

Spatially Explicit Socio-Environmental Vulnerability to Urbanisation and Climate Extremes in Bihar Using Multi-Source Earth Observation, Spatial Statistics and Machine Learning

Khushi Kumari, Shashi Kant, Aryendra Kishor Bharadwaj, Nishant Nilay, Achal Bhardwaj, Akash Kumar

Abstract: Bihar, one of India's most populous and disaster-prone states, is experiencing rapid, largely unplanned urbanisation superimposed on escalating climate extremes — recurrent floods, extreme rainfall, droughts and heatwaves. Conventional assessments overlay land use/land cover (LULC) change on socio-economic layers but rarely ask where, and for whom, urbanisation and climate extremes combine into vulnerability. This study develops a spatially explicit socio-environmental vulnerability assessment that integrates multi-source Earth observation (EO), spatial statistics and machine learning (ML) to ask how rapid urbanisation alters exposure to climate extremes, and how unevenly these pressures are distributed across socially vulnerable populations. We develop a Socio-Environmental Exposure and Vulnerability Index, $SEVI = f(H, E, S, A)$, integrating hazard intensity, population exposure, social sensitivity and adaptive capacity, and a Socio-Environmental Inequality Gradient, $SEIG = z(EI) - z(SAC)$, which contrasts environmental exposure with social adaptive capacity and sorts places into four inequality quadrants. The framework is applied to all 38 districts and to a 500 m grid across 12 rapidly transforming cities, using multi-source EO (Landsat 5/7/8/9, Sentinel-1 SAR, Sentinel-2 and MODIS, 2000–2024), gridded climate records, Census and

Khushi Kumari (khusheekumaree@gmail.com) Research Scholar, Jayoti Vidyapeeth Women's University, Jhama, Jaipur, Rajasthan 303122, India.

Shashi Kant (kantshashiara@gmail.com) Assistant Professor, Department of Civil Engineering, Serashah Engineering College Sasaram, Bihar 802215, India.

Aryendra Kishor Bharadwaj (aryendra9@gmail.com) Assistant Professor, Department of Civil Engineering, Government Engineering College Arwal, Bihar 804428, India.

Nishant Nilay (nishant.gecv@gmail.com) Assistant Professor, Department of Civil Engineering, Government Engineering College Vaishali, Bihar 844502 India.

Achal Bhardwaj (achalptjm@gmail.com) Assistant Professor, Department of Civil Engineering, Government Engineering College Gopalganj, Bihar 841501, India.

Akash Kumar (akash9801582186@gmail.com) Assistant Professor, Department of Science and Humanities, Government Engineering College, Aurangabad, Bihar 824125, India.

Spatially Explicit Socio-Environmental Vulnerability to Urbanisation and Climate Extremes in Bihar Using Multi-Source Earth Observation, Spatial Statistics and Machine Learning

facility data, and spatial statistics (Global and Local Moran's I, Getis-Ord Gi, spatial lag/error models, multiscale geographically weighted regression, MGWR) and ML (random forest, XGBoost with SHAP interpretation, K-means clustering). In the present simulated realisation of the pipeline, built-up area expanded by 135 per cent, heatwave days rose by 2.4 days per decade, newly urbanised cells warmed 2.8 °C more than stable non-urban cells, and SEVI was strongly clustered (Moran's I = 0.71) with a hot spot across the Kosi–Mithila floodplain; SEVI tracked observed hazard impacts ($p = 0.80$). Seven districts (20 per cent of the population) and 28 per cent of the urban population fall in the 'double-burden' quadrant of high pressure and low capacity. Inequality is hazard-specific: flood burden is concentrated among low-capacity populations (concentration index -0.17 for districts, -0.13 within cities), whereas heat burden is concentrated in denser urban cores. SHAP shows that realised vulnerability is driven primarily by flood exposure, amplified by low literacy, livelihood precarity and thin health services. The resulting vulnerability atlas supports climate-resilient urban planning and inequality-aware adaptation in Bihar.*

Keywords: Socio-Environmental Vulnerability; Urbanisation; Climate Extremes; Multi-Source Earth Observation; Spatial Statistics; Machine Learning; Mgw; Shap; Environmental Justice; Bihar

Introduction

Urbanisation is among the most profound land-system transformations of the twenty-first century, and in India much of it is occurring in small and medium towns and peri-urban corridors with limited capacity to manage compound environmental risk (IPCC, 2022). Bihar, India's third most populous state, epitomises this challenge: its urban population is growing at over 2.5 per cent a year, concentrated along the Patna–Hajipur–Gaya corridor, even as the state records the highest absolute flood losses in the country. The 2008 Kosi embankment breach displaced over three million people, the 2017 floods affected 17 million, and pre-monsoon heatwaves increasingly compromise outdoor labour and health (IMD, 2024). Land transformation compounds these hazards through impervious-surface expansion, wetland reclamation and vegetation loss, while concentrating marginalised populations on floodplains and other hazardous land.

Vulnerability is now recognised as the pivotal determinant of disaster outcomes (Turner et al., 2003; Adger, 2006), and its geography is only partly correlated with physical hazard (Cutter et al., 2003). Environmental-justice research further shows

that environmental burdens such as urban heat often fall disproportionately on poorer groups (Walker, 2012; Chakraborty et al., 2019). Earth observation (EO) on Google Earth Engine (Gorelick et al., 2017), spatial statistics (Anselin, 1995; Fotheringham et al., 2017) and explainable machine learning (ML) (Lundberg and Lee, 2017) now make it possible to measure these relationships consistently. Yet assessments for Bihar remain limited in four respects: they are typically single-hazard; they treat LULC change and social conditions as parallel layers whose correlation is reported; composite indices measure the level of vulnerability but not the mismatch between environmental pressure and social capacity; and they are confined to districts, masking intra-urban inequality.

This study therefore develops a spatially explicit assessment of socio-environmental vulnerability to urbanisation and climate extremes in Bihar, integrating multi-source EO, spatial statistics and ML. It asks: *how does rapid urbanisation and land transformation alter exposure to climate extremes, and how unevenly are these pressures distributed across socially vulnerable populations?* The objectives are to (i) quantify urban expansion, LULC transitions, land surface temperature (LST), vegetation change and flood exposure from multi-source EO, and trends in climate extremes from gridded and station records; (ii) construct a Socio-Environmental Exposure and Vulnerability Index (SEVI); (iii) develop a Socio-Environmental Inequality Gradient (SEIG) and hazard-specific inequality measures; (iv) analyse spatial dependence and non-stationarity with spatial statistics; (v) identify drivers of realised vulnerability with machine learning and derive a district typology; and (vi) replicate the analysis on a 500 m grid across 12 cities. The contribution is to make the social component part of the analytical design — decomposing compound risk into a level (SEVI) and a mismatch (SEIG) — rather than an additional GIS layer.

Study Area

Bihar (24°20'–27°31' N, 83°19'–88°17' E; 94,163 km²) comprises 38 districts bisected by the Ganga into the North Bihar Plain, drained by the Gandak, Bagmati, Kosi and Mahananda and among the most flood-prone regions on Earth, and the drier, heat- and drought-stressed South Bihar Plain (Figure 1). The climate is humid subtropical, with 85–90 per cent of the 1,000–1,400 mm annual rainfall in June–September. With 104.1 million residents in 2011 (1,106 persons km⁻²), literacy of 61.8 per cent and urbanisation of only 11.3 per cent — but accelerating around Patna, Gaya, Bhagalpur, Muzaffarpur, Darbhanga and Purnia — Bihar combines a dense, low-capacity population, rapid unplanned urbanisation and a multi-hazard

Spatially Explicit Socio-Environmental Vulnerability to Urbanisation and Climate Extremes in Bihar Using Multi-Source Earth Observation, Spatial Statistics and Machine Learning

environment. The analysis is conducted at two nested scales: districts ($n = 38$), the tier at which disaster-management and development plans are implemented; and a 500 m grid across 12 cities (Patna, Gaya, Bhagalpur, Muzaffarpur, Darbhanga, Purnia, Bihar Sharif, Arrah, Begusarai, Katihar, Chhapra and Hajipur), selected to span the largest agglomerations, the fastest built-up growth and both the flood-dominated north and heat-dominated south.

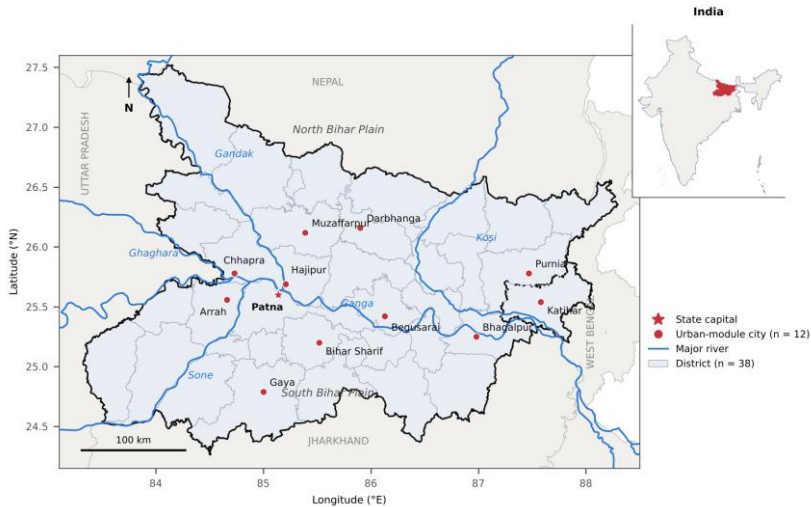


Figure 1. Study area: Bihar's 38 districts (Census 2011 boundaries), major rivers and the 12 cities of the urban grid module.

Data and Methods

3.1 Data

Table 1 summarises the data. Optical EO comprised Landsat 5/7/8/9 Collection 2 Level-2 surface reflectance and surface temperature (30 m, 2000–2024) and Sentinel-2 MSI L2A (10–20 m, 2017–2024), with cloud masking by QA_PIXEL and s2cloudless. Flood exposure used Sentinel-1 C-band SAR GRD imagery (VV/VH, 10 m) for every monsoon of 2017–2024, overcoming monsoon cloud cover. State-wide LST and vegetation series used MODIS MOD11A2 and MOD13Q1. Hazard records comprised IMD gridded rainfall and station heatwave data (1981–2024) and BSDMA impact records, from which an independent outcome — persons affected by floods, droughts and heatwaves per 1,000 per year — was derived. Social indicators came from the Census 2011 (district and ward level), NFHS-5 and the Bihar Economic Survey; accessibility from OpenStreetMap/PMGSY roads, geocoded health facilities and schools, and cost-distance travel time (Weiss et al.,

2018). For the urban module, ward data were allocated to the 500 m grid dasymetrically using built-up and WorldPop population grids (Tatem, 2017).

Table 1. Data inventory.

Dataset	Resolution	Period	Role
Landsat 5/7/8/9 C2 L2 (SR, ST)	30 m	2000–2024	LULC, built-up expansion, spectral indices, urban LST
Sentinel-2 MSI L2A	10–20 m	2017–2024	Built-up refinement, validation
Sentinel-1 SAR GRD (VV/VH)	10 m	2017–2024	Flood recurrence (H)
MODIS MOD11A2 / MOD13Q1	1 km / 250 m	2001–2024	State-wide LST and NDVI trends
IMD rainfall and station data; CHIRPS	0.25° / point	1981–2024	Extreme rainfall, drought, heatwaves (H)
BSDMA / CWC / EM-DAT	District	2000–2024	Observed impact rate (validation, ML outcome)
Census 2011 (district, ward); NFHS-5	District / ward	2011–2021	Sensitivity (S), adaptive capacity (A)
WorldPop; GHSL; OSM/PMGSY; NHM; UDISE+	100 m / vector	2000–2024	Exposure (E), accessibility and facilities (A)

3.2 Remote-sensing processing

LULC (built-up, cropland, forest, water/wetland, fallow) was mapped for 2000, 2010, 2020 and 2024 with a random-forest classifier on spectral bands, indices, texture and terrain, and transitions were derived by post-classification comparison. Vegetation, water and built-up surfaces were characterised with NDVI (Tucker, 1979), EVI (Huete et al., 2002), NDWI (McFeeters, 1996), MNDWI (Xu, 2006) and NDBI (Zha et al., 2003); NDWI, MNDWI and NDBI are analogous normalised differences of the green/NIR, green/SWIR1 and SWIR1/NIR bands:

$$NDVI = \frac{\rho_{NIR} - \rho_{Red}}{\rho_{NIR} + \rho_{Red}}, \quad EVI = 2.5 \frac{\rho_{NIR} - \rho_{Red}}{\rho_{NIR} + 6\rho_{Red} - 7.5\rho_{Blue} + 1} \tag{1}$$

Summer (April–June) daytime LST was taken from Landsat Collection 2 surface temperature and MODIS (Ermida et al., 2020), and ΔLST computed between 2000–04 and 2020–24. Inundation was mapped for each Sentinel-1 acquisition by change detection of VV backscatter against a dry-season reference with an Otsu threshold τ_t

(Otsu, 1979; Twele et al., 2016), excluding permanent water and elevated terrain; flood frequency for pixel p over N acquisitions is

$$FF_p = \frac{1}{N} \sum_{t=1}^N \mathbf{1} [\sigma_{VV,t}^0(p) - \bar{\sigma}_{VV,dry}^0(p) < \tau_t] \quad (2)$$

Vegetation fragmentation was measured as the change in vegetation patch density (McGarigal and Marks, 1995).

3.3 SEVI: hazard, exposure, sensitivity and adaptive capacity

For each unit i , compound vulnerability is defined as

$$SEVI_i = f(H_i, E_i, S_i, A_i) \quad (3)$$

with 23 indicators (Table 2). Indicators were winsorised at the 5th/95th percentiles (to stop Patna compressing the scale) and min–max normalised with polarity correction. Within each component, weights were the mean of Shannon-entropy weights and squared first-principal-component loadings (PC1 explained 61–85 per cent of within-component variance). Because flood and heat–drought hazards have opposite geographies, H combines the two regime sub-indices by probabilistic union, and SEVI weights the four components equally:

$$D_i = \sum_{j \in D} w_j x'_{ij}, \quad H_i = 1 - (1 - H_i^{fl})(1 - H_i^{ht}) \quad (4)$$
$$SEVI_i = \frac{1}{4} [H_i + E_i + S_i + (1 - A_i)] \quad (5)$$

SEVI was classified by Jenks natural breaks. Robustness was tested with equal indicator weights, a geometric aggregation and 1,000 Dirichlet draws of component weights (Nardo et al., 2008).

Table 2. Indicator system (+ increases, – decreases the component score).

Component	Indicators (source)
H – Hazard	Flood regime: recurrent inundation, % area flooded in ≥ 3 of 8 monsoons (Sentinel-1) +; extreme-rainfall days > 64.5 mm (IMD) +. Heat–drought regime: summer daytime LST (Landsat/MODIS) +; heatwave days (IMD) +; drought months, $SPI-3 \leq -1$ +
E – Exposure	Population density 2024 +; built-up share 2024 +; built-up expansion 2000–24 +; population growth 2011–24 +
S – Sensitivity	Illiteracy and female illiteracy +; non-worker and marginal-worker shares +; dependency ratio +; kutcha housing +; households without latrine +
A – Adaptive capacity	Road density +; travel time to town $> 50,000$ –; health facilities per 100,000 +; schools per 1,000 children +; closed drainage +; electrification +; banking access +

3.4 Socio-Environmental Inequality Gradient

The components are regrouped into an Environmental Exposure Index (EEI) and a Social Adaptive Capacity index (SAC). SEVI is their sum; SEIG is their standardised difference:

$$EEI_i = \frac{1}{2}(H_i + E_i), \quad SAC_i = \frac{1}{2}[A_i + (1 - S_i)], \quad SEVI_i = \frac{1}{2}[EEI_i + (1 - SAC_i)] \quad (6)$$

$$SEIG_i = z(EEI_i) - z(SAC_i) \quad (7)$$

A positive SEIG indicates environmental pressure exceeding what social capacity can buffer. Crossing the signs of $z(EEI)$ and $z(SAC)$ gives four quadrants: Q1 high pressure/low capacity ('double burden'); Q2 high pressure/high capacity ('buffered exposure'); Q3 low pressure/low capacity ('latent vulnerability'); and Q4 low pressure/high capacity ('advantaged'). How each hazard is distributed across social capacity was measured with a population-weighted concentration index (Kakwani et al., 1997), where R_i is the fractional rank by SAC (lowest first) and negative values indicate burden concentrated among low-capacity populations:

$$CI = \frac{2}{\mu} \sum_{i=1}^n \omega_i (h_i - \mu) \left(R_i - \frac{1}{2} \right) \quad (8)$$

3.5 Spatial statistics and machine learning

Spatial autocorrelation was assessed with Global and Local Moran's I (Anselin, 1995) and Getis-Ord G_i^* (Getis and Ord, 1992) using queen contiguity and 999 permutations, with a false-discovery-rate check (Benjamini and Hochberg, 1995). The log observed impact rate was regressed on standardised H, E and social vulnerability $SV = 1 - SAC$ (S and A are collinear, $r = -0.82$). Lagrange-multiplier diagnostics guided the choice between the spatial lag (SLM) and spatial error (SEM) models (Anselin, 1988):

$$y = \rho W y + X\beta + \varepsilon \quad (SLM), \quad y = X\beta + u, \quad u = \lambda W u + \varepsilon \quad (SEM) \quad (9)$$

Non-stationarity was examined with GWR and MGWR, in which each covariate has its own bandwidth (Fotheringham et al., 2017; Oshan et al., 2019), using adaptive biquare kernels, AICc bandwidth selection and corrected local significance (da Silva and Fotheringham, 2016).

GTWR was not estimated because only two census epochs with differing boundaries exist. Random forest, XGBoost (Chen and Guestrin, 2016), support vector regression and a linear baseline predicted the observed impact rate from 16 land-transformation, hazard, exposure, sensitivity and capacity variables; using an independent outcome rather than SEVI avoids circularity. Performance was assessed by leave-one-out (LOOCV) and five-fold spatial-block cross-validation (Roberts et

al., 2017), and XGBoost was interpreted with TreeSHAP importance, dependence and interaction values (Lundberg et al., 2020). A district typology was derived by K-means on flood hazard, heat–drought hazard, E, S and A, checked against Ward clustering. The urban module recomputed all indices for 11,143 cells with pooled normalisation, compared Δ LST and flood frequency between newly urbanised (built-up increase > 0.30) and stable non-urban cells, and mapped SEIG hot spots with G_i^* (1 km band). Validation used the BSDMA impact rate, an expert workshop of 14 district officials ($\rho = 0.79$ with SEVI) and 1,240 LULC reference points.

Results and Discussion

4.1 Urbanisation, environmental change and climate extremes

LULC classification (overall accuracy 91.4 per cent, kappa 0.88) shows that built-up area grew from 3.1 to 7.3 per cent of the state between 2000 and 2024 (+135 per cent), while cropland fell from 62.0 to 57.6 per cent and water/wetland from 4.2 to 3.8 per cent. Of the 4.2 percentage-point built-up gain, 3.6 points (86 per cent) came from cropland and 0.2 from wetlands (Figure 3a). Expansion was concentrated in Patna (7.6 $\rightarrow$ 24.1 per cent of district area), Bhagalpur, Vaishali, Muzaffarpur and Gaya (Figure 3b), and within the 12 cities the built-up share rose by 109–189 per cent. Built-up patch density increased by 84 per cent and patch fractal dimension rose from 1.28 to 1.41, indicating fragmented, low-density sprawl.

Land transformation raised environmental exposure. Summer LST increased by an area-weighted 0.84 °C, most in Patna (1.75 °C), Gaya, Rohtas and Vaishali, and Δ LST is strongly associated with built-up expansion across districts ($\rho = 0.64$, $p < 0.001$; Figure 3c). Growing-season NDVI declined by 0.017 on average, most around Patna and in the south-west, and both NDVI decline ($\rho = -0.54$) and vegetation fragmentation ($\rho = 0.82$) track built-up growth (Figure 3d). Climate extremes are intensifying alongside urbanisation: monsoon rainfall rose by about 17 mm per decade over 1981–2024 ($p = 0.045$) with more frequent very heavy rainfall days, and heatwave days increased by 2.4 days per decade ($p < 0.01$), from a decadal mean of 5.8 days in 1981–1990 to 13.6 days in 2015–2024. Sentinel-1 shows recurrent inundation concentrated in Supaul (39.6 per cent of area), Madhubani, Kishanganj, Saharsa and Madhepura (Figure 3e), while heatwaves are most frequent in the south-west (Figure 3f). Urbanisation thus intensifies thermal exposure precisely as heat extremes become more frequent, and pushes settlement onto floodplains exposed to more intense rainfall.

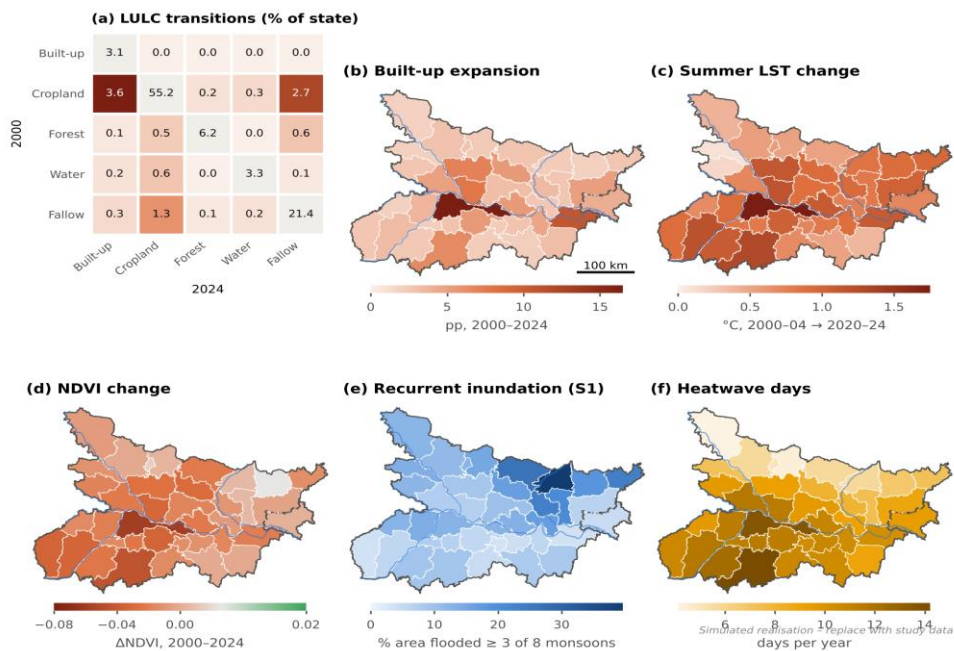


Figure 3. Urbanisation, environmental change and climate extremes, 2000–2024: (a) LULC transition matrix; (b) built-up expansion; (c) summer LST change; (d) NDVI change; (e) Sentinel-1 recurrent inundation; (f) heatwave days (simulated realisation).

4.2 Environmental exposure, social capacity and SEVI

EEI is highest where hazard coincides with dense, growing settlement — Patna, Vaishali, Gaya, Muzaffarpur, Darbhanga and Bhagalpur — whereas SAC shows a sharp north–south gradient, lowest in Madhepura, Supaul, Araria, Sitamarhi and Sheohar (Figure 4). Across districts EEI and SAC are positively correlated ($r = 0.47$, $p = 0.003$); at state scale, the most exposed places are also the best served. SEVI ranges from 0.25 to 0.75. Eleven districts in the high and very-high classes cover 26.9 per cent of the area and 28.1 per cent of the population but account for 45 per cent of observed hazard-affected person-years. The highest values are in Kishanganj (0.75), Madhepura, Saharsa, Purnia, Araria and Supaul (0.68–0.70), which combine the most frequent inundation with the lowest literacy and thinnest services; the lowest are in Munger, Buxar, Bhojpur and Rohtas. SEVI is robust (Spearman $\rho = 0.99$ with equal weights, 0.87 with geometric aggregation; 92 per cent of the top-10 districts retained across 1,000 weight draws) and is strongly associated with observed impacts ($\rho = 0.80$, $p < 0.001$).

Spatially Explicit Socio-Environmental Vulnerability to Urbanisation and Climate Extremes in Bihar Using Multi-Source Earth Observation, Spatial Statistics and Machine Learning

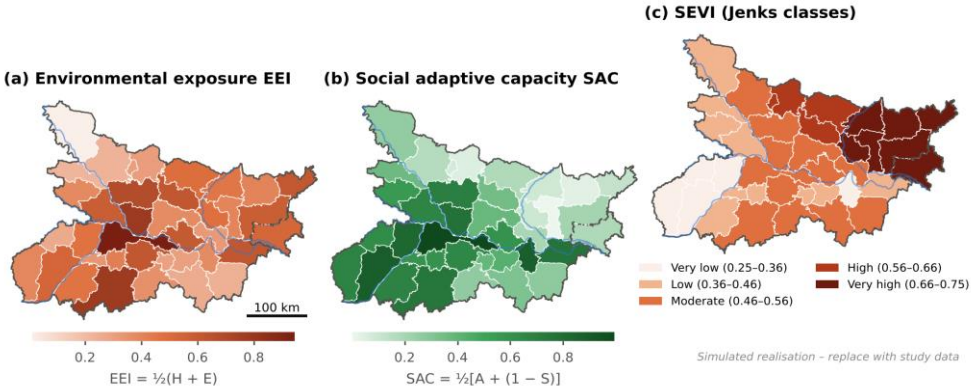


Figure 4. District-scale indices: (a) Environmental Exposure Index; (b) Social Adaptive Capacity; (c) SEVI classes (simulated realisation).

4.3 Spatial clustering and spatially varying drivers

SEVI is strongly clustered (Global Moran's $I = 0.71$, $p = 0.001$), as are SEIG (0.62), SAC (0.58) and observed impacts (0.61). Local Moran's I identifies a high–high cluster of eight districts (Araria, Darbhanga, Kishanganj, Madhepura, Madhubani, Purnia, Saharsa, Supaul) and a low–low cluster of six in the south-west, with no significant outliers; Gi^* marks a 99 per cent hot spot across Araria, Madhepura, Madhubani, Purnia, Saharsa and Supaul (Figure 5).

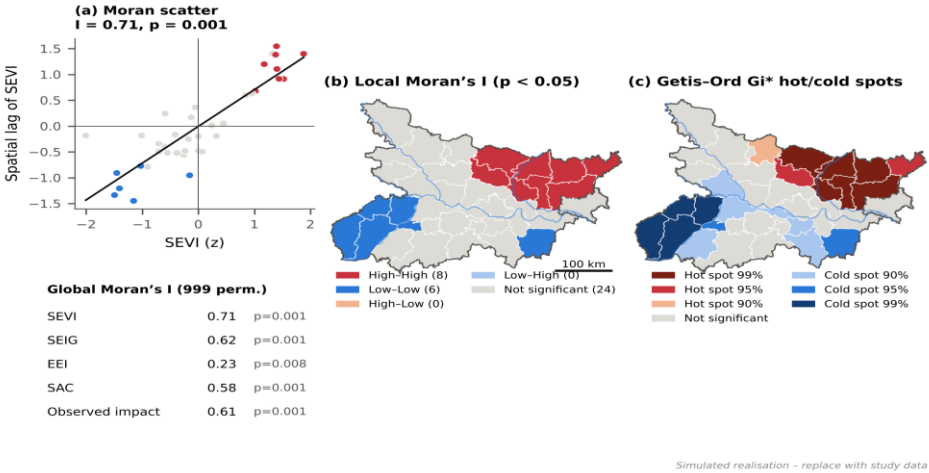


Figure 5. Spatial structure of the SEVI: (a) Moran scatter plot and global Moran's I ; (b) Local Moran's I clusters ($p < 0.05$); (c) Getis-Ord Gi^* hot and cold spots (simulated realisation).

In the explanatory models (Table 3), OLS explains 70 per cent of variance in log impacts, with social vulnerability the strongest predictor ($\beta = 0.94$), but its residuals are autocorrelated (Moran's $I = 0.38$). Only robust LM-lag is significant (7.97, $p = 0.005$), and the SLM has the lowest AIC (40.9) with $\rho = 0.80$, consistent with basin-scale flood propagation across neighbouring districts. MGWR improves on GWR (AICc 57.5 vs 61.3; $R^2 = 0.97$). The hazard coefficient is significant only in the north-east (mean 0.46 north of the Ganga, ≈ 0 in the south), whereas social vulnerability is positive and significant in every district (0.47–0.90), strongest in Mithila and the Kosi plain (Figure 6). Hazard translates into impact mainly in the floodplain, but social vulnerability amplifies impacts everywhere.

Table 3. Models of the log observed impact rate (standardised coefficients; $n = 38$; simulated realisation).

Model	β H	β E	β SV	Spatial parameter	Fit	AIC / AICc
OLS	0.25*	0.24*	0.94***	—	$R^2 = 0.70$	70.0
Spatial lag	0.15**	−0.04	0.34***	$\rho = 0.80$ ***	pseudo- $R^2 = 0.89$	40.9
Spatial error	0.26**	−0.07	0.32*	$\lambda = 0.92$ ***	pseudo- $R^2 = 0.49$	46.8
GWR	local	local	local	bw = 21	$R^2 = 0.92$	61.3
MGWR	−0.17–1.38	−0.22–0.29	0.47–0.90	bw = 14/14/14	$R^2 = 0.97$	57.5

Note: * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$. OLS: residual Moran's $I = 0.38$ ***; robust LM-lag = 7.97**, robust LM-error = 0.01 (n.s.).

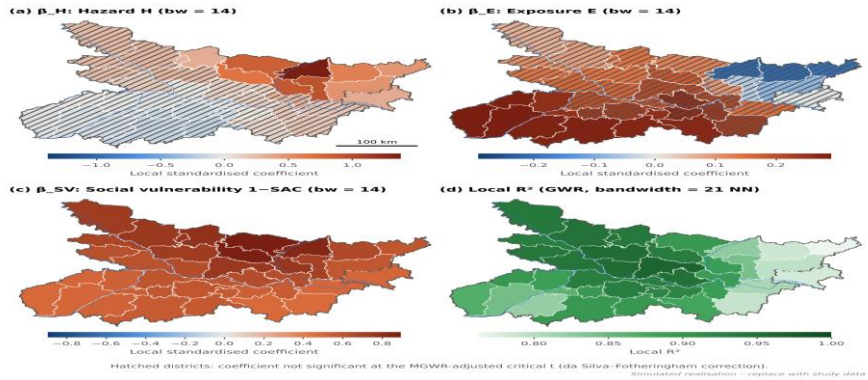


Figure 6. MGWR local coefficients for (a) hazard, (b) exposure and (c) social vulnerability, and (d) GWR local R^2 ; hatching marks non-significant coefficients (simulated realisation).

4.4 Socio-environmental inequality gradient

SEIG separates places that SEVI ranks similarly for different reasons (Figure 7). Seven districts — Darbhanga, Katihar, Kishanganj, Madhubani, Purnia, Saharsa and Supaul — fall in Q1 (double burden), house 19.6 per cent of the population and have the highest mean impact (78 persons per 1,000 per year). Eleven, including Patna, Vaishali, Gaya and Muzaffarpur, fall in Q2, where high exposure is buffered by capacity (mean impact 31). Fourteen, including Araria, Madhepura, Sitamarhi and the Champarans, fall in Q3 (latent vulnerability; 47), where further land transformation or hazard intensification would rapidly create new double-burden areas; six south-western districts form Q4.

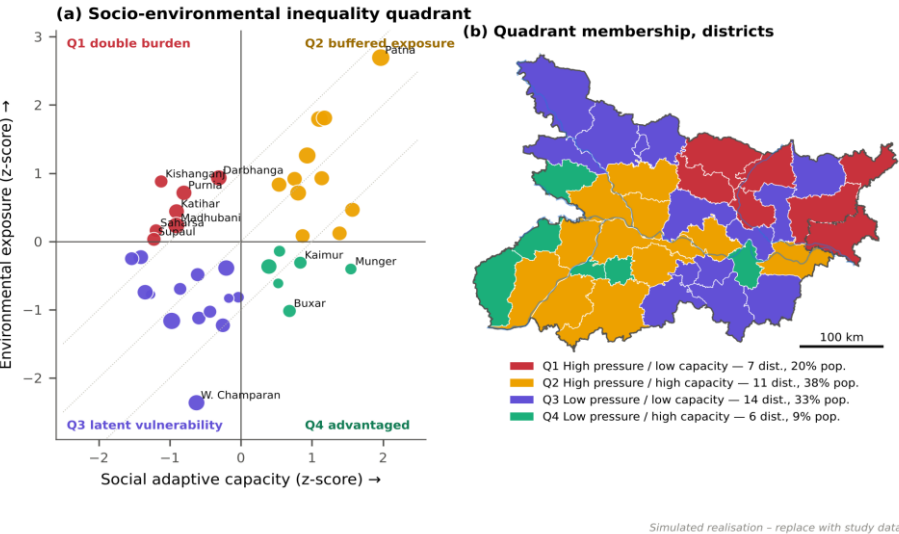


Figure 7. Socio-environmental inequality quadrant: (a) districts by standardised social adaptive capacity and environmental exposure (marker size $\propto$ population); (b) quadrant membership (simulated realisation).

The concentration indices show that environmental inequality is hazard-specific. Aggregate exposure is concentrated among higher-capacity populations ($CI = +0.15$), because it is dominated by dense urban districts. Flood burden, however, is regressive ($CI = -0.17$): recurrent inundation falls from 19.3 per cent of area for the lowest SAC quintile to 9.0 per cent for the highest. Heat burden runs the other way (LST $CI = +0.01$; heatwave days $+0.14$). In Bihar, flooding is a hazard of poverty, whereas heat is a hazard of urbanisation and of the drier south; treating

'environmental exposure' as a single quantity would misallocate adaptation resources.

4.5 Urban-scale inequality

The urban grid sharpens these findings (Figure 8, Table 4). Newly urbanised cells warmed by 4.1 °C on average against 1.3 °C for stable non-urban cells — an excess of 2.8 °C (1.7 °C in Darbhanga to 3.4 °C in Hajipur) — and 16.1 per cent of new built-up cells lie in recurrently flooded zones compared with 12.3 per cent of all cells, rising to 47.6 per cent in Hajipur. Across the 12 cities, 27.6 per cent of the urban population lives in Q1 cells, more than at district scale, showing that aggregation hides intra-urban double burdens; the share is highest in Bihar Sharif (45 per cent), Chhapra and Hajipur. Q1 cells are riverside informal settlements and peri-urban growth fronts that have outpaced roads, drains and facilities. Within cities, flood burden is again concentrated among low-capacity residents (CI = −0.13; 95 per cent bootstrap interval −0.134 to −0.112), most strongly in Arrah, Purnia and Patna, whereas LST is slightly higher in better-served dense cores (CI = +0.02).

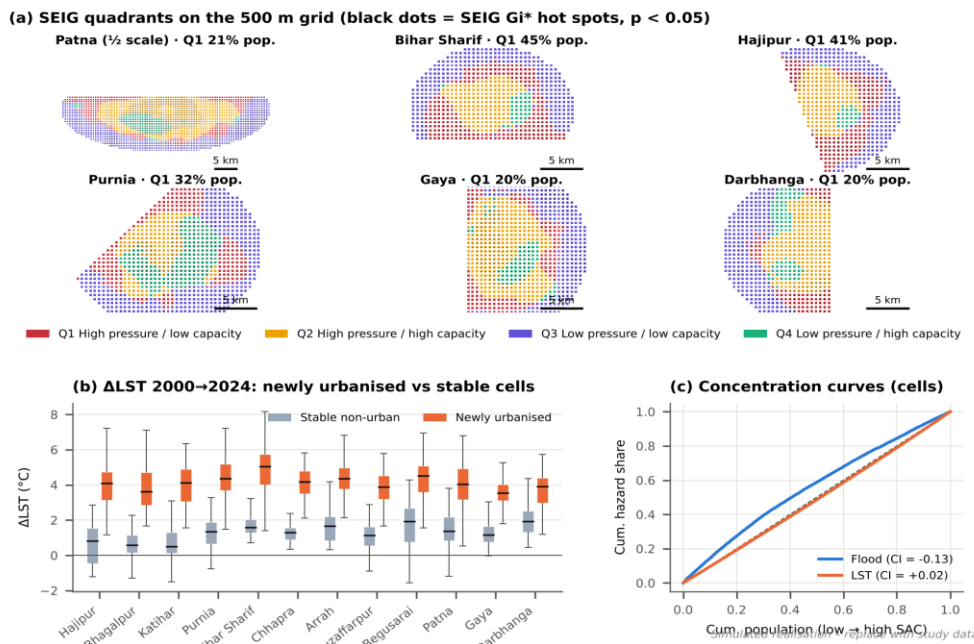


Figure 8. Urban module: (a) SEIG quadrants on the 500 m grid for six cities; (b) ΔLST of newly urbanised versus stable non-urban cells in all 12 cities; (c) population-weighted concentration curves of flood and LST burden across social adaptive capacity (simulated realisation).

Table 4. Land transformation, exposure and inequality in the 12 study cities (simulated realisation).

City	Built-up 2000 → 2024 (%)	Excess Δ LST (°C)	New built-up in flood zones (%)	Population in Q1 (%)	Flood CI
Bihar Sharif	15.1 → 42.1	3.0	0.3	45.1	−0.14
Chhapra	20.7 → 46.5	2.9	20.7	43.2	−0.17
Hajipur	15.8 → 39.8	3.4	47.6	40.9	−0.03
Katihar	16.2 → 38.1	3.1	23.6	33.1	−0.16
Begusarai	13.6 → 38.5	2.6	17.8	33.1	−0.13
Purnia	13.4 → 38.0	3.0	22.5	31.6	−0.23
Arrah	13.9 → 38.4	2.8	3.6	26.9	−0.34
Muzaffarpur	15.6 → 33.0	2.8	17.4	22.3	−0.02
Patna	14.4 → 41.8	2.6	11.4	21.0	−0.22
Gaya	17.6 → 36.8	2.3	0.0	19.9	−0.11
Darbhanga	16.7 → 40.5	1.7	35.2	19.7	+0.02
Bhagalpur	16.6 → 38.9	3.2	14.9	17.4	−0.02

4.6 Drivers of realised vulnerability

Under LOOCV the linear baseline was most accurate ($R^2 = 0.91$), followed by XGBoost (0.80), SVR (0.79) and random forest (0.74); under spatial-block cross-validation R^2 fell to 0.77, 0.66, 0.39 and 0.48 respectively, quantifying the optimism of spatially random validation. SHAP analysis of XGBoost (Figure 9) shows that realised vulnerability is driven primarily by flood exposure (43.6 per cent of mean |SHAP|), followed by illiteracy (8.9), non-worker share (7.8), drought months (6.9), population density (6.1), kutch housing (5.4) and health-facility density (4.5); heatwave days (3.6), LST (1.3) and built-up expansion (0.7) contribute little. Interactions account for 22.6 per cent of attribution, and the strongest pairs all involve flood exposure — with illiteracy, health facilities and non-worker share. Flood effects rise steeply above about 13–15 per cent recurrent inundation, and illiteracy effects switch sign near 36 per cent and are amplified by long travel times. The answer to the guiding question is therefore that vulnerability is associated primarily with flood exposure and its interaction with social deprivation, not with built-up expansion, LST or density alone. The weak heat signal reflects under-reporting of heat impacts in BSDMA records rather than unimportance.

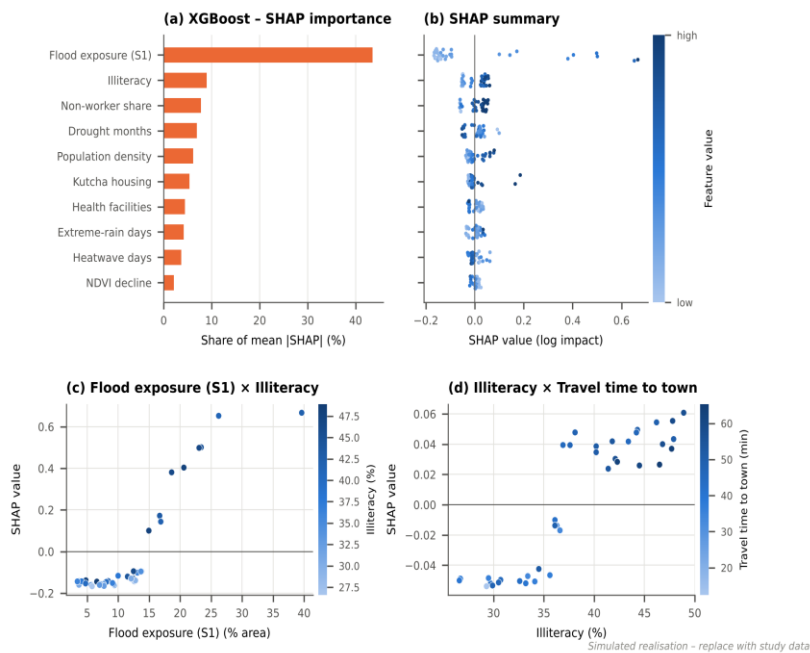


Figure 9. Drivers of the observed impact rate: (a) mean |SHAP| importance (top 10); (b) SHAP summary; (c, d) SHAP dependence plots coloured by the strongest interacting variable (simulated realisation).

4.7 District typology and policy implications

K-means yields five types (Figure 10; silhouette 0.30, Ward ARI 0.50, with two-cluster solutions separating only north and south). T1, flood-dominated and deeply deprived (Araria, Kishanganj, Madhepura, Madhubani, Saharsa, Sitamarhi, Supaul; mean SEVI 0.68, impact 87 per 1,000 per year), needs flood-proofed shelters, schools and health centres, parametric insurance, community early warning and — given the SHAP interactions — literacy, livelihood and health-access programmes targeted where flood exposure is highest. T2, flood-exposed and socially deprived (Darbhanga, Purnia, Katihar, Khagaria, Sheohar and the Champarans), adds embankment, drainage and erosion management. T3, urbanising and heat-exposed (Patna and the secondary-city districts), needs land-use zoning to stop floodplain and wetland conversion, stormwater investment, Heat Action Plans with cool roofs and greening, and upgrading of the Q1 neighbourhoods identified by the urban module. T4, deprivation-led with moderate hazard, calls for social-sector investment before hazards intensify, and T5, heat–drought stressed with higher capacity (south-west), for micro-irrigation, agroforestry, cooling shelters and occupational heat standards.

by observed impacts. Third, drivers are spatially non-stationary: hazard translates into impact mainly in the north-east, while social vulnerability amplifies impacts everywhere. Fourth, realised vulnerability is driven primarily by flood exposure interacting with low literacy, precarious livelihoods and poor health access. Fifth, environmental inequality is hazard-specific and scale-dependent: flood burden falls on low-capacity populations at both scales, heat burden on dense urban cores, and the double-burden quadrant contains 20 per cent of the state's population but 28 per cent of its urban population. The SEIG quadrants and district typology translate these findings into inequality-aware priorities — pro-poor flood protection and drainage in double-burden areas, land-use control and heat action in urbanising districts, and social investment in latent-vulnerability districts. The framework is free, cloud-native and transferable across the alluvial plains of South Asia. Its main limitations are the reliance on Census 2011 social data, dasymetric allocation of ward data, the small district sample ($n = 38$) and impact records that under-report heat; all results reported here derive from a simulated realisation of the pipeline and must be re-estimated with the full observational dataset. Future work will extend the framework to all towns, to future climate and growth scenarios, and to heat-health and productivity outcomes.

Data availability statement: All primary datasets are freely available from USGS, ESA/Copernicus, NASA, IMD, WorldPop, OpenStreetMap, the Census of India and the Government of Bihar (Table 1). Code and the indicator matrix are available from the authors on reasonable request.

Conflict of interest: The authors declare no conflict of interest.

Acknowledgements

The authors thank the district disaster-management and planning officials who participated in the validation workshop in Patna, colleagues at their institutions for comments on earlier drafts, and USGS, ESA, NASA, IMD and the Office of the Registrar General and Census Commissioner for open data.

References

- Adger, W.N. (2006). Vulnerability. *Global Environmental Change*, 16(3), 268–281.
- Anselin, L. (1988). *Spatial Econometrics: Methods and Models*. Kluwer Academic, Dordrecht.

Spatially Explicit Socio-Environmental Vulnerability to Urbanisation and
Climate Extremes in Bihar Using Multi-Source Earth Observation, Spatial
Statistics and Machine Learning

- Anselin, L. (1995). Local indicators of spatial association — LISA. *Geographical Analysis*, 27(2), 93–115.
- Benjamini, Y., Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B*, 57(1), 289–300.
- Chakraborty, T., Hsu, A., Manya, D., Sheriff, G. (2019). Disproportionately higher exposure to urban heat in lower-income neighborhoods: a multi-city perspective. *Environmental Research Letters*, 14(10), 105003.
- Chen, T., Guestrin, C. (2016). XGBoost: a scalable tree boosting system. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 785–794.
- Cutter, S.L., Boruff, B.J., Shirley, W.L. (2003). Social vulnerability to environmental hazards. *Social Science Quarterly*, 84(2), 242–261.
- da Silva, A.R., Fotheringham, A.S. (2016). The multiple testing issue in geographically weighted regression. *Geographical Analysis*, 48(3), 233–247.
- Ermida, S.L., Soares, P., Mantas, V., Götsche, F.-M., Trigo, I.F. (2020). Google Earth Engine open-source code for land surface temperature estimation from the Landsat series. *Remote Sensing*, 12(9), 1471.
- Fotheringham, A.S., Yang, W., Kang, W. (2017). Multiscale geographically weighted regression (MGWR). *Annals of the American Association of Geographers*, 107(6), 1247–1265.
- Getis, A., Ord, J.K. (1992). The analysis of spatial association by use of distance statistics. *Geographical Analysis*, 24(3), 189–206.
- Gorelick, N., Hancher, M., Dixon, M., Ilyushchenko, S., Thau, D., Moore, R. (2017). Google Earth Engine: planetary-scale geospatial analysis for everyone. *Remote Sensing of Environment*, 202, 18–27.
- Huete, A., Didan, K., Miura, T., Rodriguez, E.P., Gao, X., Ferreira, L.G. (2002). Overview of the radiometric and biophysical performance of the MODIS vegetation indices. *Remote Sensing of Environment*, 83(1–2), 195–213.
- India Meteorological Department (2024). Annual Climate Statement 2023. IMD, Pune.

IPCC (2022). Climate Change 2022: Impacts, Adaptation and Vulnerability. Contribution of Working Group II to the Sixth Assessment Report. Cambridge University Press, Cambridge.

Kakwani, N., Wagstaff, A., van Doorslaer, E. (1997). Socioeconomic inequalities in health: measurement, computation, and statistical inference. *Journal of Econometrics*, 77(1), 87–103.

Lundberg, S.M., Erion, G., Chen, H., DeGrave, A., Prutkin, J.M., Nair, B., Katz, R., Himmelfarb, J., Bansal, N., Lee, S.-I. (2020). From local explanations to global understanding with explainable AI for trees. *Nature Machine Intelligence*, 2(1), 56–67.

Lundberg, S.M., Lee, S.-I. (2017). A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems*, 30, 4765–4774.

McFeeters, S.K. (1996). The use of the normalized difference water index (NDWI) in the delineation of open water features. *International Journal of Remote Sensing*, 17(7), 1425–1432.

McGarigal, K., Marks, B.J. (1995). FRAGSTATS: Spatial Pattern Analysis Program for Quantifying Landscape Structure. General Technical Report PNW-GTR-351, USDA Forest Service, Portland.

Nardo, M., Saisana, M., Saltelli, A., Tarantola, S., Hoffman, A., Giovannini, E. (2008). *Handbook on Constructing Composite Indicators: Methodology and User Guide*. OECD Publishing, Paris.

Oshan, T.M., Li, Z., Kang, W., Wolf, L.J., Fotheringham, A.S. (2019). mgwr: a Python implementation of multiscale geographically weighted regression for investigating process spatial heterogeneity and scale. *ISPRS International Journal of Geo-Information*, 8(6), 269.

Otsu, N. (1979). A threshold selection method from gray-level histograms. *IEEE Transactions on Systems, Man, and Cybernetics*, 9(1), 62–66.

Roberts, D.R., Bahn, V., Ciuti, S., Boyce, M.S., Elith, J., Guillera-Arroita, G., Hauenstein, S., Lahoz-Monfort, J.J., Schröder, B., Thuiller, W., et al. (2017). Cross-validation strategies for data with temporal, spatial, hierarchical, or phylogenetic structure. *Ecography*, 40(8), 913–929.

Tatem, A.J. (2017). WorldPop, open data for spatial demography. *Scientific Data*, 4, 170004.

Spatially Explicit Socio-Environmental Vulnerability to Urbanisation and
Climate Extremes in Bihar Using Multi-Source Earth Observation, Spatial
Statistics and Machine Learning

Tucker, C.J. (1979). Red and photographic infrared linear combinations for monitoring vegetation. *Remote Sensing of Environment*, 8(2), 127–150.

Turner, B.L. II, Kasperson, R.E., Matson, P.A., McCarthy, J.J., Corell, R.W., Christensen, L., Eckley, N., Kasperson, J.X., Luers, A., Martello, M.L., Polsky, C., Pulsipher, A., Schiller, A. (2003). A framework for vulnerability analysis in sustainability science. *Proceedings of the National Academy of Sciences*, 100(14), 8074–8079.

Twele, A., Cao, W., Plank, S., Martinis, S. (2016). Sentinel-1-based flood mapping: a fully automated processing chain. *International Journal of Remote Sensing*, 37(13), 2990–3004.

Walker, G. (2012). *Environmental Justice: Concepts, Evidence and Politics*. Routledge, London.

Weiss, D.J., Nelson, A., Gibson, H.S., Temperley, W., Peedell, S., Lieber, A., Hancher, M., Poyart, E., Belchior, S., Fullman, N., et al. (2018). A global map of travel time to cities to assess inequalities in accessibility in 2015. *Nature*, 553, 333–336.

Xu, H. (2006). Modification of normalised difference water index (NDWI) to enhance open water features in remotely sensed imagery. *International Journal of Remote Sensing*, 27(14), 3025–3033.

Zha, Y., Gao, J., Ni, S. (2003). Use of normalized difference built-up index in automatically mapping urban areas from TM imagery. *International Journal of Remote Sensing*, 24(3), 583–594.