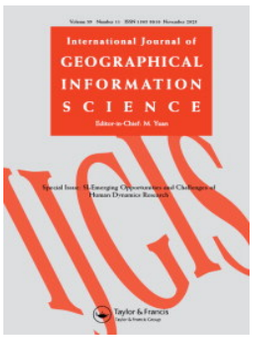

International Journal of Geographical Information Science

ISSN: 1365-8816 (Print) 1362-3087 (Online) Journal homepage: www.tandfonline.com/journals/tgis20

A model to identify causality for geographic patterns

Zuopei Zhang & Jinfeng Wang

To cite this article: Zuopei Zhang & Jinfeng Wang (06 Nov 2025): A model to identify causality for geographic patterns, International Journal of Geographical Information Science, DOI: 10.1080/13658816.2025.2581207

To link to this article: https://doi.org/10.1080/13658816.2025.2581207

Published online: 06 Nov 2025.

INTERNATIONAL JOURNAL OF GEOGRAPHICAL INFORMATION SCIENCE
https://doi.org/10.1080/13658816.2025.2581207

Taylor & Francis
Taylor & Francis Group

RESEARCH ARTICLE

A model to identify causality for geographic patterns

Zuopei Zhang[a,b] and Jinfeng Wang[a,b]

[a]State Key Laboratory of Resources and Environmental Information System, Institute of Geographic Sciences and Natural Resources Research, Chinese Academy of Sciences, Beijing, China; [b]University of Chinese Academy of Sciences, Beijing, China

ABSTRACT

Identifying causal relationships is essential for understanding the mechanisms through which natural and anthropogenic factors interact within Earth systems. However, in spatial cross-sectional data, the absence of temporal ordering poses significant challenges to traditional causal inference methods. This study proposes a novel Geographical Pattern Causality (GPC) model to detect positive, negative, dark causality and its strength between variables in spatial data. Grounded in dynamical systems theory and generalized embedding principles, the method transforms spatial neighbourhoods into lagged sequences, reconstructs the phase space, and compares symbolic trajectories to assess predictability and consistency in pattern changes—thereby inferring both the direction and type of causality. Case studies demonstrated that, compared to correlation analysis and Linear Non-Gaussian Acyclic Model (LiNGAM), the GPC model could reveal latent causal relationships among weakly correlated variables in geographical systems and capture diverse causal patterns. Despite limitations, such as sensitivity to noise and potential biases from proxy variables, the GPC model provides a novel framework for causal inference based on spatial observations, and it advances both the methodological and theoretical development of causality analysis in complex geographical systems.

ARTICLE HISTORY
Received 15 July 2025
Accepted 23 October 2025

KEYWORDS
Causal inference; spatial cross-sectional data; Geographical Pattern Causality (GPC); symbolic dynamics; nonlinear systems

1. Introduction

Causal inference occupies a foundational position in the Earth sciences. By elucidating the intricate interactions between natural and anthropogenic drivers, causal inference enables researchers to identify causal linkages among events and understand their underlying mechanisms, offering a rigorous basis for interpreting the evolution of geoscientific processes (Liu *et al.* 2007, Runge *et al.* 2023). Importantly, correlation does not imply causation; causal relationships not only involve the strength of association but also possess directionality and temporal ordering.

Although natural and social experiments—such as randomized controlled trials (RCTs)—are widely recognized as the "gold standard" for establishing causal relationships (Rubin 1974), their implementation is often hindered by limitations in data

CONTACT Jinfeng Wang wangjf@lreis.ac.cn

availability and logistical feasibility, particularly within large-scale spatiotemporal contexts (Liu *et al.* 2007, Runge *et al.* 2019a, 2023). Given the challenges in implementing RCTs at large spatiotemporal scales, researchers have increasingly relied on causal inference using observational data. While statistical approaches historically dominated the analysis of observational data, it was not until the 1980s that the concept of causal reasoning began to gain formal acceptance (Pearl and Mackenzie 2018). In recent decades, causal inference methodologies have advanced rapidly and now constitute a pivotal theoretical and methodological cornerstone for scientific inquiry in disciplines such as geography, climate science, and ecology (Kathpalia *et al.* 2022, Runge 2018, 2023).

At present, the theoretical landscape of causal inference can be broadly categorized into five major paradigms: (1) Granger causality theory, (2) Structural Causal Models (SCM), (3) Potential Outcome Framework (POF), (4) information-theoretic approaches centered on information flow, and (5) complex systems–oriented causal models. Granger causality, introduced by Granger (1969), infers causal relationships based on the predictive capacity of variables in time series data. Specifically, if the past values of variable X significantly improve the prediction of the current values of variable Y, then X is said to Granger-cause Y. This method was initially applied extensively in economics and was later adopted in natural science domains (Mosedale *et al.* 2006, Runge *et al.* 2019b). The Granger causality assumes system stationarity and is inherently limited to detecting linear relationships, which limits its effective to capturing nonlinear or structurally complex causal dynamics. To address these limitations, scholars have incorporated nonlinear analytical techniques into the development of extended methods, such as non-parametric regression approaches, local linear predictors, and Partial Directed Coherence (Baccalá and Sameshima 2001, Bell *et al.* 1996, Chen *et al.* 2004).

SCM developed by Pearl and colleagues (Dechter *et al.* 1991, Elhorst 2014, Pearl 2009), emphasizes the integration of causal graph structures with statistical reasoning to uncover causal dependencies among variables. This theoretical framework models causal structures among variables using Directed Acyclic Graphs (DAGs) and employs do-calculus to perform interventional inference. Building upon this theoretical foundation, researchers have developed a suite of graph-based causal discovery algorithms, including the Inductive Causation Algorithm (Pearl 2009), Peter-Clark algorithm (Spirtes and Glymour 1991), Linear Non-Gaussian Acyclic Model (LiNGAM) (Shimizu *et al.* 2006), and the Greedy Equivalence Search algorithm (Chickering 2003). Bayesian Belief Networks (BBNs) also use DAGs to represent directional relationships among variables. With appropriate assumptions or prior knowledge, BBNs can capture direct causal effects and provide probabilistic inference (Pearl 1982). These methods all require explicit modeling of the causal structure and typically assume the absence of unobserved confounders. Their practical implementation relies heavily on expert knowledge to construct reliable causal diagrams—an assumption that often proves challenging in complex systems such as ecosystems.

POF (Imbens and Rubin 2015, Rubin 1974) infers causal effects through counterfactual comparisons; that is, by positing the outcomes that a given observational unit would have exhibited under alternative treatment conditions. Grounded in counterfactual logic,

the POF evaluates the impact of a treatment variable on an outcome variable. In practice, this framework employs techniques, such as matching and weighted regression to adjust for confounding, with causal effects estimated by comparing actual outcomes to their corresponding counterfactual counterparts. The POF constitutes a foundational framework in econometrics and has given rise to a range of methodological variants tailored to different research scenarios. These include Propensity Score Matching (Rosenbaum and Rubin 1983), Instrumental Variables (Angrist *et al.* 1996, Angrist and Krueger 2001), Difference-in-Differences (Card and Krueger 2000, Heckman 1976), and Regression Discontinuity (Imbens and Lemieux 2008, Trochim 2001, Trochim 1984), all of which have found widespread applications across fields such as economics, sociology, and public health (Imbens and Rubin 2015, Wager and Athey 2018). A key assumption of the POF is that treatment assignment is ignorable or effectively random. However, in geoscientific contexts, this assumption is frequently violated due to phenomena, such as spatial spillover effects and spatial autocorrelation, which complicate the estimation of unbiased causal effects (Gao *et al.* 2022a).

Information-theoretic approaches to causal inference analyze the flow of information between variables to uncover nonlinear and asymmetric causal relationships. A representative method is Transfer Entropy (TE), which quantifies whether information flows from variable X to variable Y by assessing the deviation between their joint probability distribution and the product of their marginal distributions (Schreiber 2000). Under the assumption of Gaussian-distributed time series, TE is equivalent to one half of the F-statistic used in Granger causality, which indicates theoretical consistency between the two under certain conditions (Barnett *et al.* 2009). TE does not rely on linear assumptions and is thus more sensitive to detecting nonlinear and asymmetric patterns of information flow. In addition, Conditional Mutual Information is another widely used information-theoretic method for causal inference (Paluš *et al.* 2001). It identifies causal relationships based on the conditional shared information between variables and has been shown to be mathematically equivalent to TE under specific conditions (Paluš *et al.* 2001, Paluš and Vejmelka 2007). Information-theoretic methods impose stringent requirements on data quality and sample size. When applied to high-dimensional or complex systems, the effectiveness of these methods can be compromised by data sparsity and noise, which in turn reduce the stability and reliability of the inferred results.

According to dynamical systems theory, if two time series X and Y are causally related, they coexist on a shared attractor—an embedded representation of their joint dynamical system (Sugihara *et al.* 2012). Consequently, each variable implicitly contains information about the state of the other (Deyle and Sugihara 2011, Takens 1981). Convergent Cross Mapping (CCM) is an applied implementation of this theory (Sugihara *et al.* 2012). It determines the direction of causality between variables based on their ability to cross-map within a reconstructed state space. In other words, if changes in one variable consistently correspond to states of another variable in this reconstructed space, a directional causal relationship can be inferred. Based on Takens' embedding theorem (Takens 1981), CCM reconstructs the underlying attractor using time series data and identifies causal relationships by evaluating the success of cross-mapping predictions between variables. Several other models adhering to similar

principles have also been developed, including Cross-Mapping Smoothness (Diao *et al.* 2017) and Partial Cross Mapping (Egger and Lassmann 2015). The CCM series of methods can identify causal relationships and directions between variables and address the challenge of discovering causality in complex systems. However, CCM methods can detect the existence and direction of causal relationships, but they based on numerical predictions in the state space. They ignore the patterns of individual points and cannot explain specific types of causality. Stavroglou *et al.* (2019) used state space reconstruction, cross-prediction, and symbolic dynamics to analyze financial time series and proposed three distinct causal modes: forward causality, negative causality, and "dark causality." This approach has also been applied to ecological and neural system data to uncover potential causal mechanisms (Stavroglou *et al.* 2020). The Pattern Causality model builds upon CCM by also emphasizing their forms to enhance causal analysis and address some of the limitations of CCM.

Although the above models effectively address the problem of causal inference in time series data, they are not directly applicable to spatial cross-sectional data. Spatial cross-sectional data is a commonly used data type in Earth sciences, which typically consist of observational data from a specific point in time to record the spatial distribution and interactions of variables across different geographic locations. This type of data possesses a distinct spatial structure that can reveal the spatial ordering relationships between variables, which is crucial for understanding causal relationships in geographic processes.

The application of causal inference methods to spatial cross-sectional data has been widely researched. Herrera *et al.* (2016) extended Granger causality analysis by incorporating symbolic dynamics, enhancing its ability to identify nonlinear and asymmetric causal relationships. Gao *et al.* (2022a) transformed spatial cross-sectional data into spatial lags, applied the generalized embedding theory for state-space reconstruction (Deyle and Sugihara 2011), extended the application of CCM to spatial cross-sectional data, and developed the Geographical Convergent Cross Mapping (GCCM) (Gao *et al.* 2022b). This allowed for the identification of the existence, strength, and direction of causal relationships in geographic data, which has been validated in typical cases involving geographic spatial causal relationships, such as Net Primary Productivity and climate, soil heavy metals and human activity, and population and climate. The spatially transposed units with varying attributes (STUVA) problem refers to the challenge that heterogeneity among spatial units can lead to biased or confounded causal inferencing. By extracting the spatial lag sequences from spatial cross-sectional data, CCM can overcome common issues in causal inference, including the STUVA problem and spatial spillover effects and be widely applicable to social cross-sectional data (Akbari *et al.* 2023) for subsequent spatial causal inference based on spatial cross-sectional data. However, GCCM can only detect the causal directions and strengths. For specific variable interaction patterns, such as in the case of X→Y, where the causal relationship exists, GCCM does not detail whether changes in X positively or negatively influence Y, or whether there is no explicit promotion or suppression.

Given the challenges of causal inference for spatial cross-sectional data and the limitations of existing spatiotemporal causal inference models, we propose a Geographic Pattern Causality model (GPC). This model aims to identify positive, negative, and dark

causal relationships in data for Earth systems. GPC is based on dynamical system theory (Whitney 1936, Sauer *et al.* 1991), generalized embedding theory (Deyle and Sugihara 2011, Schiff *et al.* 1996), and symbolic dynamics theory (Morse and Hedlund 1938) to detect causal relationships between different geographic variables through state-space mapping of spatial cross-sectional data. The core idea of GPC is to transform each raster variable into a spatial lag sequence; that is, by aggregating information from a specific grid point and its surrounding neighborhood to form a spatial trajectory resembling a time series. This is achieved through constructing spatial lag sequences for state-space reconstruction, which allows each spatial unit to be represented as a multidimensional state point. Subsequently, we performed cross-mapping of the change patterns and corresponding symbols of points in the state space. By comparing the predicted strength and the degree of symbol matching, we determined the causal direction and type between the variables. This method integrates the phase embedding mechanism of spatial sequences and cross-mapping, and for the first time makes it possible to identify causal relationship types—positive, negative, or dark—from spatial data.

The remainder of this paper is organized as follows: Section 2 introduces the concept and framework of the GPC model; Section 3 demonstrates the results of two case studies from natural science research. Section 4 discusses the methodological contributions of the GPC, and Section 5 concludes the study.

2. GPC model

Causation involves the temporal succession and spatial contiguity of two similar entities (Hume 1985), which suggests that the spatial distribution of variables is just as critical as temporal ordering in uncovering causal linkages. The spatial propagation of objects inherently encodes temporal information, and spatial cross-sectional data capture these propagation patterns and their interactions. Consequently, understanding spatial ordering is essential for uncovering spatial causal relationships.

Previous studies have also attempted causal inference using spatial cross-sectional data and spatial distributions. For example, Wang *et al.* (2010) identified a causal relationship between nutrient deficiencies and neural tube defects through spatial analysis, with subsequent nutrient supplementation in the population leading to a significant decline in incidence (Chen *et al.* 2008). Darwin's pioneered work is a classic of this approach; by analyzing the spatial distribution of existing species, he inferred a causal link between environmental factors and biological evolution. However, extracting causal relationships from complex systems, such as geographic processes and ecosystems remains a significant challenge.

Dynamical systems theory offers a powerful toolkit for extracting causal structures within complex systems, particularly in identifying nonlinear relationships. According to Takens' theorem (Takens 1981), when the trajectory of a dynamical system converges onto an attractor—defined as a bounded and invariant manifold M—a functional mapping can be established between the system and its attractor. If two time series are causally connected, they share a common attractor, with each variable embedding information about the state of the other (Deyle and Sugihara 2011,

Takens 1981). Through further cross-prediction, one can examine the localized spatio-temporal interactions between M_x and M_y and the reconstructed attractor spaces of the respective variables.

Spatial cross-sectional data can be regarded as snapshots of a dynamical system. Each snapshot captures its state from a different perspective. According to the theory of generalized embedding, such data can be used to reconstruct the system's state space (Wells 2017). Accordingly, spatial lags can be transformed into spatial lag sequences to enable phase space reconstruction (see Figure 1 Step1).

In our proposed method, symbolic representations are employed to encode patterns of change in each variable. If the symbolic evolution of M_x, can accurately predict that of M_y, and this prediction exhibits either similarity, opposition, or neutral alignment, then a positive, negative, or dark causality from X to Y is inferred (Stavroglou *et al.* 2019). Specifically, positive causality denotes a congruent trend between the independent and dependent variables—such as simultaneous increases or decreases—while negative causality indicates inverse dynamics, where an increase in the independent variable corresponds to a decrease in the dependent variable, or vice versa. Dark causality refers to situations in which accurate predictions are achievable, yet no clear directional alignment, positive or negative, is observed. Hence, dark causality suggests the presence of latent coupling mechanisms between the variables (Stavroglou *et al.* 2019). The symbolic schemes employed in this study are detailed in Table 1.

In spatial cross-sectional data, for a given central pixel s_i and its L order spatial lags, the average (or alternative summary statistic) of all pixel values within each lag order is computed to derive the corresponding spatial lag values. These values are denoted as $h_s(x), h_{s(1)}(x), \ldots\ldots, h_{s(L-1)}(x)$.

According to the theory of generalized embedding, the resulting spatial lag sequence can be mapped onto a differentiable manifold and treated as an embedding $\psi(x,s)$, expressed as:

$$\psi(x,s) =< h_s(x), h_{s(1)}(x), \ldots\ldots, h_{s(L-1)}(x) > \tag{1}$$

Given the spatial sequence, and a specified embedding dimension E, we can construct a spatial shadow manifold where $L = 2E + 1$:

$$M_x = \begin{bmatrix} \psi(x,s) \\ \psi(x,s_1) \\ \ldots\ldots \\ \psi(x,s_n) \end{bmatrix} = \begin{bmatrix} h_s(x), h_{s(1)}(x), \ldots\ldots, h_{s(L-(E-1))}(x) \\ h_{s_1}(x), h_{s_1(1)}(x), \ldots\ldots, h_{s_1(L-(E-1))}(x) \\ \ldots\ldots \\ h_{s_n}(x), h_{s_n(1)}(x), \ldots\ldots, h_{n(L-(E-1))}(x) \end{bmatrix} \tag{2}$$

Following the same procedure, we can construct an analogous manifold M_y based on the dependent variable y (Figure 1 step2). Next, we compute the pairwise distances between all points in M_x, resulting in a distance matrix D_x:

$$D_x = \begin{bmatrix} d[\psi(x,s),\psi(x,s)] & \ldots\ldots & d[\psi(x,s),\psi(x,s_n)] \\ d[\psi(x,s_1),\psi(x,s) & \ldots\ldots & d[\psi(x,s_1),\psi(x,s_n) \\ \ldots\ldots & & \ldots\ldots \quad \ldots\ldots \\ d[\psi(x,s_n),\psi(x,s)] & \ldots\ldots & d[\psi(x,s_n),\psi(x,s_n)] \end{bmatrix} \tag{3}$$

Here, $d(*)$ denotes the distance between two points on the manifold, which can be computed using various distance metrics, such as:

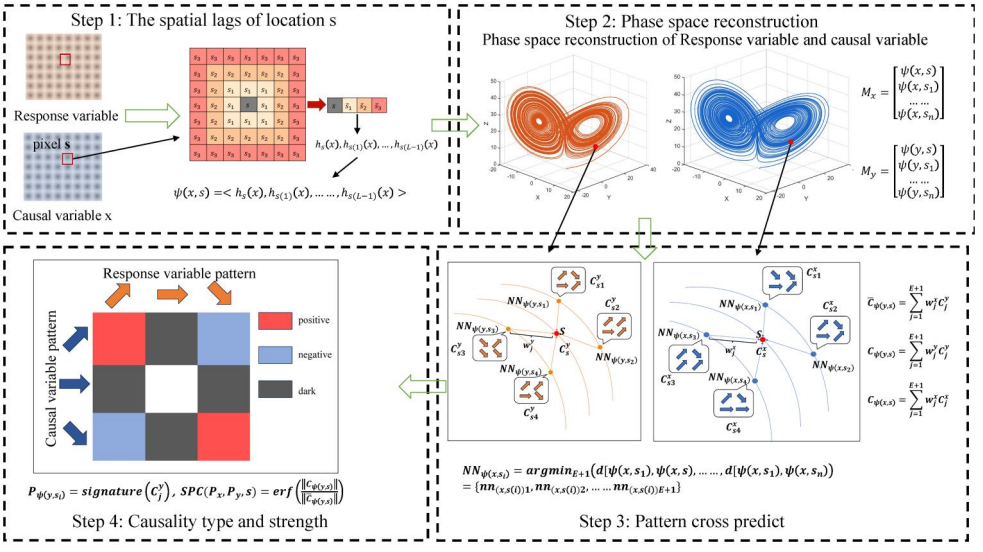

Figure 1. Workflow of GPC model.

Table 1. Symbols and definitions.

Symbols	Meaning
$\mathbf{x}, \mathbf{y}$	Spatial sequences of the independent and dependent variables
$\mathbf{E}$	Embedding dimension
$\mathbf{L}$	Lag order, $L = 2E + 1$
$\mathbf{s_i(l)}$	The l-th order spatial lag of the i-th spatial sequence center pixel
$\mathbf{h_s(x)}$	Local neighborhood average at location s for variable x
$\boldsymbol{\psi}(\mathbf{x}, \mathbf{s}),\ \boldsymbol{\psi}(\mathbf{y}, \mathbf{s})$	Spatial embedding of variable x and y at pixel s_i
$\mathbf{M_x}, \mathbf{M_y}$	Reconstructed embedding (shadow) manifold for variables x and y
$\mathbf{D_x}, \mathbf{D_y}$	Distance matrices in the reconstructed manifolds
$\mathbf{NN_{\psi(x,s_i)}}.\mathbf{NN_{\psi(y,s_i)}}$	Nearest neighbor sets (based on spatial distance)
$\mathbf{w_i^x},\ \mathbf{w_i^x}$	Weights of neighbors for variables x and y
$\mathbf{C_i^x}, \mathbf{C_i^y}, \mathbf{\bar{C}_i^y}$	Local variation pattern of x, y, and predicted pattern of y at s_i
$\mathbf{Signature}(*)$	Symbolic transformation function
$\mathbf{P_x}, \mathbf{P_y}$	Causal signatures

$$
\begin{aligned}
d_{L_1}(x_1, x_2) &= \sum |x_1 - x_2| \\
d_{L_2}(x_1, x_2) &= \sqrt{\sum (x_1 - x_2)^2}
\end{aligned}
\tag{4}
$$

Subsequently, based on the bounded simplex theory, the $E+1$ nearest neighbors of each point are identified:

$$
\begin{aligned}
NN_{\psi(x,s_i)} &= argmin_{E+1}\left(d[\psi(x,s_1), \psi(x,s), \ldots\ldots, d[\psi(x,s_1), \psi(x,s_n)\right) \\
&= \{nn_{(x,s(i))1}, nn_{(x,s(i))2}, \ldots\ldots nn_{(x,s(i))E+1}\} \\
NN_{\psi(y,s_i)} &= argmin_{E+1}\left(d[\psi(y,s_1), \psi(y,s), \ldots\ldots, d[\psi(y,s_1), \psi(y,s_n)\right) \\
&= \{nn_{(y,s(i))1}, nn_{(y,s(i))2}, \ldots\ldots nn_{(y,s(i))E+1}\}
\end{aligned}
\tag{5}
$$

Since the manifolds M_x and M_y are explicitly constructed, we can compute the nearest-neighbor sequences for all points in both manifolds. Moreover, by applying the same procedure to the spatial lags of order L, we can use M_x to predict the nearest-neighbor structure of M_y. By comparing the predicted symbolic causal patterns

with the actual symbolic sequences in M_y, we can determine whether M_x exerts a positive, negative, or ambiguous (dark) causal influence on M_y.

We then compute the symbolic change patterns, which describe how the directional variation of spatially lagged sequences within each embedded vector. These pattern captures the relative increase or decreases between adjacent elements in the embedding sequence, transforming the numerical series into a qualitative representation of its local variation trend. For the embedded vector $\psi(y, s_i) = (h_s(y), h_{s(1)}(y), \ldots\ldots, h_{s(L-(E-1))}(y))$, the change pattern C_j^y is calculated as:

$$C_j^y = \left(\frac{h_{sj(1)}(y) - h_{sj}(y)}{h_{sj}(y)}, \frac{h_{sj(2)}(y) - h_{sj(1)}(y)}{h_{sj(1)}(y)}, \ldots\ldots, \frac{h_{sj(L-(E-1))}(y) - h_{sj(L-(E-2))}(y)}{h_{sj(L-(E-2)}(y)} \right) \tag{6}$$

As shown by the above computations, both $\psi(x, s)$ and $\psi(y, s)$ have E + 1 nearest neighbors.

To classify different variation patterns, we define:

$$P_{\psi(y, s_i)} = signature\left(C_j^y\right), C_j^y \in \mathbb{R}^E \tag{7}$$

The signature (*) function converts a numerical change pattern into a symbolic representation. For embedding dimension E = 3, the pattern consists of two directional symbols (Figure 1 step3). For example, for a value at point s in variable X, denoted as $h_s(x)$, along with its neighboring values $h_{s(1)}(x)$ and $h_{s(2)}(x)$, if $h_{s(2)}(x) < h_{s(1)}(x) < h_s(x)$, this indicates a consistent increase from $h_{s(2)}(x)$ to $h_s(x)$, which is represented by the symbol ⬈⬈. A more detailed notation is provided in Table 2.

Once the variation patterns are computed, the final pattern at each center pixel is estimated by weighted averaging over its nearest neighbors (Figure 1 step 3):

$$\begin{aligned} C_{\psi(y,s)} &= \sum_{j=1}^{E+1} w_j^y C_j^y, w_j^y \in [0,1], C_j^y \in \mathbb{R}^E, for\ all\ NN_{\psi(y,s_i)} \\ \overline{C}_{\psi(y,s)} &= \sum_{j=1}^{E+1} w_j^x C_j^y, w_j^x \in [0,1], C_j^y \in \mathbb{R}^E, for\ all\ NN_{\psi(x,s_i)} \\ C_{\psi(x,s)} &= \sum_{j=1}^{E+1} w_j^x C_j^x, w_j^x \in [0,1], C_j^x \in \mathbb{R}^E, for\ all\ NN_{\psi(x,s_i)} \end{aligned} \tag{8}$$

where the weights are computed as:

$$\begin{aligned} w_j^x &= \frac{e^{-d\left(\psi(x,s),\psi(x,sj)\right)}}{\sum_j e^{-d\left(\psi(x,s),\psi(x,sj)\right)}} \\ w_j^y &= \frac{e^{-d(\psi(y,s),\psi(y,sj))}}{\sum_j e^{-d(\psi(y,s),\psi(y,sj))}} \end{aligned} \tag{9}$$

Furthermore, since both M_x and M_y are known, we compute the following three types of change pattern signatures: (1) The causal signature derived from the nearest-neighbor sequences in M_x; (2) The actual causal signature of each point in M_y; (3) The predicted causal signature of M_y based on the dynamics of M_x.These are denoted as follows:

Table 2. Symbol transformation table.

Symbol	Relation
	$h_{s(2)}(x) < h_{s(1)}(x) < h_s(x)$
	$h_{s(2)}(x) = h_{s(1)}(x) < h_s(x)$
	$h_{s(2)}(x) > h_{s(1)}(x) < h_s(x)$
	$h_{s(2)}(x) < h_{s(1)}(x) = h_s(x)$
	$h_{s(2)}(x) = h_{s(1)}(x) = h_s(x)$
	$h_{s(2)}(x) > h_{s(1)}(x) = h_s(x)$
	$h_{s(2)}(x) < h_{s(1)}(x) > h_s(x)$
	$h_{s(2)}(x) = h_{s(1)}(x) > h_s(x)$
	$h_{s(2)}(x) > h_{s(1)}(x) > h_s(x)$

$$\begin{aligned} P_x &= signature(C_{\psi(x,s)}) \\ P_y &= signature(C_{\psi(y,s)}) \\ \overline{P}_y &= signature(\overline{C}_{\psi(y,s)}) \end{aligned} \tag{10}$$

If the actual causal pattern P_y matches the predicted pattern $\overline{P}_y$, the prediction is considered valid. In this case, the influence strength at spatial point s, denoted as SPC (Spatial Pattern Causality), is defined and normalized using the Gaussian error function as follows:

$$SPC(P_x, P_y, s) = \text{erf}\left(\frac{\|C_{\psi(y,s)}\|}{\|\overline{C}_{\psi(y,s)}\|}\right) \tag{11}$$

where erf function was computed as:

$$\text{erf}(x) = \frac{1}{\sqrt{\pi}} \int_{-x}^{x} e^{-t^2} dt \tag{12}$$

For a given spatial point in M_y, let P_y denote its neighborhood's causal signature, and P_x denote the corresponding signature in M_x. If P_y and P_x have the same direction, a positive causal relationship is inferred; if they are opposite, it indicates a negative causal relationship; if they are neither similar nor opposite, the relationship is considered dark causality.

For each point in the phase space of Y, the average mode estimated from the dynamics of X can be compared with the true average mode determined by its own nearest neighbors in Y. Thus, the strength of the causal relationship from X to Y is measured by the degree of consistency between these two modes.

Based on the average modal signatures, a causal heatmap is constructed. The first column represents the true average modes of X, while the first row corresponds to those of Y. Each cell in the matrix indicates the proportion of agreement between the estimated Y mode from X's dynamics and the actual Y mode (Figure 1 step 4).

According to the definition of modal causality, the diagonal elements of the CS matrix represent consistency between the average modes of X and Y, indicating positive causality. The anti-diagonal elements indicate that the average modes of X and Y are opposite, indicating negative causality. The remaining elements indicate that the average modes of X and Y are neither the same nor opposite, representing dark causality. The overall causal strength can be calculated from the heatmap as follows:

Positive causal strength is the weighted sum of the diagonal elements of the CS matrix.

$$CS(positive) = \frac{1}{n}\sum\nolimits_{main}(matrix(CS)) \tag{13}$$

Negative causal strength is the weighted sum of the anti-diagonal elements of the CS matrix:

$$CS(negative) = \frac{1}{n}\sum\nolimits_{back}(matrix(CS)) \tag{14}$$

Dark causal strength is the weighted sum of the other elements of the CS matrix:

$$CS(dark) = \frac{1}{n}\sum\nolimits_{other}(matrix(CS)) \tag{15}$$

The final causal type between the two variables is determined as:

$$Cs = \max(CS(positive), CS(negative), CS(dark)) \tag{16}$$

After calculating the causality from $x \rightarrow y$, the variable order is reversed to compute $y \rightarrow x$, in order to determine whether the causal relationship between variables x and y is unidirectional or bidirectional. If the causal strengths in the two directions differ significantly, the causality is considered unidirectional, meaning that x causes y; in this case, the historical information of x is embedded in y, and the modal pattern of points on y's attractor can accurately predict the corresponding pattern on x's attractor. When both directions show strong causal strengths, it indicates a bidirectional causal relationship, denoted as $x \leftrightarrow y$. In the following section, we present case studies to demonstrate the application of the GPC method in practical scenarios.

3. Empirical evaluation of the GPC model through case studies

In this section, we present two case studies to demonstrate the implementation and interpretability of the GPC model, thereby evaluating its effectiveness across different scenarios. The first case investigates the causal relationship between soil heavy metal concentrations and nighttime lights—used as a proxy for human activity—while the second explores the causality between the Normalized Difference Vegetation Index (NDVI) and land surface temperature (LST). These causal links have been previously corroborated by earlier studies. Additionally, we apply Pearson correlation analysis and the LiNGAM causal discovery algorithm to the same datasets for comparison.

3.1. Extracted causations between soil pollution and multiple influencing factors

Soil heavy metal pollution is a major global environmental issue and has attracted widespread concern, particularly regarding its impact on human health (Zhang *et al.* 2018). Heavy metals in soil, such as lead, cadmium, copper, and mercury, can enter the human body through the food chain and drinking water, potentially causing a range of health problems, including poisoning, cancer, and neurological disorders (Feng *et al.* 2020, Kumar *et al.* 2020, Shi *et al.* 2023). The distribution of heavy metal

pollution varies significantly across regions, with areas experiencing high levels of urbanization and industrialization being more severely affected (Feng *et al.* 2020, Kumar *et al.* 2020, Shi *et al.* 2023). Soil heavy metal concentrations tend to remain stable over long periods, and large-scale surveys are often difficult to conduct. As a result, the temporal intervals of such data are typically 10 to 20 years, making it difficult to carry out attribution analyses based on time-series data in soil pollution research (Qin *et al.* 2021).

The sources and influencing factors of soil heavy metals are diverse, and their formation mechanisms are relatively complex. The types and causes of heavy metal pollution near roads, residential areas, and industrial zones vary significantly (Fang *et al.* 2025, Sun *et al.* 2025), but the existence of causal relationships has been demonstrated by receptor models and statistical analyses (Huang *et al.* 2018, Yuan *et al.* 2021). Soil heavy metal concentrations also exhibit distinct spatial characteristics, and the correlations between influencing factors and pollutants are often weak. This makes it difficult for traditional statistical methods to effectively analyze the relationships between soil heavy metals and their drivers. Therefore, accurately identifying the causal relationship between heavy metals and population density can serve as an important indicator of the reliability of a causal inference model.

In this study, heavy metals (Cu, Cd, Mg, and Pb) were selected as indicators of soil pollution, while nighttime light data were used as a proxy for population aggregation (Qin *et al.* 2021, Tan *et al.* 2018, Wang *et al.* 2018). Although population density does not directly determine heavy metal levels, areas with higher nighttime light intensity often coincide with stronger industrial, transportation, and residential activities that may contribute to soil pollution. The data were obtained from measurements of industrial pollutant concentrations and residential density in Illinois and Indiana, USA (Gao *et al.* 2023), with the spatial distributions of four soil heavy metal and nighttime light shown in Figure 2.

We sequentially applied the GPC to perform pairwise calculations between nighttime light data and four heavy metals (Cu, Mg, Pb, and Cd), aiming to examine their causal strength and pattern types. The results are summarized in Table 3. along with comparisons to traditional correlation analysis (Pearson) and structural causal discovery (LiNGAM).

According to Table 3, we found that the causal pathways from nighttime light to heavy metal concentrations showed consistently moderate to high causal strengths across all four metals (ranging from 0.40 to 0.44). Among them, the causal strength from nighttime light to Cu was the highest at 0.44, followed by Mg (0.42) and Pb (0.40). The causal pattern for these three metals was identified as *positive*, indicating that increased human activity intensity significantly contributes to higher concentrations of these heavy metals in soil, exhibiting a synergistic upward trend. For Cd, the causal strength was 0.41, but the causal pattern was classified as *dark*, possibly due to the more dispersed spatial distribution of cadmium in the soil.

In the reverse causal direction, i.e. from heavy metals to nighttime light, the overall causal strength decreased significantly. The causal strengths of Pb $\rightarrow$ nightlight and Cd $\rightarrow$ nightlight were only 0.14 and 0.04, respectively, which can be considered negligible. The causal strength of Cu $\rightarrow$ nightlight was 0.17, representing a weak

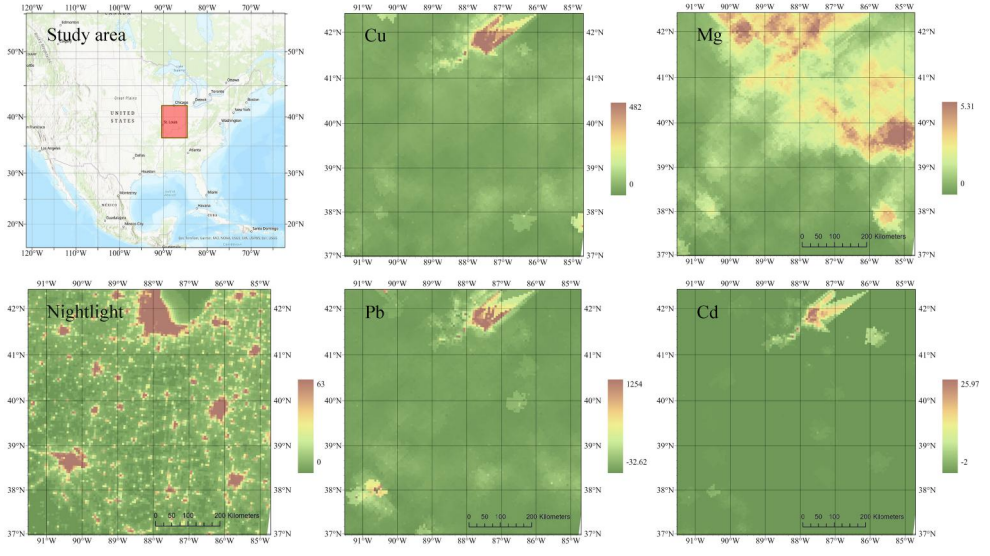

Figure 2. Location of the study area in Illinois and Indiana, USA, and spatial distributions of nightlight and four soil heavy metals. Heavy metal data are derived from soil surveys; nightlight data reflect industrial emissions and residential density.

Table 3. Causal patterns and strengths between nighttime light and heavy metals based on the GPC model, with comparisons to Pearson correlation and LiNGAM results.

	GPC model			
Variable	Causality strength	Causality type	Pearson	LiNGAM
nightlight → cu	0.44	Positive	0.22	0.15
cu → nightlight	0.17	Dark		0
nightlight → mg	0.42	Positive	0.24	0
mg → nightlight	0.24	Positive		0.013
nightlight → pb	0.4	Dark	0.16	0.05
pb → nightlight	0.14	Positive		0
nightlight → cd	0.41	Dark	0.16	0.15
cd → nightlight	0.04	Dark		0

association with a *dark* causal pattern, suggesting the absence of a stable directional structure. Mg → nightlight had a causal strength of 0.24. Although the causal strengths in the reverse direction are not zero, they do not indicate genuine causal relationships. This phenomenon may be attributed to the effect of *enslaved association*, where information from the true causal variable lingers in the evolution of the response variable, resulting in weak yet non-significant predictive power in the reverse mapping. Such patterns are commonly observed in state space reconstruction methods and should not be interpreted as evidence of causality without considering trend prediction, convergence behavior, and statistical significance.

In contrast, the Pearson correlation coefficients for all variable pairs remained low, ranging from 0.15 to 0.24. The nightlight–Cu pair showed the highest Pearson value (0.22), slightly above the others but still far below the causal strengths identified by the GPC method. This highlights the limitations of correlation analysis in uncovering dynamic dependencies within spatial data, particularly in systems characterized by nonlinearity and heterogeneity.

Additionally, we applied the LiNGAM method to conduct causal analysis on the dataset, using the results as a benchmark to compare with the outputs of the GPC model. As shown in Table 3, LiNGAM detected no causal relationships for most paths, with only weak signals identified for *nightlight* → *Cu* (0.15) and *nightlight* → *Cd* (0.15). For the causal directions involving *Pb* and *Mg*, all coefficients were zero. Notably, in the reverse directions (e.g. *Cd* → *nightlight*, *Pb* → *nightlight*), LiNGAM reported zero causal coefficients across the board, indicating its limited capacity to identify weakly coupled or nonlinearly dependent causal structures.

3.2. Extracted causations between NDVI and LST

The interaction between Land Surface Temperature (LST) and the Normalized Difference Vegetation Index (NDVI) is a central topic in research on urban climate regulation and ecosystem responses (Rasul *et al.* 2017, Zhou *et al.* 2015). LST serves as a key indicator of the urban heat island effect, while NDVI is widely used to represent vegetation cover and health. Numerous studies have shown that increasing urban green space helps to reduce surface temperatures and mitigate heat islands. Conversely, elevated temperatures can suppress vegetation growth through mechanisms such as soil drought and heat stress (Lin *et al.* 2020, Yuan and Bauer 2007). Although the correlation between LST and NDVI is well documented, the direction and structural mechanisms of causality remain debated (Gao *et al.* 2022a, 2022b). In particular, under nonlinear coupling and spatial heterogeneity often observed in urban environments, traditional methods face significant challenges in effectively identifying causal relationships (Sugihara *et al.* 2012).

To investigate the causal relationship between LST and NDVI in depth, Beijing and its surrounding areas (116.0°E–117.5°E, 39.0°N–41.0°N) were selected as the study region (Figure 3). MODIS product data were obtained from the Google Earth Engine (GEE) platform, extracting time series of Land Surface Temperature (MOD11A2) and NDVI (MOD13A2) with a spatial resolution of 1 km and a temporal resolution of 16 days. After standardizing the data, the GPC model was applied to identify causal patterns and assess the strength of bidirectional causal pathways between NDVI and LST. The results were then compared with those obtained from Pearson correlation analysis and the LiNGAM method.

The GPC analysis results indicate that the causal strength from NDVI to LST is 0.37, while the causal strength from LST to NDVI is 0.28. Both show moderate but relatively weak causal associations, and the causal type for both directions is Negative. This suggests that, on one hand, increased vegetation coverage significantly suppresses the rise of land surface temperature; on the other hand, high temperatures adversely affect vegetation growth. This bidirectional negative feedback mechanism represents a typical self-regulating pattern in urban ecosystems, reflecting the complex and stable coupling relationship between green spaces and temperature.

Traditional correlation analysis shows that the Pearson correlation coefficient between NDVI and LST is −0.64, indicating a significant negative correlation. Further analysis using the GPC method identified a clear negative causal relationship from NDVI to LST, indicating the cooling effect of vegetation on land surface temperature

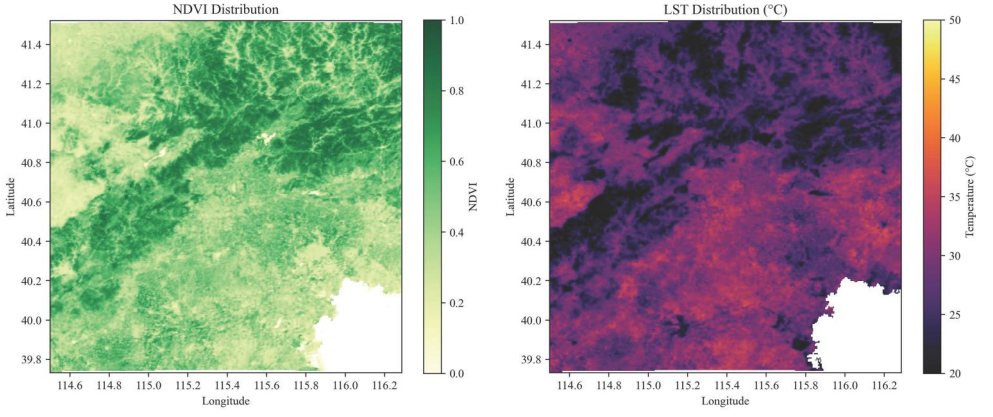

Figure 3. Spatial distributions of NDVI and LST in Beijing and surrounding areas. where NDVI serves as a proxy for vegetation density (e.g. tree coverage), and LST is used as a proxy for urban temperature. NDVI and LST jointly reflect the spatial heterogeneity of urban green infrastructure and the intensity of the urban heat island effect.

in the study area. This causal type not only reflects the statistical negative correlation but also indicates the underlying mechanism between the variables. In contrast, the LiNGAM method failed to identify any significant causal paths within the same variable pair, with causal coefficients equal to zero. This result is inconsistent with known physical mechanisms and expert knowledge and demonstrates LiNGAM's limited ability to detect nonlinear or bidirectional negative feedback structures. Compared to LiNGAM, GPC more effectively characterizes the implicit nonlinear causal relationships in complex systems, providing a novel approach for understanding driving mechanisms in ecological processes, and highlighting the limitations of traditional linear methods in practical applications.

In summary, the GPC model accurately identifies both the causal strength and the bidirectional negative pattern between NDVI and LST, showing stronger sensitivity and explanatory power compared to traditional methods. It is particularly suitable for revealing complex dynamic structural relationships among variables in urban ecosystems. The findings emphasize the important role of urban greening in regulating the microclimate and offer data support and methodological basis for future urban heat island mitigation strategies.

4. Discussion

4.1. The principle of the GPC model

Traditional causal inference methods often rely on the temporal order of variables to identify causal directions, such as Granger causality and CCM. However, cross-sectional spatial data lack temporal information and, therefore, are difficult to directly apply these time-based causal analysis approaches. To address this challenge, the GPC model draws on dynamical systems theory by introducing the concept of state space reconstruction. It extends the embedding logic of time series into the spatial domain to enable the identification of causal relationships among spatial variables.

In complex systems—whether dominated by deterministic or stochastic processes—the spatial or spatiotemporal evolution of variables jointly shapes specific attractor structures. According to the generalized embedding theorem, if two variables, X and Y, are causally related, they will coexist on the same attractor of the underlying dynamical system. This attractor represents the coupled spatiotemporal dynamics between the studied variables—for instance, the vegetation–temperature system in the NDVI–LST case and the human activity–pollution system in the nightlight–heavy metal case. However, mere coexistence on the same attractor is insufficient to determine causal direction. Deeper causal structures are reflected in the predictability between local trajectory patterns of the two variables. In other words, if X is a cause of Y, then X participates in the generative process of Y, and thus the attractor reconstructed from Y should contain information about X. This implies that one can assess whether Y contains effective predictive information about X through a process of cross-prediction, thereby inferring the causal direction.

GPC is built precisely on this foundation. Rather than relying on numerical correlation between variables, it examines whether the local trajectory of one variable can effectively predict the trajectory pattern of another. If $X \rightarrow Y$, then the local spatial neighborhood in M_y should be able to reliably recall the trajectory structure in M_x. This predictive capability reflects the internal transmission of system information from X to Y. Specifically, if $X \rightarrow Y$, the local neighborhood trajectories in M_y should accurately reconstruct corresponding patterns in M_x, signifying the directional propagation of information within the system. GPC moves beyond the use of numerical correlation coefficients to assess variable relationships. Instead, it relies on the similarity and directional mapping of trajectory evolutions within the embedded space to establish causal inference. Although it belongs to the class of predictive causality frameworks, its logic is grounded in cross-mapping within reconstructed phase space, aligning it theoretically with methods such as CCM, Pattern Causality, and GCCM.

4.2. The advantage of the GPC model

Case studies based on real environmental data demonstrate that the GPC method effectively identifies both direct and indirect causal relationships between variables and can capture directionally asymmetric bidirectional causal effects. Compared to traditional spatial correlation models, GPC not only enhances causal interpretability in nonlinear systems but also complements conventional spatial and spatiotemporal inference by revealing directional and even hidden causal interactions underlying spatial processes. GPC presents the following advantages:

GPC constructs spatial "trajectories" analogous to time series by jointly embedding geographical variables and their spatial neighborhood values. According to the generalized embedding theorem, this approach is equivalent to mapping the manifold of a dynamical system using multiple observation functions, enabling system state reconstruction without requiring true temporal dimensions. This design allows GPC to capture causal relationships even when time series data are absent or signals are weak.

GPC exhibits strong adaptability when handling nonlinear coupling and weakly correlated systems, making it suitable for various scales and types of geographical data,

and particularly advantageous for analyzing observational, non-experimental data. As illustrated in the aforementioned case studies, LiNGAM failed to effectively identify clear causal directions in both studies, possibly because it assumes linear relations, which are often violated in spatially coupled systems. In contrast, GPC successfully detected positive or dark causal relationships that align with established physical mechanisms (e.g. vegetation regulating surface temperature and human activities influencing soil pollution), demonstrating robustness in handling spatial coupling and nonlinear weak dependencies. Notably, in the heavy metal pollution case, although Pearson correlation coefficients were low, GPC was still able to uncover causal links in weakly correlated data, highlighting its applicability to causal discovery in weakly coupled systems. The identified causal links were further cross-validated with expert knowledge and known mechanisms to reduce potential false positives.

4.3. Limitations and future works

Despite the strong applicability and effectiveness of GPC, several limitations remain. Bartsev *et al.* (2021) pointed out that CCM is sensitive to periodic time series and noise, which may lead to incorrect causal inferences. However, Gao *et al.* (2023) argued that periodicity is rare in spatial data and therefore has a limited impact on causal inference results. Furthermore, GPC requires the effect variable to be somewhat "enslaved" by the cause variable, meaning the pattern of the cause variable should be reflected in the effect variable. Although GPC performs well on weakly correlated data, its causal inference capability diminishes when both variables are dominated by a common hidden factor or exhibit extremely weak coupling.

The choice of proxy variables can influence conclusions, a challenge common to all causal models (Chen *et al.* 2007, Prentice 1989). For example, nighttime light data typically correlate highly with socioeconomic development, but in unlit industrial areas or unconventional residential zones, actual human activity and pollution may be high while light intensity remains low (Bruederle and Hodler 2018). In such cases, the causal patterns identified by GPC may represent apparent causality, where the detected relationship reflects limitations of the proxy rather than a true underlying mechanism, and should therefore be interpreted cautiously with domain knowledge.

At the implementation level, GPC faces challenges in selecting spatial embedding dimensions and lag orders. A too low embedding dimension may overlook critical dependency structures, while too high a dimension may introduce noise or cause overfitting. Although GPC has relatively low parameter dependence compared to some machine learning methods, spatial scale and the choice of distance functions still impact the analysis results. Currently, three fixed distance metrics (Euclidean, Manhattan, and Chebyshev.) are used for weighting. Future work could adapt distance functions to the intrinsic properties of data and explore more informative definitions of spatial neighborhoods to further enhance the model's ability to capture complex spatial dependencies. Modal causal analysis relies on trajectory pattern recognition and matching, which can be sensitive to spatial noise and outliers, potentially causing unstable pattern recognition or bias in causal classification. Moreover, the current approach does not fully account for spatiotemporal causal interactions. While spatial

embedding simulates quasi-temporal behavior, it still insufficiently models causal dynamics, such as time delays and lag effects in real spatiotemporal sequences. Future improvements may involve integrating CCM, spatiotemporal convolutional networks, and other techniques to enhance the model's dynamic adaptability. Additionally, the current model assumes that spatial lags adequately capture dependencies among spatial processes which may not hold in more complex systems. Future work could investigate the underlying mechanisms linking spatial lags with spatial processes.

Future research can extend the GPC method to the temporal dimension, developing causal inference models suitable for spatiotemporal data to more comprehensively reveal the dynamic causal relationships among variables. In addition, combining complex dynamical system approaches with causal graph techniques can leverage the strengths of dynamical systems in identifying causal directions to generate more reliable causal networks. Integrating Structural Causal Models (SCM), research can focus on constructing causal graphs and estimating causal effects tailored to spatial cross-sectional data, thereby improving the accuracy and applicability of spatial causal inference. Moreover, future work could explore more informative definitions of spatial neighborhood to further enhance the model's ability to capture complex spatial dependencies and investigate the underlying mechanisms linking spatial lags with spatial processes.

5. Conclusion

In this study, we developed the GPC model to unveil causal relationships within spatial cross-sectional data. This addresses a key gap in causal inference methodologies, which often rely on temporal data and struggle to capture asymmetric, nonlinear, and heterogeneous causal structures in high-dimensional spatial systems. By introducing phase space reconstruction with spatial lag terms and symbolic dynamics, the GPC model enables robust detection of causal directions and types without requiring time-series data.

Our method was validated through two typical case studies, demonstrating its practical effectiveness in identifying meaningful causal patterns between environmental variables. These results provide a new paradigm for causal inference in geographical and ecological research, offering an analytical framework to unravel complex spatial causal mechanisms, deepen understanding of environmental processes, and support informed decision-making.

Despite its strengths, GPC has limitations, including sensitivity to spatial noise, challenges in selecting appropriate embedding parameters, and the current inability to fully model temporal dynamics such as time delays and lagged effects. Future research should focus on extending GPC to integrate spatiotemporal data, improving robustness against noise and outliers, and combining dynamical system-based causality with structural causal models to construct comprehensive causal networks tailored for spatial data.

Acknowledgement

We thank the anonymous reviewers for their constructive comments and suggestions. We also thank Dr. Zhangqi Yu for reviewing the literature citations and overall formatting of the manuscript.

Author contributions

Zuopei Zhang's contributions include conceptualization, methodology design, formal analysis, visualization, original draft preparation, and manuscript review and editing.
Jinfeng Wang was responsible for project administration, supervision, and manuscript review and editing.
All authors have read and agreed to the published version of the manuscript.

Disclosure statement

No potential conflict of interest was reported by the author(s).

Declaration of generative AI in scientific writing

This study used ChatGPT in the writing process to improve the readability and language of the manuscript.

Funding

This research was supported by the following funding sources: the National Natural Science Foundation of China [42471451] for the project "Spatial differentiation information transfer and effects of geoscience modeling" [2025.1–2028.12]; National Natural Science Foundation of China [42071375] for the project "Spatial stratification model selection theory of Geodetector" [2021.1–2024.12]; National Natural Science Foundation of China [41531179] for the project "Based on the theory of "Trinity" spatial sampling and the construction of the binary lookup table "[2021.1–2024.12]; National Key R & D Program of China [2022YFC3600800] for the project "Atlas and evolutionary trajectory of healthy life expectancy in China" [2025.1–2028.12]. We would like to express our sincere gratitude for the financial support provided by these esteemed institutions, which made this research possible.

Notes on contributors

Zuopei Zhang is a doctoral student at the Institute of Geographic Sciences and Natural Resources Research, Chinese Academy of Sciences.

Jinfeng Wang is a professor in Geographical Information Sciences at the Institute of Geographic Sciences and Natural Resources Research, Chinese Academy of Sciences. His research focuses on spatial statistics.

ORCID

Jinfeng Wang http://orcid.org/0000-0002-6687-9420

Data availability statement

The data and codes that support the findings of the present study are available on Figshare (https://doi.org/10.6084/m9.figshare.29607422). All model implementation and analyses were conducted in a Python 3.7 environment using Jupyter Notebook as the primary programming platform. The code relies on scientific computing and remote sensing data processing libraries, including NumPy, pandas, scikit-learn, and Rasterio.

References

Akbari, K., Winter, S., and Tomko, M., 2023. Spatial causality: a systematic review on spatial causal inference. *Geographical Analysis*, 55 (1), 56–89.

Angrist, J.D., Imbens, G.W., and Rubin, D.B., 1996. Identification of causal effects using instrumental variables. *Journal of the American Statistical Association*, 91 (434), 444–455.

Angrist, J.D., and Krueger, A.B., 2001. Instrumental variables and the search for identification: from supply and demand to natural experiments. *Journal of Economic Perspectives*, 15 (4), 69–85.

Baccalá, L.A., and Sameshima, K., 2001. Partial directed coherence: a new concept in neural structure determination. *Biological Cybernetics*, 84 (6), 463–474.

Barnett, L., Barrett, A.B., and Seth, A.K., 2009. Granger causality and transfer entropy are equivalent for gaussian variables. *Physical Review Letters*, 103 (23), 238701.

Bartsev, S., *et al.*, 2021. Imperfection of the convergent cross-mapping method. *IOP Conference Series: Materials Science and Engineering*, 1047 (1), 012081.

Bell, D., Kay, J., and Malley, J., 1996. A non-parametric approach to non-linear causality testing. *Economics Letters*, 51 (1), 7–18.

Bruederle, A., and Hodler, R., 2018. Nighttime lights as a proxy for human development at the local level. *PloS One*, 13 (9), e0202231.

Card, D., and Krueger, A.B., 2000. Minimum wages and employment: a case study of the fast-food industry in New Jersey and Pennsylvania: reply. *American Economic Review*, 90 (5), 1397–1420.

Chen, G., *et al.*, 2008. Prevention of NTDs with periconceptional multivitamin supplementation containing folic acid in China. *Birth Defects Research. Part A, Clinical and Molecular Teratology*, 82 (8), 592–596.

Chen, H., Geng, Z., and Jia, J., 2007. Criteria for surrogate end points. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 69 (5), 919–932.

Chen, Y., *et al.*, 2004. Analyzing multiple nonlinear time series with extended Granger causality. *Physics Letters A*, 324 (1), 26–35.

Chickering, D.M., 2003. Optimal structure identification with greedy search. *The Journal of Machine Learning Research*, 3, 507–554.

Dechter, R., Meiri, I., and Pearl, J., 1991. Temporal constraint networks. *Artificial Intelligence*, 49 (1-3), 61–95.

Deyle, E.R., and Sugihara, G., 2011. Generalized theorems for nonlinear state space reconstruction. *PloS One*, 6 (3), e18295.

Diao, M., Leonard, D., and Sing, T.F., 2017. Spatial-difference-in-differences models for impact of new mass rapid transit line on private housing values. *Regional Science and Urban Economics*, 67, 64–77.

Egger, P.H., and Lassmann, A., 2015. The causal impact of common native language on international trade: evidence from a spatial regression discontinuity design. *The Economic Journal*, 125 (584), 699–745.

Elhorst, J.P., 2014. *Spatial econometrics: from cross-sectional data to spatial panels*. Springer Berlin, Heidelberg.

Fang, S., *et al.*, 2025. Combined pollution of soil by heavy metals, microplastics, and pesticides: Mechanisms and anthropogenic drivers. *Journal of Hazardous Materials*, 485, 136812.

Feng, L., *et al.*, 2020. The systematic exploration of cadmium-accumulation characteristics of maize kernel in acidic soil with different pollution levels in China. *The Science of the Total Environment*, 729, 138972.

Gao, B., *et al.*, 2022a. Causal inference in spatial statistics. *Spatial Statistics*, 50, 100621.

Gao, B., *et al.*, 2022b. Temporally or spatially? Causation inference in Earth System Sciences. *Science Bulletin*, 67 (3), 232–235.

Gao, B., *et al.*, 2023. Causal inference from cross-sectional earth system data with geographical convergent cross mapping. *Nature Communications*, 14 (1), 5875.

Granger, C.W.J., 1969. Investigating causal relations by econometric models and cross-spectral methods. *Econometrica*, 37 (3), 424–438.

Heckman, J., 1976. The common structure of statistical models of truncation, sample selection and limited dependent variables and a simple estimator for such models. *Annals of Economic and Social Measurement*, 5 (4), 475–492.
Herrera, M., Mur, J., and Ruiz, M., 2016. Detecting causal relationships between spatial processes. *Papers in Regional Science*, 95 (3), 577–595.
Huang, Y., *et al.*, 2018. A modified receptor model for source apportionment of heavy metal pollution in soil. *Journal of Hazardous Materials*, 354, 161–169.
Hume, D., 1985. *A treatise of human nature*. London, England: Penguin Classics.
Imbens, G.W., and Lemieux, T., 2008. Regression discontinuity designs: A guide to practice. *Journal of Econometrics*, 142 (2), 615–635.
Imbens, G.W., and Rubin, D.B., 2015. *Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction*. Cambridge: Cambridge University Press.
Kathpalia, A., Manshour, P., and Paluš, M., 2022. Compression complexity with ordinal patterns for robust causal inference in irregularly sampled time series. *Scientific Reports*, 12 (1), 14170.
Kumar, A., *et al.*, 2020. Lead toxicity: health hazards, influence on food chain, and sustainable remediation approaches. *International Journal of Environmental Research and Public Health*, 17 (7), 2179.
Lin, M., *et al.*, 2020. Vegetation feedbacks during drought exacerbate ozone air pollution extremes in Europe. *Nature Climate Change*, 10 (5), 444–451.
Liu, J., *et al.*, 2007. Complexity of coupled human and natural systems. *Science (New York, N.Y.)*, 317 (5844), 1513–1516.
Morse, M., and Hedlund, G.A., 1938. Symbolic dynamics. *American Journal of Mathematics*, 60 (4), 815–866.
Mosedale, T.J., *et al.*, 2006. Granger causality of coupled climate processes: ocean feedback on the North Atlantic oscillation. *Journal of Climate*, 19 (7), 1182–1194.
Paluš, M., *et al.*, 2001. Synchronization as adjustment of information rates: detection from bivariate time series. *Physical Review E*, 63 (4), 046211.
Paluš, M., and Vejmelka, M., 2007. Directionality of coupling from bivariate time series: how to avoid false causalities and missed connections. *Physical Review E*, 75 (5), 056211.
Pearl, J., 1982. Reverend Bayes on inference engines: A distributed hierarchical approach. In *Proceedings of the Second AAAI Conference on Artificial Intelligence*, 133–136. AAAI Press.
Pearl, J., 2009. *Causality: models, reasoning and inference*. Cambridge, UK: Cambridge University Press.
Pearl, J., and Mackenzie, D., 2018. *The book of why: the new science of cause and effect*. New York, NY: Basic Books.
Prentice, R.L., 1989. Surrogate endpoints in clinical trials: definition and operational criteria. *Statistics in Medicine*, 8 (4), 431–440.
Qin, G., *et al.*, 2021. Soil heavy metal pollution and food safety in China: effects, sources and removing technology. *Chemosphere*, 267, 129205.
Rasul, A., *et al.*, 2017. A review on remote sensing of urban heat and cool islands. *Land*, 6 (2), 38.
Rosenbaum, P.R., and Rubin, D.B., 1983. The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70 (1), 41–55.
Rubin, D.B., 1974. Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, 66 (5), 688–701.
Runge, J., 2018. Causal network reconstruction from time series: from theoretical assumptions to practical estimation. *Chaos (Woodbury, N.Y.)*, 28 (7), 075310.
Runge, J., *et al.*, 2019a. Detecting and quantifying causal associations in large nonlinear time series datasets. *Science Advances*, 5 (11), eaau4996.
Runge, J., *et al.*, 2019b. Inferring causation from time series in Earth system sciences. *Nature Communications*, 10 (1), 2553.
Runge, J., *et al.*, 2023. Causal inference for time series. *Nature Reviews Earth & Environment*, 4 (7), 487–505.
Sauer, T., *et al.*, 1991. Embedology. *Journal of Statistical Physics*, 65 (3-4), 579–616.

Schiff, S.J., *et al.*, 1996. Detecting dynamical interdependence and generalized synchrony through mutual prediction in a neural ensemble. *Physical Review. E, Statistical Physics, Plasmas, Fluids, and Related Interdisciplinary Topics*, 54 (6), 6708–6724.

Schreiber, T., 2000. Measuring information transfer. *Physical Review Letters*, 85 (2), 461–464.

Shi, J., *et al.*, 2023. Spatiotemporal variation of soil heavy metals in China: the pollution status and risk assessment. *The Science of the Total Environment*, 871, 161768.

Shimizu, S., *et al.*, 2006. A linear Non-Gaussian acyclic model for causal discovery. *The Journal of Machine Learning Research*, 7, 2003–2030.

Spirtes, P., and Glymour, C., 1991. An algorithm for fast recovery of sparse causal graphs. *Social Science Computer Review*, 9 (1), 62–72.

Stavroglou, S.K., *et al.*, 2019. Hidden interactions in financial markets. *Proceedings of the National Academy of Sciences of the United States of America*, 116 (22), 10646–10651.

Stavroglou, S.K., *et al.*, 2020. Unveiling causal interactions in complex systems. *Proceedings of the National Academy of Sciences of the United States of America*, 117 (14), 7599–7605.

Sugihara, G., *et al.*, 2012. Detecting causality in complex ecosystems. *Science (New York, N.Y.)*, 338 (6106), 496–500.

Sun, S., *et al.*, 2025. Source-risk-driver analysis of heavy metal pollution in karst soils: An integrated assessment in eastern Yunnan, China. *Ecological Indicators*, 176, 113699.

Takens, F 1981. Detecting strange attractors in turbulence. *In*: D. Rand, *et al.*, ed. *Dynamical systems and turbulence, Warwick 1980*. Berlin, Heidelberg: Springer, 366–381.

Tan, M., *et al.*, 2018. Modeling population density based on nighttime light images and land use data in China. *Applied Geography*, 90, 239–247.

Trochim, W., 2001. Regression discontinuity design. *International Encyclopedia of the Social and Behavioral Sciences*, 19, 12940–12945.

Trochim, W.M.K., 1984. *Research design for program evaluation: the regression-discontinuity approach*. Beverly Hills: Sage.

Wager, S., and Athey, S., 2018. Estimation and inference of heterogeneous treatment effects using random forests. *Journal of the American Statistical Association*, 113 (523), 1228–1242.

Wang, J.F., *et al.*, 2010. Geographical detectors-based health risk assessment and its application in the neural tube defects study of the Heshun Region, China. *International Journal of Geographical Information Science*, 24 (1), 107–127.

Wang, L., *et al.*, 2018. Mapping population density in China between 1990 and 2010 using remote sensing. *Remote Sensing of Environment*, 210, 269–281.

Wells, R., 2017. *Differential and complex geometry: origins, abstractions and embeddings*. Cham, Switzerland: Springer International Publishing AG.

Whitney, H., 1936. Differentiable manifolds. *The Annals of Mathematics*, 37 (3), 645–680.

Yuan, F., and Bauer, M.E., 2007. Comparison of impervious surface area and normalized difference vegetation index as indicators of surface urban heat island effects in Landsat imagery. *Remote Sensing of Environment*, 106 (3), 375–386.

Yuan, X., Xue, N., and Han, Z., 2021. A meta-analysis of heavy metals pollution in farmland and urban soils in China over the past 20 years. *Journal of Environmental Sciences (China)*, 101, 217–226.

Zhang, X., *et al.*, 2018. Spatial distribution of metal pollution of soils of Chinese provincial capital cities. *The Science of the Total Environment*, 643, 1502–1513.

Zhou, D., *et al.*, 2015. The footprint of urban heat island effect in China. *Scientific Reports*, 5 (1), 11160.