a thousand... exposures: opportunities and challenges of deep learning image analyses in exposure science and environmental epidemiology. *Environment international*, 122:3–10, 2019.

Ander Wilson, Hsiao-Hsien Leon Hsu, Yueh-Hsiu Mathilda Chiu, Robert O Wright, Rosalind J Wright, and Brent A Coull. Kernel machine and distributed lag models for assessing windows of susceptibility to environmental mixtures in children's health studies. *The annals of applied statistics*, 16(2):1090, 2022.

S. Wilson and Jon Wakefield. Pointless spatial modeling. *Biostatistics*, 21(2):e17–e32, 2018.

Jeffrey M Wooldridge. *Econometric analysis of cross section and panel data*. MIT press, 2010.

Appendix A. Comparison of Exposure-Outcome Modeling Frameworks

We provide a visualization in Figure A1 to compare the strategy and structure of different exposure-outcome modeling frameworks.

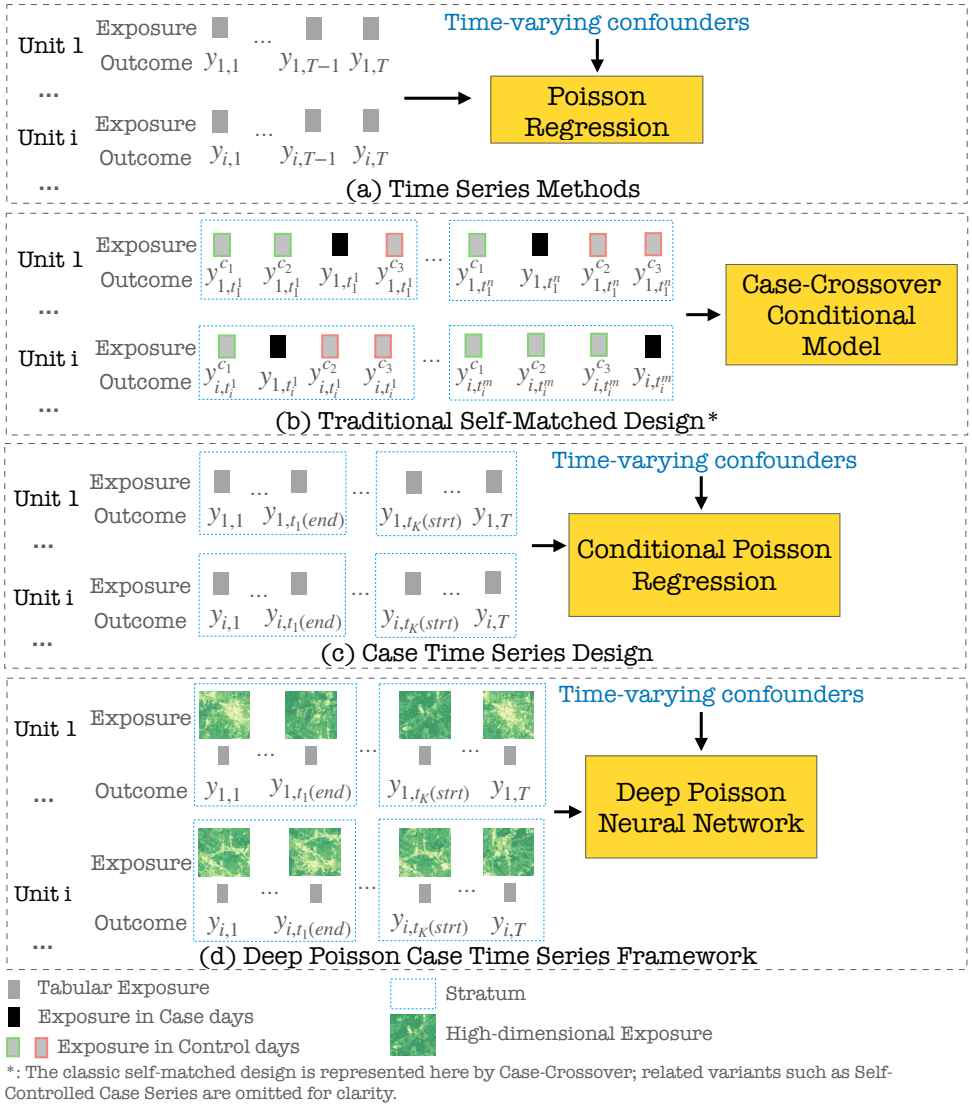

Figure A1: Comparison of different exposure-outcome modeling frameworks. $Y_{i,t}$ denotes the outcome for unit i at time t ($t = 1, \ldots, T$). In (b), $t_{i,m}$ denotes the case day in the mth stratum for unit i, and $y_{i,t_i^m}^{c_p}$ ($p = 1, 2, 3$) denotes the outcome on control day c_p in the mth stratum for unit i. In (c) and (d), K is the number of strata for each unit, and $y_{i,t_k(\text{strt})}$ and $y_{i,t_k(\text{end})}$ denote outcomes on the first and last days of stratum k.

Appendix B. Assumptions of CTS Model

To protect the validity of the CTS design, the key assumptions previously given (Gasparrini, 2021) are:

A1 (outcome distributional assumptions) *The collection* $\{Y_{it}\}_{i,t}$ *consists of conditionally independent Poisson random variables, with*

$$Y_{it} \,\big|\, x_{it}, z_{it} \sim \text{Poisson}(\mu_{it}).$$

Note: For our setup, there is a slight modification to include the high-dimensional information h_{it},

$$Y_{it} \,\big|\, x_{it}, h_{it}, z_{it} \sim \text{Poisson}(\mu_{it}).$$

A2 (outcome-independent follow-up period) *The period of observation for each case* i *must be independent of the outcome* Y_{it}.

A3 (outcome-independent exposure distribution) *The probability of the exposure* x_{it} *must be independent of the outcome history before time* t*, i.e., the exposure process does not depend on past outcomes beyond what is captured by covariates.*

A4 (constant baseline risk) *Baseline risk is constant conditionally on measured time-varying predictors (after including* $\xi_{i(k)}$*,* $s(t)$*, and* z_{it}*, residual baseline drift is negligible).*

For the causal interpretation used here, we additionally assume:

B1 (Consistency) *The observed outcome equals the potential outcome under the observed exposure; no interference across units:*

$$Y_{it} = Y_{it}\big(E_{i,t-L:t}\big)$$

when $E_{i,t-L:t}$ *is the observed exposure history for unit* i*, and*

$$Y_{it}(\cdot) \text{ does not depend on } E_{i',t-L:t} \quad \text{for } i' \neq i.$$

B2 (Positivity) *Each unit exhibits sufficient within-unit exposure variation across time, i.e., for all histories with nonzero density, there is positive probability of each relevant exposure level.*

B3 (No unmeasured time-varying confounding) *Conditional on* $\xi_{i(k)}$*, shared calendar* $s(t)$*, and observed time-varying covariates* z_{it}*, the exposure history is as-if random over time within unit:*

$$E_{i,t-L:t} \perp Y_{i,\cdot}(\cdot) \,\big|\, \xi_{i(k)}, s(t), z_{it}.$$

Intuitively, after removing unit baselines and shared calendar patterns (and adjusting for measured covariates), the remaining exposure-history fluctuations are unconfounded.

B4 (Correct mean model for counts) *For count outcomes, the log-link Poisson mean is correctly specified for identification of* f *via maximum likelihood (up to usual regularity conditions).*

B5 (finite window) Acute effects are dominated by a finite lag window L; long-range drift is delegated to $\mu_{it}^{(0)}$ or explicit calendar features inside $\mathcal{F}_\theta$.

Appendix C. Additional Methodological Discussion

C.1. Discussion of the area-ID auxiliary loss

Area ID is not supplied as an explicit outcome covariate; however, its auxiliary loss shapes the shared exposure representation and can therefore influence outcome predictions indirectly. We assess this dependence with the no-area-ID ablation. The stratum offset, residual penalty, and within-stratum evaluation reduce, but do not eliminate, the possibility that stable spatial information enters the residual.

This component is optional and exposure-specific. In our synthetic experiments, the area-ID loss is not used, yet DPCTS still outperforms comparison models. This suggests that the main performance gains do not depend on area-ID supervision, while the loss can serve as a useful regularizer in satellite-imagery applications.

C.2. Negative-binomial extension for overdispersed counts

DPCTS currently uses a Poisson likelihood, which is natural for count outcomes and aligns with the CTS motivation. In addition, under conditional independence, sums of Poisson counts remain Poisson, with the aggregated mean equal to the sum of component means. This property makes the current Poisson formulation convenient for the multi-scale aggregation objective.

However, real-world health count data may exhibit overdispersion due to unmeasured heterogeneity, correlated events, or other sources of extra-Poisson variation. A negative-binomial extension is therefore a promising direction for future work. One possible extension is to replace the Poisson observation model with a negative-binomial likelihood while retaining the same stratum-specific baseline and neural residual structure, allowing the model to account for extra-Poisson variability.

Such an extension would require additional care for multi-scale supervision. Unlike conditionally independent Poisson counts, a generic sum of negative-binomial counts is not automatically negative-binomial under arbitrary dispersion parameterizations. Therefore, a negative-binomial DPCTS would require an explicit dispersion and dependence parameterization for aggregated outcomes rather than a direct drop-in replacement of the Poisson likelihood. We leave this extension to future work.

C.3. Relationship between MR-AUC and rank-based metrics

MR-AUC is related to established rank-based measures such as Kendall's τ and Spearman's rank correlation, since all evaluate whether predicted risks preserve the ordering of observed outcomes. Kendall's τ summarizes concordance and discordance over observation pairs, while Spearman's rank correlation measures the correlation between observed and predicted ranks.

MR-AUC uses a related pairwise-ordering idea but formulates it as a discrimination task over oriented within-stratum pair-difference examples. For each stratum, we form pairs of time points with a fixed orientation, including pairs with tied observed counts. Each pair receives a binary label indicating whether the first time point has a higher observed count than the second, and a score given by the corresponding difference in predicted risks. The

resulting AUC measures how well these pair-difference scores discriminate positive from negative within-stratum comparisons.

We use MR-AUC as a complementary diagnostic because it aligns with the self-matched acute-risk setting. The relevant discrimination task is not whether a model separates high-risk areas from low-risk areas, but whether it captures temporal risk ordering within the same matched stratum. Unlike global AUC, MR-AUC removes between-stratum baseline differences from the discrimination task. Unlike snRMSE or Stratified Deviance R^2, MR-AUC is less sensitive to global scale calibration and instead summarizes within-stratum temporal ordering.

C.4. snRMSE in low-count strata

The goal of snRMSE is to measure prediction error relative to each stratum's own outcome scale. This avoids a common limitation of global RMSE, where high-count strata can dominate the metric and obscure performance on lower-count strata. Since both the numerator and denominator are on the count scale, snRMSE can be interpreted as a within-stratum relative prediction error.

However, relative-error metrics can have higher variability in very sparse Poisson strata. Even when the model is well calibrated, observed counts have larger relative random variation when the expected count in a stratum is small. Therefore, low-mean strata may contribute noisier relative errors to snRMSE.

For this reason, we interpret snRMSE together with Stratified Deviance R^2 and MR-AUC rather than as a standalone performance criterion. Stratified Deviance R^2 evaluates likelihood-based improvement over the stratum-mean baseline, snRMSE evaluates scale-normalized prediction error, and MR-AUC evaluates within-stratum temporal ordering. Together, these metrics provide complementary views of model performance under the self-matched design.

Appendix D. Experiment Details

D.1. Synthetic Experiment

D.1.1. Synthetic dataset

Data generation We simulate daily areal count outcomes for $N_{\text{blk}} = 200$ spatial units over $N_{\text{yrs}} = 3$ years ($T = 365N_{\text{yrs}}$ days per unit). For each block i, draw a block-specific intercept $\alpha_i \sim \mathcal{N}(0, 0.3^2)$. We optionally include shared calendar effects by defining a stratum index $s(t)$ (month×day-of-week) and sampling $\gamma_{s(t)} = \gamma^{\text{mo}}_{\text{month}(t)} + \gamma^{\text{dow}}_{\text{dow}(t)}$, where $\gamma^{\text{mo}} \in \mathbb{R}^{12}$ and $\gamma^{\text{dow}} \in \mathbb{R}^7$ are sampled as mean-zero Gaussian vectors and then centered.

Exposures and signal For each (i, t), generate continuous exposure channels $x^{(1)}_{i,t} \sim \mathcal{N}(0, 1)$, $x^{(2)}_{i,t} \sim \mathcal{N}(0, 1)$, and an interaction $x^{(3)}_{i,t} = x^{(1)}_{i,t} x^{(2)}_{i,t}$. We define a nonlinear risk component

$$f(\mathbf{x}_{i,t}) \;=\; 0.3x^{(1)}_{i,t} \;+\; \big(-0.15(x^{(2)}_{i,t})^2 + 0.5\big) \;+\; 0.2x^{(3)}_{i,t}.$$

Additional nuisance channels In addition to the exposure channels that enter the outcome model, we generate two nuisance covariates that are provided to all models as extra inputs but have no causal or generative effect on $Y_{i,t}$. First, we draw an i.i.d. noise channel $u_{i,t} \sim \mathcal{N}(0,1)$. Second, we include a deterministic linear time trend $\text{trend}_t \in [-1,1]$ shared across blocks. These channels act as irrelevant features that test robustness to spurious inputs and help ensure performance gains are not driven by trivial time indexing.

Heat-wave Indicator (Setting (c) only) We generate a seasonal temperature series

$$\text{temp}_{i,t} = 60 + 20\sin\left(2\pi\frac{t-200}{365}\right) + \delta_i + \epsilon_{i,t}, \quad \delta_i \sim \mathcal{N}(0, 1.5^2),\ \epsilon_{i,t} \sim \mathcal{N}(0, 3^2),$$

and inject two non-overlapping 4-day heat-wave events per year near the seasonal peak. We define $\text{HW}_{i,t} = 1$ if, within the previous $W = 7$ days (excluding day t), there exists a run of at least $k = 3$ consecutive days with temp $\geq 85°\text{F}$; otherwise $\text{HW}_{i,t} = 0$.

Experimental settings We evaluate three DGMs: (i) **Block FE only** ($\gamma_{s(t)} = \beta_{\text{hw}} = 0$); (ii) **Block FE + calendar FE** ($\beta_{\text{hw}} = 0$); (iii) **Block FE + calendar FE + heat-wave** ($\beta_{\text{hw}} = 1.20$ with $\text{HW}_{i,t}$ included). In all settings, we set r as a global log-rate intercept (r=0.5 in our experiments), and we have:

$$Y_{i,t} \sim \text{Poisson}(\mu_{i,t}), \qquad \log \mu_{i,t} = r + \alpha_i + \gamma_{s(t)} + f(\mathbf{x}_{i,t}) + \beta_{\text{hw}}\text{HW}_{i,t}.$$

D.1.2. DPCTS model form

Let $s(i,t)$ denote a discrete stratum index that encodes the design-controlled baseline (e.g., unit×month×day-of-week). From training data, we compute a fixed baseline mean for each stratum s, denoted $\mu_s^{(0)} > 0$, and use it as an offset. We then model deviations from this baseline in log space:

$$Y_{i,t} \sim \text{Poisson}(\mu_{i,t}), \qquad \log \mu_{i,t} = \log \mu_{s(i,t)}^{(0)} + r_{i,t}, \qquad r_{i,t} = g_\theta(\mathbf{z}_{i,t}),$$

where $\mathbf{z}_{i,t}$ stacks the exposure channels and g_θ is an MLP. Predicted means are $\hat{\mu}_{i,t} = \mu_{s(i,t)}^{(0)} \exp(g_\theta(\mathbf{z}_{i,t}))$.

We train g_θ by minimizing the Poisson negative log-likelihood:

$$\mathcal{L}_{\text{Pois}} = \frac{1}{|\mathcal{D}|} \sum_{(i,t)\in\mathcal{D}} (\hat{\mu}_{i,t} - Y_{i,t} \log \hat{\mu}_{i,t}) + \text{const.}$$

D.2. High-dimensional and temporal exposure experiments

D.2.1. Common protocol: panel, strata baseline, and splits

Panel All experiments are panel count time series indexed by spatial unit (block or tile) $i \in \{1, \ldots, I\}$ and day $t \in \{1, \ldots, T\}$ with outcome $Y_{i,t} \in \mathbb{N}_0$ and mean $\mu_{i,t}$.

Strata baseline We define discrete strata $s(i,t)$ to absorb stable unit and calendar structure (e.g., $s(i,t)$ = block × month × day of week, or a simplified variant used in a given experiment). From the training period only, we compute a fixed baseline mean for each stratum, $\mu_s^{(0)} > 0$, and use it as an offset for validation or test without refitting. This ensures chronological leakage control.

Outcome generation (image-only) For each unit-day (i, t), we sample a single high-dimensional image exposure $h_{i,t}$ (MNIST digit or Sentinel-2 tile) and compute an image-driven scalar signal $u_{i,t} = \psi(h_{i,t})$ (e.g., engineered image statistics or a fixed feature extractor). We then form the latent log-rate as an additive model of baseline components and the current-image signal, $\eta_{i,t} = \alpha_i + \gamma_{s(t)} + w\, u_{i,t} + w_{\text{tr}}\text{trend}_t$ (with the trend term included when used), and set $\lambda_{i,t} = \exp(\eta_{i,t})$. Finally, outcomes are drawn as $Y_{i,t} \sim \text{Poisson}(\lambda_{i,t})$.

Outcome generation (image time series) For each unit-day (i, t), we generate (or select) a high-dimensional exposure $h_{i,t}$ (MNIST digits or Sentinel-2 tiles) and define a latent log-rate as the sum of a unit baseline, a calendar component, and an exposure-driven signal derived from the current image. In temporal variants, we induce dependence on exposure history by adding an AR-like term and/or a lag bank of past image signals. We exponentiate to obtain $\lambda_{i,t}$ and draw counts as $Y_{i,t} \sim \text{Poisson}(\lambda_{i,t})$.

Chronological splits All splits are forward in time. Any preprocessing that depends on exposures (e.g., image normalization or PCA) is fit on training only and then applied unchanged to validation/test.

Evaluation We report likelihood-based metrics (Poisson NLL / perplexity) and self-matched metrics computed within strata (e.g., Stratified Deviance R^2, snRMSE, and within-stratum ranking metrics).

D.2.2. MNIST exposure experiments

Data construction We simulate a panel with $I = 200$ blocks and $T = 3 \times 365$ days. For each (i, t), we sample one MNIST image $h_{i,t}$ (and its digit label $d_{i,t}$) from the pooled dataset. We form strata $s(i, t)$ (block combined with a calendar stratum), and compute $\mu_s^{(0)}$ from training counts. We consider two exposure settings: (i) *image-only*, where the risk depends on the current image; and (ii) *image time series*, where the risk depends on an image-history signal constructed from lagged images.

Preprocessing Images are standardized using training-set statistics only. For linear/spline baselines, we optionally extract low-dimensional features (engineered image summaries and/or PCA fit to flattened training images only).

D.2.3. Sentinel-2 exposure experiments

Dataset construction Each observation is a 6-channel Sentinel-2 tile $h_{i,t} \in \mathbb{R}^{6\times H\times W}$ paired with a daily count outcome. We build a balanced tile-day panel by selecting dates and tiles with consistent coverage (and dropping irregular boundary tiles). Strata $s(i, t)$ are defined according to the fixed-effect design used in the experiment (e.g., tile and day-of-week, or tile and a finer calendar stratum), and $\mu_s^{(0)}$ is estimated on training only. An example image tile in Durham County is provided in Figure D2.

Missingness handling Masked or missing pixels are forward-filled over time within each tile and channel, and all channels are then standardized using training-set per-channel mean and standard deviation. An example image tile imputation in Durham County is provided in Figure D3.

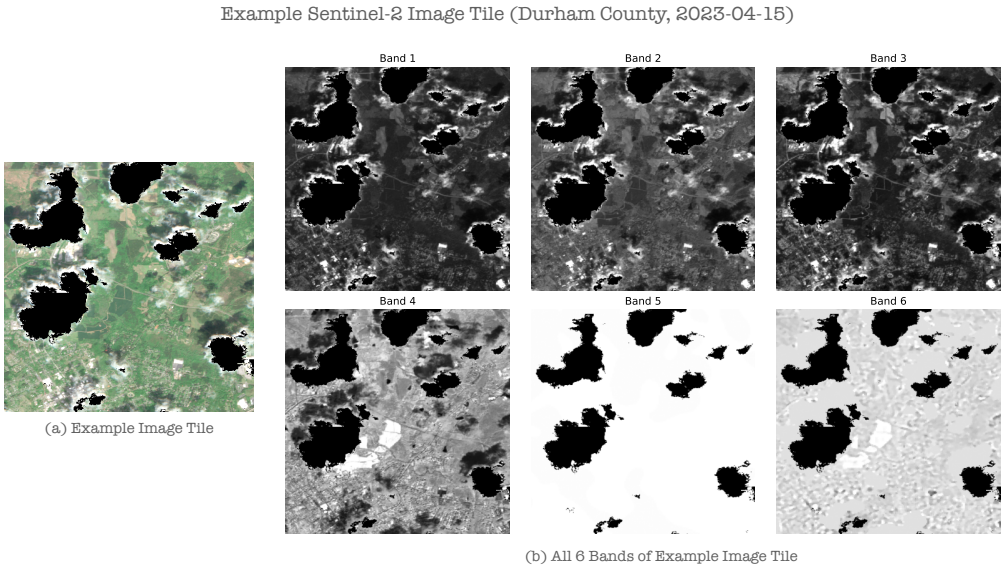

Figure D2: Example satellite-image tile in Durham County on April 15, 2023. Cloud-covered pixels are treated as missing and shown in black. (a) An RGB visualization of the example tile. (b) The six input channels used for the example tile.

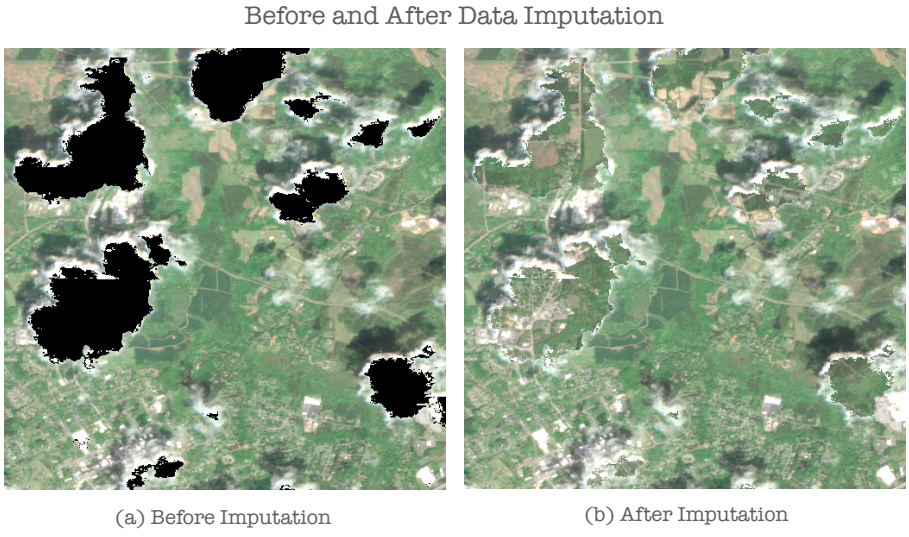

Figure D3: Example satellite-image tile imputation in Durham County on April 15, 2023: (a) before imputation, with cloud-covered pixels shown in black; and (b) after imputation.

Temporal windows For the image-time-series setting, we form rolling windows of length W within each tile time series. If fewer than W prior observations exist, we left-pad with the earliest available index and include a binary mask to ignore padded steps.

D.3. Experiment on multimodal exposures

D.3.1. Data construction and preprocessing

Inputs Each observation (i, t) contains (i) tabular weather features and (ii) a 6-channel Sentinel-2 tile. Tabular covariates include current temperature, an explicit lag bank of length L=6, and smooth annual seasonality via $\sin(2\pi\,\mathrm{doy}/365)$ (doy = day of year) and $\cos(2\pi\,\mathrm{doy}/365)$ (plus a linear trend). All tabular covariates are standardized using training-set means and standard deviations; the resulting transformation is then applied to the validation and test sets.

Satellite tiles We use six Sentinel-2 channels: four surface-reflectance bands (B2, B3, B4, and B8) and two Level-2A auxiliary maps (aerosol optical thickness and water vapour). Missing pixels are handled by the same forward-fill procedure used in the image-only Sentinel-2 experiments, and the tiles are then standardized per-channel using training-set pixel mean and standard deviation. For non-neural baselines, we optionally fit PCA on flattened training tiles (32 components) and append PCA scores as additional covariates; deep models use raw standardized tiles.

Strata baseline (training-only) We define discrete strata $s(i, t)$ to capture stable unit and calendar structure (e.g., block×day of week in this experiment). From training data, we compute a fixed baseline mean $\mu_s^{(0)} > 0$ for each stratum and map it to all rows in validation/test. This baseline is used as the offset in the residual DPCTS formulation below.

Outcome generation For each tile-day (i, t), we form a latent log-rate by combining block and calendar components with two smoothed history signals: one from temperature (current value and lags) and one from image-derived scalars computed from the Sentinel-2 tile, plus a cross-modal interaction between the two histories. To avoid unrealistically spiky dynamics, we apply temporal smoothing in log space (AR-like low-pass filtering and a clamp on day-to-day changes) before exponentiating to obtain $\lambda_{i,t} = \exp(\eta_{i,t})$. We then apply tail control and a global (plus smooth month-wise) calibration to match target marginal scale and upper-tail behavior, and finally sample outcomes as $Y_{i,t} \sim \mathrm{Poisson}(\lambda_{i,t})$.

D.3.2. Multimodal simulator

We simulate a balanced tile-day panel over selected dates and a fixed subset of tiles with consistent coverage. The data-generating rate combines (i) latent calendar variation, (ii) smoothed histories derived from temperature and image summaries, and (iii) a mild interaction between modalities. Counts are sampled as $Y_{i,t} \sim \mathrm{Poisson}(\lambda_{i,t})$.

Appendix E. More Experimental Results

E.1. Synthetic fixed-effects Poisson study

We conducted a controlled synthetic study to assess whether DPCTS better captures nonlinear exposure-outcome relationships with fixed effects (FE) under multiple scenarios. Daily counts are generated for $N_{\text{blk}} = 200$ blocks over three years. The outcome follows a Poisson model with block-level fixed effects and, depending on the setting, additional calendar effects and an injected heat-wave-like mechanism:

$$Y_{i,t} \sim \text{Poisson}(\lambda_{i,t}) \tag{27}$$

$$\log \lambda_{i,t} = \alpha_i + \gamma_{s(t)} + f(\mathbf{x}_{i,t}) + \beta_{\text{hw}} \text{HW}_{i,t}. \tag{28}$$

Here α_i is a block fixed effect, $\gamma_{s(t)}$ is a shared calendar effect (month×day-of-week), $f(\cdot)$ includes nonlinear and interaction terms over continuous exposures, and $\text{HW}_{i,t}$ is a heat-wave-like indicator (defined only in the heat-wave setting).

We evaluate three settings: (a) Block FE only ($\gamma_{s(t)} = \beta_{\text{hw}} = 0$); (b) Block FE + calendar FE ($\beta_{\text{hw}} = 0$); (c) Block FE + calendar FE + heat-wave (all terms active). We compare DPCTS against DLNM, neural-linear and neural-spline CTS models as baseline models, ensuring all methods receive the same inputs in each setting. Performance is reported under a forward holdout split using Poisson NLL, Stratified Deviance R^2, snRMSE, and Multi-Rank AUC; for setting (c) we additionally report a heat-wave intervention relative risk by forcing $\text{HW}_{i,t} \leftarrow 1$ on the test set. Results for all settings are summarized in Table 4; full details are in Appendix D.1. We also report the metrics with the ground-truth values, which provide the theoretically best values under the substantial uncertainty in the Poisson model.

As shown in Table 4, DPCTS outperforms all other comparison models in terms of evaluation metrics we used. In the heat-wave scenario (c), our model demonstrates superior performance in estimating the relative risk of heat-wave exposure. Compared with the ground truth (GT), our model achieves an error rate of 2.4%, substantially lower than DLNM (6.3%), Spline-CTS (7.5%), and Linear-CTS (13.0%). This indicates that our model more effectively captures the complex, non-linear relationship between exposure and health outcomes, which is an advantage when modeling extreme environmental stressors like heat waves.

E.2. Image-only setting for high-dimensional exposures

In the image-only setting, we compare against the Case Crossover (CCO) design using both linear and spline feature encoders. As the baseline methods cannot directly use imagery, all CCO methods employ engineered image statistics and PCA features as their exposure representations. Results are summarized in Table 5.

E.3. Ablation study for model components

To better understand the contribution of key components in the proposed framework, we conduct an ablation study focusing on the multi-scale aggregation loss, area-ID auxiliary loss and the post-hoc affine calibration step. The multi-scale loss targets accuracy across

Table 4: Synthetic study results across settings. Values are averages over five runs. Arrows indicate preferred direction. In setting (c), $\widehat{\mathrm{RR}}_{\mathrm{HW}}$ is the heat-wave intervention RR; GT uses true μ.

Setting	Model	↓ NLL	↑ Stratified Dev R^2	↓ snRMSE	↑ MR-AUC	$\widehat{\mathrm{RR}}_{\mathrm{HW}}$
(a) Block FE	DPCTS	**0.501** ± 0.003	**0.315** ± 0.004	**0.014** ± 0.002	**0.716** ± 0.002	—
	DLNM	0.504 ± 0.004	0.311 ± 0.005	0.014 ± 0.002	0.712 ± 0.003	—
	Linear-CTS	0.513 ± 0.005	0.307 ± 0.006	0.014 ± 0.003	0.708 ± 0.003	—
	Spline-CTS	0.506 ± 0.004	0.310 ± 0.005	0.014 ± 0.002	0.710 ± 0.003	—
	GT	0.476	0.319	0.014	0.718	—
(b) Block + Cal. FE	DPCTS	**0.525** ± 0.004	**0.324** ± 0.005	**0.016** ± 0.002	**0.723** ± 0.002	—
	DLNM	0.545 ± 0.006	0.312 ± 0.006	0.017 ± 0.002	0.722 ± 0.003	—
	Linear-CTS	0.581 ± 0.008	0.263 ± 0.009	0.017 ± 0.003	0.704 ± 0.004	—
	Spline-CTS	0.552 ± 0.006	0.296 ± 0.007	0.017 ± 0.002	0.723 ± 0.003	—
	GT	0.505	0.351	0.016	0.724	—
(c) Block + Cal. FE + HW	DPCTS	**0.627** ± 0.005	**0.451** ± 0.006	**0.020** ± 0.003	**0.758** ± 0.003	**3.24** ± 0.08
	DLNM	0.636 ± 0.006	0.394 ± 0.008	0.021 ± 0.004	0.727 ± 0.004	3.11 ± 0.10
	Linear-CTS	0.642 ± 0.007	0.281 ± 0.010	0.022 ± 0.005	0.715 ± 0.005	2.89 ± 0.13
	Spline-CTS	0.638 ± 0.006	0.408 ± 0.008	0.021 ± 0.004	0.737 ± 0.004	3.07 ± 0.11
	GT	0.617	0.492	0.019	0.764	3.32

Table 5: Image-only exposure results on MNIST and Sentinel-2 datasets. All values are averages over 5 independent runs. (GT = ground truth model that uses true μ for predicting outcomes.)

Dataset	Model	↑ Stratified Dev R^2	↓ snRMSE(×1e-3)	↑ MR-AUC
MNIST	DPCTS	**0.160** ± 0.003	**1.049** ± 0.005	**0.692** ± 0.002
	CCO-Linear	0.109 ± 0.006	1.089 ± 0.008	0.655 ± 0.002
	CCO-Spline	0.122 ± 0.007	1.074 ± 0.009	0.679 ± 0.003
	GT	0.170	1.043	0.696
Sentinel-2	DPCTS	**0.775** ± 0.004	**0.504** ± 0.006	**0.897** ± 0.002
	CCO-Linear	0.753 ± 0.006	0.531 ± 0.008	0.881 ± 0.003
	CCO-Spline	0.761 ± 0.005	0.523 ± 0.007	0.886 ± 0.003
	GT	0.804	0.493	0.906

temporal resolutions, the area-ID loss regularizes the representation, and affine calibration corrects systematic mean-scale bias.

We perform the ablation study using the Sentinel-2 dataset in the time-series setting. All models are trained and evaluated using the same protocol and metrics as in the main experiments to ensure a fair comparison. For the ablation study, we consider four variants: (1) DPCTS (full model), which includes all components; (2) DPCTS w/o multi-scale, which removes the multi-scale aggregation loss; (3) DPCTS w/o area-ID, which removes the area-ID auxiliary loss; and (4) DPCTS w/o calibration, which removes the post-hoc affine calibration step. We then report performance using the same stratification-aware metrics defined in the main text.

The ablation results in Table 6 show that all components contribute to performance. In particular, removing the multi-scale aggregation loss leads to the largest degradation, while the area-ID loss and calibration step provide consistent gains, with calibration mainly improving numerical accuracy.

Table 6: Ablation study on model components. Experiments are run on the Sentinel-2 satellite-image dataset in the image-time-series setting.

Model	↑ Stratified Dev R^2	↓ snRMSE	↑ MR-AUC
DPCTS	**0.842** $\pm$ 0.004	**0.721** $\pm$ 0.006	**0.926** $\pm$ 0.002
w/o multi-scale	0.819 $\pm$ 0.007	0.961 $\pm$ 0.008	0.903 $\pm$ 0.003
w/o area-ID	0.825 $\pm$ 0.007	0.922 $\pm$ 0.007	0.919 $\pm$ 0.002
w/o calibration	0.823 $\pm$ 0.005	0.938 $\pm$ 0.007	**0.926** $\pm$ 0.002

Table 7: Comparison of imputation strategies.

Imputation	↑ Dev R^2	↓ snRMSE	↑ MR-AUC
None	0.798 $\pm$ 0.006	0.780 $\pm$ 0.009	0.900 $\pm$ 0.004
Forward fill	0.804 $\pm$ 0.005	0.768 $\pm$ 0.008	0.904 $\pm$ 0.003
Linear interp.	0.814 $\pm$ 0.005	0.755 $\pm$ 0.007	0.910 $\pm$ 0.003
None + marker	0.822 $\pm$ 0.005	0.749 $\pm$ 0.008	0.919 $\pm$ 0.003
Forward fill + marker	**0.842** $\pm$ 0.004	**0.721** $\pm$ 0.006	**0.926** $\pm$ 0.002

E.4. Missing data strategies

We evaluate different missing data handling strategies using Sentinel-2 satellite imagery in the image time-series setting; results are reported in Table 7.

E.5. Loss-weight sensitivity

We conducted a one-factor-at-a-time sensitivity analysis on loss weights using the Sentinel-2 image time-series setup. We scaled α_{agg}, α_{area}, and λ_{res} to $0.5\times$ and $2\times$ their default values while keeping all other hyperparameters fixed. Results are reported in Table 8.

The results show that performance changes only modestly, indicating that DPCTS does not rely on a narrowly tuned set of loss weights.

E.6. Post-hoc affine calibration

We also report the learned post-hoc affine calibration parameters used in DPCTS under the Sentinel-2 image time-series setup. Calibration is estimated only on the validation set and then fixed for test evaluation. Across runs, the learned parameters are close to the identity transformation, with $\hat{\alpha} = -0.04 \pm 0.02$ and $\hat{\beta} = 1.06 \pm 0.03$. This indicates that calibration performs only a minor global scale correction rather than substantially changing the learned exposure-risk structure. Since affine calibration preserves within-stratum ordering, MR-AUC remains unchanged.

E.7. Counterfactual temperature-response analysis

To interpret the learned real-world exposure-response pattern, we performed a counterfactual analysis on the cardiovascular ED dataset. Holding all other covariates fixed, we replaced the observed temperature history with representative histories at different percentiles and estimated the predicted residual relative risk against the median temperature-history level. Results are shown in Table 9.

Table 8: Loss-weight sensitivity analysis on the Sentinel-2 image time-series setting. Each row varies one loss weight while keeping all other hyperparameters fixed.

Loss-weight setting	**Dev R^2 ↑**	**snRMSE(×1e-3) ↓**	**MR-AUC ↑**
Default	**0.842** ± 0.004	**0.721** ± 0.006	**0.926** ± 0.002
$\alpha_{\text{agg}} = 0.5\times$	0.835 ± 0.005	0.746 ± 0.011	0.922 ± 0.003
$\alpha_{\text{agg}} = 2\times$	0.831 ± 0.006	0.759 ± 0.013	0.920 ± 0.004
$\alpha_{\text{area}} = 0.5\times$	0.839 ± 0.004	0.733 ± 0.009	0.924 ± 0.002
$\alpha_{\text{area}} = 2\times$	0.836 ± 0.005	0.741 ± 0.010	0.923 ± 0.003
$\lambda_{\text{res}} = 0.5\times$	0.836 ± 0.004	0.735 ± 0.010	0.924 ± 0.002
$\lambda_{\text{res}} = 2\times$	0.840 ± 0.005	0.748 ± 0.011	0.921 ± 0.003

Table 9: Counterfactual temperature-response analysis on the real-world cardiovascular ED dataset. Predicted residual relative risks are reported relative to the median temperature-history level.

Temperature history	**Residual RR**	**% CVD ED risk change**
10th percentile	0.994 ± 0.006	−0.6%
50th percentile	1.000	Reference
90th percentile	1.006 ± 0.007	+0.6%
95th percentile	1.014 ± 0.009	+1.4%

The counterfactual results show a modest but monotonic increase in predicted CVD ED risk under higher temperature histories. This pattern is consistent with the sparse, low-signal nature of the real-world setting and with prior evidence on heat-related cardiovascular morbidity (Cleland et al., 2023).