Supporting Information

Gatto et al. 10.1073/pnas.1217567109

SI Text

Derivation of Onset Conditions. To analyze stability, we consider the Jacobian of the linearized system evaluated at the disease-free equilibrium $\boldsymbol{X_0}$ ($S_i = H_i, I_i = 0$, $B_i = 0$) for all i, which is given by

$$\boldsymbol{J} = \begin{bmatrix} j_{11} & 0 & j_{13} \\ 0 & j_{22} & j_{23} \\ 0 & j_{32} & j_{33} \end{bmatrix},$$

where

$$\begin{aligned} j_{11} &= -\mu \boldsymbol{U_n} \\ j_{13} &= -m_S \boldsymbol{HQ\beta} - (1-m_S)\boldsymbol{H\beta} \\ j_{22} &= -\phi \boldsymbol{U_n} \\ j_{23} &= m_S \boldsymbol{HQ\beta} + (1-m_S)\boldsymbol{H\beta} \\ j_{32} &= \frac{m_I}{K}\boldsymbol{pW^{-1}Q^T} + \frac{1-m_I}{K}\boldsymbol{pW^{-1}} \\ j_{33} &= -(\mu_B + l)\boldsymbol{U_n} + l\boldsymbol{W^{-1}P^TW}. \end{aligned}$$

Note that the variables for pathogen have been scaled as $B_i^* = B_i/K$. Because of its block-triangular structure, the Jacobian has obviously n eigenvalues equal to $-\mu$; therefore, instability is determined by the eigenvalues of the block matrix

$$\boldsymbol{J^*} = \begin{bmatrix} j_{22} & j_{23} \\ j_{32} & j_{33} \end{bmatrix}.$$

$\boldsymbol{J^*}$ is a proper Metzler matrix (1); namely, its off-diagonal entries are all nonnegative and at least one diagonal entry is negative. Thus its eigenvalue with maximal real part (dominant eigenvalue) is real. If the union of the graphs associated with matrices $\boldsymbol{P}$ and $\boldsymbol{Q}$ is strongly connected, then the graph associated with $\boldsymbol{J^*}$ is also strongly connected. Therefore one can apply the Perron–Frobenius theorem (2) for irreducible matrices and state that the dominant eigenvalue is a simple real root of the characteristic polynomial. The condition for the transcritical bifurcation of the disease-free equilibrium is that the dominant eigenvalue crosses the imaginary axis at zero; namely, the determinant of $\boldsymbol{J^*}$ is zero (3). Actually, when the disease-free equilibrium is stable, all the eigenvalues have negative real parts and $\det(\boldsymbol{J^*})$ is positive because $\boldsymbol{J^*}$ is a matrix of order $2n$. So the disease-free equilibrium becomes unstable when $\det(\boldsymbol{J^*})$ switches from positive to negative or equivalently the dominant eigenvalue becomes zero. For block matrices of the kind

$$\begin{bmatrix} \boldsymbol{A} & \boldsymbol{B} \\ \boldsymbol{C} & \boldsymbol{D} \end{bmatrix},$$

in which all blocks are square and matrix $\boldsymbol{A}$ commutes with matrix $\boldsymbol{C}$, the following equality holds (4):

$$\det\begin{bmatrix} \boldsymbol{A} & \boldsymbol{B} \\ \boldsymbol{C} & \boldsymbol{D} \end{bmatrix} = \det(\boldsymbol{AD} - \boldsymbol{CB}).$$

As $\boldsymbol{U_n}$ obviously commutes with any matrix, we have

$$\begin{aligned} \det(\boldsymbol{J^*}) = \det\Big[& \phi(\mu_B + l)\boldsymbol{U_n} - \phi l \boldsymbol{W^{-1}P^TW} + \\ & - \frac{m_S m_I}{K}\boldsymbol{pW^{-1}Q^THQ\beta} + \\ & - \frac{m_I(1-m_S)}{K}\boldsymbol{pW^{-1}Q^TH\beta} + \\ & - \frac{(1-m_I)m_S}{K}\boldsymbol{pW^{-1}HQ\beta} + \\ & - \frac{(1-m_I)(1-m_S)}{K}\boldsymbol{pW^{-1}H\beta}\Big]. \end{aligned}$$

Let us now introduce the basic reproduction number of each community i when isolated from the others. For waterborne diseases this quantity reads as (5)

$$R_{0i} = \frac{p_i H_i \beta_i}{W_i K \mu_B \phi}.$$

Although originally derived via stability analysis, the previous expression can also be obtained with standard epidemiological arguments [e.g., typical of susceptible-infected-recovered (SIR)-like models (6)]. In fact, after the introduction of an infected individual into a completely susceptible population, the number of new infections per unit time is $\beta_i B_i S_i/(K + B_i)$. At the disease onset $S_i \approx H_i$ and both B_i and I_i are small, so the rate of new infections is approximately $\beta_i B_i H_i/K$. To find the number of secondary infections produced by one infected during the infectious period, a relationship between B_i and I_i must be found. To this end, one can assume equilibrium in the bacterial dynamics ($dB_i/dt = 0$), which gives $B_i = p_i I_i/(\mu_B W_i)$. The rate of new infections produced by one infected at the disease onset is thus given by $p_i H_i \beta_i/(\mu_B W_i K)$. This expression has to be multiplied by the average length of the infectious period $1/(\mu + \alpha + \gamma) = 1/\phi$, thus providing the basic reproduction number for waterborne diseases introduced above. We can now define the matrix

$$\boldsymbol{R_0} = \begin{bmatrix} R_{01} & 0 & \cdots & 0 \\ 0 & R_{02} & \cdots & 0 \\ \cdots & & & \\ 0 & 0 & \cdots & R_{0n} \end{bmatrix} = \frac{\boldsymbol{p}}{K\mu_B\phi}\boldsymbol{H\beta W^{-1}}.$$

Because $\boldsymbol{H}$, $\boldsymbol{\beta}$, and $\boldsymbol{W^{-1}}$ are diagonal, thus commuting, matrices, we can also state that $\boldsymbol{R_0} = \frac{\boldsymbol{p}}{K\mu_B\phi}\boldsymbol{W^{-1}H\beta}$ and rework the determinant of $\boldsymbol{J^*}$ as

$$\begin{aligned} \det(\boldsymbol{J^*}) = [\phi(\mu_B + l)]^n \det\Big[& \boldsymbol{U_n} - \frac{l}{\mu_B + l}\boldsymbol{W^{-1}P^TW} + \\ & - \frac{\mu_B}{\mu_B + l}\frac{m_S m_I}{K\mu_B\phi}\boldsymbol{pW^{-1}Q^THQ\beta} + \\ & - \frac{\mu_B}{\mu_B + l}\frac{m_I(1-m_S)}{K\mu_B\phi}\boldsymbol{pW^{-1}Q^TH\beta} + \\ & - \frac{\mu_B}{\mu_B + l}\frac{(1-m_I)m_S}{K\mu_B\phi}\boldsymbol{pW^{-1}HQ\beta} + \\ & - \frac{\mu_B}{\mu_B + l}(1-m_I)(1-m_S)\boldsymbol{R_0}\Big]. \end{aligned}$$

As the determinant of a product of square matrices is the product of determinants and the determinant of an inverse matrix is the inverse of the determinant, we can write

$$\det(\boldsymbol{J}^*) = [\phi(\mu_B + l)]^n \det\Bigg\{\boldsymbol{W}\Bigg[\boldsymbol{U_n} - \frac{l}{\mu_B + l}\boldsymbol{W}^{-1}\boldsymbol{P}^T\boldsymbol{W} + - \frac{\mu_B}{\mu_B + l}\Bigg((1-m_I)(1-m_S)\boldsymbol{R_0} + + \frac{m_S m_I}{K\mu_B\phi}\boldsymbol{p}\boldsymbol{W}^{-1}\boldsymbol{Q}^T\boldsymbol{H}\boldsymbol{Q}\boldsymbol{\beta} + + \frac{m_I(1-m_S)}{K\mu_B\phi}\boldsymbol{p}\boldsymbol{W}^{-1}\boldsymbol{Q}^T\boldsymbol{H}\boldsymbol{\beta} + + \frac{(1-m_I)m_S}{K\mu_B\phi}\boldsymbol{p}\boldsymbol{W}^{-1}\boldsymbol{H}\boldsymbol{Q}\boldsymbol{\beta}\Bigg)\Bigg]\boldsymbol{W}^{-1}\Bigg\}.$$

Therefore, because $\boldsymbol{p}$ and $\boldsymbol{W}$ are diagonal and thus commuting matrices, the condition $\det(\boldsymbol{J}^*) = 0$ is given by

$$\det\Bigg\{\boldsymbol{U_n} - \frac{l}{\mu_B + l}\boldsymbol{P}^T - \frac{\mu_B}{\mu_B + l}\Bigg[(1-m_I)(1-m_S)\boldsymbol{R_0} + + \frac{m_S m_I}{K\mu_B\phi}\boldsymbol{p}\boldsymbol{Q}^T\boldsymbol{H}\boldsymbol{Q}\boldsymbol{\beta}\boldsymbol{W}^{-1} + + \frac{m_I(1-m_S)}{K\mu_B\phi}\boldsymbol{p}\boldsymbol{Q}^T\boldsymbol{H}\boldsymbol{\beta}\boldsymbol{W}^{-1} + + \frac{(1-m_I)m_S}{K\mu_B\phi}\boldsymbol{p}\boldsymbol{H}\boldsymbol{Q}\boldsymbol{\beta}\boldsymbol{W}^{-1}\Bigg]\Bigg\} = 0.$$

In addition to the matrix $\boldsymbol{R_0} = \frac{\boldsymbol{p}}{K\mu_B\phi}\boldsymbol{H}\boldsymbol{\beta}\boldsymbol{W}^{-1}$ we can now introduce three other matrices of reproduction numbers; namely,

$$\boldsymbol{R_0^I} = \frac{\boldsymbol{p}\boldsymbol{Q}^T\boldsymbol{H}\boldsymbol{\beta}\boldsymbol{W}^{-1}}{K\mu_B\phi}, \quad \boldsymbol{R_0^S} = \frac{\boldsymbol{p}\boldsymbol{H}\boldsymbol{Q}\boldsymbol{\beta}\boldsymbol{W}^{-1}}{K\mu_B\phi},$$

and

$$\boldsymbol{R_0^{IS}} = \frac{\boldsymbol{p}\boldsymbol{Q}^T\boldsymbol{H}\boldsymbol{Q}\boldsymbol{\beta}\boldsymbol{W}^{-1}}{K\mu_B\phi},$$

corresponding to metacommunities with infectives only being mobile, susceptibles only being mobile, and both infectives and susceptibles being mobile, respectively. If we account for the different probabilities of movement in the metacommunity, we can define a transmission matrix averaged over nonmobile individuals, mobile infectives, and mobile susceptibles as

$$\boldsymbol{T_0} = (1-m_I)(1-m_S)\boldsymbol{R_0} + m_S m_I \boldsymbol{R_0^{IS}} + m_I(1-m_S)\boldsymbol{R_0^I} + (1-m_I)m_S\boldsymbol{R_0^S}.$$

Therefore, the bifurcation of the disease-free equilibrium corresponds to the condition

$$\det\left(\boldsymbol{U_n} - \frac{l}{\mu_B + l}\boldsymbol{P}^T - \frac{\mu_B}{\mu_B + l}\boldsymbol{T_0}\right) = 0.$$

Equivalently, the dominant eigenvalue Λ_0 of the matrix

$$\boldsymbol{G_0} = \frac{l}{\mu_B + l}\boldsymbol{P}^T + \frac{\mu_B}{\mu_B + l}\boldsymbol{T_0},$$

which is a convex combination of $\boldsymbol{P}^T$ and $\boldsymbol{T_0}$, must equal unity. Actually, the disease-free equilibrium switches from being stable to being a saddle, thus triggering the start of the disease, whenever the dominant eigenvalue of $\boldsymbol{J}^*$ switches from positive to negative, and hence whenever Λ_0 switches from being less than 1 to being larger than 1.

Geography of Disease Onset. The geography of disease onset, i.e., the spatial localization of the sites that are hit with more strength during the early phase of the epidemic, is determined by the dominant eigenvector of the Jacobian matrix $\boldsymbol{J}^*$. The eigenvector lies in the subspace (of dimension $2n$) $S_i - H_i = 0$ $(i = 1, n)$ and has strictly positive components I_i and B_i according to the Perron–Frobenius theorem for nonnegative matrices (2). The dominant eigenvector of $\boldsymbol{J}^*$ can be computed by solving

$$\boldsymbol{J}^*\begin{bmatrix}\boldsymbol{i}\\ \boldsymbol{b}\end{bmatrix} = \lambda\begin{bmatrix}\boldsymbol{i}\\ \boldsymbol{b}\end{bmatrix},$$

where λ is the dominant eigenvalue of $\boldsymbol{J}^*$, and $\boldsymbol{i}$ and $\boldsymbol{b}$ are the components of the dominant eigenvalue corresponding, respectively, to infectives and pathogens. Writing again $\boldsymbol{J}^*$ as

$$\boldsymbol{J}^* = \begin{bmatrix}\boldsymbol{A} & \boldsymbol{B}\\ \boldsymbol{C} & \boldsymbol{D}\end{bmatrix}$$

we get

$$\boldsymbol{A}\boldsymbol{i} + \boldsymbol{B}\boldsymbol{b} = \lambda\boldsymbol{i}$$

$$\boldsymbol{C}\boldsymbol{i} + \boldsymbol{D}\boldsymbol{b} = \lambda\boldsymbol{b}.$$

Because close to the transcritical bifurcation through which the disease-free equilibrium loses stability the dominant eigenvalue λ of $\boldsymbol{J}^*$ is equal to 0, from the first equation we have

$$\boldsymbol{i} = -\boldsymbol{A}^{-1}\boldsymbol{B}\boldsymbol{b};$$

therefore, the second equation can be written as

$$-\boldsymbol{C}\boldsymbol{A}^{-1}\boldsymbol{B}\boldsymbol{b} + \boldsymbol{D}\boldsymbol{b} = 0.$$

Because $\boldsymbol{A}$ is a diagonal matrix with equal diagonal entries, with simple algebraic manipulations we can write the previous equation as

$$(\boldsymbol{A}\boldsymbol{D} - \boldsymbol{C}\boldsymbol{B})\boldsymbol{b} = 0.$$

From the previous section we already know that

$$\boldsymbol{A}\boldsymbol{D} - \boldsymbol{C}\boldsymbol{B} = \phi(\mu_B + l)\left(\boldsymbol{U_n} - \boldsymbol{W}^{-1}\boldsymbol{G_0}\boldsymbol{W}\right)$$

and hence

$$\boldsymbol{G_0}\boldsymbol{W}\boldsymbol{b} = \boldsymbol{W}\boldsymbol{b}.$$

Therefore, we can conclude that, close to the transcritical bifurcation of the disease-free equilibrium, where the dominant eigenvalue Λ_0 of $\boldsymbol{G_0}$ is equal to 1, the dominant eigenvector $\boldsymbol{g_0}$ of matrix $\boldsymbol{G_0}$ corresponds to the pathogens' components of the dominant eigenvalue of $\boldsymbol{J}^*$ multiplied by the volumes of the corresponding water reservoirs ($\boldsymbol{g_0} = \boldsymbol{W}\boldsymbol{b}$). The infectives' components $\boldsymbol{i}$ of the Jacobian matrix can thus be computed as

$$\boldsymbol{i} = \frac{m_S\boldsymbol{H}\boldsymbol{Q}\boldsymbol{\beta} + (1+m_S)\boldsymbol{H}\boldsymbol{\beta}}{\phi}\boldsymbol{W}^{-1}\boldsymbol{g_0}$$

and they can be used to effectively portray the geography of disease onset. Note, however, that this simple relationship between

the dominant eigenvector of $\boldsymbol{G_0}$ and the infectives' components of the dominant eigenvector of $\boldsymbol{J}^*$ holds only close to the transcritical bifurcation of the disease-free equilibrium. In general, for parameter combinations for which the dominant eigenvalue of $\boldsymbol{G_0}$ is significantly larger than 1, the study of the geography of disease onset requires the computation of the eigenvalues and the eigenvectors of matrix $\boldsymbol{J}^*$.

Pathways for Pathogen Propagation. ***Hydrological transport.*** In the studies of the Haiti and Thukela epidemics as well as in the analysis of theoretical landscapes, the hydrological interconnections subsumed into matrix $\boldsymbol{P}$ are assumed to conform to the following mechanism. The fraction P_{ij} of pathogens that move between two nodes of the hydrological network (say from i to j) is given by

$$P_{ij} = \begin{cases} \dfrac{P_{\text{out}}}{d_{\text{out}}(i)P_{\text{out}} + d_{\text{in}}(i)P_{\text{in}}} & \text{if } i \to j \\ \dfrac{P_{\text{in}}}{d_{\text{out}}(i)P_{\text{out}} + d_{\text{in}}(i)P_{\text{in}}} & \text{if } i \leftarrow j \\ 0 & \text{if } i \nleftrightarrow j, \end{cases}$$

with $j \neq i$ ($P_{ii} = 0$). P_{out} (P_{in}) is the fraction of pathogens moving along an outward (inward) edge and d_{out} (d_{in}) is the out degree (in degree) of node i, that is, the number of outward (inward) edges. Note that the quantity P_{ij} can be derived from the discretization of the standard advection-dispersion equation for water flow or, equivalently, from a biased random-walk process on an oriented graph (7). The transport process is assumed to be conservative, i.e., $\sum_{j\in N_i} P_{ij} = 1$, where N_i is the set of neighbors connected to node i [of cardinality $d(i) = d_{\text{out}}(i) + d_{\text{in}}(i)$]. The bias of hydrological transport along the river (which is related to downstream velocity) can thus be defined as $b = P_{\text{out}} - P_{\text{in}} = 2P_{\text{out}} - 1$. Note also that matrix $\boldsymbol{P}$ must account for proper boundary conditions (BCs) for the leaves and the outlet of the river network. At the outlet (labeled as node 1), in particular, absorbing BCs can be used to characterize river basins with coastal regions where pathogens are not found in interepidemic periods, whereas reflecting BCs are better suited for cholera endemic areas where brackish water represents a reservoir for pathogens. To properly define BCs, we introduce a fictitious node 0 downstream of the network outlet (i.e., node 0 is connected with node 1 only), so that $\sum_{j\in N_1} P_{1j} = 1$, with N_1 including node 0. Absorbing BCs thus correspond to setting $P_{01} = 0$ and $P_{00} = 1$, whereas (purely) reflecting BCs can be obtained by imposing $P_{01} = P_{10}$ and $P_{00} = 1 - P_{10}$. In the same way we can define fictitious nodes upstream of the network leaves. All the numerical examples described in the main text are obtained with reflecting BCs for the leaves and absorbing BCs for the outlet of the river network, to mimic nonendemic settings.

Human mobility. Local communities are linked through human mobility, which is described by means of matrix $\boldsymbol{Q}$. We assume that the entries of $\boldsymbol{Q}$ can be estimated through a functional choice (8, 9) in which attractivity of a given destination site is supposed to be proportional to its size and decrease exponentially with its distance from the home site,

$$Q_{ij} = \frac{H_j \exp(-d_{ij}/D)}{\sum_{k\neq i}^{n} H_k \exp(-d_{ik}/D)},$$

where $j \neq i$ ($Q_{ii} = 0$). In the previous expression d_{ij} is the distance between node i and node j and D is the average travel distance. Although such a model is clearly not expected to fully capture the complexity of real human movement patterns, gravity-like models have been widely applied in the epidemiological literature to describe the impact of human mobility on the emergence of a suite of human diseases, including influenza, HIV, measles, and recently, cholera (10). Therefore we apply gravity models both in the analysis of waterborne disease epidemics spreading in theoretical landscapes (Peano networks) and in the study of real-word outbreaks (cholera epidemics in Haiti and KwaZulu-Natal, South Africa). Note also that we assume that the disease does not significantly impair human mobility ($m_S = m_I = m$). Extensions to different human mobility models based on small-world and scale-free graphs are briefly analyzed in the case of theoretical landscapes.

Modeling Specifications in Case Studies. ***Analysis of the Haiti epidemic.*** As a first case study, we have applied our theoretical framework for the definition of onset conditions for epidemics of waterborne disease to the cholera outbreak that struck Haiti in 2010 and is still ongoing (8, 11). Population distribution (Fig. S1*A*), river and mobility network structures (Fig. S1*B*), and parameter values (Table S1) have been borrowed from a recent study (11) based on a slightly modified version of the epidemiological model presented in *Materials and Methods* of the main text. The model used in that paper (11) does in fact account also for immunity loss along years (which can be obviously neglected in the study of onset conditions and in the analysis of the disease course immediately following the epidemic peak, as explained above) and for the role of rainfall as driver of increased water contamination. Specifically, regarding rainfall, Rinaldo and colleagues (11) assumed that the contamination rate can be expressed as

$$p_i(t) = p_0(1 + \psi R_i(t)),$$

where p_0 is the baseline contamination rate, $R_i(t)$ is precipitation intensity at site i and time t, and ψ is a suitable proportionality constant. Here, for consistency with the epidemiological model, we have dropped time dependency and assumed the following expression for rainfall-driven water contamination,

$$p_i = p_0\left(1 + \psi \overline{R}_i\right),$$

where $\overline{R}_i$ represents the average rainfall intensity recorded at site i from November 2010 to May 2011 [data from the National Aeronautics and Space Administration–Japan Aerospace Exploration Agency's Tropical Rainfall Measuring Mission (11)].

The computation of eigenvalues and eigenvectors has been performed at the scale of the river/mobility network nodes, which also represent the computational units for the simulations presented in the previous analysis (11). However, because epidemiological data and population distribution are, respectively, available at coarser (10 Haitian administrative departments; data available online through the website of Ministère de la Santé Publique et de la Population, Republique d'Haiti, http://www.mspp.gouv.ht) or finer (LandScan data at 1-kmd^2 pixel resolution, available online at the Oak Ridge National Laboratory website, http://www.ornl.gov/sci/landscan) spatial scales, we have suitably downscaled (or upscaled) incidence data and eigenvector components to these two spatial resolutions (results shown in Fig. 1 in the main text).

We have tested the sensitivity of Λ_0 to changes of single parameter values. To that end, for each model parameter (say θ) we computed

$$\theta' = \theta_{\boldsymbol{0}}(1 + \delta),$$

where $\theta_{\boldsymbol{0}}$ is the reference value of the parameter (reported in Table S1) and δ sets the scale of parameter variability ($-0.99 \leq \delta \leq 0.99$), and repeated the computation of the dominant eigenvalue of matrix $\boldsymbol{G_0}$ (Fig. 1*D* in the main text).

To test also for the robustness of the spatial patterns prescribed by the components of the dominant eigenvector of $\boldsymbol{G_0}$, we have performed a sensitivity analysis of the results presented in Fig. 1 *C* and *E* in the main text with respect to random variations of the parameter values. Specifically, for each parameter θ we independently computed a stochastic value θ' as

$$\theta' = \theta_0(1 + \xi\delta),$$

where ξ is a random variable drawn from a uniform distribution $\mathcal{U}(-1, 1)$ and δ sets the scale of parameter uncertainty. Eigenvalue and eigenvector computation was then performed with the randomized parameter values. For each value of δ ($0 \le \delta \le 0.99$) we have repeated 200 times the procedure just outlined and recorded mean and standard deviation of the distribution of coefficients of determination resulting from the comparison of predicted vs. observed epidemic spatial patterns (Fig. 1*F* in the main text).

Analysis of Thukela epidemic. As a second case study, we have reexamined the cholera outbreak that occurred in the KwaZulu-Natal (KZN) province of South Africa, specifically in the Thukela river basin, in 2000–2001 (10, 12). To this end, we have made use of (*i*) local epidemiological and demographic data, (*ii*) data about availability and distribution of drinking water resources and toilet facilities, and (*iii*) information on local hydrological networks. The relevant epidemiological data (*i*) were provided by the KZN Health Department (http://www.kznhealth.gov.za/). Data consist of a record of cholera cases including information about date and location (i.e., health subdistrict) of each hospitalized case (Fig. S2 *A* and *B*); the dataset also includes the record of local population size for each health subdistrict. Georeferenced data on availability of piped drinking water and improved toilet facilities (*ii*) have been retrieved from the Geographic Information System (GIS) set up for the 2001 South African census (http://www.statssa.gov.za; Fig. S2 *C* and *D*). Hydrological data about the river networks of the KZN province (*iii*) were derived from the GIS provided by the South African Department of Water Affairs and Forestry (http://www.dwaf.gov.za/).

Perennial rivers and channel endpoints have been considered as edges and nodes of the hydrological network, respectively (7). Demographic, epidemiological, hydrological, and sanitary data have thus been interpolated from health subdistricts to network nodes. Specifically, nearest-neighbor interpolation has been used, with distances being computed from subdistrict centroids. Interpolated population abundances and great-circle node-to-node distances have been used for the gravity model of population mobility as well (see above). Water resources were assumed to be related to local population size according to the relation $W_i = cH_i$, with c being a proportionality constant (12). Sanitary data have been used to impose a plausible spatial distribution for the exposure ($\boldsymbol{\beta}$) and contamination ($\boldsymbol{p}$) parameters (10). Specifically, we have assumed $\beta_i = \beta_m\omega_i$ ($p_i = p_m\tau_i$), where β_m (p_m) is the maximum exposure (contamination) rate and ω_i (τ_i) is the local fraction of households without access to piped water (improved toilet facilities).

Several model parameters have been drawn from the literature or from demographic/epidemiological data. In particular, the mortality rate of the population (μ) has been computed as the inverse of the average human lifetime in the KZN region [about 60 y (12)], and hence $\mu = 4.6 \times 10^{-5}$ (d^{-1}). We assumed (5, 12) that people can be exposed to contaminated water or food at most once a day in the worst-case scenario ($\omega = 1$); thus we set $\beta_m = 1.0$ (d^{-1}). The recovery rate γ from cholera can be evaluated as the inverse of the average duration of the disease in infected individuals, which is approximatively 5 d (5); therefore $\gamma = 0.20$ (d^{-1}). On the basis of the count of lethal cholera cases recorded during the Thukela-KZN epidemic, Bertuzzo and colleagues (12) estimated the value of the additional mortality rate due to cholera as $\alpha = 4.0 \times 10^{-4}$ (d^{-1}). The same study suggested the numerical value of the mortality rate μ_B of free-living vibrios in the Thukela basin; i.e., $\mu_B = 0.23$ (d^{-1}).

The remaining parameters (namely p_m, l, b, $m_S = m_I = m$, and D) could not be derived from literature data and had thus to be numerically tuned through proper techniques. Parameter estimation has been performed by combining extensive numerical simulations of the epidemiological model presented in *Materials and Methods* (with pathogen concentration suitably rescaled as $B_i^* = B_i/K$) with the exploration of the parameter space through Markov chain Monte Carlo (MCMC) sampling (13), implemented in the DREAM$_{ZS}$ algorithm (14, 15). The goodness of fit of each single simulation has been computed as the residual sum of squares between weekly hospitalized cholera cases per hydrological unit as evaluated from the KZN epidemiological record and from model simulations (namely assuming a hospitalized-to-infected ratio of 0.2). The MCMC algorithm has been initialized with broad flat prior distributions for parameter values and let run until convergence (about 15,000 iterations). Differently from the Haiti case study, here we wanted to test the predictive ability of our framework; therefore, only the first 120 d of epidemiological data have been used to tune model parameters. Parameter values are reported in Table S2.

Before evaluating the performance of the dominant eigenvector of matrix $\boldsymbol{G_0}$ associated with the Thukela epidemic, it is interesting to assess whether correlations exist between the georeferenced data available for this case study (i.e., spatial distributions of population $\boldsymbol{H}$; drinking water availability, measured by parameter $\boldsymbol{\beta}$; and sanitation conditions, measured by parameter $\boldsymbol{p}$) and the distribution of cholera cases. Because the spatial resolution of the model is relatively fine-grained (287 computational units for the Thukela basin), we have partitioned network nodes into clusters to increase the robustness of our spatial analyses. Clustering has been performed on the basis of the geographical positions of the nodes via the standard k-means algorithm (16). We have thus generated $k = 30$ clusters (average cluster radius $r \approx 10$ km), properly upscaled the georeferenced variables, and performed a linear correlation analysis. The results show that there is basically no correlation between the population distribution and the spatial pattern of cholera cases recorded during the onset of the epidemic ($R^2 = 0.02$, $P = 0.45$), whereas cholera cases are slightly correlated to the availability of drinking water ($R^2 = 0.22$, $P < 0.01$) and sanitation conditions ($R^2 = 0.20$, $P < 0.05$).

Because of the availability of georeferenced data, it is also possible to determine the spatial distribution of the local basic reproduction numbers of the disease (R_{0i}). Although cholera cases are obviously correlated to this quantity ($R^2 = 0.30$, $P < 0.01$), contrasting the map of local basic reproduction numbers to that of cholera cases (Fig. S3) reveals that the R_{0i} cannot be considered a satisfactory measure of the likelihood of local outbreaks. As a matter of fact, data show that there are regions where the disease emerged even if the local basic reproduction number is less than 1, whereas in other regions the disease hardly proliferated despite R_{0i} being larger than 1. These observations strongly support the importance of accounting for spatial dynamics (and, in particular, for connectivity structures) in the definition of onset conditions for waterborne diseases. The dominant eigenvector of matrix $\boldsymbol{G_0}$ does indeed outperform the value of R_{0i} as a predictive tool for the localization of disease cases during the onset of the epidemic ($R^2 = 0.86$, $P < 0.001$).

Quantitative results do obviously depend on the details of spatial clustering. However, we have repeated eigenvector computations and the related spatial analyses for different numbers of clusters (namely $k = 15$, $r \approx 15$ km; $k = 60$, $r \approx 5$ km)

and found no significant qualitative differences with respect to the results shown here and in the main text for $k = 30$ (Figs. S4 and S5).

Disease Spread in Theoretical Metapopulations. We have also studied theoretical landscapes, namely characterized by a Peano-like hydrological network. The Peano construct is generated iteratively from a basic prefractal, a cross seeded in a corner of the square domain, and develops into an iconic exactly self-similar structure spanning the entire plane [as iterations tend to infinity, the total network length behaves like an area (17)]. Iterations consist of cutting in half each branch to reproduce the prefractal on four equal subbasins. Peano's topological measures (like Horton's numbers and the Tokunaga cyclicity) match closely those of real river networks (18), but fail to satisfy the statistics of aggregation and upstream/downstream distances. However, geometrical constraints imposed by a fractal network on transport of species, populations, or pathogens [known to imply strong corrections on the speed of traveling waves (19)] chiefly depend on topological features because transmission fronts are dominantly affected by the bifurcation structure encountered along the backbone of the network (7). Hence topology, rather than the fine structure of the subpaths, dominates the process, making a case for the generalized use of Peano's network, also in view of its exact solvability.

The study of a theoretical landscape requires some assumptions on the spatial distribution of the population. The simplest choice is a uniform distribution ($H_i = h\ \forall i$). The results shown in Fig. 3 *A* and *B* in the main text (and in Fig. S6 *A* and *B* for different parameter settings) refer to this assumption. However, a uniform population distribution may represent a somewhat crude simplification of the observed spatial arrangement of human communities. Empirical observations show in fact that a much more realistic model for the size of human settlements is given by the so-called Zipf's law (20, 21), according to which the size distribution of human communities can be well represented by a power-law distribution, namely by $\text{prob}(H_i) \propto H_i^{-2}$. Therefore, we have also performed our analysis by sampling the size of each local community from a power-law distribution. To allow comparison with the results of the homogeneous case, we have imposed the normalizing constraint $\sum_{i=1}^{n} H_i = nh$. The results shown in Fig. 3 *C* and *D* in the main text (as well as in Fig. S6 *C* and *D*) have been obtained with a Zipf-like population distribution. Note that, in this case, for each analyzed parameter setting we have extracted 500 independent realizations of sample size n from the population distribution and we have distributed the n population sizes randomly in the landscape. Colors in Fig. 3 *C* and *D* and Fig. S6 *C* and *D* thus code the fraction of different realizations for which onset conditions are met.

Our theoretical framework allows for the description of the geography of disease onset in theoretical landscapes as well. In particular, the predictive ability of the dominant eigenvector of matrix $\boldsymbol{G_0}$ can be tested against numerical simulations of the epidemiological model implemented on a Peano-like network. To initialize the simulation, we have assumed that the outbreak starts close to the outlet of the river network (bottom-right corner of the embedding domain), as is often observed in real-world epidemics (10, 12), with 1‰ of the local population representing the initial infective pool. Fig. S7 shows the dominant eigenvector of $\boldsymbol{G_0}$ and the simulation of epidemic onset in a Peano network with either homogeneous (Fig. S7 *A–D*) or Zipf-like (Fig. S7 *E–H*) population distribution. Note that in these theoretical examples it is possible to detect the "emergent" phase of the outbreak, i.e., the occurrence of the very first cases of the disease—something that cannot be usually done in real-world applications (such as the Haiti and Thukela cholera epidemics analyzed here). Specifically, disease emergence has been identified numerically from model simulations as the week after which daily incidence and its first and second time derivatives exceed 5‰ of the respective maximum values recorded in the simulation.

Not only different population distributions (Fig. S7), but also different models of human mobility can be applied and analyzed in our framework. Small-world graphs (22, 23) represent an alternative connectivity model to the gravity-like approach used elsewhere in the main text. Starting from an adjacency structure characterized by local connections (like, e.g., in nearest-neighbor coupling), a possible way to create a small-world network is that of randomly rewiring some existing edges and thus introducing long-range shortcuts whose effect is that of significantly decreasing the average path lengths (small-world effect). Specifically, each link $i \leftrightarrow j$ of the preexisting connectivity graph is removed with a probability r and a new link $i \leftrightarrow k$ is created. Node k can be randomly chosen either from a uniform distribution or via some other ad hoc procedure accounting, for instance, for preferential attachment. In the latter case the probability with which node k is chosen is a function of its degree. Note that both procedures lead to the formation of small-world networks and that preferential attachment also leads (at least in the limit of infinite networks) to the formation of scale-free networks, i.e., networks with power-law degree distributions (23, 24). Once connectivity structure has been defined, mobility fluxes can be then computed either according to a gravity model or assuming that the strength of the connections is independent of travel distance, so that if a node i is connected to k other nodes, then $Q_{ij} = 1/k$ for each of its neighbors [the so-called "propagule rain" scenario (25)]. Here we have first defined a local connectivity structure by linking each human community to its nearest Peano neighbors; then, we have performed either random rewiring or random rewiring with preferential attachment; in both cases we have applied spatially homogeneous distributions for population density and local basic reproduction numbers and assumed the propagule rain scenario for population fluxes; finally, we have evaluated onset conditions as a function of model parameters, performing several realizations of the rewiring processes to account for the stochasticity involved in the generation algorithms (200 independent realizations for each parameter setting). Results are reported in Fig. S8 and show that small-world connectivity can greatly favor the emergence of subthreshold epidemics, especially in the presence of preferential attachment.

1. Farina L, Rinaldi S (2000) *Positive Linear Systems: Theory and Applications* (Wiley Interscience, New York).
2. Gantmacher F (1959) *Theory of Matrices* (AMS Chelsea Publishing, Providence, RI).
3. Kuznetsov Y (1995) *Elements of Applied Bifurcation Theory* (Springer, New York).
4. Silvester J (2000) Determinants of block matrices. *Math Gazette* 84:460–467.
5. Codeço CT (2001) Endemic and epidemic dynamics of cholera: The role of the aquatic reservoir. *BMC Infect Dis* 1:1.
6. Anderson R, May R (1991) *Infectious Diseases of Humans: Dynamics and Control* (Oxford Univ Press, Oxford).
7. Bertuzzo E, Casagrandi R, Gatto M, Rodriguez-Iturbe I, Rinaldo A (2010) On spatially explicit models of cholera epidemics. *J R Soc Interface* 7(43):321–333.
8. Bertuzzo E, et al. (2011) Prediction of the spatial evolution and effects of control measures for the unfolding Haiti cholera outbreak. *Geophys Res Lett* 38:L06403.
9. Mari L, et al. (2012) On the role of human mobility in the spread of cholera epidemics: Towards an epidemiological movement ecology. *Ecohydrology* 5(5):531–540.
10. Mari L, et al. (2012) Modelling cholera epidemics: The role of waterways, human mobility and sanitation. *J R Soc Interface* 9(67):376–388.
11. Rinaldo A, et al. (2012) Reassessment of the 2010–2011 Haiti cholera outbreak and multi-season projections via inclusion of rainfall and waning immunity. *Proc Natl Acad Sci USA* 109:6602–6607.
12. Bertuzzo E, et al. (2008) On the space-time evolution of a cholera epidemic. *Water Resour Res* 44:W01424.

13. Gilks WR, Richardson S, Spiegelharter DJ (1995) *Markov Chain Monte Carlo in Practice* (Chapman & Hall, New York).
14. ter Braak C, Vrugt J (2008) Differential evolution Markov chain with snooker updater and fewer chains. *Stat Comput* 18:435–446.
15. Vrugt J, et al. (2009) Accelerating Markov chain Monte Carlo simulation by differential evolution with self-adaptive randomized subspace sampling. *Int J Nonlin Sci Numer Simul* 10:273–290.
16. Lloyd S (1982) Least square quantization in pcm. *IEEE Trans Inf Theory* 28:129–137.
17. Mandelbrot B (1983) *The Fractal Geometry of Nature* (Freeman, New York).
18. Rodriguez-Iturbe I, Rinaldo A (1997) *Fractal River Basins: Chance and Self-Organization* (Cambridge Univ Press, New York).
19. Bertuzzo E, Maritan A, Gatto M, Rodriguez-Iturbe I, Rinaldo A (2007) River networks and ecological corridors: Reactive transport on fractals, migration fronts, hydrochory. *Water Resour Res* 43:W04419.
20. Zipf GK (1949) *Human Behavior and the Principle of Least Effort: An Introduction to Human Ecology* (Addison-Wesley, Cambridge, MA).
21. Newman MEJ (2005) Power laws, Pareto distributions and Zipf's law. *Contemp Phys* 46:323–351.
22. Watts DJ, Strogatz SH (1998) Collective dynamics of 'small-world' networks. *Nature* 393(6684):440–442.
23. Albert R, Barabási A-L (2002) Statistical mechanics of complex networks. *Rev Mod Phys* 74:4797.
24. Barabási AL, Albert R (1999) Emergence of scaling in random networks. *Science* 286 (5439):509–512.
25. Gotelli N (1991) Metapopulation models: The rescue effects, the propagule rain, and the core-satellite hypothesis. *Am Nat* 138:768–776.

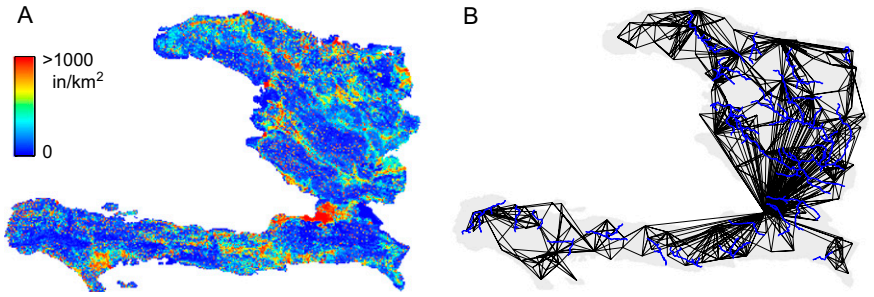

Fig. S1. Population distribution, hydrological connectivity, and human mobility network in Haiti. (*A*) Spatial distribution of population density obtained by remote sensing (LandScan data by Oak Ridge National Laboratory), which is translated into a georeferenced spatial distribution of nodes *i* endowed with population H_i (11). (*B*) Main Haitian rivers (blue) and a relevant subset of the network of human mobility (black), here portrayed synthetically by the four largest outbound connections for each node.

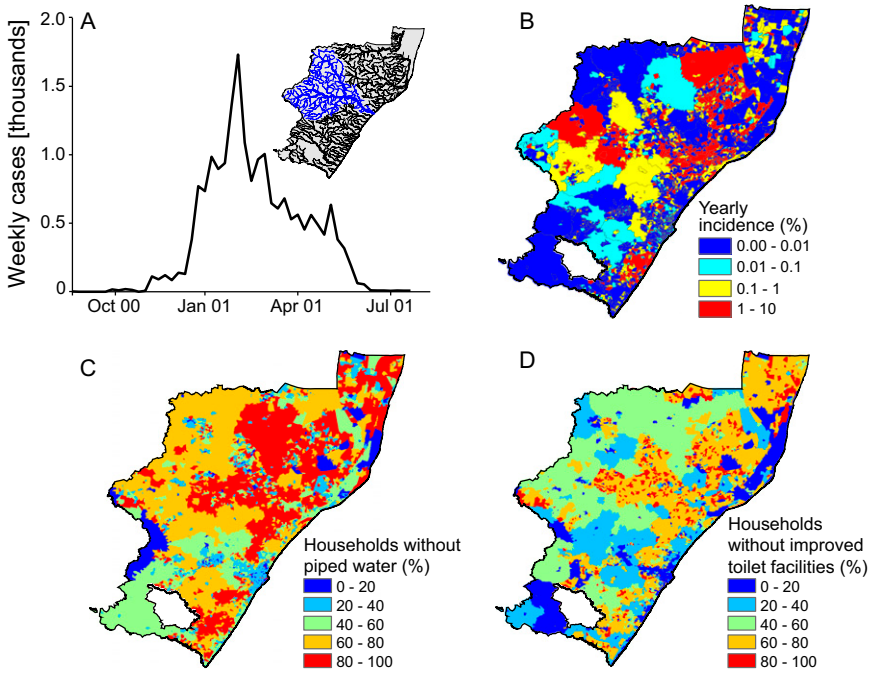

Fig. S2. Epidemiological, demographic, and sanitation data for the Thukela-KZN cholera epidemic. (*A*) Total incidence data (weekly cases) recorded in the Thukela river network from October 2000 to July 2001. *Inset* shows Thukela river network. (*B*) Yearly cholera incidence, evaluated as number of reported cases in each health district from October 2000 to September 2001 divided by population size. (*C*) Spatial distribution of households without access to piped water. (*D*) Spatial distribution of households without access to improved toilet facilities.

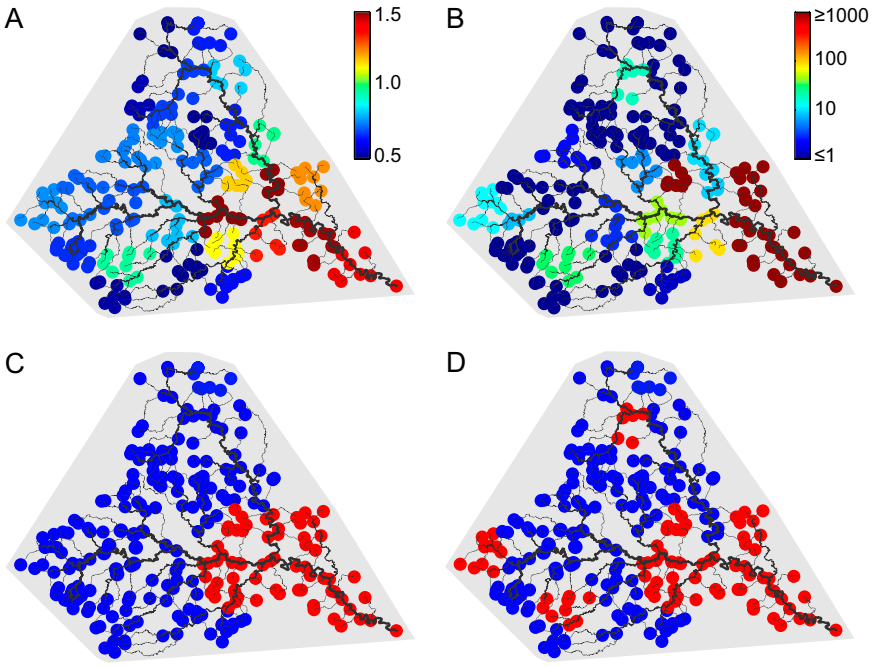

Fig. S3. (*A* and *B*) Spatial distribution of local basic reproduction numbers (*A*) and cholera cases reported during the onset phase of the Thukela epidemic (*B*). For a better visual comparison, *C* and *D* show, respectively, the localization of regions with $R_{0i} > 1$ and more than 10 cholera cases (red).

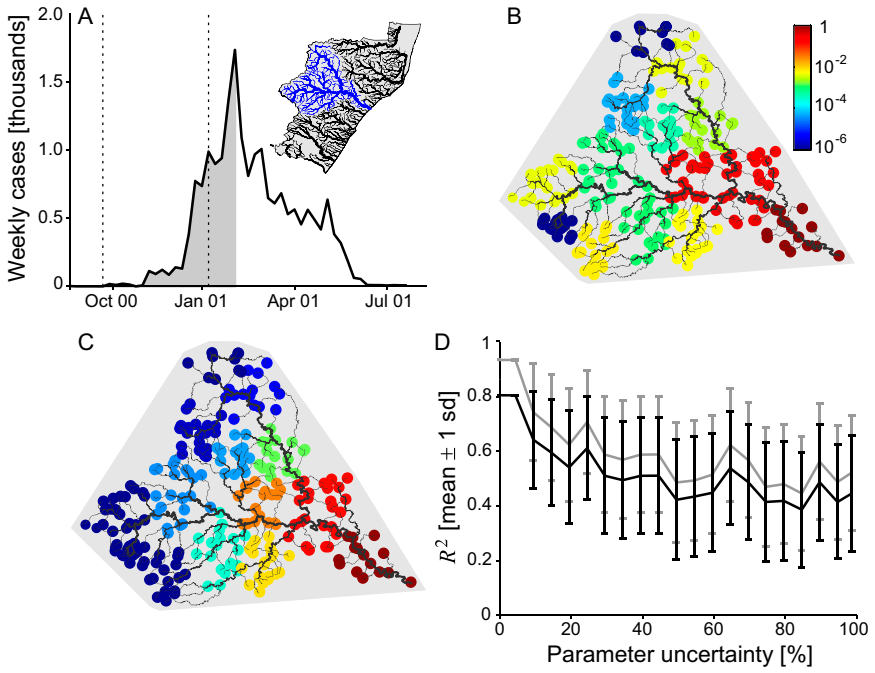

Fig. S4. (*A–D*) Data and model predictions of Thukela epidemic, as in Fig. 2 *A–C* and *F* in the main text with $k = 15$ clusters (instead of $k = 30$).

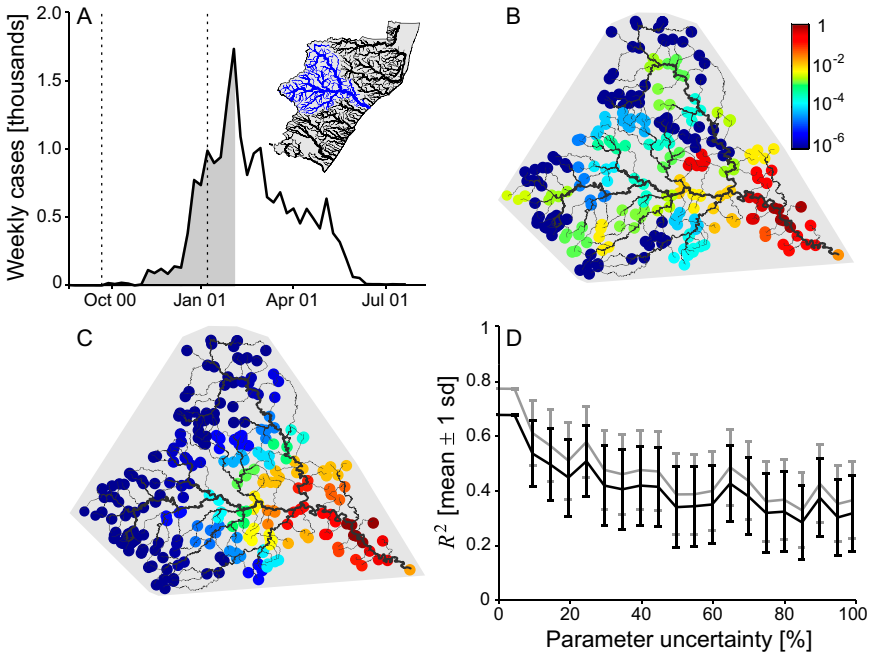

Fig. S5. Data and model predictions of Thukela epidemic, as in Fig. 2 *A–C* and *F* in the main text with $k = 60$ clusters (instead of $k = 30$).

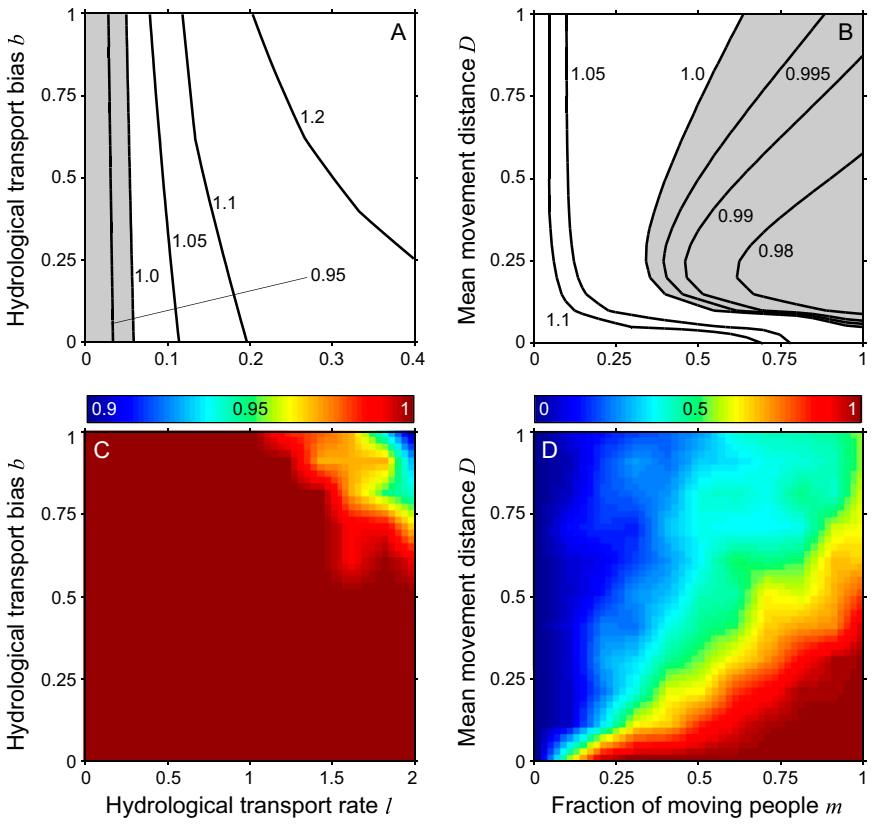

Fig. S6. Onset conditions of a waterborne disease epidemic in a Peano network with gravity-like mobility, as in Fig. 3 in the main text with $m_S = m_I = 0.5$ (*A* and *C*), $D = 0.01$ (*A* and *C*), $l = 1$ (*B* and *D*), and $b = 0.5$ (*B* and *D*).

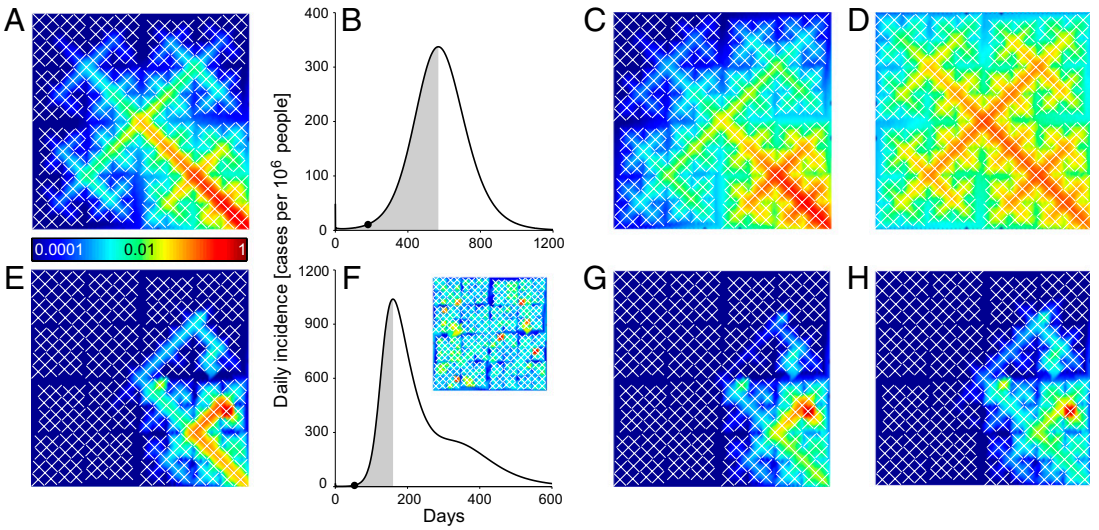

Fig. S7. Dominant eigenvector of $\mathbf{G_0}$ and simulation of epidemic onset in a Peano network with homogeneous (*A–D*) or Zipf-like (*E–H*) population distribution. (*A*) Dominant eigenvector (associated to $\Lambda_0 = 1.001$). (*B*) Simulated temporal pattern of daily incidence. (*C*) Simulated spatial distribution of weekly cases during the emerging phase of the outbreak (black dot in *B*). (*D*) Simulated spatial distribution of cumulative cases up to disease peak (gray shading in *B*). (*E–H*) As in *A–D*, with Zipf-like population distribution (*F*, *Inset*). The dominant eigenvalue is $\Lambda_0 = 1.11$. Peano construct is here used at the fifth stage of iteration (*SI Text*). Network outlet is located in the lower-right corner of the spatial domain. Parameter values: $\beta = 1$, $\alpha = \mu = 0$, $\mu_B = 0.23$, $l = 1$, $b = 0.5$, $m_S = m_I = 0.125$, $D = 0.05$, $R_{0i} = 1.2$.

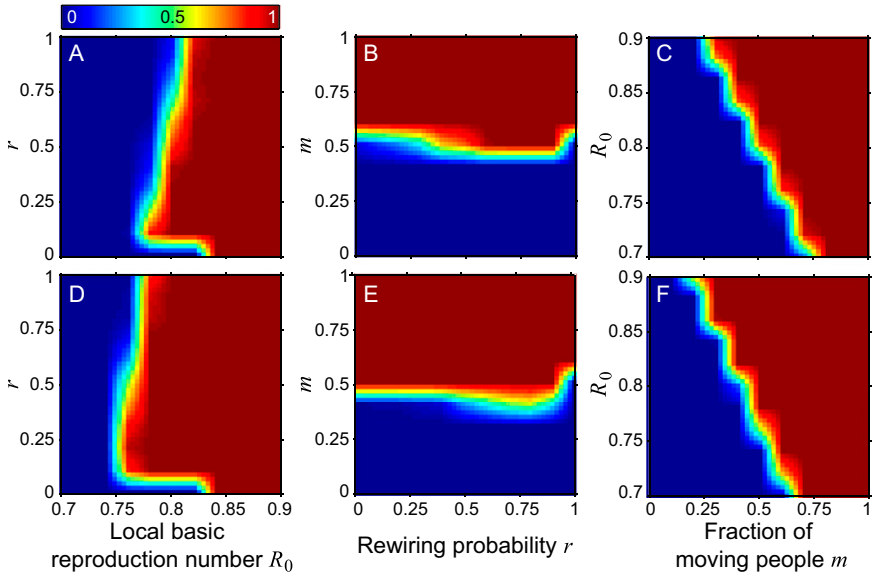

Fig. S8. Epidemic onset conditions in a Peano network with homogeneous population distribution, hydrological transport, and different models of human mobility. (*A–C*) Random rewiring. (*D–F*) Random rewiring with preferential attachment (details in main text). The color scale indicates the fraction of realizations in which the onset condition based on the dominant eigenvalue of $\mathbf{G_0}$ is met. Parameter values: $\beta = 1$, $\alpha = \mu = 0$, $\mu_B = 0.23$, $l = 0.5$, $b = 0.2$, $m_S = m_I = 0.5$ (*A* and *D*); $R_{0i} = 0.8$ (*B* and *E*); and $r = 0.5$ (*C* and *F*).

Table S1. Parameter values for the Haiti cholera epidemic

Parameter	Units	Value
μ	d^{-1}	4.5×10^{-5}
β	d^{-1}	1.0
γ	d^{-1}	0.20
α	d^{-1}	4.0×10^{-3}
μ_B	d^{-1}	0.20
$p_0/(Kc)$	—	0.14
ψ	$d \cdot mm^{-1}$	4.8×10^{-2}
l	d^{-1}	1.8
b	—	1.0
m	—	0.69
D	km	100

Table S2. Estimated parameter values for the Thukela cholera epidemic

Parameter	Units	Value
$p_m/(Kc)$	—	0.11
l	d^{-1}	0.10
b	—	0.04
m	—	0.27
D	km	3.1