Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Borrowing strength across space in distributed lag non-linear models: a geographically weighted approach

Joni Downs

downs@usf.edu

University of South Florida

Jaryd Hinch

University of South Florida

Hansamali Abeysinghe Mudiyanselage

University of South Florida

Research Article

Keywords: Exposure-lag-response, Spatial non-stationarity, Small-area estimation, Spatial heterogeneity, Temperature and mortality, Hierarchical models

Posted Date: November 3rd, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11527913/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Borrowing strength across space in distributed lag non-linear models: a geographically weighted approach

Joni Downs*, Jaryd Hinch, Hansamali Abeysinghe Mudiyanselage

[1] School of Geosciences, University of South Florida, Tampa, FL, USA

* Corresponding author: downs@usf.edu

Statements and Declarations

Funding. No external funding was received.

Competing interests. The authors have no competing interests to declare.

Data and code availability. Code is available to reviewers at https://zenodo.org/records/23570999?preview=1&token=eyJhbGciOiJIUzUxMiJ9.eyJpZCI6IjYxNWZhOTA1LTI5YzUtNGRhYy04NTc3LTg5ODliODY4NjJmMyIsImRhdGEiOnt9LCJyYW5kb20iOiJhOTQ3NGI3MzEyZTJmYzQ0M2ZjZDAyY2Y5ZDIyYmVkYyJ9.4CCEIW_QpxTRSImic5Wp2zt_VFpkyeFR4U_0TcMy6O7aaDrk4CdEPINN_gQBnTN0JHcXlSROI0QBYCBmAYimyw and will be deposited in Zenodo with a DOI upon acceptance. The mortality data for the application are publicly available from the Brazilian Mortality Information System (SIM, DATASUS), and the temperature data from ERA5-Land (Copernicus Climate Change Service).

Acknowledgements

This work used Bridges-2 at the Pittsburgh Supercomputing Center through allocation MTH260031 from the Advanced Cyberinfrastructure Coordination Ecosystem: Services & Support (ACCESS) program, which is supported by U.S. National Science Foundation grants #2138259, #2138286, #2138307, #2137603, and #2138296.

Author contributions

J.D.: Conceptualization, Methodology, Software, Formal analysis, Investigation, Visualization, Supervision, Project administration, Writing – original draft, Writing – review & editing. J.H.: Conceptualization, Validation, Writing – review & editing. H.A.M.: Conceptualization, Validation, Writing – review & editing. All authors read and approved the final manuscript.

Borrowing strength across space in distributed lag non-linear models: a geographically weighted approach

Distributed lag non-linear models (DLNMs) are widely used to estimate exposure-response associations across geographic areas. Because units with few events can give unstable estimates in independent models, spatial methods are designed to borrow strength from nearby units. We propose a geographically weighted DLNM (GW-DLNM), in which each unit's first-stage estimates are shrunk toward a local mean formed from its neighbors, weighted by distance and precision. We compared GW-DLNM with Pooled, Two-stage, Tensor, and spatial Bayesian (SB-DLNM) approaches in 1,600 simulated scenarios with 48,000 replicates. The design varied region shape, number and size inequality of units, spatial pattern and magnitude of heterogeneity, exposure-response and lag-response shapes, effect size, and event rate. Accuracy depended mainly on effect heterogeneity and event rate. Pooled was most accurate when all units shared one effect. When the effect varied, SB-DLNM had the lowest median error, and GW-DLNM was most often the most accurate method under high heterogeneity. The first-stage heterogeneity statistic I^2, together with event rate, indicated which method was likely to be most accurate. An application to temperature and mortality in the 853 municipalities of Minas Gerais, Brazil, illustrates the methods. These results provide practical guidance for choosing among methods when exposure-response associations vary geographically.

Key words: Exposure-lag-response; Spatial non-stationarity; Small-area estimation; Spatial heterogeneity; Temperature and mortality; Hierarchical models

1. INTRODUCTION

Disease occurrence often shows associations with environmental conditions, and often these effects are non-linear and delayed in time. Distributed lag non-linear models (DLNMs) offer a framework to describe these relationships jointly (Gasparrini, Armstrong, and Kenward 2010). A DLNM builds a cross-basis from two sets of functions: one for the exposure dimension and one for the lag dimension (Gasparrini 2011). The result is a bidimensional exposure–lag–response surface, from which overall cumulative associations and lag-specific effects can be derived (Gasparrini 2014). DLNMs are routinely used in environmental epidemiology to study infectious, vector-borne, and non-communicable diseases, as well as sources of mortality. Common predictors include atmospheric, hydrological, land surface, and socio-environmental variables, while classic applications include temperature-related mortality (Guo et al. 2011; Gasparrini et al. 2015), air pollution and hospital admissions (Zhang et al. 2019; Chen et al. 2020), and climatic drivers of mosquito-borne diseases (Chien and Yu 2014; Miley et al. 2020).

DLNMs are often applied to geographic data covering many spatial units, such as cities, counties, or districts. In the simplest application, a separate model is fitted to each spatial unit. The disadvantage is that units with few events can give unstable and imprecise estimates, and nothing is learned from similar neighboring units. Instead, a Pooled model can be fitted to all units, which gives stable estimates but assumes one exposure–response curve shared by every unit and hides any geographic variation (Gasparrini 2021, 2022). More commonly, a Two-stage design is used: a DLNM is fitted to each unit, and the unit-specific coefficients are then combined by multivariate meta-analysis, which estimates an average curve and the between-unit variance (Gasparrini, Armstrong, and Kenward 2012; Sera et al. 2019; Masselot and Gasparrini 2025). Although both Pooled and Two-stage models are widely used, neither accounts for spatial autocorrelation or for the possibility that exposure–response relationships vary across space.

Several extensions add spatial structure to DLNMs. One group fits all spatial units in a single model. The spatial Bayesian DLNM (SB-DLNM) gives each unit its own cross-basis coefficients and links them through conditional autoregressive priors (Quijal-Zamorano et al. 2024). Related Bayesian formulations combine Laplacian P-splines with conditional autoregressive priors (Rutten et al. 2025) allow spatial effect modification (Rutten et al. 2026b) and scale to nationwide small-area data by fitting the model region by region (Rutten et al. 2026a). The tensor-product (Tensor) approach adds the spatial coordinates to the exposure and lag dimensions in a single smooth (Economou et al. 2024), clustering approaches group units that share a common curve (Briz-Redón et al. 2025), and spatially varying coefficient models allow linear distributed lag effects to change across space (Warren, Luben, and Chang 2020; Hadianfar, Karimi, and Saki 2026). A second group keeps the two-stage design and adds spatial structure to the second stage, through conditional autoregressive dependence between units (Wang et al. 2023) or spatially correlated residuals in the meta-regression (Sim et al. 2018), later extended to functional meta-predictors (Yu et al. 2021). Geographically weighted regression has also been applied to

first-stage DLNM estimates to map how vulnerability factors modify the exposure effect across space (Yang et al. 2026).

Each of these methods has been evaluated in settings that match its motivating application. Simulation studies have typically been built around a single study region and its areal units, such as municipalities in Sicily (Rutten et al. 2026a), or around the specific data structure a method was designed for, such as binary outcomes in one city (Hadianfar, Karimi, and Saki 2026). This focus is natural when a method is developed for a specific problem, and it shows that each method works well in the setting for which it was designed. It leaves open how the methods compare with one another under common conditions, and how their relative performance depends on features of the study region that vary between applications: the shape of the region, the number and size of its areal units, the strength and spatial pattern of heterogeneity in the true effects, and the amount of data in each unit. Applied researchers choosing a method for a new region therefore have little guidance on which approach suits their data.

Geography offers a complementary spatial perspective. A relationship that differs from place to place is an instance of spatial non-stationarity, a central concern of quantitative geography. Geographically weighted regression addresses it by fitting a local model at each location, in which nearby observations receive more weight than distant ones and a bandwidth determines how far each local model borrows (Brunsdon, Fotheringham, and Charlton 1996; Fotheringham, Brunsdon, and Charlton 2009). In DLNM analyses, geographic weighting has so far served a descriptive purpose, to map how vulnerability factors modify first-stage estimates across space (Yang et al. 2026). The same principle can also serve as the estimator itself. Each unit's whole exposure-lag-response curve can be shrunk toward a local mean of its neighbors, in proportion to the precision of its own estimate and of the local mean. Compared with conditional autoregressive structures, which link each unit to its adjacent units with a fixed neighborhood, geographic weighting lets the extent of borrowing adapt to how quickly the relationship changes across space. Neighborhoods range from very local when effects vary sharply to nearly the whole region when they are shared. Because neighborhoods are defined by distance, they do not depend on the number or shape of bordering units, which vary widely when units differ in size. Compared with a tensor-product surface, which smooths all units onto one function of the coordinates, a local mean retains each unit's own deviation where its data support it.

The same principle can be applied to the second stage of a DLNM analysis. Three features distinguish it from applying geographically weighted regression to first-stage estimates. The local mean excludes the unit itself, so it acts as an external prior built only from neighbors. The number of neighbors is chosen by leave-one-out predictive likelihood, which rewards predictions that are both accurate and well calibrated. The uncertainty of the local mean is carried into each unit's interval, so the method yields full curves with standard errors for every unit, including units whose first-stage model could not be fitted. Each spatial unit's exposure–response curve can borrow strength from a neighborhood of nearby units, with the size of the neighborhood estimated from the data. Compared with conditional autoregressive structures, which link each unit to its adjacent units with a fixed

neighborhood, geographic weighting lets the extent of borrowing adapt to how quickly the relationship changes across space, from very local neighborhoods when effects vary sharply to nearly the whole region when they are shared. Because neighborhoods are defined by distance, they do not depend on the number or shape of bordering units, which vary widely when units differ in size. Unlike Tensor's smoothing over coordinates, fitting a single surface to all units, geographic weighting combines a local mean with each unit's own deviation, so each estimate reflects both its neighborhood and its own data in proportion to their precision.

We propose a geographically weighted DLNM (GW-DLNM), in which each unit's first-stage coefficients are shrunk toward a local mean of its nearest neighbors, with the neighborhood size chosen from the data. We compare it with Pooled, Two-stage, Tensor, and SB-DLNM in a simulation study with four scenario blocks: no effect, an effect shared by all units, and effects that vary between units with moderate or high heterogeneity. Within these blocks we vary the region's shape, the number and size of units, the spatial pattern of heterogeneity, and the amount of data. We determine which method performs best under which conditions. We illustrate the methods with temperature and mortality data for the 853 municipalities of Minas Gerais, Brazil. We conclude with recommendations for selecting a method in practice based on data characteristics.

2. GEOGRAPHICALLY WEIGHTED DLNM

The GW-DLNM estimates an exposure–lag–response curve for every spatial unit in two stages. In the first stage, a DLNM is fitted separately to each unit's time series, which gives a set of coefficients describing that unit's curve and their uncertainty. Units with few events have imprecise first-stage estimates. In the second stage, each unit's estimate is combined with those of its neighbors. Following the principle of geographically weighted regression, nearby units contribute more than distant ones and precise estimates more than imprecise ones, and the size of the neighborhood is chosen from the data. Each unit's final curve is a weighted compromise between its own estimate and the local mean: units with abundant data keep curves close to their own estimate, and units with sparse data borrow more from their neighbors. When the curves vary sharply across space, the selected neighborhood is small; when they are similar everywhere, it grows to cover most of the region, approaching the conventional two-stage model. Sections 2.1 to 2.4 describe each step.

2.1 First stage: unit-specific DLNMs

Let y_{st} be the count of events in areal unit $s = 1, \dots, N$ in period t, with population P_s and exposure series x_{st}. In the first stage a DLNM is fitted separately to each unit,

$$y_{st} \sim \text{Poisson}(\mu_{st}), \qquad \log\mu_{st} = \log P_s + \alpha_s + \boldsymbol{c}_{st}^{\top}\boldsymbol{\beta}_s,$$

where $\boldsymbol{c}_{st}$ is the row of the cross-basis for unit s at time t, built from exposure values $x_{s,t-\ell}$ at lags $\ell = 0, \dots, L$. The cross-basis combines a natural cubic spline in the exposure dimension and a natural cubic spline in the lag dimension, giving p coefficients. The same cross-basis, with knots placed at percentiles of the exposure pooled over all units, is used for every unit, so that the coefficient vectors $\boldsymbol{\beta}_s$ are comparable across units. Each fit gives

an estimate $\widehat{\boldsymbol{\beta}}_s$ with covariance matrix $\boldsymbol{S}_s$, the inverse of the observed information with the intercept α_s profiled out. In applications, the first-stage model can also contain adjustment terms, such as splines for trend and season.

2.2 Second stage: geographically weighted random-effects model

The second stage treats the first-stage estimates as noisy observations of unit-specific true coefficients,

$$\widehat{\boldsymbol{\beta}}_s = \boldsymbol{\beta}_s + \boldsymbol{\varepsilon}_s, \qquad \boldsymbol{\varepsilon}_s \sim N(\mathbf{0}, \boldsymbol{S}_s),$$
$$\boldsymbol{\beta}_s = \boldsymbol{m}_s + \boldsymbol{u}_s, \qquad \boldsymbol{u}_s \sim N(\mathbf{0}, \boldsymbol{\Psi}).$$

In the standard two-stage design $\boldsymbol{m}_s$ is a single global mean, or a linear function of unit-level covariates. Here $\boldsymbol{m}_s$ is a local mean, the average coefficient vector of the units surrounding s. It is estimated by kernel-weighted generalized least squares over the other units,

$$\widehat{\boldsymbol{m}}_s(k) = \boldsymbol{A}_s^{-1} \sum_{j \neq s} w_{sj}\,(k)\, \boldsymbol{W}_j \widehat{\boldsymbol{\beta}}_j, \qquad \boldsymbol{A}_s = \sum_{j \neq s} w_{sj}\,(k)\, \boldsymbol{W}_j, \qquad \boldsymbol{W}_j = (\boldsymbol{S}_j + \boldsymbol{\Psi})^{-1},$$

with bisquare kernel weights on the distance d_{sj} between unit centroids,

$$w_{sj}(k) = \begin{cases} \left[1 - \left(d_{sj}/h_s(k)\right)^2\right]^2 & d_{sj} < h_s(k) \\ 0 & \text{otherwise,} \end{cases}$$

where the bandwidth $h_s(k)$ is the distance from unit s to its $(k+1)$-th nearest unit, so that exactly k neighbors receive positive weight. This adaptive bandwidth widens neighborhoods where units are large and sparse and narrows them where units are small and dense, the standard choice in geographically weighted regression for areal units of unequal size (Fotheringham, Brunsdon, and Charlton 2009).

Each neighbor's contribution is weighted both by proximity and by precision, so that units with few events contribute less. Unit s is excluded from its own local mean, which therefore acts as an external prior for unit s built only from its neighbors, and serves for bandwidth selection (Section 2.3). Treating the weights as fixed, the variance of the local mean is

$$\boldsymbol{V}_s = \mathrm{Var}(\widehat{\boldsymbol{m}}_s) = \boldsymbol{A}_s^{-1} \left(\sum_{j \neq s} w_{sj}^2\, \boldsymbol{W}_j \right) \boldsymbol{A}_s^{-1}.$$

The number of neighbors k controls how local the borrowing is. Small k lets the local mean follow local structure. At the largest values of k nearly every unit contributes to every local mean, with weights that still decline with distance, so the method approaches the conventional two-stage model with a global mean.

Between-unit covariance. Given the local means, the residuals $\boldsymbol{r}_s = \widehat{\boldsymbol{\beta}}_s - \widehat{\boldsymbol{m}}_s$ satisfy approximately $\boldsymbol{r}_s \sim N(\boldsymbol{0}, \boldsymbol{S}_s + \boldsymbol{V}_s + \boldsymbol{\Psi})$. $\boldsymbol{\Psi}$ is estimated from these residuals by maximum likelihood with an expectation–maximization algorithm. Because the weights $\boldsymbol{W}_j$ depend on $\boldsymbol{\Psi}$, the local means and $\boldsymbol{\Psi}$ are updated alternately, starting from the between-unit covariance of a conventional two-stage meta-regression on the unit coordinates. A single $\boldsymbol{\Psi}$ is estimated for the whole map, so the strength of shrinkage is the same everywhere while its target, the local mean, varies from place to place.

2.3 Bandwidth selection

The number of neighbors k is chosen by leave-one-out predictive likelihood. For each candidate k, the local means and $\boldsymbol{\Psi}$ are estimated as in Section 2.2. Since $\widehat{\boldsymbol{m}}_s$ excludes unit s, each unit's estimate is then predicted from its neighbors alone, with predictive covariance $\boldsymbol{C}_s = \boldsymbol{S}_s + \boldsymbol{V}_s + \boldsymbol{\Psi}$, and the criterion

$$\mathrm{CV}(k) = \sum_{s=1}^{N} [\log|\boldsymbol{C}_s| + \boldsymbol{r}_s^{\top} \boldsymbol{C}_s^{-1} \boldsymbol{r}_s]$$

equals minus twice the Gaussian log predictive likelihood, up to a constant, so the selected k minimizes $\mathrm{CV}(k)$. Candidate values are 15 numbers spaced geometrically from 4 to the number of units with a first-stage estimate minus two, which gives finer resolution for small neighborhoods, where the criterion changes most. The criterion favors neighborhood sizes whose predictions of each unit's estimate are both accurate and well calibrated, and it requires no refitting of the first-stage models. When the selected k is the largest candidate, the data borrow from nearly the whole region, as expected when the exposure–response relationship is shared across units.

2.4 Unit-level estimates

With the selected $\hat{k}$ and covariance $\widehat{\boldsymbol{\Psi}}$, each unit's coefficients are estimated by the best linear unbiased predictor, which shrinks the unit's own estimate toward its local mean,

$$\widetilde{\boldsymbol{\beta}}_s = \widehat{\boldsymbol{m}}_s + \boldsymbol{K}_s(\widehat{\boldsymbol{\beta}}_s - \widehat{\boldsymbol{m}}_s), \qquad \boldsymbol{K}_s = \widehat{\boldsymbol{\Psi}}(\boldsymbol{S}_s + \widehat{\boldsymbol{\Psi}})^{-1},$$

with covariance

$$\mathrm{Var}(\widetilde{\boldsymbol{\beta}}_s - \boldsymbol{\beta}_s) = \boldsymbol{K}_s \boldsymbol{S}_s + (\boldsymbol{I} - \boldsymbol{K}_s)\, \boldsymbol{V}_s\, (\boldsymbol{I} - \boldsymbol{K}_s)^{\top}.$$

The first term is the posterior variance of $\boldsymbol{\beta}_s$ when the local mean is known, and the second adds the uncertainty of the local mean. Units with precise first-stage estimates keep most of their own information, and units with imprecise estimates are drawn toward their neighborhood. For a unit whose first-stage model cannot be fitted, the estimate is the local mean itself, with covariance $\widehat{\boldsymbol{\Psi}} + \boldsymbol{V}_s$. Exposure–response and lag–response curves for each unit are obtained from $\widetilde{\boldsymbol{\beta}}_s$ and its covariance, relative to a reference exposure. These variances treat the kernel weights, $\widehat{\boldsymbol{\Psi}}$ and $\hat{k}$ as fixed.

2.5 Implementation

The method is implemented in R (R Core Team 2026). First-stage models are fitted by Poisson regression with the unit intercept profiled out, using cross-bases constructed with the dlnm package (Gasparrini 2011). In the second stage, $\boldsymbol{\Psi}$ is estimated by an expectation–maximization algorithm for maximum likelihood, run to convergence at each step and initialized at the between-unit covariance of a two-stage meta-regression on unit coordinates fitted by restricted maximum likelihood with the mixmeta package (Sera et al. 2019). Local means and $\boldsymbol{\Psi}$ are updated alternately for four rounds at each candidate k. When the conventional two-stage model is fitted to the same first-stage estimates, its meta-regression is the starting model and is reused, which roughly halves the computing time without changing results. The candidate values of k are independent of one another and can be evaluated in parallel.

3. METHODS

3.1 Simulation aims and design

The simulation study has two aims: to evaluate how well the GW-DLNM recovers unit-specific exposure–lag–response relationships, and to compare it with pooled, two-stage, tensor-product, and SB-DLNM models under a wide range of geographic and epidemiological conditions. These conditions are set by five geographic factors (region shape, number of counties, county size inequality, spatial pattern of heterogeneity, and temperature layout) and five epidemiological factors (exposure–response shape, lag structure, relative risk, effect spread, and event rate), each with four levels (Table 1). Because a full factorial design would require 4^{10} scenarios, the factors are combined in a 4^{10-5} fractional factorial design with 1,024 scenarios, in which all main effects and two-factor interactions can be estimated (Skrondal 2000). The study follows the ADEMP structure (Morris, White, and Crowther 2019).

Table 1. Factors and levels of the simulation design.

Category	Factor	Levels
Geography	Region shape	Square; elongated; L-shaped; T-shaped
	Number of counties	60; 120; 240; 480
	County size inequality	Equal; mild; moderate; extreme
	Temperature layout	North–south; cool ridge; warm river valleys; warm edges
	Spatial pattern	Independent; local; regional; boundary
Epidemiology	Exposure–response	J-shaped; linear; threshold; unimodal
	Lag–response	Decaying; delayed; short; harvesting
	Effect size (RR)	1.0 (null); 1.2; 1.4; 1.6
	Effect spread	0.0; 0.15; 0.3; 0.5
	Event rate	0.125; 0.5; 2; 8

For analysis, the scenarios are grouped into four blocks according to the true between-county heterogeneity (Figure 1). In the null control block (256 scenarios), no

county has an association between temperature and the outcome. In the no-heterogeneity control block (192 scenarios), every county has the same exposure–lag–response relationship. The moderate heterogeneity block contains the remaining 576 scenarios, and the high heterogeneity block contains 576 scenarios with the same settings and twice the between-county variation, giving 1,600 scenarios in total. Each scenario has 30 replicates, for 48,000 in total. One replicate in the high heterogeneity block was not run, so 47,999 replicates were analyzed. Results are reported separately for each block.

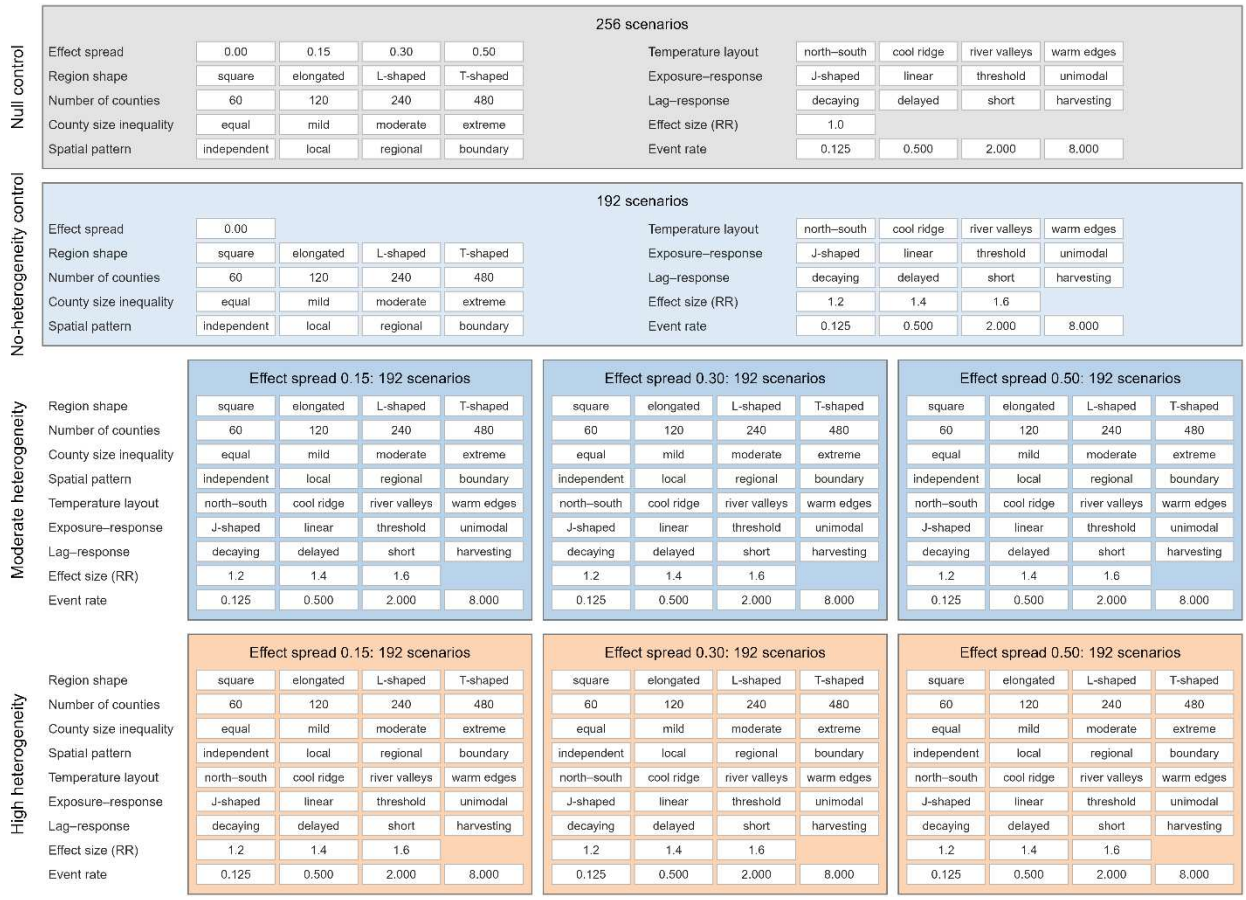


Figure 1. Scenario blocks of the simulation study. Each box lists the levels of every design setting that occur in the block. The null control block (256 scenarios) has RR = 1.0 and the no-heterogeneity control block (192 scenarios) has an effect spread of 0. The moderate and high heterogeneity blocks each contain 576 scenarios, 192 at each effect spread of 0.15, 0.30 and 0.50, with the same settings in both blocks.

3.2 Geography

Region shape. Each study region is a grid of 3,844 square cells, grouped by random tiling into 961 tetrominoes, shapes made of four cells joined edge to edge (Golomb 1954). The four region shapes are tetromino shapes themselves: square, elongated, L-shaped and T-shaped regions are each built from four equal blocks, so all have the same area. Every

region has the same total population of 28.1 million, distributed among two large cities, six major cities with suburbs, 16 mid-sized cities and a rural background, with a sparsely populated inland area covering 18% of the region (Online Resource 1, Text S1).

Number of counties. Counties are formed by grouping the four-cell pieces. For each shape, 16 maps cross the four numbers of counties with the four levels of county size inequality, giving 64 maps (Figures 2 and S1). In equal maps counties have nearly equal area (coefficient of variation of area 0.02 to 0.14). In mild, moderate and extreme maps, urban counties are small and rural counties merge increasing and more variable numbers of pieces (coefficients of variation 0.48 to 0.55, 0.68 to 0.77 and 0.91 to 1.25), and in extreme maps the sparse inland area is divided into a few very large counties. Within a shape, all maps share the same landscape, population and true effect surface, so they differ only in how the region is divided into counties. County population is the sum over its pieces, and county coordinates are centroids.

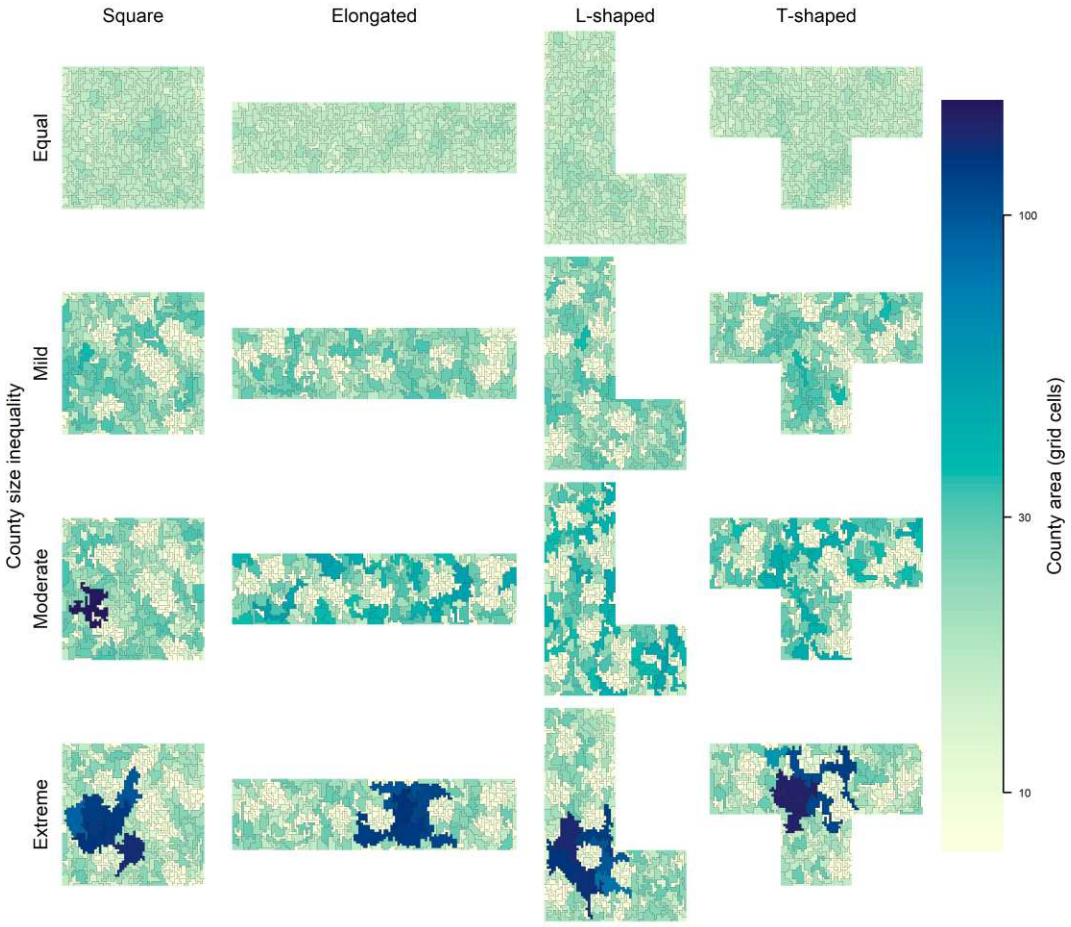


Figure 2. Simulation geography. Columns show the four region shapes (square, elongated, L-shaped, T-shaped), all of equal area and drawn at the same scale. Rows show the four levels of county size inequality, each with 240 counties. Color gives county area in grid cells (log scale). Lines are county boundaries. The full set of maps, for 60 to 480 counties, is shown in Figure S1.

Spatial pattern. Where the effect varies is set by a surface z defined on the tetrominoes. The surface is standardized to mean 0 and standard deviation 1 over the region and then truncated at ±2.5 to avoid extreme effects. The four patterns therefore differ in spatial arrangement but not in the amount of variation, which is set separately by the effect spread (Section 3.3). In the independent pattern, z is independent standard normal on each tetromino, so neighboring areas are unrelated. The local and regional patterns are Gaussian random fields with correlation $\exp(-(d/r)^2)$ and ranges of 7 and 17 cells, about 2.5 and 6 county widths in the maps with 480 counties. The boundary pattern is a step between the two sides of an irregular north–south line placed at a random position between 35% and 65% of the region's width (Figure S2). Each county's value $\bar{z}_s$ is the population-weighted mean of the surface over its tetrominoes. The surface depends only on the region shape, the pattern and the replicate, so all 16 maps of a shape share the same underlying effect and differ only in how it is aggregated into counties. Because averaging cancels unrelated values, the independent pattern produces the smallest between-county variation.

Temperature. The exposure is a weekly temperature series for each county over 520 weeks, with 8 additional weeks for the lag structure. Temperature of county *s* in week *t* is

$$T_{st} = \bar{T}_s + A_s \cos(2\pi(t - 30)/52.18) + a_{s,y(t)} + 2.5\, w_{st} + e_{st}.$$

The county mean $\bar{T}_s$ is the population-weighted mean of one of four mean-temperature layouts, each spanning 10 °C across the region: a north–south gradient with a warm south, the same gradient with a cool ridge, warm river valleys, and warm north and east edges (Figure S3). An urban heat island adds up to 2 °C in large cities and their suburbs. The seasonal amplitude A_s increases from 8 °C in the south to 12 °C in the north. Year-to-year anomalies $a_{s,y}$ (SD 1 °C) and weekly anomalies w_{st} (first-order autoregressive with coefficient 0.6) are spatially correlated Gaussian fields with correlation $\exp(-(d/r)^2)$ and ranges of about 22 and 13 cells, fixed in physical distance on every map. County-level noise e_{st} has SD 0.5 °C.

3.3 Epidemiology

Outcome model. Weekly counts are Poisson, $y_{st} \sim \text{Poisson}(\mu_{st})$, $\log \mu_{st} = \log(P_s \lambda/10^5) + m_s \sum_{l=0}^{8} g(l)\, f(T_{s,t-l})$, where P_s is the county population, λ the event rate per 100,000 population, f the exposure–response function on the log relative risk scale, g the lag–response function and m_s the county multiplier. The lag weights are normalized to sum to one, so f is also the cumulative exposure–response over lags 0 to 8 weeks. The model has no trend or seasonal component outside temperature, so seasonality in the outcome arises only through the exposure.

Exposure–response. Four shapes represent forms commonly reported for temperature and health. All equal zero at the reference temperature T_0, the 40th percentile of temperature pooled over counties and weeks, and all are scaled by the effect size RR. The J-shaped function, typical of temperature and mortality, is quadratic on both sides of T_0. It

reaches log RR at the pooled 99th percentile T_{99} and a smaller cold effect, log(1 + 0.375(RR − 1)), at the 1st percentile T_1. The linear function rises steadily through T_0 and reaches log RR at T_{99}, so temperatures below T_0 carry a lower risk than the reference. The threshold function is zero below the 75th percentile T_{75} and linear above it, reaching log RR at T_{99}, representing an effect that appears only in hot weeks. The unimodal function is a Gaussian function of temperature centered at T_{75} with standard deviation $(T_{99} - T_{75})/2$, shifted to equal zero at T_0 and scaled so the relative risk equals RR at the peak. It represents an outcome that peaks at an intermediate temperature and declines at higher temperatures, as in the transmission of some vector-borne infectious diseases. Because the peak lies at T_{75}, each county's 99th percentile falls on the declining side of the curve.

Lag–response. Four forms over lags 0 to 8 weeks represent different timings of the effect. The decaying function, $g(\ell) \propto \exp(-\ell/2)$, has its largest effect in the week of exposure, declining over the following weeks. The delayed function is a Gaussian peak at lag 3 with standard deviation 1.2, as for an outcome with an incubation period. The short function places weights 0.6 and 0.4 on lags 0 and 1 only, an acute effect. The harvesting function is positive at lags 0 to 2 and negative at lags 3 to 8 (weights 1, 0.7, 0.3, −0.1, −0.2, −0.2, −0.15, −0.1 and −0.05 before normalization), so part of the early excess is offset by a later deficit, as in mortality displacement.

Effect size (RR). RR is the relative risk at the 99th percentile of temperature (at the peak for the unimodal function) and takes values 1.0, 1.2, 1.4 and 1.6. The null control block uses RR = 1.0 to model no association in any county.

Effect spread. Each county's exposure–lag–response curve is the common curve scaled by a county multiplier, $m_s = \exp(\sigma \bar{z}_s - \sigma^2/2)$, applied to the log relative risk. The effect spread σ (0, 0.15, 0.30 or 0.50) sets how strongly the effect varies between counties. At σ = 0.30, for example, a county one standard deviation above average on the surface has its log relative risk multiplied by about 1.3, and one standard deviation below by about 0.7. The term $-\sigma^2/2$ keeps the multipliers centered near one, and σ = 0 gives the same curve in every county. Because $\bar{z}_s$ averages the surface over each county's tetrominoes, the realized between-county spread is smaller than σ on maps with fewer, larger counties, particularly for the independent pattern. The realized spread is recorded for every replicate. In the high heterogeneity block, the log multiplier has twice the standard deviation realized in the corresponding moderate heterogeneity replicate (Supplementary Text S1), with the same map, seed, temperature series and spatial pattern.

Event rate. The event rate λ takes values 0.125, 0.5, 2 and 8 events per 100,000 population per week. Over 520 weeks and a total population of 28.1 million, these give about 18,000, 73,000, 292,000, and 1.17 million expected events per replicate before the temperature effect. Per county, the expected total ranges from about 40 events to about 19,500. The lowest values occur at the lowest rate with 480 counties, and the highest at the highest rate with 60 counties. These figures assume equal county populations. In maps with unequal counties, the smallest counties have fewer events and the largest have more.

3.4 Modeling

All five methods use the same exposure series, lag range (0 to 8 weeks) and reference temperature T_0, and none includes terms beyond the exposure and county intercepts. Methods that fit a cross-basis use a natural cubic spline in temperature, with knots at the 25th, 50th and 75th percentiles of pooled temperature, and a natural cubic spline with 3 degrees of freedom in the lag.

Pooled. A single DLNM is fitted to all counties by Poisson regression, with county-specific intercepts. Every county receives the same curve, so the pooled model is the reference for no between-county variation.

Two-stage. A DLNM with the shared cross-basis is fitted separately in each county by Poisson regression. The county coefficients are then combined by multivariate meta-regression on the county centroid coordinates, fitted by restricted maximum likelihood with the mixmeta package. Each county's curve is its best linear unbiased prediction, which shrinks the county's own estimate toward a mean that varies linearly across the region. Borrowing is global, with every county contributing to every prediction. Counties without a first-stage estimate receive the meta-regression prediction at their coordinates.

Tensor. A Poisson generalized additive model with a tensor-product smooth over temperature, lag and the two coordinates, and with county intercepts, is fitted with mgcv. The temperature and lag margins are cubic regression splines with the same knots as the shared cross-basis, so they span the same functions as the other methods. The spatial margin is a thin-plate spline whose basis size grows with the number of counties, from 10 basis functions with 60 counties to 80 with 480. The exposure–lag–response therefore varies as one smooth surface across the region, with the amount of smoothing set by penalized likelihood.

Spatial Bayesian (SB-DLNM). A Poisson model in which each cross-basis coefficient has a global value plus a BYM2 random effect on the county adjacency graph, which combines a spatially structured and an unstructured component. Neighbors are counties that share a border, and the degree of smoothing is governed by hyperparameters with INLA default priors. The model is fitted with R-INLA using the Laplace approximation and grid integration over the hyperparameters. County curves are computed from 1,000 joint posterior samples.

Geographically weighted (GW-DLNM). The GW-DLNM is the method described in Section 2. It uses the same first-stage county estimates as the two-stage model and starts from the two-stage between-county covariance. Each county's estimate is shrunk toward a local mean of its k nearest neighbors, weighted by a bisquare kernel and by precision. The number of neighbors k is chosen by leave-one-out predictive likelihood from 15 values spaced geometrically between 4 and the number of counties with a first-stage estimate minus two (Section 2.3).

Estimates and intervals. For every method, county curves are computed from the estimated coefficients and their covariance matrix, relative to T_0. Intervals are 95% Wald intervals on the log relative risk scale, the estimate plus or minus 1.96 standard errors. The covariance is the model-based covariance for the pooled model, the covariance of the best linear unbiased predictor for the two-stage model and the GW-DLNM, the Bayesian posterior covariance for the tensor-product model, and the covariance of the 1,000 posterior samples for the SB-DLNM. All methods are therefore evaluated with intervals of the same form.

Failure handling. A run counted as a failure if it produced an error, non-finite estimates, a failed convergence check or a non-zero INLA mode status. Runs were also limited in time. A run counted as a failure if it did not finish within the job time limit or took more than three times the median runtime of that method on the same map within the same block. The pooled model, whose fits took under a second, was exempt from these runtime limits. Failures of the computing environment were rerun unchanged. A method failure received one retry with a fallback: the simplified Laplace approximation with empirical Bayes integration for INLA, up to 500 iterations for the tensor-product model, and the method-of-moments estimator for the meta-regression. Runs that failed the retry were recorded as failed and never replaced. Within each block, methods are compared on the replicates in which all five produced estimates. Each method's results on all of its own successful replicates, and first-attempt and final failure rates (Table 2), are also reported.

Computational environment. All analyses were carried out in R 4.6.1 (R Core Team 2026), using dlnm 2.4.10 (Gasparrini 2011), mixmeta 1.2.2 (Sera et al. 2019), mgcv 1.9.4 (Wood 2017) and R-INLA 26.8.7 (Rue, Martino, and Chopin 2009). Simulations were run on the Bridges-2 system at the Pittsburgh Supercomputing Center (Boerner et al. 2023). During this work the authors used Claude (Anthropic) to assist with writing and debugging R code for the simulation and analyses and with editing text. The authors reviewed, ran, and verified all code and take full responsibility for the content of the publication.

3.5 Estimands and performance

Estimands. The primary estimand is each county's cumulative log relative risk at its own 99th percentile of temperature, relative to T_0. It summarizes the effect of extreme heat and varies between counties through both the effect multiplier and each county's own temperature distribution. There are two secondary estimands. The first is the cumulative log relative risk at each county's own 95th percentile of temperature, a less extreme point on the curve. The second is each county's whole cumulative exposure–response curve over its own 5th to 95th percentile of temperature.

Accuracy. Accuracy at the own 99th and 95th percentiles is measured by the root mean squared error (RMSE) of the estimated log relative risk over the counties of a replicate. For the whole curve, the RMSE is computed over temperatures spanning each county's own range, pooled over counties.

Coverage. Coverage at the own 99th and 95th percentiles is the proportion of counties whose 95% interval contains the true value. For the whole curve, it is the proportion of temperatures and counties at which the interval contains the true curve.

Identifying high-risk counties. The area under the ROC curve (AUC) measures how well the estimated log relative risks at the own 99th percentile rank the 10% of counties with the highest true values above the others. The AUC is undefined in the null control block, where no county has an effect.

Summaries and comparisons. Each measure is computed within a replicate and summarized over the replicates of a block. RMSE is summarized by its median, and coverage and AUC by their means. All methods are applied to the same simulated data, so methods are also compared within replicates. These comparisons use the share of replicates in which each method has the lowest RMSE and paired ratios of RMSE between methods. The moderate and high heterogeneity blocks are compared within the 576 pairs of scenarios by the ratio of RMSE between the two members of each scenario pair.

Meta-model. The contribution of each design setting is estimated with a linear model for the log of the scenario-mean RMSE, over the 1,344 scenarios with a true effect. The model is fitted separately for the own 99th percentile and for the whole curve. It includes all design settings, an indicator for the high heterogeneity block, the interaction of each with method, and all two-way interactions among the settings and the block. Because adding the high block makes the design unbalanced, shares of variation are computed from type II sums of squares, relative to their total. Shares are reported for each term and summed over the geographic and epidemiological settings.

Spatial distribution of errors. To show where errors arise, county-level bias and RMSE at the own 99th percentile are averaged over the replicates of each scenario and mapped for every scenario and method. Counties with at least one grid cell on the edge of the region are also compared with interior counties.

Events and heterogeneity. The design settings of a real replicate are unknown to the analyst. Each replicate therefore also records two quantities an analyst can compute from their own data. The first is the median over counties of the observed total number of events per county. The second is the heterogeneity statistic I^2 of the first-stage estimates of the log relative risk at each county's own 99th percentile, computed from Cochran's Q with inverse-variance weights and set to zero when negative (Higgins and Thompson 2002). Because it is computed at each county's own 99th percentile, I^2 reflects both differences between the counties' exposure-response curves and differences in their temperature distributions, as an analyst's estimate would. Performance is related to these quantities in Section 4.

3.6 Application

We applied the five methods to the association between temperature and mortality in Minas Gerais, a state of about 21 million people in southeastern Brazil. The state is divided

into 853 municipalities that range from small rural units to the metropolitan area of Belo Horizonte, and its climate ranges from the hot, semi-arid north to the cooler highlands of the south. Individual death records were obtained from the Brazilian Mortality Information System (Sistema de Informações sobre Mortalidade, SIM) through OpenDataSUS and DATASUS. We used non-fetal deaths of residents, assigned to the municipality of residence and to epidemiological weeks by date of death. Deaths from external causes (ICD-10 V01 to Y98) and from COVID-19 (B34.2, U07.1, U07.2) were excluded, as were deaths with an unknown municipality of residence (about 0.1%). The analysis covers 2010 to 2019 and 2022 to 2024. The years 2020 and 2021 were excluded because mortality in those years was dominated by the COVID-19 pandemic. Temperature-mortality associations in Brazil, including Minas Gerais, have been examined previously, with cold-related risks generally exceeding heat-related risks (Kephart et al. 2022; Cortes et al. 2025). Daily minimum 2 m air temperature was taken from ERA5-Land on a 0.1° grid (Muñoz-Sabater et al. 2021). Municipal values are area-weighted means over the grid cells intersecting each municipality, averaged over the days of each week. Municipal population was interpolated geometrically between the 2010 and 2022 Censuses and extrapolated at the same rate to 2024.

The first stage was a quasi-Poisson DLNM for each municipality's weekly deaths, with the logarithm of population as an offset and a natural spline of time with 4 degrees of freedom per year for trend and season. Time was measured in calendar weeks, with spline knots placed within the analyzed years, and lagged temperatures were taken from the calendar weeks preceding each outcome week, including weeks in 2020 and 2021. The cross-basis was the one used in the simulation, with knots at the 25th, 50th and 75th percentiles of temperature and a natural spline with 3 degrees of freedom over lags of 0 to 8 weeks. Seven municipalities had too few deaths for a first-stage fit. Two-stage and GW-DLNM combine the first-stage estimates directly. Pooled, Tensor and SB-DLNM fit all municipalities jointly, and these models used each municipality's trend and seasonal component from its first-stage model as a fixed offset. All five methods therefore share the same control for trend and season. Tensor used a spatial basis of 142 thin-plate functions. For each municipality and method, we estimated the minimum-mortality temperature (MMT) within the municipality's own 1st to 99th percentiles of temperature, and the cumulative relative risk at its own 1st percentile (cold) and 99th percentile (heat) relative to the MMT.

4. RESULTS

4.1 Model fitting

All five methods produced estimates for 97.0% to 98.0% of the replicates in each block (Table 2). Pooled produced estimates for every replicate except one in the high heterogeneity block. Tensor failed on at most 0.01% of replicates in the null control, no-heterogeneity control, and moderate heterogeneity blocks, and on 0.22% in the high heterogeneity block. All of these fits exceeded the runtime limits. Two-stage and GW-

DLNM failed on 1.7% to 2.0% of replicates in each block. Almost all of these failures occurred in replicates with 480 counties and the lowest event rate. SB-DLNM failed on 0.2% of replicates in the null control and no-heterogeneity control blocks, 0.5% in the moderate and 1.0% in the high heterogeneity block. Most of these fits (241 of 282) exceeded the runtime limits. All comparisons that follow use the replicates for which all five methods produced estimates. Fit time increased with the number of counties for every method (Figure S4). At 480 counties, the median fit took under a second for Pooled, about 4 to 8 minutes for Two-stage, GW-DLNM and Tensor, and 13 to 16 minutes for SB-DLNM.

Table 2. Failure rates (%) by method and scenario block. Cells give the final failure rate, with the first-attempt failure rate in parentheses. Failure criteria and fallbacks are described in Section 3.4. The pooled model has no fallback and shows the final rate only. The header gives the number of replicates in each block (n, 30 per scenario).

Method	Null control (n = 7,680)	No heterogeneity (n = 5,760)	Moderate heterogeneity (n = 17,280)	High heterogeneity (n = 17,279)
Pooled	0.00	0.00	0.00	0.01
Two-stage	2.02 (2.02)	1.79 (1.79)	1.71 (1.71)	1.74 (1.75)
Tensor	0.00 (0.00)	0.00 (0.00)	0.01 (0.01)	0.22 (0.22)
SB-DLNM	0.23 (0.25)	0.17 (0.21)	0.51 (0.59)	0.96 (1.07)
GW-DLNM	2.02 (2.02)	1.79 (1.79)	1.71 (1.71)	1.83 (1.84)

4.2 Overall performance

The ranking of the methods depended on how much the effect varied between counties (Figure 3).

Null control. In the null control block, Pooled and Tensor had the lowest median RMSE (0.011 each), followed by SB-DLNM (0.018), GW-DLNM (0.029) and Two-stage (0.040). Tensor was the most accurate in 53% of replicates and Pooled in 45%. Coverage ranged from 0.95 for Pooled to 1.00 for SB-DLNM. The share of counties whose interval wrongly excluded zero was 4.9% for Pooled, 3.9% for Tensor, 3.5% for Two-stage, 1.4% for GW-DLNM and 0.1% for SB-DLNM (Table S1).

No heterogeneity. In the no-heterogeneity control block, Pooled had the lowest median RMSE (0.022), followed by Tensor and SB-DLNM (0.030 each), GW-DLNM (0.040) and Two-stage (0.052). Pooled was the most accurate in 52% of replicates and Tensor in 36%. Coverage was close to nominal for SB-DLNM (0.94), GW-DLNM (0.94), and Two-stage (0.93) but lower for Tensor (0.85) and Pooled (0.77). The low coverage of Pooled came from the threshold and unimodal shapes.

Moderate heterogeneity. In the moderate heterogeneity block, SB-DLNM had the lowest median RMSE (0.058), followed by Tensor (0.061), GW-DLNM (0.066), Pooled (0.074) and Two-stage (0.075). Tensor was most often the most accurate (35% of replicates), followed by SB-DLNM (23%) and GW-DLNM (21%). Coverage was close to nominal for SB-DLNM (0.92), GW-DLNM (0.91), and Two-stage (0.91) but lower for Tensor (0.71) and Pooled

(0.38). SB-DLNM identified the 10% of counties with the highest true effect best (mean AUC 0.94), followed by Tensor and GW-DLNM (0.92 each), Two-stage (0.89) and Pooled (0.88).

High heterogeneity. In the high heterogeneity block, errors increased over moderate heterogeneity for every method but unequally. Median RMSE increased by a factor of 1.82 for Pooled, 1.52 for Tensor, 1.38 for SB-DLNM, 1.27 for GW-DLNM, and 1.22 for Two-stage. SB-DLNM (0.080) and GW-DLNM (0.083) then had the lowest median RMSE, ahead of Tensor (0.089), Two-stage (0.092), and Pooled (0.122). GW-DLNM was most often the most accurate (31% of replicates), followed by Tensor (28%) and SB-DLNM (23%). Coverage stayed close to nominal for SB-DLNM (0.92), Two-stage (0.91), and GW-DLNM (0.91), but fell further for Tensor (0.64) and Pooled (0.24). SB-DLNM (AUC 0.94) and GW-DLNM (0.93) identified the highest-risk counties best, followed by Two-stage (0.92), Tensor (0.91), and Pooled (0.83).

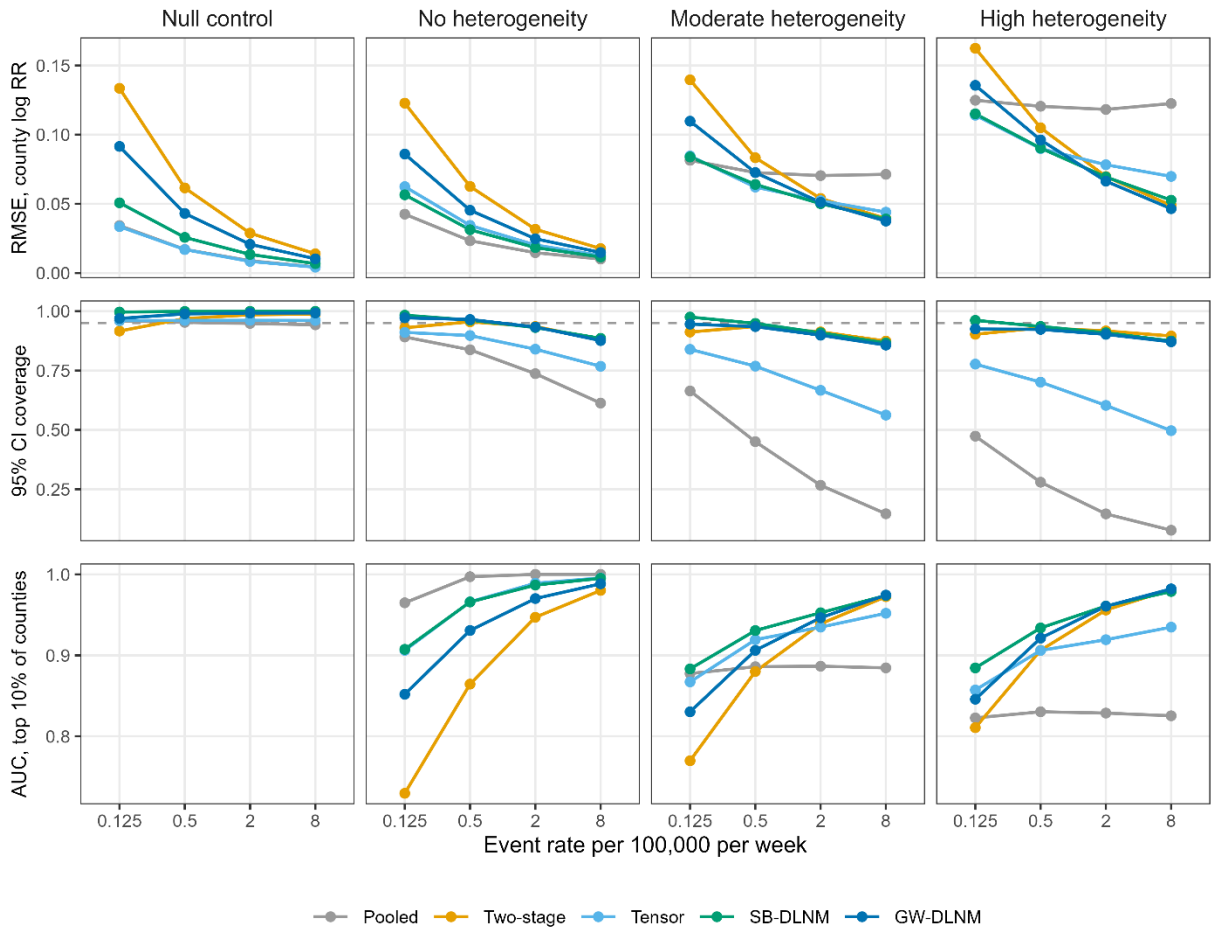


Figure 3. Performance at the 99th percentile of temperature by event rate, in the four scenario blocks (columns). Rows show the median RMSE of the cumulative log RR at each county's own 99th-percentile temperature, the mean coverage of its 95% confidence interval (dashed line at 0.95), and the mean area

under the ROC curve (AUC) for identifying the 10% of counties with the highest true log RR. AUC is undefined in the null control block, where no county has an effect.

4.3 Meta-model

The design settings and the heterogeneity block explained 93% of the variation in the log scenario-mean RMSE at the own 99th percentile (Table S2). The epidemiological settings dominated. Their main effects accounted for 61.3%, led by event rate (21.8%), effect spread (19.9%), effect size (11.2%) and exposure-response shape (8.3%). Interactions among them added 9.9%. The geographic settings accounted for 2.3%, mostly from the number of counties (1.5%) and the spatial pattern (0.7%). Region shape, county size inequality, temperature layout and lag-response each accounted for 0.1% or less. The heterogeneity block accounted for 4.9%. Terms involving method accounted for 11.7%. Most of this came from the interaction of method with the epidemiological settings (8.1%), chiefly event rate (3.6%) and effect spread (2.6%), and the main effect of method accounted for 1.9%. The ranking of the methods therefore depended mainly on the event rate and the effect spread. Results for the whole curve were similar, with the epidemiological settings accounting for 69.7% and their interaction with method for 10.1%.

4.4 Geography

Region shape, county size inequality and temperature layout had almost no influence on accuracy. Within each method, their levels changed the median RMSE by about 0.01 or less in every block. Only the number of counties and the spatial pattern of the effect changed how the methods compared (Figure 4). Results over the whole exposure-response curve were similar (Figure S5).

Number of counties. Accuracy declined as the number of counties increased. The decline differed between methods. In the high heterogeneity block, from 60 to 480 counties the median RMSE rose from 0.110 to 0.134 for Pooled, from 0.076 to 0.109 for Two-stage, from 0.071 to 0.099 for GW-DLNM and from 0.071 to 0.091 for SB-DLNM. Tensor changed least (0.088 to 0.091). Tensor was therefore most often the most accurate with many counties (35% of replicates at 480 counties) and GW-DLNM with few (36% at 60 counties).

Spatial pattern. When the effect varied smoothly over large areas or changed across a single boundary, Tensor was most often the most accurate. In the moderate heterogeneity block it was best in 47% of replicates with the regional and 46% with the boundary pattern. When neighboring counties were unrelated, in the independent pattern, GW-DLNM was most often the most accurate (26% of replicates with moderate and 42% with high heterogeneity), and Tensor had the highest median RMSE (0.096 in the high heterogeneity block, compared with 0.078 for SB-DLNM and GW-DLNM). With the local pattern, SB-DLNM and GW-DLNM had the lowest errors in both heterogeneity blocks.

4.5 Epidemiology

The event rate, effect spread, effect size and exposure–response shape had the largest effects on accuracy. The lag structure had little influence (Figure 5). Results over the whole curve were similar (Figure S6).

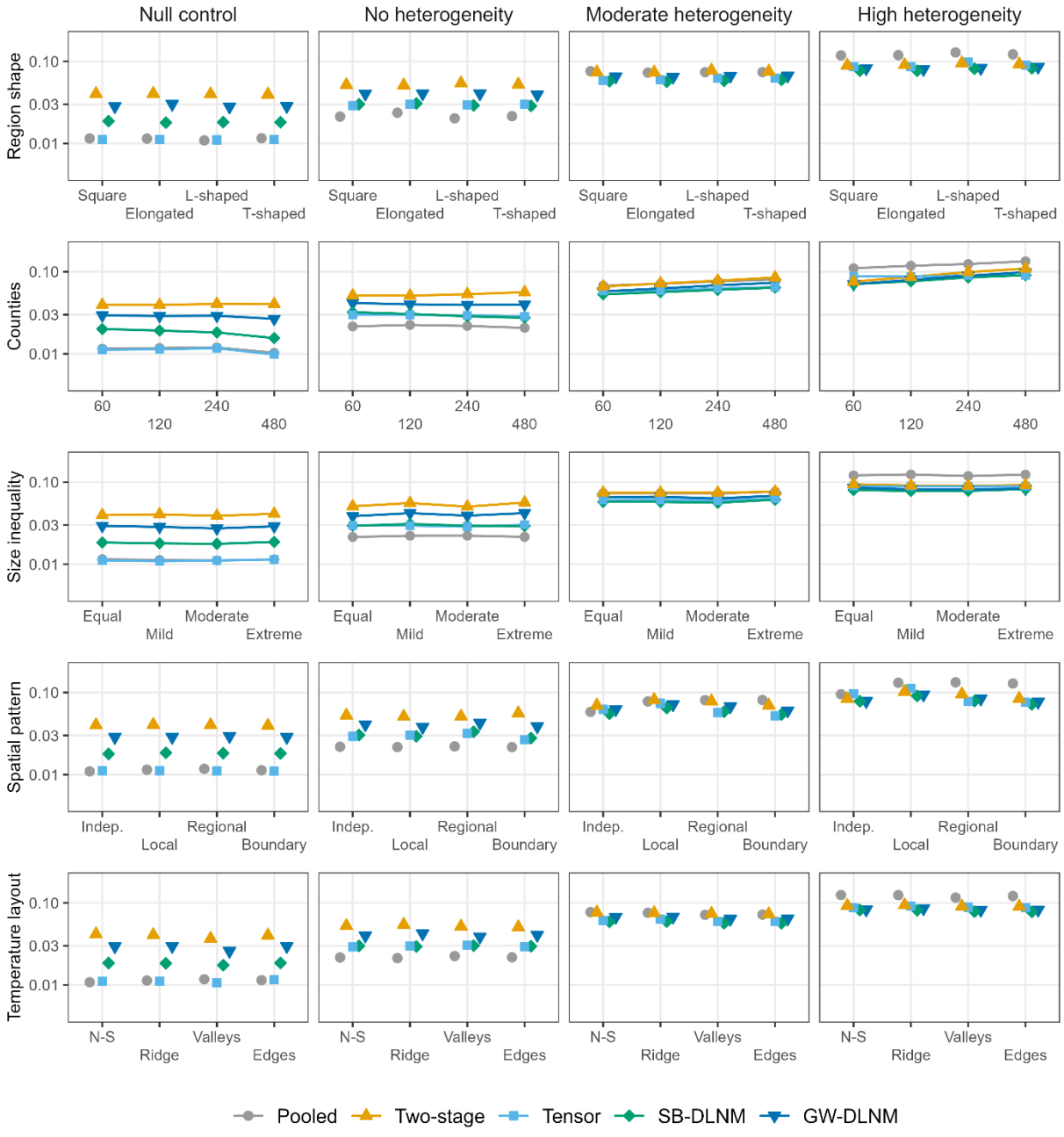


Figure 4. Accuracy across the geographic design settings in the four scenario blocks (columns). Each panel shows the median RMSE of the cumulative log RR at each county's own 99th-percentile temperature for each level of one setting, on a log scale. Settings that are fixed within a block show a single level.

Event rate. The event rate changed the ranking of the methods more than any other setting. At the lowest rate, Tensor and SB-DLNM had the lowest median RMSE (0.085 and 0.084 with moderate heterogeneity, 0.114 and 0.115 with high), and Two-stage had the highest (0.140 and 0.162). At the highest rate, GW-DLNM had the lowest median RMSE (0.038 with moderate and 0.046 with high heterogeneity) and was most often the most accurate (38% and 49% of replicates). The RMSE of Pooled changed little with the event rate (0.070 to 0.081 with moderate and 0.118 to 0.125 with high heterogeneity). Coverage declined as the event rate increased for every method. For SB-DLNM it fell from 0.98 to 0.87 and for GW-DLNM from 0.95 to 0.86 in the moderate heterogeneity block.

Effect spread and effect size. Errors increased with effect spread and effect size. In the high heterogeneity block, at the largest spread, GW-DLNM (0.103) and SB-DLNM (0.104) had the lowest median RMSE, followed by Two-stage (0.111), Tensor (0.129), and Pooled (0.194). GW-DLNM was most often the most accurate (39% of replicates). At the smallest spread in the moderate heterogeneity block, Tensor and SB-DLNM had the lowest median RMSE (0.044 each), and Tensor was most often the most accurate (40%). The effect size showed the same pattern. At RR 1.6 with high heterogeneity, GW-DLNM (0.106) and SB-DLNM (0.109) had the lowest median RMSE, and GW-DLNM was most often the most accurate (40%).

Lag-response. The lag structure changed the median RMSE by 0.01 or less within each method. With the harvesting function, GW-DLNM was more often the most accurate (30% of replicates with moderate and 41% with high heterogeneity) and SB-DLNM less often (12% and 9%) than with the other functions.

Exposure–response shape. For the J-shaped, linear and threshold functions, coverage at the own 99th percentile was 0.91 to 0.98 for Two-stage, SB-DLNM and GW-DLNM in the moderate and high heterogeneity blocks (Table S3). For the unimodal function it was 0.83 to 0.85, and the median RMSE of these methods was higher (0.091 to 0.112, compared with 0.048 to 0.091). At the own 95th percentile, the unimodal function gave coverage of 0.93 to 0.95, close to the 0.94 to 0.98 of the other shapes. Over the whole curve, coverage was 0.88 to 0.91 for the unimodal function and 0.96 to 0.99 for the others. The ranking of the methods for the unimodal function depended on the measure. At the own 99th percentile, Tensor was most often the most accurate with moderate heterogeneity (43% of replicates) and SB-DLNM rarely (4% with moderate and 6% with high heterogeneity). At the own 95th percentile and over the whole curve, GW-DLNM was most often the most accurate (26% to 41%), and SB-DLNM in 14% to 22% of replicates. In the no-heterogeneity control block, coverage of Pooled was 0.94 and 0.95 for the J-shaped and linear functions, 0.75 for the threshold, and 0.44 for the unimodal function.

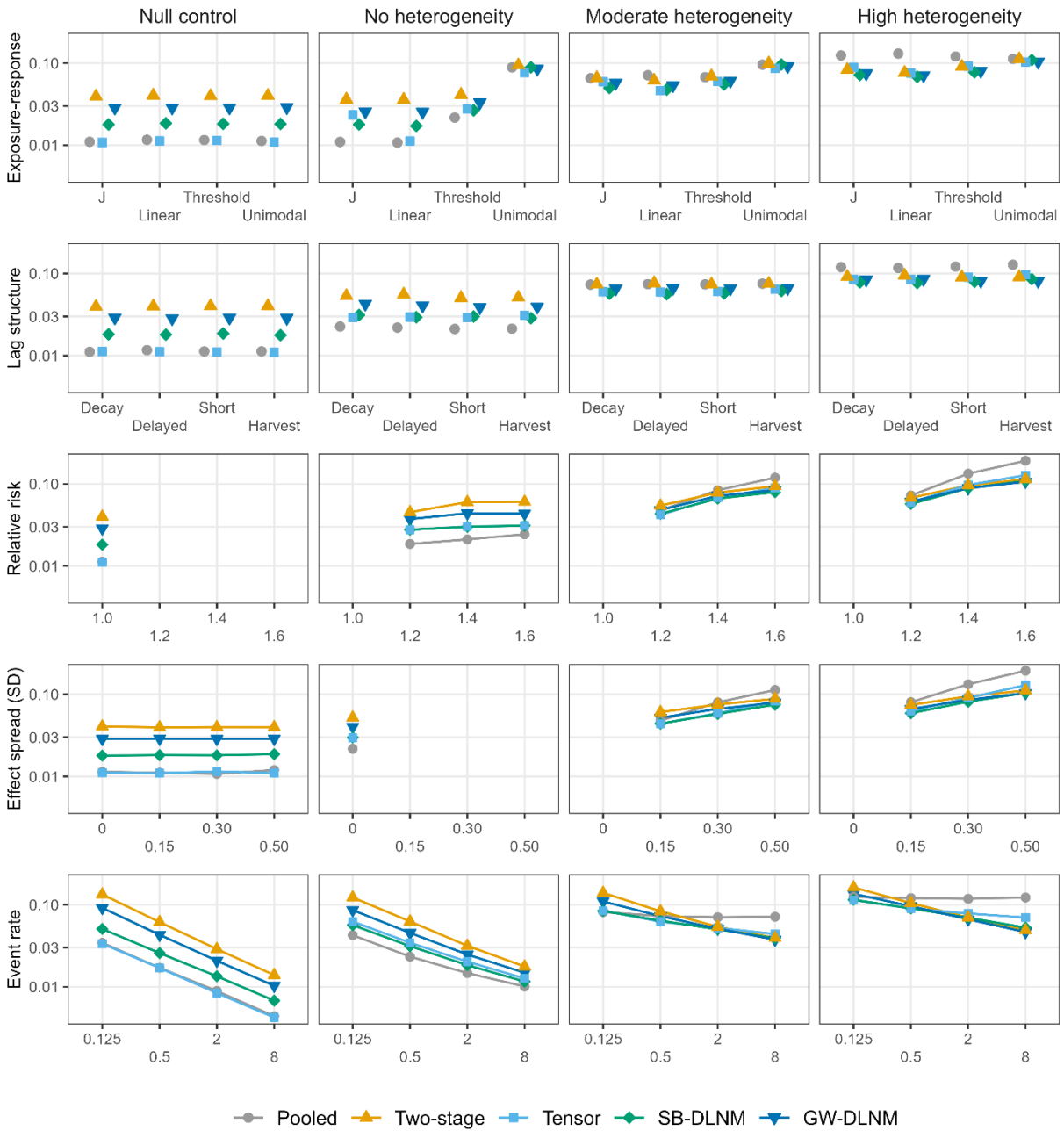

Null control
No heterogeneity
Moderate heterogeneity
High heterogeneity
Exposure-response
Lag structure
Relative risk
Effect spread (SD)
Event rate
0.10
0.03
0.01
J
Linear
Threshold
Unimodal
Decay
Delayed
Short
Harvest
1.0
1.2
1.4
1.6
0
0.15
0.30
0.50
0.125
0.5
2
8
Pooled
Two-stage
Tensor
SB-DLNM
GW-DLNM

Figure 5. Accuracy across the epidemiological design settings in the four scenario blocks (columns). Each panel shows the median RMSE of the cumulative log RR at each county's own 99th-percentile temperature for each level of one setting, on a log scale. Settings that are fixed within a block show a single level.

4.6 Error maps

In a scenario of the moderate heterogeneity block with spatially independent effect variation and high estimated heterogeneity, Pooled and Tensor produced large errors that clustered in space, with mean RMSE over counties of 0.103 and 0.105 (Figure 6). Two-stage, SB-DLNM and GW-DLNM had smaller errors with little spatial pattern (0.055, 0.056 and 0.055). Further examples across the range of I^2 are shown in Figures S7a-d, and maps for every scenario in Online Resource 2. Errors were slightly larger in border counties than in interior counties for every method (Figure S8). Over the scenarios with a true effect, the median ratio of border to interior RMSE was 1.04 for Pooled, 1.05 for GW-DLNM, 1.07 for SB-DLNM, 1.10 for Two-stage and 1.15 for Tensor. The ratio depended on the mean-temperature layout. It was largest when the warmest areas lay along the edges of the region (1.12 to 1.23) and fell below one for four of the five methods when the warmest areas followed interior river valleys (0.86 to 1.02).

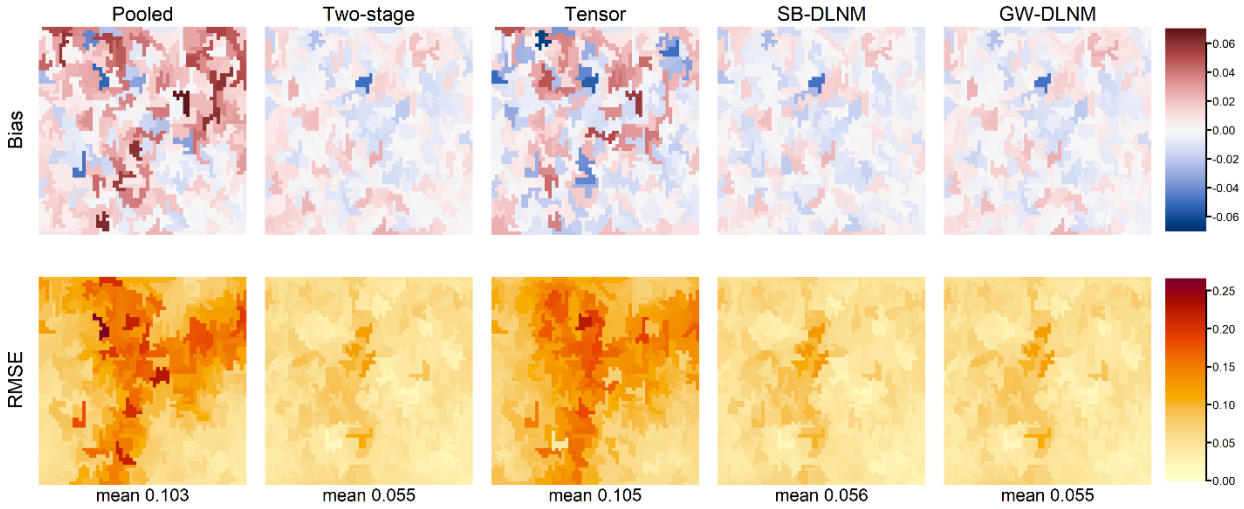


Figure 6. County-level errors in one scenario of the moderate heterogeneity block with high estimated heterogeneity (median I^2 0.80): square region, 240 counties of equal size, spatially independent effect variation (effect spread 0.5), RR 1.6 at the 99th percentile. The top row shows bias and the bottom row RMSE of the cumulative log RR at each county's own 99th-percentile temperature, averaged over each method's successful replicates.

4.7 Method selection

In the no-heterogeneity control, Pooled had the lowest RMSE at almost every I^2 and was the most accurate method in about half of the replicates (Figure 7, Table 3). In the moderate and high heterogeneity blocks, the RMSE of Pooled rose steeply with I^2. Tensor also lost

accuracy at high I^2 when counties had many events. The RMSE of Two-stage, SB-DLNM, and GW-DLNM rose more slowly, converging at high I^2. When I^2 was below 0.25, Tensor was most often the most accurate method in the moderate heterogeneity block, with shares of 0.39 to 0.43 of replicates (Table 3). In the high heterogeneity block, Tensor and SB-DLNM shared this position. When I^2 exceeded 0.5, GW-DLNM was most often the most accurate method in both heterogeneity blocks. Its share rose with the number of events per county, from 0.34 to 0.44 in the moderate heterogeneity block and from 0.39 to 0.52 in the high heterogeneity block. Within the two heterogeneity blocks, the only cell in which one method was most accurate in more than half of the replicates was high heterogeneity with I^2 above 0.5 and more than 2,000 events per county.

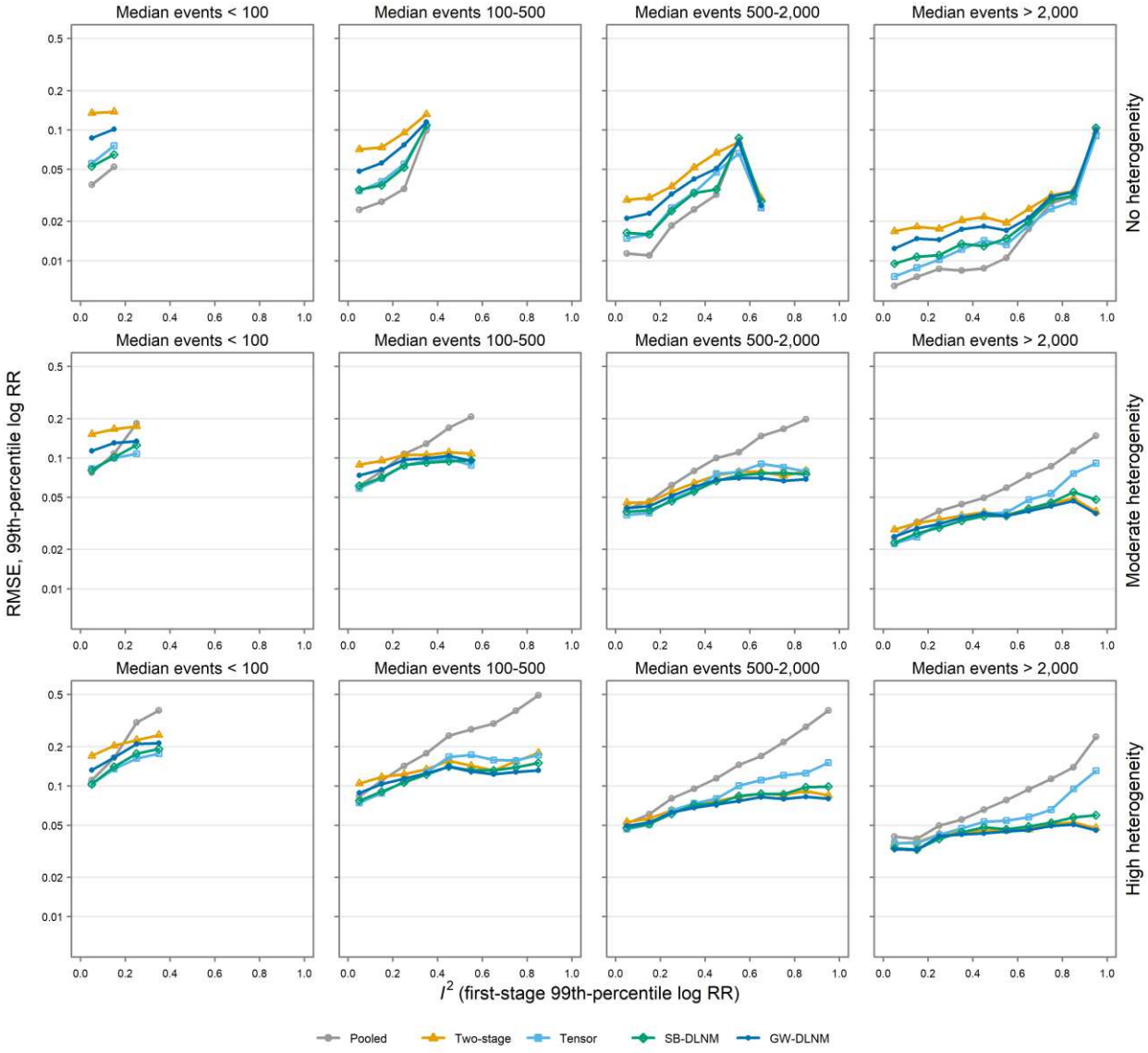


Figure 7. Accuracy against estimated heterogeneity, by scenario block. Each point is the median RMSE of the cumulative log RR at each county's own 99th-percentile temperature over the replicates in an I^2 interval of

width 0.1. I^2 is computed from the first-stage estimates of each replicate, as an analyst would compute it. Panels group replicates by the median over counties of total events per county. Rows show the three blocks with an effect. Intervals with fewer than 30 replicates are omitted. Summaries are over the replicates in which all five methods produced estimates: no heterogeneity (5,647), moderate heterogeneity (16,896) and high heterogeneity (16,754). Log scale.

Table 3. Share of replicates in which each method had the lowest RMSE of the cumulative log RR at each county's own 99th-percentile temperature, by scenario block, by I^2 of the first-stage estimates (below 0.25, 0.25 to 0.5, above 0.5) and by the median over counties of total events per county (below 500, 500 to 2,000, above 2,000). Shares in each row sum to one. The last column gives the number of replicates in each row. Summaries are over the replicates in which all five methods produced estimates: no heterogeneity (5,647), moderate heterogeneity (16,896) and high heterogeneity (16,754). The largest share in each row is in bold.

Block	I^2	Events per county	Pooled	Two-stage	Tensor	SB-DLNM	GW-DLNM	n
No heterogeneity	Below 0.25	Below 500	**0.58**	0.01	0.30	0.04	0.07	2,617
	Below 0.25	500 to 2,000	**0.56**	0.02	0.35	0.03	0.04	846
	Below 0.25	Above 2,000	**0.59**	0.01	0.37	0.02	0.01	352
	0.25 to 0.5	Below 500	**0.58**	0.01	0.21	0.03	0.18	106
	0.25 to 0.5	500 to 2,000	**0.44**	0.07	0.39	0.03	0.08	479
	0.25 to 0.5	Above 2,000	**0.54**	0.00	0.40	0.01	0.05	400
	Above 0.5	500 to 2,000	0.25	0.00	**0.52**	0.00	0.23	125
	Above 0.5	Above 2,000	0.33	0.01	**0.55**	0.01	0.11	722
Moderate heterogeneity	Below 0.25	Below 500	0.26	0.02	**0.39**	0.22	0.10	7,297
	Below 0.25	500 to 2,000	0.16	0.04	**0.43**	0.24	0.14	1,829
	Below 0.25	Above 2,000	0.11	0.05	**0.42**	0.30	0.13	414
	0.25 to 0.5	Below 500	0.09	0.03	**0.41**	0.30	0.18	821
	0.25 to 0.5	500 to 2,000	0.07	0.07	**0.34**	0.29	0.24	1,442
	0.25 to 0.5	Above 2,000	0.06	0.08	**0.41**	0.21	0.23	1,051
	Above 0.5	Below 500	0.00	0.02	**0.40**	0.24	0.34	83
	Above 0.5	500 to 2,000	0.02	0.08	0.25	0.24	**0.40**	998
	Above 0.5	Above 2,000	0.01	0.16	0.20	0.18	**0.44**	2,961
High heterogeneity	Below 0.25	Below 500	0.18	0.03	**0.40**	0.24	0.14	6,084
	Below 0.25	500 to 2,000	0.12	0.06	**0.33**	0.26	0.23	1,082
	Below 0.25	Above 2,000	0.08	0.09	0.24	**0.37**	0.22	197
	0.25 to 0.5	Below 500	0.04	0.06	0.31	**0.33**	0.26	1,516
	0.25 to 0.5	500 to 2,000	0.05	0.08	0.27	0.29	**0.31**	1,259
	0.25 to 0.5	Above 2,000	0.06	0.13	0.24	0.25	**0.32**	649
	Above 0.5	Below 500	0.00	0.09	0.24	0.28	**0.39**	507
	Above 0.5	500 to 2,000	0.02	0.11	0.18	0.20	**0.49**	1,879
	Above 0.5	Above 2,000	0.01	0.20	0.12	0.15	**0.52**	3,581

4.8 Application

The data contained 1.59 million deaths, with a median of 664 per municipality. Heterogeneity between municipalities was low. The I^2 of the first-stage log RR was 0.06 at the 99th percentile and 0.05 at the 1st percentile, computed relative to each municipality's median temperature. All five methods estimated a substantial cold effect and little heat effect in most municipalities (Figure 8). The median cold RR was 1.29 for Pooled, 1.35 for Two-stage, 1.36 for GW-DLNM, 1.48 for Tensor and 1.63 for SB-DLNM. The median heat RR was between 1.00 and 1.03 for all methods except SB-DLNM (1.09). The median MMT ranged from 19.3 °C to 20.3 °C. The methods differed mainly in how much variation between municipalities they retained. GW-DLNM and Two-stage produced smooth, closely related estimates (Spearman correlation of log heat RR 0.80). Tensor estimates were smooth at a broader scale (correlation with GW-DLNM 0.65). SB-DLNM retained much

more variation, with estimates in small municipalities close to their first-stage values. Its 95th percentile of cold RR was 5.4, compared with 1.4 to 2.4 for the other methods, and its correlation with GW-DLNM was 0.34.

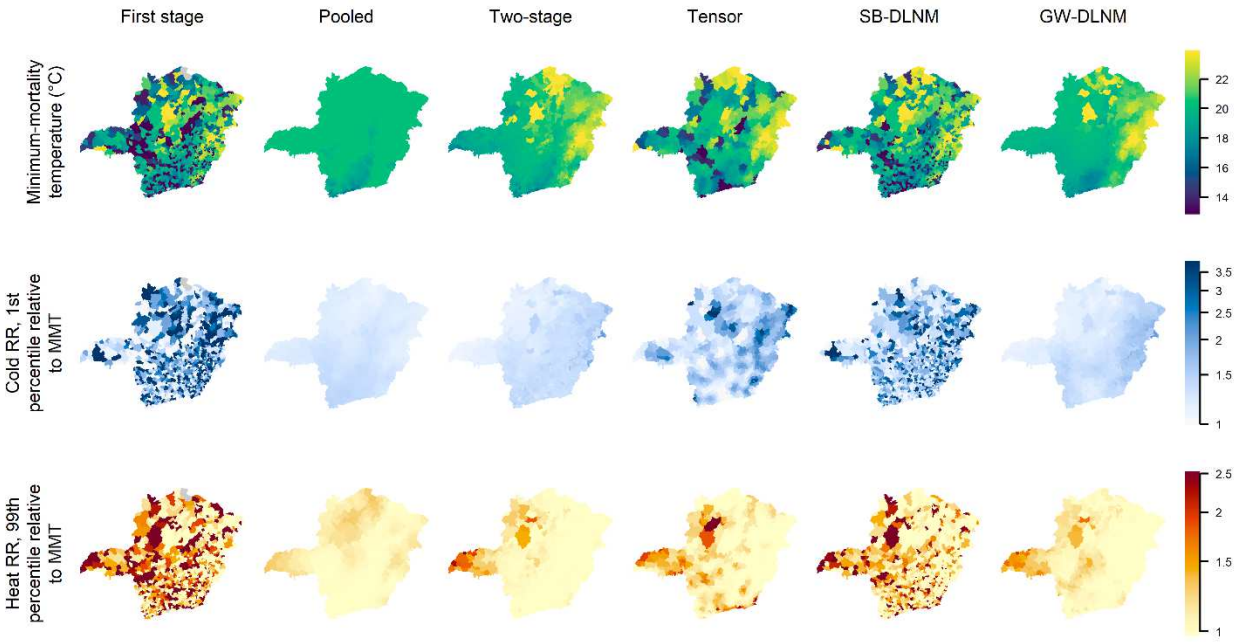


Figure 8. Temperature-mortality associations in the 853 municipalities of Minas Gerais, Brazil, 2010 to 2024 (2020 and 2021 excluded). Rows show the minimum-mortality temperature (MMT), the cumulative RR at each municipality's 1st percentile of weekly minimum temperature relative to its MMT (cold), and the RR at its 99th percentile relative to its MMT (heat). Columns show the first-stage estimates and the five methods. Gray municipalities (7) have no first-stage estimate.

5. DISCUSSION

The choice among methods for estimating unit-specific exposure-response curves depends on the structure of the effect and on the information available in each unit. Methods that borrow strength across units exchange variance for bias, and the size of that bias depends on how far each unit's true curve lies from the curve toward which it is shrunk. When effects vary between units, the target of shrinkage determines the bias. Each of the methods explored embodies a different assumption about how the curves of different units relate to one another, and each performed best where its assumption held. Pooled assumes one curve for all units and is the most efficient estimator when that assumption holds. Tensor represents the curves as a smooth surface over space and was most accurate when the effect varied smoothly over large areas or changed across a single boundary. Its penalty also smooths over differences between neighboring units that the surface cannot follow. Two-stage shrinks each unit toward a mean that varies linearly across the region and treats the remaining differences between units as exchangeable, so

it captures a broad gradient but not local spatial structure. SB-DLNM combines spatially structured and unstructured random effects, which allows it to adapt to a range of spatial patterns and is consistent with its low error across conditions. GW-DLNM shrinks each unit toward a mean built from its neighbors, which follows spatially structured variation and still stabilizes estimates in sparse units.

Accuracy and uncertainty quantification ranked the methods differently. A penalized smooth surface can achieve low average error while its intervals understate uncertainty where the true surface is rougher than the penalty allows (Wood 2017). A single pooled curve cannot represent variation between units (Park et al. 2025). Methods that estimate unit-specific curves with an explicit between-unit variance retain calibrated intervals, although coverage declines when unit estimates are very precise, because a small systematic bias then becomes large relative to the standard error (Morris 1983). SB-DLNM intervals were conservative when there was no effect, with coverage of 1.00 in the null control block. Where the purpose of an analysis is to identify high-risk areas, such as for targeting heat-health warnings, calibrated intervals are as important as point accuracy.

The weak influence of the geometry of the study region indicates that these findings transfer across regions of different shape and administrative structure. The relevant geography is the spatial structure of the effect and the size of the units in which it is observed, and unit size involves a trade-off. Smaller units can resolve finer spatial variation in the effect, but each contains fewer events, so first-stage estimates are less precise and every method relies more on borrowing from other units. Larger units contain more events, but they average over variation within each unit and reduce the heterogeneity that can be observed, an instance of the modifiable areal unit problem (Openshaw 1984). When data can be aggregated to more than one level, the choice of unit is therefore part of the choice of method.

In an applied analysis, the heterogeneity of the effect and its spatial structure are unknown, so a method must be chosen without knowledge of the underlying epidemiology (Park, Lee, and Kim 2026). The data provide two observable indicators after the first stage, the I^2 of the first-stage estimates and the number of events per unit. Together they indicate whether a pooled curve, a smooth spatial surface or local borrowing is likely to perform best. Neither is a direct measure of true heterogeneity. I^2 can be high even when all units share one effect, so it is best understood together with the number of events and prior knowledge of the exposure. When I^2 is low, Pooled or Tensor is a reasonable choice. When I^2 is high and units contain many events, GW-DLNM is recommended. SB-DLNM performs well across the range and is a robust default when heterogeneity is uncertain, although it is the most computationally demanding. The Minas Gerais application illustrates the case of low heterogeneity. The methods agreed on the main features of the temperature-mortality association, a substantial cold effect and little heat effect, consistent with previous studies of temperature and mortality in Brazil (Kephart et al. 2022; Cortes et al. 2025). They differed in how much variation between municipalities they retained. SB-DLNM retained the most, with estimates in small municipalities close to their first-stage values. Because heterogeneity was low, much of this variation is likely noise, which suggests that with

default priors SB-DLNM may borrow too little when many units have few events. When heterogeneity is low, the choice of method thus has little consequence for regional conclusions and more for the map of local effects.

This study has several limitations. The simulations considered a single exposure, weekly data and a Poisson outcome without overdispersion or seasonality beyond temperature. Overdispersion and seasonality, common in surveillance data, would mainly affect first-stage estimates in units with few events. Heterogeneity was limited to the amplitude of the exposure-response curve, and shifts in the curve, such as those arising from acclimatization, were not simulated. All methods used a common natural spline basis, which limited every method for the unimodal curve. SB-DLNM was affected most and was rarely the most accurate method for that curve. Stronger heterogeneity was represented only by doubling the between-county standard deviation, and heavy-tailed or outlying unit effects were not examined. Each method was applied with a single pre-specified configuration, and data-specific tuning could alter relative performance. For SB-DLNM, this included the INLA default priors. Its intervals were computed from the posterior covariance, in the same Wald form as for the other methods, and may differ from quantile-based credible intervals when posteriors are skewed. Finally, the application lacks a known truth and therefore illustrates the behavior of the methods without ranking them, although the low I^2 suggests methods with stronger smoothing are likely closer to the truth.

Despite these limitations, the simulations span a wide range of geographic and epidemiological conditions, and the results provide practical guidance for choosing among methods. No single method was best in every setting. The heterogeneity of the effect and the information available per unit should guide the choice, and when heterogeneity is high and units contain adequate data, GW-DLNM offers accurate estimates with calibrated uncertainty at modest computational cost.

Statements and Declarations

Funding. No external funding was received.

Competing interests. The authors have no competing interests to declare.

Data and code availability. Code is available to reviewers at https://zenodo.org/records/23570999?preview=1&token=eyJhbGciOiJIUzUxMiJ9.eyJpZCI6IjYxNWZhOTA1LTI5YzUtNGRhYy04NTc3LTg5ODliODY4NjJmMyIsImRhdGEiOnt9LCJyYW5kb20iOiJhOTQ3NGI3MzEyZTJmYzQ0M2ZjZDAyY2Y5ZDIyYmVkYyJ9.4CCEIW_QpxTRSImic5Wp2zt_VFpkyeFR4U_0TcMy6O7aaDrk4CdEPINN_gQBnTN0JHcXlSROI0QBYCBmAYimyw and will be deposited in Zenodo with a DOI upon acceptance. The mortality data for the application are publicly available from the Brazilian Mortality Information System (SIM, DATASUS), and the temperature data from ERA5-Land (Copernicus Climate Change Service).

Acknowledgements

This work used Bridges-2 at the Pittsburgh Supercomputing Center through allocation MTH260031 from the Advanced Cyberinfrastructure Coordination Ecosystem: Services & Support (ACCESS) program, which is supported by U.S. National Science Foundation grants #2138259, #2138286, #2138307, #2137603, and #2138296.

Author contributions

J.D.: Conceptualization, Methodology, Software, Formal analysis, Investigation, Visualization, Supervision, Project administration, Writing – original draft, Writing – review & editing. J.H.: Conceptualization, Validation, Writing – review & editing. H.A.M.: Conceptualization, Validation, Writing – review & editing. All authors read and approved the final manuscript.

REFERENCES

Boerner, T. J., Deems, S., Furlani, T. R., Knuth, S. L., and Towns, J. (2023), "Access: Advancing innovation: Nsf's advanced cyberinfrastructure coordination ecosystem: Services & support," in Practice and experience in advanced research computing 2023: Computing for the common good, pp. 173–176.

Briz-Redón, Á., Corberán-Vallet, A., Iftimi, A., and Íñiguez, C. 2025. A mixture of distributed lag non-linear models to account for spatially heterogeneous exposure-lag-response associations. arXiv.

Brunsdon, C., Fotheringham, A. S., and Charlton, M. E. (1996), "Geographically weighted regression: a method for exploring spatial nonstationarity," *Geographical analysis*, 28 (4), 281–298.

Chen, T.-T., Zhan, Z.-Y., Yu, Y.-M., Xu, L.-J., Guan, Y., and Ou, C.-Q. (2020), "Effects of hourly levels of ambient air pollution on ambulance emergency call-outs in Shenzhen, China," *Environmental Science and Pollution Research*, 27 (20), 24880–24888.

Chien, L.-C., and Yu, H.-L. (2014), "Impact of meteorological factors on the spatiotemporal patterns of dengue fever incidence," *Environment international*, 73, 46–56.

Cortes, T. R., Pescarini, J. M., Ranzani, O. T., Paixão, E. S., Neves, D. J. D., Rebouças, P., Ribeiro-Silva, R. C., Ferreira, A. J. F., Galvão, L. A. C., Brickley, E. B., Smeeth, L., Jacobson, L. V., Silveira, I. H., and Barreto, M. L. (2025), "Ambient temperature and non-accidental mortality: A nationwide space-time-stratified case-crossover study within the 100 million Brazilian Cohort," *Environ Int*, 205, 109892. DOI: 10.1016/j.envint.2025.109892.

Economou, T., Parliari, D., Tobias, A., Dawkins, L., Stoner, O., Steptoe, H., Lowe, R., Athanasiadou, M., Sarran, C., and Lelieveld, J. 2024. A unifying modelling approach for hierarchical distributed lag models. arXiv.

Fotheringham, A. S., Brunsdon, C., and Charlton, M. (2009), "Geographically weighted regression," *The Sage handbook of spatial analysis*, 1, 243–254.

Gasparrini, A. (2011), "Distributed lag linear and non-linear models in R: the package dlnm," *Journal of statistical software*, 43, 1–20.

--- (2014), "Modeling exposure–lag–response associations with distributed lag non-linear models," *Statistics in medicine*, 33 (5), 881–899.
--- (2021), "The case time series design," *Epidemiology (Cambridge, Mass.)*, 32 (6), 829.
--- (2022), "A tutorial on the case time series design for small-area analysis," *BMC Medical Research Methodology*, 22 (1), 129.
Gasparrini, A., Armstrong, B., and Kenward, M. G. (2010), "Distributed lag non-linear models," *Statistics in medicine*, 29 (21), 2224–2234.
Gasparrini, A., Armstrong, B., and Kenward, M. G. (2012), "Multivariate meta-analysis for non-linear and other multi-parameter associations," *Statistics in medicine*, 31 (29), 3821–3839.
Gasparrini, A., Guo, Y., Hashizume, M., Kinney, P. L., Petkova, E. P., Lavigne, E., Zanobetti, A., Schwartz, J. D., Tobias, A., and Leone, M. (2015), "Temporal variation in heat–mortality associations: a multicountry study," *Environmental health perspectives*, 123 (11), 1200–1207.
Golomb, S. W. (1954), "Checker boards and polyominoes," *The American Mathematical Monthly*, 61 (10), 675–682.
Guo, Y., Barnett, A. G., Pan, X., Yu, W., and Tong, S. (2011), "The impact of temperature on mortality in Tianjin, China: a case-crossover design with a distributed lag nonlinear model," *Environmental health perspectives*, 119 (12), 1719.
Hadianfar, A., Karimi, O., and Saki, A. (2026), "Spatially varying coefficient generalized distributed lag models for binary response with MCEM estimation," *BMC Medical Research Methodology*, 26, 166. DOI: 10.1186/s12874-026-02886-1.
Higgins, J. P., and Thompson, S. G. (2002), "Quantifying heterogeneity in a meta-analysis," *Statistics in medicine*, 21 (11), 1539–1558.
Kephart, J. L., Sánchez, B. N., Moore, J., Schinasi, L. H., Bakhtsiyarava, M., Ju, Y., Gouveia, N., Caiaffa, W. T., Dronova, I., and Arunachalam, S. (2022), "City-level impact of extreme temperatures and mortality in Latin America," *Nature medicine*, 28 (8), 1700–1705.
Masselot, P., and Gasparrini, A. (2025), "Modelling extensions for multi-location studies in environmental epidemiology," *Statistical Methods in Medical Research*, 34 (3), 615–629.
Miley, K. M., Downs, J., Beeman, S. P., and Unnasch, T. R. (2020), "Impact of the Southern Oscillation Index, temperature, and precipitation on Eastern equine encephalitis virus activity in Florida," *Journal of Medical Entomology*, 57 (5), 1604–1613.
Morris, C. N. (1983), "Parametric empirical Bayes inference: theory and applications," *Journal of the American statistical Association*, 78 (381), 47–55.
Morris, T. P., White, I. R., and Crowther, M. J. (2019), "Using simulation studies to evaluate statistical methods," *Statistics in medicine*, 38 (11), 2074–2102.
Muñoz-Sabater, J., Dutra, E., Agustí-Panareda, A., Albergel, C., Arduini, G., Balsamo, G., Boussetta, S., Choulga, M., Harrigan, S., and Hersbach, H. (2021), "ERA5-Land: A state-of-the-art global reanalysis dataset for land applications," *Earth system science data discussions*, 2021, 1–50.
Openshaw, S. (1984), "The modifiable areal unit problem," *Concepts and techniques in modern geography*.

Park, J., Sim, K., Yang, D., Kim, Y., Hashizume, M., and Chung, Y. (2025), "Nonparametric Bayesian Poisson hurdle random effects model: an application to temperature–suicide association study," *Environmental and Ecological Statistics*, 32 (2), 579–601.
Park, S., Lee, D., and Kim, M. (2026), "Locally sparse spatially varying coefficient models with altitude effects," *Environmental and Ecological Statistics*, 1–27.
Quijal-Zamorano, M., Martinez-Beneito, M. A., Ballester, J., and Marí-Dell'Olmo, M. (2024), "Spatial Bayesian distributed lag non-linear models (SB-DLNM) for small-area exposure-lag-response epidemiological modelling," *International journal of epidemiology*, 53 (3), dyae061. DOI: 10.1093/ije/dyae061.
R Core Team, R. 2026. R: A language and environment for statistical computing. Vienna, Austria.
Rue, H., Martino, S., and Chopin, N. (2009), "Approximate Bayesian Inference for Latent Gaussian models by using Integrated Nested Laplace Approximations," *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 71 (2), 319–392. DOI: 10.1111/j.1467-9868.2008.00700.x.
Rutten, S., Neyens, T., Duarte, E., and Faes, C. 2026a. Scaling Spatially Varying Distributed Lag Nonlinear Models for Nationwide Small-Area Data: A Divide-and-Conquer Approach. arXiv.
Rutten, S., Neyens, T., Duarte, E., Gasparrini, A., and Faes, C. (2026b), "Distributed lag non-linear models with spatial effect modification using Laplacian P-splines," *Environmental and Ecological Statistics*, 1–26.
Rutten, S., Sumalinab, B., Gressani, O., Neyens, T., Duarte, E., Hens, N., and Faes, C. (2025), "Penalized distributed lag non-linear models for small area data using Laplacian-P-splines," *Statistics and Computing*, 36 (1), 38. DOI: 10.1007/s11222-025-10790-9.
Sera, F., Armstrong, B., Blangiardo, M., and Gasparrini, A. (2019), "An extended mixed-effects framework for meta-analysis," *Statistics in medicine*, 38 (29), 5429–5444.
Sim, G., Kim, H., Zanobetti, A., Schwartz, J., and Chung, Y. (2018), "Non-Parametric Bayesian Multivariate Metaregression: An Application in Environmental Epidemiology," *Journal of the Royal Statistical Society Series C: Applied Statistics*, 67 (4), 881–896. DOI: 10.1111/rssc.12256.
Skrondal, A. (2000), "Design and analysis of Monte Carlo experiments: Attacking the conventional wisdom," *Multivariate Behavioral Research*, 35 (2), 137–167.
Wang, W., Liao, F., Yin, F., and Ma, Y. (2023), "Mapping the short-term exposure–response relationships between environmental factors and health outcomes and identifying the causes of heterogeneity: A multivariate-conditional-meta-autoregression-based two-stage strategy," *Spatial Statistics*, 53, 100720. DOI: https://doi.org/10.1016/j.spasta.2022.100720.
Warren, J. L., Luben, T. J., and Chang, H. H. (2020), "A spatially varying distributed lag model with application to an air pollution and term low birth weight study," *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 69 (3), 681–696. DOI: 10.1111/rssc.12407.
Wood, S. N. (2017), *Generalized additive models: an introduction with R*: chapman and hall/CRC.

Yang, H.-Y., Chen, S.-C., Chang, H.-Y., and Wang, Y.-Y. (2026), "Climate vulnerability factors for temperature-related respiratory mortality: a nationwide two-stage time-series study from 2008 to 2021," *Environmental Health*, 25 (1), 63. DOI: 10.1186/s12940-026-01309-4.
Yu, J., Park, J., Choi, T., Hashizume, M., Kim, Y., Honda, Y., and Chung, Y. (2021), "Nonparametric Bayesian Functional Meta-Regression: Applications in Environmental Epidemiology," *Journal of Agricultural, Biological and Environmental Statistics*, 26 (1), 45–70. DOI: 10.1007/s13253-020-00409-z.
Zhang, Y., Ni, H., Bai, L., Cheng, Q., Zhang, H., Wang, S., Xie, M., Zhao, D., and Su, H. (2019), "The short-term association between air pollution and childhood asthma hospital admissions in urban areas of Hefei City in China: a time-series study," *Environmental Research*, 169, 510–516.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- OnlineResource4metamodelfull.csv
- OnlineResource3errormapindex.csv
- OnlineResource2countyerrormaps.pdf
- OnlineResource1.pdf