Challenges in identifying simple pattern-forming mechanisms in the development of settlements using demographic data

Bartosz Prokop [*] and Lendert Gelens
Laboratory of Dynamics in Biological Systems, Department of Cellular and Molecular Medicine, KU Leuven, Leuven 3000, Belgium

Peter F. Pelz and John Friesen
Chair of Fluid Systems, TU Darmstadt, 64287 Darmstadt, Germany

(Received 4 August 2021; revised 14 April 2023; accepted 9 May 2023; published 7 June 2023)

The rapid increase of population and settlement structures in the Global South during recent decades has motivated the development of suitable models to describe their formation and evolution. Such settlement formation has been previously suggested to be dynamically driven by simple pattern-forming mechanisms. Here, we explore the use of a data-driven white-box approach, called SINDy, to discover differential equation models directly from available spatiotemporal demographic data for three representative regions of the Global South. We show that the current resolution and observation time of the available data are insufficient to uncover relevant pattern-forming mechanisms in settlement development. Using synthetic data generated with a generic pattern-forming model, the Allen-Cahn equation, we characterize what the requirements are for spatial and temporal resolution, as well as observation time, to successfully identify possible model system equations. Overall, the study provides a theoretical framework for the analysis of large-scale geographical and/or ecological systems, and it motivates further improvements in optimization approaches and data collection.

DOI: 10.1103/PhysRevE.107.064305

I. INTRODUCTION

The fraction of the population living in urban or settlement structures has grown exponentially over the last several decades, especially in the Global South [1,2]. This trend poses one of the main challenges in our world [3] as the rising population in such structures is in need of vital infrastructure [4] while simultaneously affecting (mostly negative) climatic developments [5,6]. Consequently, there is an urgent need to understand underlying processes of urbanization and anticipate the emergence of these structures.

Urbanization and development of settlement structures depends on several mechanisms based on repulsion and attraction [7,8]. Such interactions can lead to three major settlement distributions (see Fig. 1). Their existence has been confirmed in recent settlement pattern studies of different regions in the Global South, in which regular distributions are dominating [9–13].

The existence of regularly patterned distributions in settlements, and in other spatial systems, can be an indicator for the existence of instability-driven pattern-forming mechanisms [14]. A similar concept of linking spatial distributions with specific driving mechanisms has been successfully applied in the field of plant and animal ecology for decades [15–17]. To understand the emergence of these patterns from underlying interactions, different modeling approaches have been developed. For example, urban development can be modeled by agent- or cellular-automata-based approaches [18,19] which include detailed interactions at the level of individuals. Despite their accuracy in specific cases, such models have several drawbacks. They lack generalization, require large and detailed data sets to be fit on, and can become computationally expensive. Another approach is to use reaction-diffusion models. Such systems consisting of differential equations are simpler and directly interpretable, yet they can also lead to highly complex patterns and dynamics [20,21]. In the context of urban structures, Pelz *et al.* [22] have developed a theoretical framework describing the formation of informal settlements (so-called *slums*) in the Global South. Furthermore, this framework was extended to describe the morphogenesis of urban systems in the United States as a reaction-diffusion system in Ref. [23].

Deriving such models is typically done by suggesting sets of possible models from experimental measurements of specific interactions paired with scientific intuition and determining which one provides the best fit to measured data. However, as data structures of urban systems are complex, suggesting certain model structures can be difficult. One way to address this is by using recently developed data-driven model discovery approaches, such as the system identification approach called "Sparse Identification of Nonlinear Dynamics" (SINDy) [24]. SINDy has recently gained attention in many fields, such as engineering [25], physics [26], chemistry [27], and biology [28]. The method has shown success in identifying interpretable models in the form of ordinary or partial (through the extension PDE-FIND [29]) differential equations from synthetic data of, e.g., pattern-forming mechanisms in the form of reaction-diffusion equations [29,30]. However, the literature on model discovery from real

[*]Corresponding author: bartosz.prokop@kuleuven.be

 ©2023 American Physical Society

TABLE I. Investigated regions with their geopolitical location and attributes.

Region	Attributes
Punjab, India	Northwest India, border region with Pakistan, agricultural region called Granary of India
Nile delta, Egypt	North Egypt, densely populated, fast growing agricultural region
Kano State, Nigeria	North Nigeria, border region to Niger, agricultural region in one of the fastest growing economies

temporal or spatiotemporal data with SINDy is scarce (e.g., on generic benchmark problems in Ref. [31] or Ref. [32]), as the method struggles with commonly encountered high-noise and low-data situations.

In this work, we theoretically and practically investigate if pattern-forming processes can be responsible for settlement development in the Global South and attempt to identify such mechanisms using SINDy directly from existing spatiotemporal data of population patterns. In Sec. II, we motivate the potential relevance of pattern-forming equations from data analysis of satellite images. In Sec. III, we apply the SINDy algorithm to satellite images and try to identify mathematical equations in the form of one-component pattern-forming mechanisms. As model identification from the real data sets turns out to be challenging, we identify and study two main challenges, namely, data availability and quality and, determine their influence on model discovery from spatiotemporal data in Sec. IV. Lastly, in Sec. V, we discuss our findings and elaborate on these challenges and what is required to overcome them.

II. SETTLEMENT DEVELOPMENT AS A PATTERN-FORMING PROCESS

We start our study by selecting three representative regions of emerging countries—the Punjab region in India, the Nile Delta in Egypt, and the Kano State region in Nigeria (see Table I). The regions are chosen as they lie in countries which can be considered representative of the Global South: all countries have had steady population as well as steady economic growth over the last 20 years [1]. Despite and because of the great cultural differences, all three societies are in transition from agricultural countries to industrialized nations.

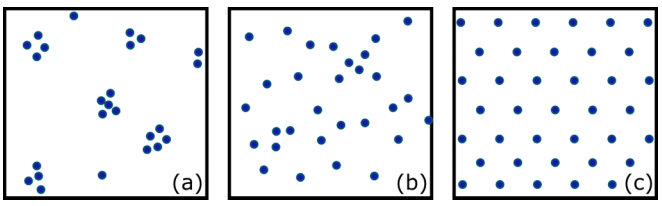

FIG. 1. Possible settlement arrangements following the phenomenological interpretation as point processes, resulting in clustered (a), random (b), or regular (c) distributions.

Furthermore, the spatial distribution of settlements has been studied in these regions and their regularity has been characterized on different spatial scales [10,12,13]. This allows us to select subregions [here called regions of interest (ROI)] which show a regular distribution. This is illustrated for the Punjab region in India in Fig. 2(a). For a detailed explanation of how we used satellite data, in combination with the *Global Artificial Impervious Area* (GAIA, temporal resolution $\Delta t = 1\ \text{year}^{-1}$, spatial resolution $\Delta x = \Delta y \approx 30$ m) [33,34] data set, to select the ROIs, we refer to Appendix D. For each ROI, we then track the settlement evolution over a period of about 15 years using the data set *WorldPop* [35] depicting spatial population distributions [see Fig. 2(b)]. The *WorldPop* data set is used later for the model identification process as described in Fig. 2(c) [36].

Previous work has shown that the observed emergence of rural settlement structures could be caused by simple reaction-diffusion pattern-forming mechanisms [22]. Following the example of Pelz *et al.* [22], we divide the system into a population living in the rural settlements and a supply potential of agricultural land which is complementary to the population. In other words, areas that are in agricultural use are uninhabited, and vice versa. We also assume that there is a limit to agricultural exploitability and urban densification. The population density at a spatial point $\boldsymbol{x} = (x, y)$ and time t is given by $u(\boldsymbol{x}, t)$ and the corresponding supply potential is given by $v(\boldsymbol{x}, t)$.

The change of the concentrations of u and v is defined by three complementary global contributions: (i) birth or death of the population or the cultivation or sealing of agricultural space within a domain of size A, (ii) migration to and from other cities outside of A leading to a decrease or an increase in supply potential, and (iii) migration to areas with higher supply potential over the boundary C of A where a settlement exists or is created. These mechanisms lead to the formulation of a typical reaction-diffusion model (for more detail, see Appendix A and Ref. [22]):

$$\begin{aligned} u_t &= \nabla^2 u + Rf(u, v), \\ v_t &= D\nabla^2 v + Rg(u, v), \\ \text{with} \quad & (\boldsymbol{n} \cdot \boldsymbol{\nabla}) \begin{pmatrix} u \\ v \end{pmatrix} = 0 \text{ on } C, \end{aligned} \tag{1}$$

with D and R being the diffusion and reaction coefficients alongside the respective reaction terms f and g, which are being evaluated under no-flux boundary conditions.

The linear stability analysis of the homogeneous system (no diffusion) provides the Jacobian matrix $\boldsymbol{J}$. It is well known that specific sign-combinations of components of the Jacobian allow for an initial homogeneous distribution of population and supply potential to be stable—either through substrate inhibition or as an activator-inhibitor system [37]. In the case of settlement structures, the only physically sensible choice is substrate inhibition, as the resulting concentration patterns are out of phase [22,37]:

$$\boldsymbol{J} = \begin{pmatrix} f_u|_0 & f_v|_0 \\ g_u|_0 & g_v|_0 \end{pmatrix} = \begin{pmatrix} + & + \\ - & - \end{pmatrix}. \tag{2}$$

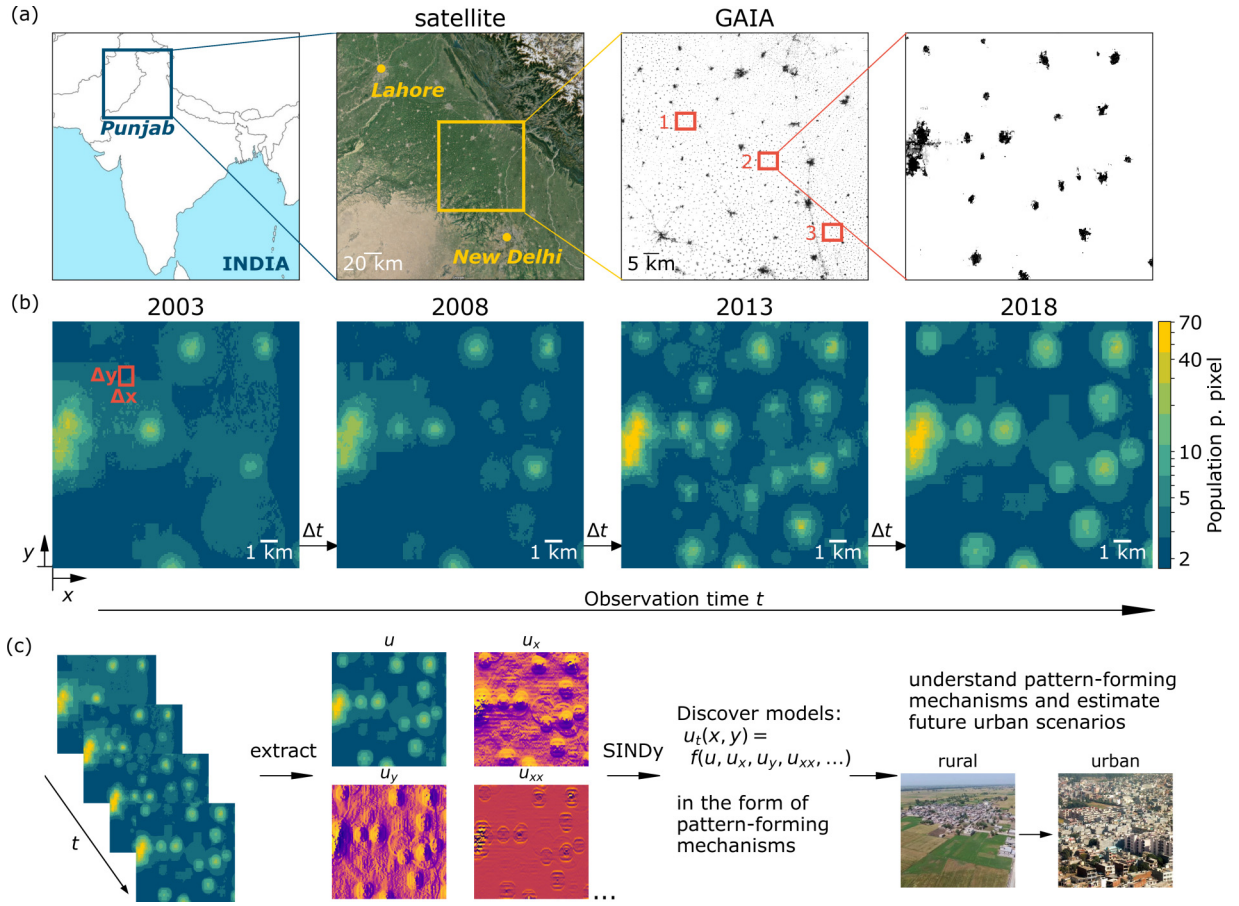

FIG. 2. Overview of the workflow of this study. (a) Selection of ROIs, starting from the identification of suitable regions of the Global South, here Punjab, India, selection of an excerpt of the region in the data set GAIA and determination of ROIs from regular settlement patterns with the average nearest neighbor (ANN, see Appendix D). (b) Data of population density patterns (on logarithmic scale) from the data set *WorldPop* from ROI 2 of Punjab, India, for four time points with a spatial resolution of Δx, $\Delta y \approx 100$ m (enlarged in figure) at the equator and a temporal resolution of $\Delta t = 1$ year^{-1}. (c) Workflow to investigate possible pattern-forming mechanisms in the development of settlement structures. Starting from the time series of population density patterns, we extract spatiotemporal features and apply SINDy to discover models in the form of one-component partial differential equations. This can help us understand the role of such mechanisms in settlement structures and estimate future urban scenarios.

- $f_u|_0 > 0$, ***people attract other people*****:** The population u increases due to self-reproduction of u in an environment of sufficient sustenance.
- $f_v|_0 > 0$, ***supply attracts people*****:** The amount of available agricultural area v attracts people, increasing the population u of the settlement.
- $g_u|_0 < 0$, ***people inhibit supply*****:** The higher the population u, the less agricultural area v is available, especially due to the limitation by the agricultural needs of surrounding settlements. This is the case until the maximum amount of agricultural area is used and does not suffice to supply the population of a settlement, leading to a decrease of population u.
- $g_v|_0 < 0$, ***supply inhibits additional supply*****:** Due to limited resources, the supply production decreases when agricultural efficiency plateaus.

In the presence of diffusion, the linear stability shows that an equally dispersed population and supply potential densities can destabilize to form spatial patterns when the following condition is met (see Appendix B):

$$f_u|_0 > -\frac{g_v|_0}{D}. \tag{3}$$

This is the case when the attraction of people to A dominates the inhibition of supply potential due to the emergence of settlements. If the domain size A and the diffusion coefficient D of the system are sufficiently large, different settlements can emerge that constantly compete against each other over the supply potential, eventually leading to a regular distribution of settlements.

Assuming that the total population density u and the supply potential v are conserved, $u + v = c_{\max}$, the system reduces to a simpler one-component equation [38,39]:

$$\begin{aligned} u_t(x,y,t) &= \check{D}\nabla^2 u(x,y,t) + R\check{f}[u(x,y,t)], \\ \text{with} \quad \check{D} &= \frac{D+1}{2}, \\ \text{and} \quad \check{f} &= \tfrac{1}{2}\{f[u(x,y,t)] + g[u(x,y,t)]\}. \end{aligned} \tag{4}$$

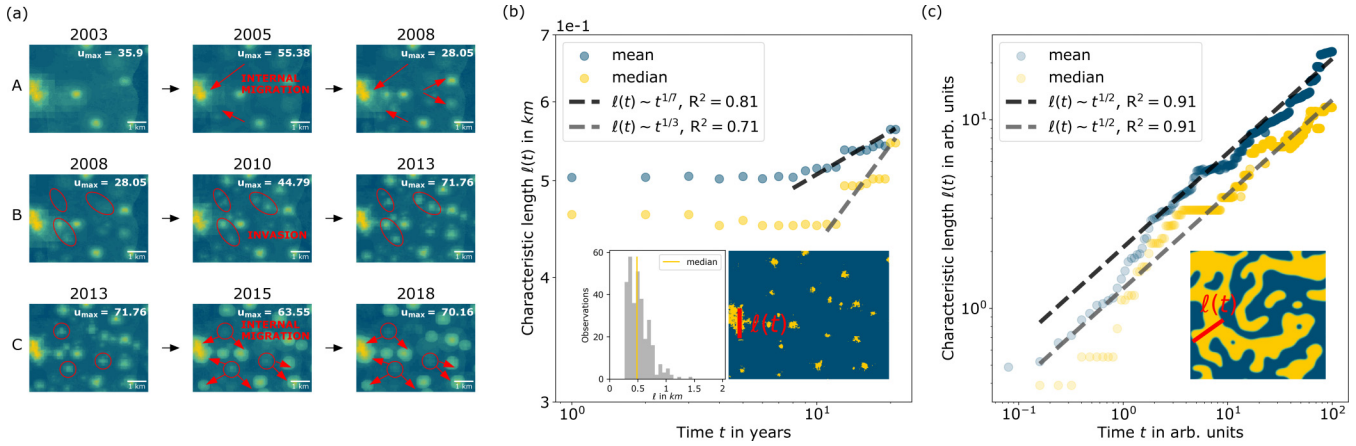

FIG. 3. (a) Example of multiple effects found in spatiotemporal data sets of population densities indicating the existence of suggested reaction-diffusion equations (population density data from data set *WorldPop* [35], ROI 2, Punjab, India). The observable behaviors are internal migration into existing settlements (A), invasion of not occupied agricultural space (B), or local migration triggered by competition over available agricultural space (C). (b) Calculated characteristic length of settlement patterns resulting from the mean and the median heavy-sided feature size distribution (shown in inlet, ROI 2, Punjab, India, in 2018). The characteristic lengths obey the power law when growing, with $\ell(t) \sim t^{1/7}$ and $\ell(t) \sim t^{1/3}$, respectively (with R^2 scores). (c) Calculated characteristic length of the pattern (an example of a feature size is shown in the inset) resulting from the Allen-Cahn equation. Here, both the mean and the median characteristic length obey the power law with $\ell(t) \sim t^{1/2}$ when growing (with R^2 scores).

Based on this proposed reaction-diffusion description consistent with pattern formation [22], we first set out to see whether the satellite data [Fig. 2(b)] contained any clear signatures of such developing rural settlement patterns. Visual inspection of the *WorldPop* data set indicates the possible presence of three key processes: local migration through attraction of bigger settlements [e.g., between 2003 and 2008, Fig. 3(a)/A], invasion or occupation of available agricultural space [e.g., between 2008 and 2013, Fig. 3(a)/B], and local migration induced by competition of settlements over available space leading to changes in settlement patterns [e.g., between 2013 and 2018, Fig. 3(a)/C].

We then analyzed the characteristic lengths $\ell(t)$ of settlements [see Fig. 3(b)] using the GAIA data set [33,34]. We chose the characteristic length to be the size of the respective features, evaluating the distribution feature sizes and determining the predominant size following Refs. [40] and [41] [see the inset in Fig. 3(b); for more details see Appendix E]. This analysis shows that when the characteristic size of settlements is growing, this growth is well described by a power law ct^{β}, with an exponent of $\beta = 1/7$ for the mean and $\beta = 1/3$ median characteristic lengths.

Interestingly, from the literature we know that the time evolution of characteristic lengths of patterns driven by coarsening mechanisms also follow such a power law. For example, for the Allen-Cahn equation, it was theoretically shown that the change of $\ell(t)$ is described by the power law with $\beta = 1/2$ [40–42], while for the Cahn-Hilliard equation [38], $\beta = 1/3$ [40,41]. Indeed, we calculated the characteristic length for a simulated pattern for the Allen-Cahn equation of the form

$$u_t(x, y, t) = \alpha \nabla^2 u + \beta u + \gamma u^2 - u^3, \tag{5}$$

using the same method as for the settlement patterns, which revealed that the mean and the median characteristic length $\ell(t)$ follow the power law with $\beta = 1/2$ [see Fig. 3(b)]. Even though the exponent β of the power law for settlement patterns is not the same as that for the Allen-Cahn or the Cahn-Hilliard equation, it indicates that settlement patterns could be a product of a simple, coarsening pattern-forming mechanism in the proposed form of Eq. (4).

III. MODEL IDENTIFICATION FROM SETTLEMENT DATA

A. Model identification using SINDy

The SINDy method to identify differential equation models from spatiotemporal data sets has been increasingly applied in many fields, i.e., in fluid mechanics [24]. However, it has, to our knowledge, never been used for studying large scale, geo-sociological questions. Therefore, we asked ourselves whether we could use this method to discover a partial differential equation (PDE) that provides a good description of the measured time evolution of rural settlements in the Global South. If successful, such a PDE would also provide new insights in potential pattern-forming mechanisms in settlement development.

The main idea behind the SINDy method is the assumption that dynamic systems can be described through either ordinary or, in our case, partial differential equations (using PDE-FIND) [29] with sparse structure in the following form:

$$u_t = N[u(\boldsymbol{x}, t), u_x, u_y, \ldots, \boldsymbol{x}, \boldsymbol{\xi}]. \tag{6}$$

The temporal change of u, u_t, is a function of the variable u itself, its spatial derivatives, and a set of coefficients $\boldsymbol{\xi}$. Differential equations of this form can be linearly combined:

$$u_t = \xi_1 + \xi_2 u + \xi_3 u^2 + \xi_4 u_x + \xi_5 u_{xx} + \cdots. \tag{7}$$

This equation can be rewritten as a row vector containing all combinations and derivatives of the quantity, called the term library, and a coefficient vector $\boldsymbol{\xi}$ containing all coefficients:

$$u_t = (1 \quad u \quad u^2 \quad u_x \quad u_{xx} \quad \cdots) \cdot \boldsymbol{\xi}. \tag{8}$$

TABLE II. Terms included in the library for the one-component, two-dimensional equation sorted by combinations of u and its derivatives.

	Terms
Combinations	$1, u, u^2, u^3$
Derivatives	$u_x, u_y, u_{xx}, u_{yy}, u_{xy}, u_{xxx}, u_{yyy}, u_{xxy}, u_{yyx}, u_{xxxx}, u_{yyyy}, u_{xxyy}, u_{xxxy}, u_{yyyx}$

The values of each term in the library can be calculated from a single shot at a given point in time. If this system is extended to all available time points, a linear system of equations with the unknown parameter vector $\boldsymbol{\xi}$ and the term library matrix $\boldsymbol{\Theta}$ is formed:

$$(\boldsymbol{u}_t) = (1 \quad u \quad u^2 \quad u_x \quad u_{xx} \quad \cdots) \cdot (\boldsymbol{\xi}) = \boldsymbol{\Theta} \cdot \boldsymbol{\xi}. \tag{9}$$

We assume that the settlement evolution could be captured by a PDE similar to Eqs. (4) and (5). Therefore, we use the term library given in Table II, which includes derivatives up to the fourth order.

This system poses an overdetermined optimization problem for values of $\boldsymbol{\xi}$ and can be solved using regression algorithms (for more detailed information on regression algorithms for SINDy, see Ref. [43]). In contrast to the original work in which the method PDE-FIND was introduced [29], we apply a sparsity-promoting algorithm with the SR3 algorithm developed by Zheng *et al.* [44]. This method includes the additional variable $\boldsymbol{w}$, which is forced to be close to the coefficient parameter and therefore relaxes the optimization problem:

$$\min_{\boldsymbol{\xi},\boldsymbol{w}} \frac{1}{2}\|\boldsymbol{u}_t - \boldsymbol{\Theta}\boldsymbol{\xi}\|_2^2 + \lambda\|\boldsymbol{w}\|_1 + \frac{\alpha}{2}\|\boldsymbol{w} - \boldsymbol{\xi}\|_2^2, \quad \text{with} \quad \lambda = \frac{l^2}{2\alpha}. \tag{10}$$

Here, two hyperparameters of the optimization have to be set: the threshold l and the parameter of the optimization α, which provides the penalizing parameter λ of the regularization.

After model identification, we analyze the discovered models with the Akaike information criterion (AIC). The AIC is a measure of parsimony [45]. It compares the goodness of fit of a given model to other proposed models and weighs it with the model's complexity aiming to maximize the information provided by the simplest-as-possible model. For our analysis, we apply the corrected formulation for finite sample sizes of the AIC (AIC_c) proposed by Ref. [46], resulting from Ref. [47]:

$$\mathrm{AIC}_c = \mathrm{AIC} + \frac{2(k+1)(k+2)}{m-k-2}. \tag{11}$$

The AIC is described by the likelihood function of average error over time and space ϵ as follows:

$$\begin{aligned} \mathrm{AIC} &= m \ln(\epsilon/m) + 2k, \\ \text{with } \epsilon &= \frac{\left|\sum_{i=1}^{m_t} y_i - N(x_i, \boldsymbol{\xi})\right|}{m} \\ \text{and } m &= m_s n_{\mathrm{ROI}}. \end{aligned} \tag{12}$$

The AIC depends on the number of observations m (size of region $m_s = XY$ and amount of included ROIs $n_{\mathrm{ROI}} = 3$, where we interpret an observation as the time series at every spatial point) and the number of terms (k) describing the complexity of an identified model.

Using this approach, we determine the most parsimonious model among all potential models and study its properties. In order to compare the identified models, we further normalize the AIC by the minimal value of the respective analysis $\mathrm{AIC}_{\mathrm{min}}$. Here, following Refs. [46] and [47], a model that has an $\mathrm{AIC}_c - \mathrm{AIC}_{\mathrm{min}} < 2$ has strong support for being the correct underlying system, while the ones with $\mathrm{AIC}_c - \mathrm{AIC}_{\mathrm{min}} < 8$ have weak support. Hereafter, we always refer to the corrected AIC_c when the AIC is mentioned.

B. Application to settlement data

Using the outlined approach, we look for potential models to describe the available settlement data from the *WorldPop* data set in the different regions (India, Egypt, and Nigeria). We scan sets of thresholds from $l = 10^{-6}$ to $l = 10^{2}$ and optimization hyperparameters $\alpha = 10^{-3}$ to $\alpha = 10^{3}$. The SR3 algorithm was applied with a tolerance of 10^{-2} and using 200 iterations (for details see Ref. [44]).

From the parameter scan, we determine for which combinations of l and α the identified models provide an AIC < 2. In order to preselect possible solutions, we only evaluate equations with this AIC at overlapping parameter pairs of l and α. The reason for this is that, if the dynamics in the whole region follow the same rules (or dynamical behavior), all best identified models should have the same mechanistic form. Therefore, in Fig. 4(a), we show for which sets of (l, α) the low-AIC regions of each ROI overlap (in yellow).

Next, we selected unique model equations at their lowest parameter values for l and α, respectively. The parameter combinations for unique solutions of each ROIs are shown in Fig. 4(a). These unique solutions are further compared with the use of the AIC, where we also depict the error and the complexity of the found equations in Fig. 4(b). The optimal selected equations are highlighted by markers with red borders and the values of coefficients are shown in Fig. 4(c).

For the identified models, we analyze if they indeed reproduce the same spatiotemporal dynamics and patterns as present in the original data they were trained on. Figure 5(a) shows a representative time evolution of the original and simulated data for ROI 2 in India. One can see that, while the spatial patterns become more pronounced in the original data, this is not the case in the simulated model. This is also seen by explicitly plotting the difference between the original data and the simulation. These observations are general for all analyzed ROIs (full analysis is in the shared repository).

To better understand why the identified model equations do not accurately capture the original data, we look more closely at the coefficients of each model term, as well as the overall contribution of each term. We find that the assigned coefficients vary over multiple orders of magnitude, ranging from 10^{-3} for some production terms up to 10^{6} for some higher-order derivatives of the diffusive terms. We calculated the actual contribution c_j of each term as the value of each term θ_{ij} at a set time point t (we arbitrarily chose $t = 10$) and

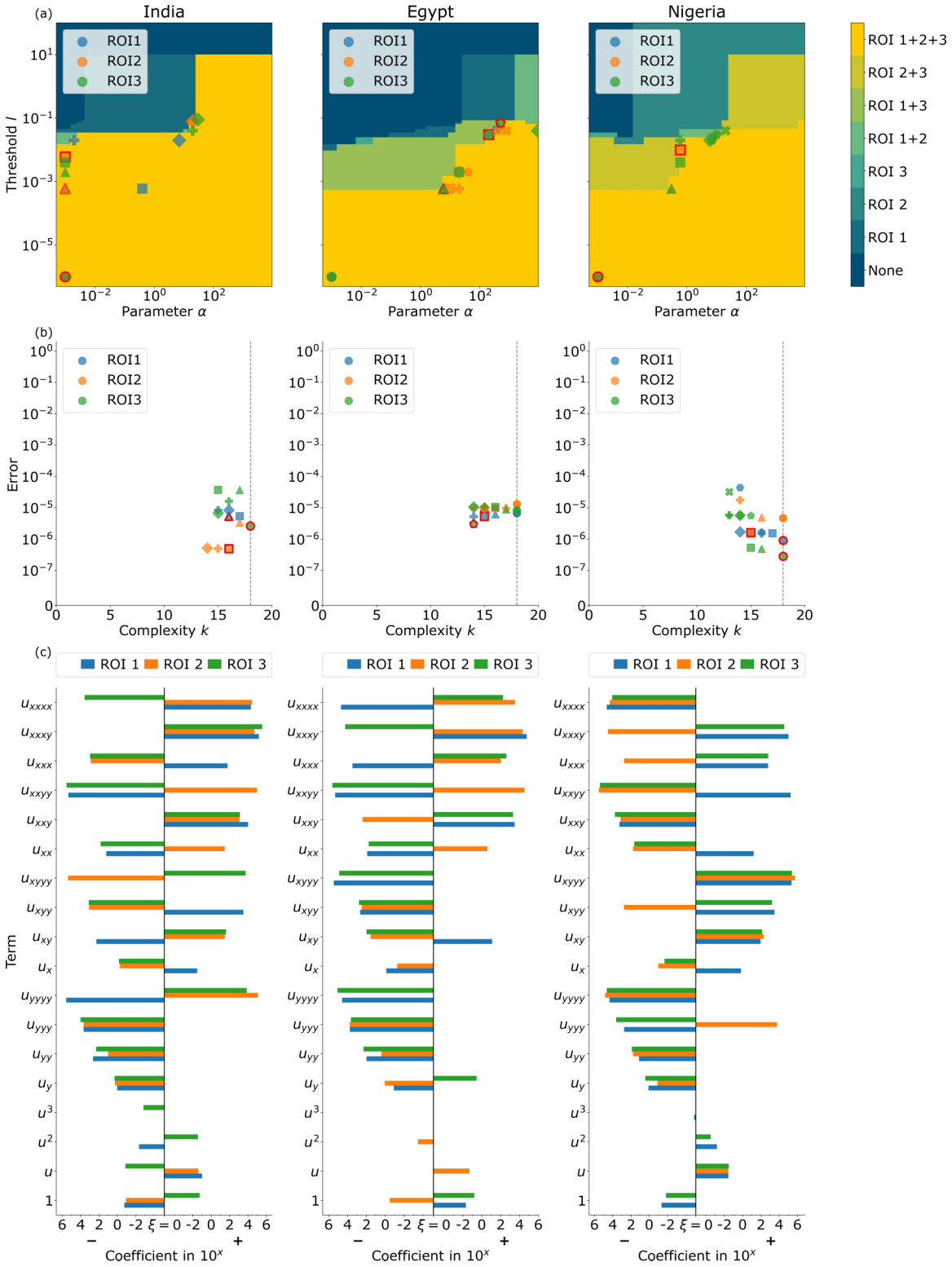

FIG. 4. Parameter sweep of all ROIs and regions for sets of threshold l and optimization parameter α. (a) The analysis of the ROIs shows different combinations of parameters where the AIC falls beneath 2 for the respective ROI. As we search for a regionally valid solution, we select only unique solutions at combinations of l and α where the AIC < 2 for all ROIs overlap. The selected unique equations and their respective optimization parameter sets where they were found first are shown. Red bordered markers depict the best identified models of each ROI. We see that the AIC provides the best equations with the lowest error and complexity (except for ROI 2, from which we show later through analysis of contributions that only the production terms are significant). (b) The respective error and complexity of the found equations are depicted. The red bordered markers show the best model for each ROI in each region following the AIC. (c) The identified coefficients $\boldsymbol{\xi}$ [as in Eq. (9)] are shown for each equation. All identified models contain most of the derivatives (diffusive terms) with large coefficient values. If found, the coefficient values of the reaction terms are multiple magnitudes smaller than those of the diffusive terms.

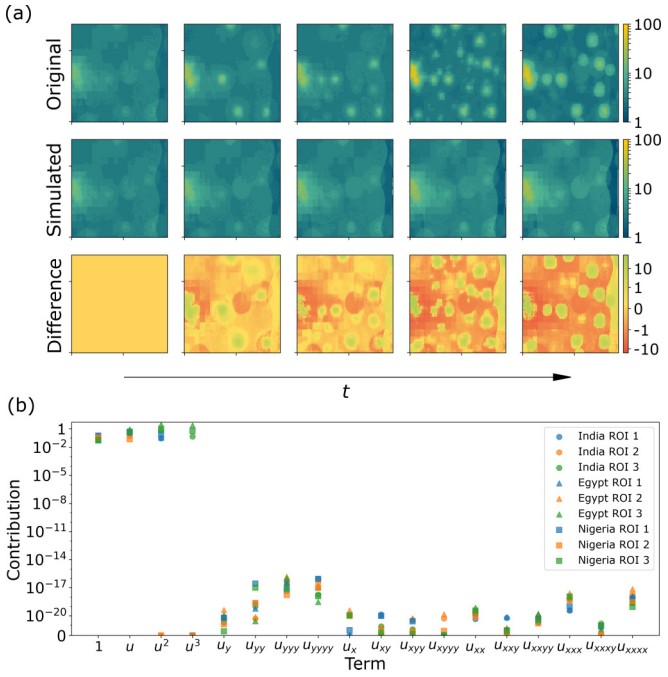

FIG. 5. (a) The simulation of found models shows that the concentration in the whole region increases with time, but no patterns form. Despite being trained on the data set, the models are not able to reproduce the training data (here ROI 2, India). (b) Calculated contributions to the change of u. Even though the magnitude of coefficients is on the order of 10^6 for derivatives, their contribution is negligible and lies between the orders of 10^{-17} and 10^{-21}.

averaged over all spatial points, and we then multiplied with the respective coefficient ξ_j,

$$c_j = |\xi_j \cdot \bar{\theta}_j|, \quad \text{with} \quad \bar{\theta}_j = \frac{\sum_{x=0}^{X} \sum_{y=0}^{Y} \theta_{i=10\,j}(x,y)}{XY}. \quad (13)$$

Figure 5(b) shows that, even though the coefficients of derivative terms were multiple orders of magnitude larger than those of the production terms, their contribution is, in fact, negligible compared to production terms. The dominance of the terms proportional to $(1, u, u^2, u^3)$ over all terms with spatial derivatives leads to the observed overall increase in concentration while preserving the initial pattern. In conclusion, the identified models do not capture any pattern-forming mechanism as the interaction of reaction and diffusion processes is crucial (see Sec. II).

IV. LOW-DATA LIMITS IN MODEL IDENTIFICATION OF PATTERN-FORMING PROCESSES

One possible reason is that the SINDy method was unable to recover a reaction-diffusion model that correctly describes the settlement data due to a lack of spatiotemporal resolution and/or insufficient observation time. Indeed, it is known that model discovery with SINDy is dependent on the amount of temporal points and the size of the time step Δt [44,48]. Therefore, we decided to study the limits of the SINDy method in recovering a reaction-diffusion model for low spatial and/or temporal resolution, as well as short observation times. As we found that the observed settlements followed

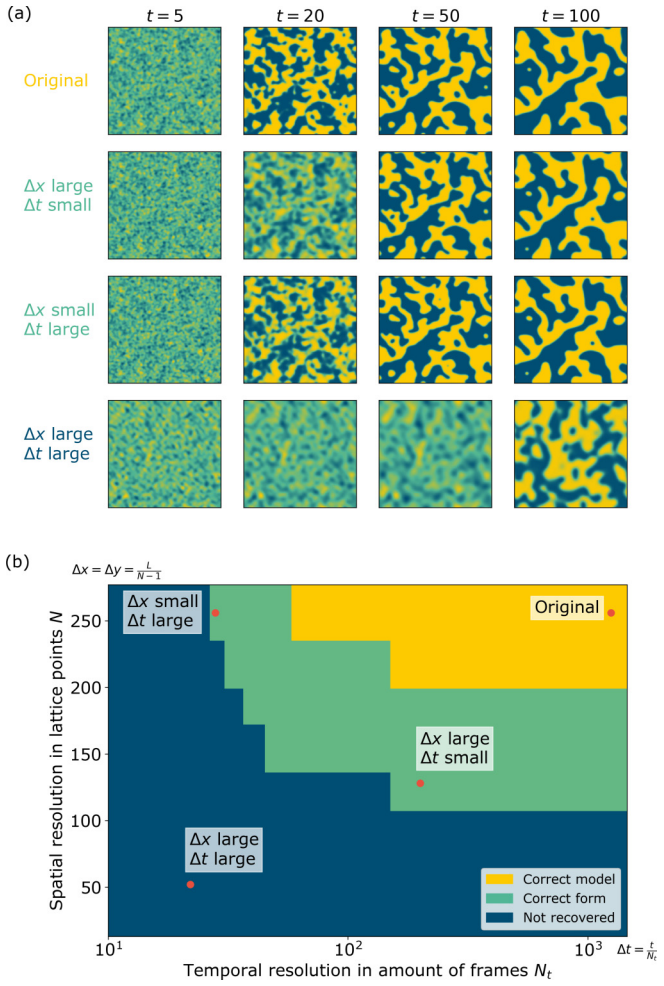

FIG. 6. Recovery of the AC equation with SINDy: sensitivity to spatiotemporal resolution. (a) Time simulation of the AC equation, subsampled with different spatial and temporal resolutions, as indicated in panel (b). (b) Diagram showing for which resolution SINDy is able to recover the AC equation. Recovery is only successful for sufficiently high spatial and temporal resolution (214×214, $N_t > 80$ frames) of the simulated data set of $t = 100$.

coarsening dynamics (see Sec. II), we decided to study the recovery of the Allen-Cahn (AC) equation (5, [39,49]) using SINDy in the low-data limit:

$$u_t(x,y,t) = \alpha \nabla^2 u + \beta u + \gamma u^2 - u^3,$$
$$\text{with} \quad \alpha = 0.1, \ \beta = 0.5, \ \gamma = -0.01. \quad (14)$$

With this set of coefficients, the initial condition $u_{\text{init}}(x,y,t=0) \sim \mathcal{N}(0, 0.01)$ (which was created once and used for all simulations), and zero-flux boundary conditions, Eq. (14) creates coarsening labyrinth patterns as shown in Fig. 6(a) (Original).

We then first subsampled this data set generated by simulating the AC equation by imposing a spatial resolution of $\Delta x = \Delta y = 0.39$ [$N_x = N_y = N = 256$ lattice points with $\Delta x = L/(N-1)$ and $L = 100$] and a temporal resolution of $\Delta t = 0.08$ ($N_t = 1250$ frames with $N_t = t/\Delta t$ and $t = 100$). This subsampled data set was then used as input to the SINDy algorithm, which was able to recover the original AC

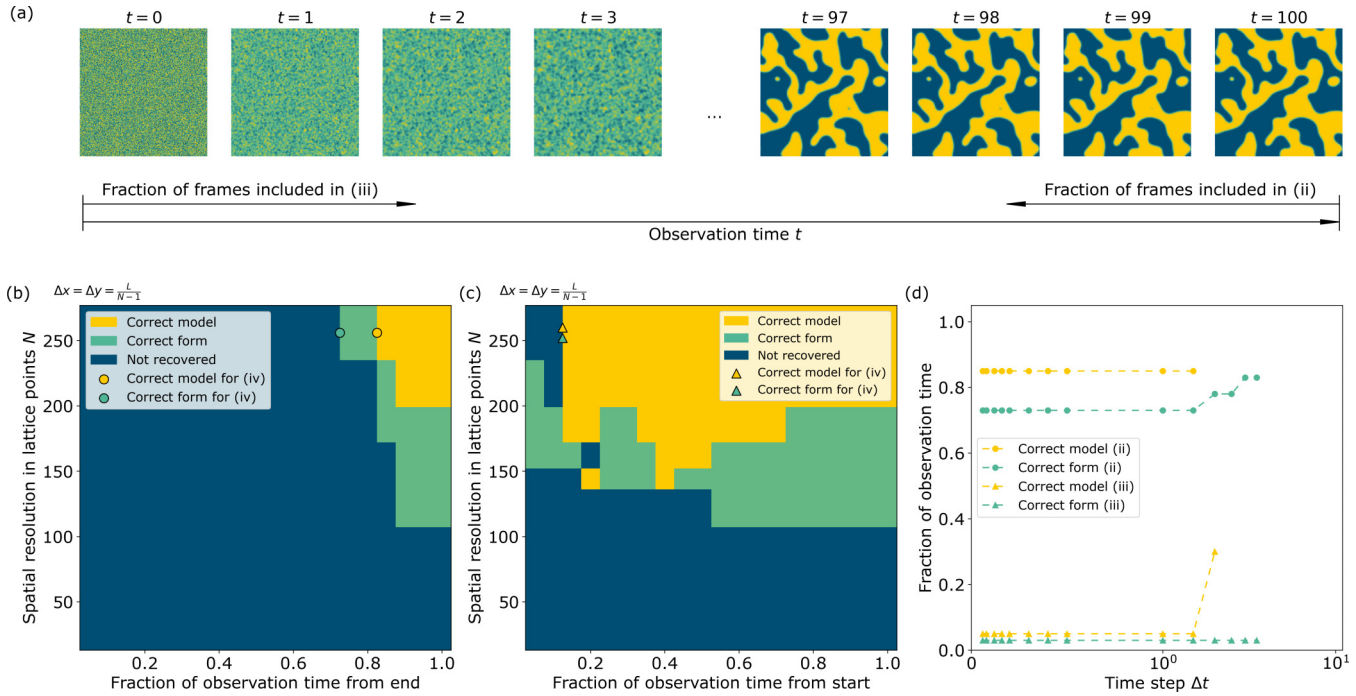

FIG. 7. Recovery of the AC equation with SINDy: sensitivity to observation time. (a) Time series of the simulated AC equation with temporal resolution of $\Delta t = 0.16$ and observation time $t = 100$. (b) and (c) Diagrams showing for which resolution SINDy is able to recover the AC equation with changing observation time and spatial resolution. The observed time window used for model recovery is chosen at the end (b) or start (c) of the time series in panel (a). (d) For a spatial resolution of $N = 256$ lattice points, varying the temporal resolution shows that identification is less sensitive to resolution than to the amount of available data [for case (c), the transition points are plotted with offset for visualization].

equation (14) in the correct form with a maximum error of 3% in the coefficients.

Next, we further subsampled the data temporally and spatially in order to identify the limits of the SINDy approach. We decreased the temporal resolution from $\Delta t = 0.08$ ($N_t = 1250$ frames) in 26 steps to $\Delta t = 10$ ($N_t = 10$ frames). Similarly, the spatial resolution was reduced from $\Delta x = \Delta y = 0.39$ ($N = 256$ lattice points) in 18 steps to $\Delta x = \Delta y = 7$ ($N = 14$ lattice points). We then compared the identified model to the original AC model, characterizing if the form of the model (correct terms) was correct and whether it had identified the correct coefficients (see Fig. 6). This analysis shows that the SINDy algorithm is sensitive to spatial and temporal resolution. Only for sufficiently high spatial ($\Delta x = \Delta y < 0.46$, $N > 214$ lattice points) and temporal resolution ($\Delta t < 1.25$, $N_t > 80$ frames) we correctly recover the AC equation with the proper coefficients [yellow region in Fig. 6(b)]. For lower resolution, we are still able to recover the mechanistic form of the equation (but not the correct coefficients) in the green region in Fig. 6(b). In Fig. 6(a) we show how the detected mechanistic forms in this region are still able to recover the overall dynamics of the system (for low-spatial and high-temporal resolution and for high-spatial and low-temporal resolution). However, the reduced resolution leads to coefficients smaller than those in the original AC equation, which slows down the dynamics. When decreasing the resolution even further, the optimization algorithm includes and overestimates higher-order derivatives. Such models are no longer able to capture the dynamics of a system accurately [blue region in Fig. 6(b); Fig. 6(a), low-spatial and low-temporal resolution].

We then wondered how the observation time affects the possibility to recover the correct model. This question is especially relevant as we only observe relatively slow changes of settlement structures over 20 years using the *WorldPop* data set. We repeated the study of the recovery of the AC equation (Fig. 6) for a fixed time resolution ($\Delta t = 0.16$, $N_t = 625$ frames) for which we could successfully identify the AC equation (provided the spatial resolution was sufficiently high). We then redid this analysis for varying spatial resolution and observation time. The observation time was progressively reduced by only considering the first or last fraction of the original time series as input to the SINDy optimization [see sketch in Fig. 7(a)]. Note that in this way we are also reducing the total amount of input data. The results of this analysis are shown in Figs. 7(b) and 7(c).

As expected, the AC equation can no longer be recovered when reducing the observation time and/or spatial resolution below a certain threshold. Interestingly, we find that the system is very sensitive to observation time when using the later stages of the dynamical evolution [Fig. 7(b)], while this is much less the case when using the initial part of the dynamical evolution of the AC equation [Fig. 7(c)]. This illustrates that the observation time required to correctly recover the underlying model equation strongly depends on when one measures the system dynamics. In particular, our analysis shows that it is best to capture as much as possible of the dynamical changes at the relevant timescales. In this case, much of the

initial patterns form quickly at the start, while later the patterns coarsen only slowly.

Finally, we then also investigated how sensitive these findings related to observation time were with respect to time resolution. We fixed the spatial resolution at $\Delta x = \Delta y = 0.39$ or $N = 256$ lattice points and determined the critical thresholds in terms of observation time for correct model identification for varying time resolution Δt. Figure 7(d) shows that successful model recovery is less sensitive to time resolution (can be varied over 2 orders of magnitude) than to observation time (both duration and exact timing).

This also shows that there is a limit to the additional information that can be provided by higher spatial and temporal resolution if the duration and/or timing of the observed time window is not well chosen for successful model recovery. In the case of the settlement data set under study here, this analysis suggests that it is plausible that the observed changes in population density in the provided data set are inadequate for proper model identification due to the too short observation time compared to the relevant timescales over which settlements develop.

V. DISCUSSION AND CONCLUSION

The goal of this work was to not only theoretically describe the possible role of simple pattern-forming mechanisms in the development of urban structures (in our case settlements) of the Global South, as has been done before in Refs. [22,23], but to provide an unbiased approach to identify such models directly from data.

In order to do this, we selected three representative regions of the Global South from India, Egypt, and Nigeria and analyzed the occurrence of regularity in settlement structures in these. Using these data, we selected smaller ROIs in the spatiotemporal data set *WorldPop* [35] of population density distributions.

Following this, we extended the ideas of Ref. [22], motivating and providing an alternative theoretical point of view on pattern-forming mechanisms in rural, agriculturally dominated settlement structures. We argued that together with features of regularity (as suggested by Refs. [12,14,17]), pattern-forming mechanisms could be responsible for the emergence of settlement structures. This extension can be a starting point in critically evaluating urban modeling approaches that strive for more complexity over generalization. Here, we substantiated this claim as we observed the suggested behavior in actual population density patterns, while also showing that the characteristic length of patterns resulting from settlements follows a power law, similarly to coarsening patterns in, e.g., Allen-Cahn or Cahn-Hilliard models.

We then introduced the SINDy [24,29] method together with the AIC [45,46], allowing us to derive and investigate spatiotemporal models for the dynamics of population density patterns. However, using the SINDy method, we were not able to identify simple pattern-forming mechanisms directly from selected ROIs of regions in the Global South. The found equations were neither sparse nor represented known pattern-forming mechanisms from literature. The assigned coefficients differed in multiple orders of magnitude between production terms (10^{-3}) and diffusion terms (10^{6}).

As a result, it seems necessary to change the target of optimization from solely evaluating the coefficients to targeting the actual contribution of terms, which, e.g., has been recently suggested by Ref. [50]. Additionally, the configuration of our SINDy approach does not include any time or space dependency of parameters, as suggested by Ref. [51], which can prevent us from capturing important dynamical behavior of settlement systems in the Global South. The found models show that the models are not able to recreate training data. An analysis of term contributions has revealed that the model dynamics are dominated by the production terms and cannot be understood as pattern-forming mechanisms.

Following this unsuccessful application, we identified and studied challenges of model identification in spatiotemporal data sets considering the quality and availability of data. We suggested that the used data set had too low spatial and temporal resolution or an insufficient observation time. Subsequently we studied this question with a sensitivity study of SINDy towards low-data limits. We show that SINDy (here PDE-FIND) is sensitive to spatial and temporal resolution while identifying that observation time and the observed dynamics have a significant influence on the recovery of underlying dynamics. We see that, when fast dynamics of pattern formation are captured, lower spatial and temporal resolution and a shorter observation time are required to correctly identify the model. When only observing slow dynamics (as seen in our settlement data) model identification is more sensitive towards limited spatial resolution and requires longer observations times in order to correctly identify the AC equation. Hence, we need to closely follow the rapid improvements in data acquisition with satellite imagery, which would provide us with sufficiently good data in resolution, while simultaneously and more importantly providing us with longer observation times. If such challenges are overcome, this work provides a ready-to-use framework to discover pattern-forming mechanisms in settlement development.

Moreover, the structure of the data should be adjusted. Currently the *WorldPop* data set does not allow for uninhabited areas with a population density of 0. Here, *WorldPop* itself is developing an improved data set, where population densities and built-up areas are mapped. At the moment of publication this data set only contains a single time point but, when extended, it will provide new opportunities to study our question. Furthermore, the available data sets limit us to only a single observed variable, introducing a strong assumption when considering models. Here, the application of methods coming from Koopman theory, e.g., delay embedding [43,52], could pose an interesting line of work that could provide "hidden' variables, adding information for the optimization and resulting in models of higher dimensions (as has been recently attempted for synthetic data from a shallow-water model [53] or a spatiotemporal Lotka-Volterra model [54]).

In conclusion, we have provided an initial framework for the evaluation and identification of the role of simple pattern-forming mechanisms in the development of settlement structures. So far, the efforts to provide model equations describing such mechanisms have been unfruitful. However, we developed a possible theoretical motivation and were able to identify the major challenges of model identification in low-data limits in pattern formation.

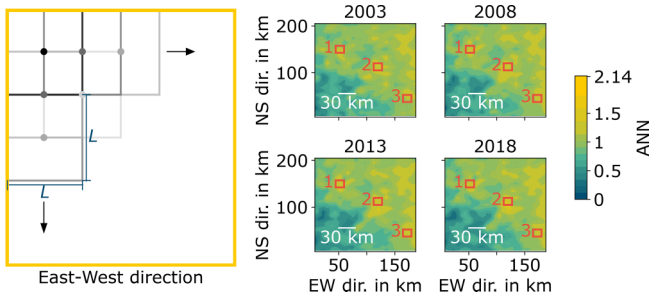

FIG. 8. Moving window method for ANN calculation with changing L, the window side length, and a part of the results for the Punjab region in India from year 2003. Similar figures for all regions, years, and ten window sizes between $L = 5$ km and $L = 50$ km can be found in the repository.

All calculations, simulations, and graphs are done in PYTHON. For SINDy, we use the package PYSINDY [55,56], and for simulations, we developed a simple forward-Euler solver. All algorithms are available in our GITLAB repository [57]. Furthermore, raw data and algorithms are archived via RDR by KU Leuven under the link provided in Ref. [58].

ACKNOWLEDGMENTS

The work of J.F. is funded by the LOEWE Program of Hesse State Ministry for Higher Education, Research and the Arts within the project "Uniform detection and modeling of slums to determine infrastructure needs." We also want to thank Nikita Frolov for his input and constructive discussions.

APPENDIX A: FORMULATION OF REACTION-DIFFUSION EQUATIONS

With the definitions from Sec. II we formulate balance equations for the respective agents u' and v':

$$\dot{N}'_u = \frac{\partial}{\partial t}\int_A u' dA = \int_A \hat{U} R\, f(u', v') - \oint_C \boldsymbol{J}'_u \cdot \boldsymbol{n}\, dC,$$
$$\dot{N}'_v = \frac{\partial}{\partial t}\int_A v' dA = \int_A \hat{V} R\, g(u', v') - \oint_C \boldsymbol{J}'_v \cdot \boldsymbol{n}\, dC. \quad \text{(A1)}$$

Here N describes the amount of population or agriculturally used area in the finite area A and accordingly $\dot{N}$ describes the change in the whole area, whereas u' and v' describe local changes. The long-distance effects are a product of the reaction terms $f(u, v)$ or $g(u, v)$ and the reaction rates $\hat{U}R$ and $\hat{V}R$. Here, $u := u'/\hat{U}$ and $v := v'/\hat{V}$ are dimensionless by division with reference or maximum densities $\hat{U}$ and $\hat{V}$.

Similarly to Ref. [22], the short-distance effects are also driven by a density gradient which can be modeled with Fick's first law. By applying Gauss' theorem, we get the following two reaction diffusion equations:

$$u_t = \hat{U} R\, f(u, v) + D_u \nabla^2 u,$$
$$v_t = \hat{V} R\, g(u, v) + D_v \nabla^2 v. \quad \text{(A2)}$$

With the additional dimensionless transformations $t := Rt'$, $\boldsymbol{x} = \boldsymbol{x}'\sqrt{R/D_u}$, and $D := D_v/D_u$, we derive the dimensionless standard form of reaction-diffusion equations:

$$u_t = \nabla^2 u + R f(u, v),$$
$$v_t = D\nabla^2 v + R g(u, v). \quad \text{(A3)}$$

APPENDIX B: LINEAR STABILITY ANALYSIS

As done in Ref. [22] and following Ref. [39], we perform a linear stability analysis around the linearized state of Eq. (1) with $u = U + \delta u$ and $u = V + \delta v$ with the homogeneous solutions U and V. With the perturbation ansatz $\delta u = \mathcal{R}[\delta \hat{u} \exp(\sigma t + i\boldsymbol{k}\boldsymbol{x})]$ or vice versa with v, we derive an eigenvalue problem with the eigenvalue σ, the Kronecker delta $\boldsymbol{\delta}$, $\boldsymbol{u} = (u, v)$, the Jacobi $\boldsymbol{J}(f, g)$, and $\boldsymbol{D} = 0$:

$$[\sigma \boldsymbol{\delta} - \boldsymbol{J}(f, g)]\delta \hat{\boldsymbol{u}} = 0, \quad \text{(B1)}$$
$$\rightarrow\ \sigma^2 - \boldsymbol{J}(f, g)\boldsymbol{I}\,\sigma + \det[\boldsymbol{J}(f, g)] = 0. \quad \text{(B2)}$$

Solving the eigenvalue problem results in two conditions for the Jacobi matrix $\boldsymbol{J}(f, g)$,

$$\boldsymbol{J} = \begin{pmatrix} f_u|_0 & f_v|_0 \\ g_u|_0 & g_v|_0 \end{pmatrix}, \quad \text{(B3)}$$

which lead to instability,

$$f_u|_0 + g_v|_0 < 0, \quad \text{(B4)}$$

$$\det[\boldsymbol{J}(f, g)] = g_v|_0\, f_u|_0 - g_u|_0\, f_v|_0 > 0. \quad \text{(B5)}$$

As described in Ref. [22] the only reasonable formulation of the Jacobi matrix is

$$\boldsymbol{J} = \begin{pmatrix} + & + \\ - & - \end{pmatrix}. \quad \text{(B6)}$$

Other forms where the columnwise signs are the same results in concentrations spatially in phase, and in the form with rowwise same signs only the shown can be suitably used as shown in Sec. II.

As Turing patterns can arise due to diffusion, we as well study the short-distance effects. We can reformulate diffusion as a product of the specific energy $k_B T$, with the Boltzmann constant k_B and the temperature T, and the mobility μ. At constant T, the ratio $D = \mu_v/\mu_v$ results in

$$\boldsymbol{B} := \boldsymbol{J}(f, g) - \boldsymbol{D}k, \quad \text{with} \quad \boldsymbol{D} = \begin{pmatrix} 1 & 0 \\ 0 & D \end{pmatrix}, \quad \text{(B7)}$$

and allows us to rewrite the eigenvalue problem in Eq. (B1) as

$$[\sigma \boldsymbol{\delta} - \boldsymbol{B}(f, g)]\delta \hat{\boldsymbol{u}} = 0. \quad \text{(B8)}$$

This results again in two conditions for the Jacobi:

$$f_u|_0 + g_v|_0 - k^2(1 + D) < 0, \quad \text{(B9)}$$

$$\det(\boldsymbol{B}) = (f_u|_0 - k^2)(g_v|_0 - k^2) - g_u|_0\, f_v|_0 > 0. \quad \text{(B10)}$$

Turing instability is achieved when the condition Eq. (B10) is violated, resulting in the necessary condition for the diffusion-induced instability:

$$D f_u|_0 + g_v|_0 > 0 \quad \rightarrow \quad f_u|_0 > -\frac{g_v|_0}{D}. \quad \text{(B11)}$$

TABLE III. Coordinates of the regions of interest.

Region		West	South	East	North
India		75.3855	28.8265	77.4804	30.6380
	ROI 1	76.0086	29.4155	75.8528	29.5319
	ROI 2	76.7096	29.7066	76.5538	29.8230
	ROI 3	77.2547	30.2305	77.0990	30.3469
Egypt		29.8650	29.9671	32.1010	31.8803
	ROI 1	30.3377	30.7896	30.1804	30.9044
	ROI 2	31.1260	30.6174	30.9686	30.7322
	ROI 3	31.5200	31.2488	31.3627	31.3636
Nigeria		7.3774	11.1881	9.3336	13.1056
	ROI 1	8.2020	11.9172	8.0646	12.0492
	ROI 2	8.2708	12.5114	8.1333	12.6434
	ROI 3	8.9580	12.1152	8.8205	12.2473

APPENDIX C: REGION INFORMATION

Here we attach the geographical data of the regions used to demonstrate the workflow of our method (see Table III). The coordinates are given in decimal degrees in the reference system WGS84.

APPENDIX D: CALCULATION OF ANN WITH MOVING WINDOWS OF DIFFERENT SIZES

In order to select suitable excerpt sizes from our selected *WorldPop* data sets from Table III, we follow a similar approach as in Ref. [12]. With varying window sizes, starting from square windows with side lengths L of 5 km up to 50 km, we scan over the respective data set calculating the ANN while moving the windows in north-south or east-west direction by $L/2$ (see Fig. 8). We calculate the ANN as in Ref. [59],

$$\mathrm{ANN} = \frac{\frac{\sum_{i=1}^{N} d_i}{N}}{\sqrt{\frac{S}{N}}}, \tag{D1}$$

with d_i being the distance of a settlement to the next-nearest settlement, and N being the total amount of settlements in a window area $S = L^2$. The ANN evaluates the regularity of a point pattern and assigns a value between 0 and 2.14 describing the distribution of settlements (0, clustered; 1, random; 2.14, regular). With this we generated contour diagrams of ANN in the selected regions over time allowing us to estimate a "characteristic" window size where regularity is dominating. By visual inspection we first selected the suitable window size to be 15 km and three respective ROIs of this size which have the most regular distribution over the observation time from our contour diagrams (see Fig. 8 for a part of the obtained results).

APPENDIX E: CALCULATION OF THE CHARACTERISTIC LENGTH $\ell(t)$

We calculated the characteristic length following Refs. [40] and [41] in which it is defined as follows:

$$\ell(t) = \frac{2\pi}{\int q p(q,t) dq}. \tag{E1}$$

Here, q describes the modes' or waves' lengths of a Fourier analysis of a spatial system with their respective probability of occurrence $p(q,t)$, evaluated in all directions or over 2π at every point. To simplify the analysis, we only scan four directions: horizontal, vertical, and two diagonal directions for every point.

[1] UN, World population prospects—Average annual rate of population change (percentage) (2019), https://population.un.org/wpp/.

[2] UN, Population facts: Policies on spatial distribution and urbanization have broad impacts on sustainable development (2020), https://www.un.org/development/desa/pd/content/policies-spatial-distribution-and-urbanization-have-broad-impacts-sustainable-development.

[3] F. Retief, A. Bond, J. Pope, A. Morrison-Saunders, and N. King, Environ. Impact Assess. Rev. **61**, 52 (2016).

[4] E. A. Adams, J. Stoler, and Y. Adams, Am. J. Hum. Biol. **32**, e23368 (2020).

[5] H. Nagendra, X. Bai, E. S. Brondizio, and S. Lwasa, Nat. Sustainability **1**, 341 (2018).

[6] S. Thacker, D. Adshead, M. Fay, S. Hallegatte, M. Harvey, H. Meller, N. O'Regan, J. Rozenberg, G. Watkins, and J. W. Hall, Nat. Sustainability **2**, 324 (2019).

[7] W. Christaller, *Die zentralen Orte in Süddeutschland (German), Central places in Southern Germany*, 3rd ed. (Wissenschaftliche Buchgesellschaft, Darmstadt, 1980).

[8] J. C. Hudson, Ann. Assoc. Ame. Geogr. **59**, 365 (1969).

[9] R. Yang, Q. Xu, and H. Long, J. Rural Stud. **47**, 413 (2016).

[10] A. A. AbouKorin, Ain Shams Eng. J. **9**, 1819 (2018).

[11] J. Friesen, H. Taubenböck, M. Wurm, and P. F. Pelz, Habitat Int. **73**, 79 (2018).

[12] K. Henn, J. Friesen, J. Hartig, and P. F. Pelz, ISPRS Int. J. Geo-Inf. **9**, 541 (2020).

[13] B. Prokop and J. Friesen, Preprints 10.20944/preprints202104.0752.v1.

[14] R. M. Pringle and C. E. Tarnita, Annu. Rev. Entomol. **62**, 359 (2017).

[15] G. Theraulaz, E. Bonabeau, S. C. Nicolis, R. V. Solé, V. Fourcassié, S. Blanco, R. Fournier, J.-L. Joly, P. Fernández, A. Grimal, P. Dalle, and J.-L. Deneubourg, Proc. Natl. Acad. Sci. USA **99**, 9645 (2002).

[16] C. Grohmann, J. Oldeland, D. Stoyan, and K. E. Linsenmair, Insectes Soc. **57**, 477 (2010).

[17] C. E. Tarnita, J. A. Bonachela, E. Sheffer, J. A. Guyton, T. C. Coverdale, R. A. Long, and R. M. Pringle, Nature (London) **541**, 398 (2017).

[18] C. Losiri, M. Nagai, S. Ninsawat, and R. Shrestha, Sustainability **8**, 686 (2016).

[19] M. Batty and R. Milton, Urban Stud. **58**, 3071 (2021).

[20] R. M. May, Nature (London) **261**, 459 (1976).

[21] A. M. Turing, Bull. Math. Biol. **52**, 153 (1990).

[22] P. F. Pelz, J. Friesen, and J. Hartig, Phys. Rev. E **99**, 022302 (2019).

[23] J. Friesen, R. Tessmann, and P. F. Pelz, Reaction-diffusion model describing the morphogenesis of urban systems in the US, in *Proceedings of the 5th International Conference on Geographical Information Systems Theory, Applications and Management* (SciTePress, Heraklion, Crete, Greece 2022).

[24] S. L. Brunton, J. L. Proctor, and J. N. Kutz, Proc. Natl. Acad. Sci. USA **113**, 3932 (2016).

[25] P. A. K. Reinbold, L. M. Kageorge, M. F. Schatz, and R. O. Grigoriev, Nat. Commun. **12**, 3219 (2021).

[26] A. V. Ermolaev, A. Sheveleva, G. Genty, C. Finot, and J. M. Dudley, Sci. Rep. **12**, 12711 (2022).

[27] M. Hoffmann, C. Fröhner, and F. Noé, J. Chem. Phys. **150**, 025101 (2019).

[28] N. M. Mangan, S. L. Brunton, J. L. Proctor, and J. N. Kutz, IEEE Transactions on Molecular, Biological and Multi-Scale Communications **2**, 52 (2016).

[29] S. H. Rudy, S. L. Brunton, J. L. Proctor, and J. N. Kutz, Sci. Adv. **3**, e1602614 (2017).

[30] H. Schaeffer, Proc. R. Soc. A **473**, 20160446 (2017).

[31] U. Fasel, J. N. Kutz, B. W. Brunton, and S. L. Brunton, Proc. R. Soc. A: **478**, 20210904 (2022).

[32] S. M. Hirsh, D. A. Barajas-Solano, and J. N. Kutz, R. Soc. Open Sci. **9**, 211823 (2022).

[33] P. Gong, X. Li, J. Wang, Y. Bai, B. Chen, T. Hu, X. Liu, B. Xu, J. Yang, W. Zhang, and Y. Zhou, Remote Sens. Environ. **236**, 111510 (2020).

[34] X. Liu, Y. Huang, X. Xu, X. Li, X. Li, P. Ciais, P. Lin, K. Gong, A. D. Ziegler, A. Chen, P. Gong, J. Chen, G. Hu, Y. Chen, S. Wang, Q. Wu, K. Huang, L. Estes, and Z. Zeng, Nat. Sustainability **3**, 564 (2020).

[35] Worldpop, Worldpop—Population density (2019), www.worldpop.org.

[36] The computation of the ANN, the resulting selection of the ROIs, and the later computation of the feature lengths was done with GAIA because it provides discrete spatial data (1, built-up area; 0, no built-up area). As such, and in contrast to *WorldPop*, which distributes population quantities across administrative regions, leading to nonexistent values of 0, the identification of settlement locations and sizes is straightforward. However, for model identification *WorldPop* is used, since most pattern-forming mechanisms describe concentration distributions of different (chemical or other) species and do not create discrete patterns [37,39], similar to *WorldPop*.

[37] J. D. Murray, *Mathematical Biology. II Spatial Models and Biomedical Applications* (Springer, Berlin, 2003), p. 811.

[38] J. W. Cahn and J. E. Hilliard, J. Chem. Phys. **28**, 258 (1958).

[39] M. C. Cross and P. C. Hohenberg, Rev. Mod. Phys. **65**, 851 (1993).

[40] S. Puri, Phase Transitions **77**, 407 (2004).

[41] B. König, O. J. Ronsin, and J. Harting, Phys. Chem. Chem. Phys. **23**, 24823 (2021).

[42] H. Christiansen, S. Majumder, M. Henkel, and W. Janke, Phys. Rev. Lett. **125**, 180601 (2020).

[43] K. Champion, P. Zheng, A. Y. Aravkin, S. L. Brunton, and J. N. Kutz, IEEE Access **8**, 169259 (2020).

[44] P. Zheng, T. Askham, S. L. Brunton, J. N. Kutz, and A. Y. Aravkin, IEEE Access **7**, 1404 (2019).

[45] H. Akaike, Information theory and an extension of the maximum likelihood principle, in *Proceedings of the 2nd International Symposium on Information Theory, Tsahkadsor, Armenia, USSR, 2-8 September, 1971*, edited by B. N. Petrov and F. Csaki (Akadémiai Kiadó, Budapest, 1973), p. 267.

[46] N. M. Mangan, J. N. Kutz, S. L. Brunton, and J. L. Proctor, Proc. R. Soc A **473**, 20170009 (2017).

[47] K. P. Burnham and D. R. Anderson, Sociological Methods Res. **33**, 261 (2004).

[48] S. Thaler, L. Paehler, and N. A. Adams, J. Comput. Phys. **397**, 108851 (2019).

[49] S. M. Allen and J. W. Cahn, Acta Metall. **27**, 1085 (1979).

[50] G. T. Naozuka, H. L. Rocha, R. S. Silva, and R. C. Almeida, Nonlinear Dyn. **110**, 2589 (2022).

[51] S. Rudy, A. Alla, S. L. Brunton, and J. N. Kutz, SIAM J. Appl. Dyn. Syst. **18**, 643 (2019).

[52] J. Bakarji, K. Champion, J. N. Kutz, and S. L. Brunton, arXiv:2201.05136.

[53] S. Ouala, S. L. Brunton, B. Chapron, A. Pascual, F. Collard, L. Gaultier, and R. Fablet, Phys. D (Amsterdam, Neth.) **446**, 133630 (2023).

[54] P. Y. Lu, J. Ariño Bernad, and M. Soljačić, Commun. Phys. **5**, 206 (2022).

[55] B. de Silva, K. Champion, M. Quade, J.-C. Loiseau, J. Kutz, and S. Brunton, J. Open Source Software **5**, 2104 (2020).

[56] A. Kaptanoglu, B. de Silva, U. Fasel, K. Kaheman, A. Goldschmidt, J. Callaham, C. Delahunt, Z. Nicolaou, K. Champion, J.-C. Loiseau, J. Kutz, and S. Brunton, J. Open Source Software **7**, 3994 (2022).

[57] https://gitlab.kuleuven.be/gelenslab/publications/settlements.git.

[58] B. Prokop, L. Gelens, P. F. Pelz, and J. Friesen, Code for Challenges in identifying simple pattern-forming mechanisms in the development of settlements using demographic data (2023), https://doi.org/10.48804/4X8IPY.

[59] P. J. Clark and F. C. Evans, Ecology **35**, 445 (1954).