Supporting Information

Reiner et al. 10.1073/pnas.1108438109

S1. Data Transformation

The original data consist of the number of cases per month per thana divided by the population of that thana. The population of each thana is estimated monthly by fitting exponential growth to the three relevant decadal censuses (1981, 1991, 2001). As mentioned in the main text, the primary analyses were conducted on the 14 y (1995–2008) of data for which only the El Tor biotype of cholera in Dhaka was present. Several of the thanas have split during this time so for continuity we use the thana boundaries as they were in 1992 and aggregate cases and population appropriately.

To use a finite-state Markov chain model, the data need to be categorized into a set of discrete variables. To form the most natural discretization possible, we binned all 0s together and denoted this state as group "0" (48.86% of the data fall into this category). We then evenly split the remaining 54.14% into two groups: a "low" group where the rate of cases per 10,000 was between 0 and 1.85 (which we denote as group "1"), and a "high" group where the rate of cases per 10,000 was >1.85 (which we denote as group "2").

Although <3% of all data were >9 per 10,000, there were four data points >20 per 10,000 (with the maximum being 40.7 per 10,000), most of them during the extreme epidemic of Fall 1998. Thus, the discretization causes the elimination of the events in the tail of the distribution of epidemic size and merges these events with less pronounced ones.

Although the transformation does remove the tail events, it does maintain the ordering of the data as can be seen in Fig. S1. The red line is the number of cases per 10,000 averaged across all of Dhaka, and the blue line is the average of the 21 thanas after the transformation, when each thana is represented by the average of its corresponding group (4.58 cases per 10,000 for the high state, and 1.11 cases per 10,000 for the low state). If we consider the ranks of each month (i.e., which month had the highest rate of cases, which had the second highest, etc.), we can see that the rankings are almost identical for the true data (Fig. S2, red line) and the transformed data (Fig. S2, blue line). The months that constitute the highest 10% of the true data correspond almost perfectly to the months that constitute the highest 10% of the transformed data, and the same can be said for the top 15%, 20%, and so on. It can also be seen by inspection that interannual variability of the two time series exhibits a strong correspondence in the timing and frequency of its cycles. Because we are interested in understanding this variability and predicting whether there will or will not be an outbreak (and not the exact number of cases for a given month), the transformed data give us essentially the same information as the true data.

S2. Model Description

The model assumes that the cholera state of each thana will change stochastically from one month to the next according to transition probabilities that depend on the current cholera state of the thana, the cholera states of its neighbors, the season, and the value of the climate covariates [El Niño southern oscillation (ENSO) and/or flooding]. Here, we give a full mathematical specification of the model via precise definitions of the transition probabilities. In the following, the state 0 corresponds to no cholera; 1, to low cholera; and 2, to high cholera.

Climate-Independent Model. We first define a model that is not influenced by ENSO or flooding. Let $X_{m,t}$ be the cholera state of thana m at time t and $\mathcal{N}(k)$ be the set of thanas neighboring thana k. Let $p_{i,j,k,t}$ be the probability that thana k goes from state i at time $t-1$ to state j at time t. We postulate that

$$p_{i,0,k,t} = \mathbb{P}_{i,0,\mathcal{D}(k)} \times \mathrm{Neigh}(i,0,\mathcal{V}(k,t-1),\mathcal{D}(k)) \times \mathrm{Seas}(i,0,t-1,\mathcal{D}(k)) \qquad \textbf{[S1]}$$

$$p_{i,2,k,t} = \mathbb{P}_{i,2,\mathcal{D}(k)} \times \mathrm{Neigh}(i,2,\mathcal{V}(k,t-1),\mathcal{D}(k)) \times \mathrm{Seas}(i,2,t-1,\mathcal{D}(k)) \qquad \textbf{[S2]}$$

$$p_{i,1,k,t} = 1 - p_{i,0,k,t} - p_{i,2,k,t}, \qquad \textbf{[S3]}$$

where $\mathcal{V}(k,t) = \max_{m\in\mathcal{N}(k)} X_{m,t}$ is the worst cholera state among the thanas neighboring thana k at time t, $\mathcal{D}(k)$ indicates whether thana k is in the core or the periphery, and the 12 constant parameters $\mathbb{P}_{i,j,d}$ represent baseline transition probabilities of moving from state i to state j for any thana in region d.

The neighborhood function is multiplicative in its effect and has the following form:

$$\mathrm{Neigh}(i,j,v,d) = \left(1+\alpha_{i,j,d}\right)^{v}, \; j = 0,2. \qquad \textbf{[S4]}$$

The 12 coefficients $\alpha_{i,j,d}$ are parameters to be estimated.

The seasonality function also enters multiplicatively. It takes the form

$$\mathrm{Seas}(i,j,t,d) = \left(1+\beta_{i,j,d}\right)^{Se(t,d)}, \; j = 0,2, \qquad \textbf{[S5]}$$

where $\mathrm{Se}(t,d)$ is periodic in t with a period of 12 mo. We impose the constraints $\mathrm{Se}(2,d) = 0$ and $\mathrm{Se}(5,d) = 1$. There are thus 32 parameters associated with seasonality (Seas). The 20 seasonality parameters $\mathrm{Se}(t,d)$ are constrained only to be positive. However, it so happened that all estimated values were also <1. This result confirms that the month of May ($t = 5$) has the largest seasonal effect and February ($t = 2$) the smallest. The fitted values are shown in Fig. S3.

Climate-Dependent Model. In the climate-dependent model, the climate driver is assumed to increase the probability of transition to state 2 (high cholera). Let $p'_{i,j,k,t}$ be the $i \to j$ transition probability for thana k at time t. We postulate that

$$p'_{i,2,k,t} = f\left(p_{i,2,k,t} \times \mathrm{Nino}(t-1,\mathcal{D}(k))\right), \qquad \textbf{[S6]}$$

where the El Niño function has the sigmoidal form

$$\mathrm{Nino}(t,d) = 1 + A_d \frac{\tan\left(\frac{h_d}{2}\cdot\frac{\mathrm{ENSO}(t-10)}{M_d}\right)}{\tan\left(\frac{h_d}{2}\right)} \qquad \textbf{[S7]}$$

and $\mathrm{ENSO}(t)$ is the value of the ENSO anomalies index at time t. There are thus six parameters associated with Nino. Fig. S4 illustrates the flexibility of this function. Because the ENSO index is standardized, only ENSO anomalies affect the transition probabilities. The cutoff function $f(x) = \min(1, \max(0, x))$ ensures that $p'_{i,2,k,t}$ lies between 0 and 1.

To enforce the total-probability constraint, we adjust the remaining two transition probabilities proportionate to their values:

$$\begin{aligned} p'_{i,0,k,t} &= \left(1-p'_{i,2,k,t}\right)\frac{p_{i,0,k,t}}{p_{i,0,k,t}+p_{i,1,k,t}} \\ p'_{i,1,k,t} &= \left(1-p'_{i,2,k,t}\right)\frac{p_{i,1,k,t}}{p_{i,0,k,t}+p_{i,1,k,t}}. \end{aligned} \qquad \textbf{[S8]}$$

Alternative Parameterization. To investigate the robustness of our statistical results to parameterization assumptions, we repeated all analyses using a different parameterization. Specifically, in the climate-independent model, we exchanged Eqs. **S1** and **S2**. That is, neighbor and seasonal effects entered directly into the expressions for the probabilities of transition to states 1 and 2; the expression for the remaining probability was taken to be the complement. Because the expressions $p_{i,j,k,t}$ enter into the climate-dependent Eqs. **S7** and **S8**, this method effectively reparameterizes the climate-dependent model as well.

The results of this reparameterization were much the same, indicating that the results we found are robust to our choice of parameterization. In particular, exactly the same effects were statistically significant. Perhaps interestingly, under the alternative parameterization, the P value for the spatially specific effect of ENSO decreased from 0.0202 to 0.0025 and the overall log-likelihood increased slightly but not significantly (from −3,156.6 to −3,154.69). The results of these hypothesis tests are summarized in Table S2.

S3. Model Fitting

We fit the models by maximizing the likelihood. In the full model, we must therefore solve an optimization problem over 62 parameters. Under the Markovian assumption of our model, the transition from one month to the next is independent of all other transitions. Therefore, our likelihood is just the product of the likelihoods of each month's transition. As an optimization problem, we could use more complex techniques such as MCMC to identify the maxima. Continued advancements in this field (ref. 1 for example) make MCMC procedures a reasonable alternative, especially if a model with many more parameters were fit. However, implementation of MCMC approaches requires considerable knowledge of statistical computing, whereas the older, more straightforward Nelder–Mead simplex algorithm for maximizing the likelihood (2) works well for our purposes. More specifically, we use the Nelder–Mead simplex algorithm instead of more complicated algorithms for two reasons: first, because we are dealing with transition probabilities, the parameter space has numerous linear constraints. As such, the likelihood function has numerous discontinuities (and is thus not everywhere differentiable). Second, and more importantly, the Nelder–Mead algorithm has the property that if every point on the initial simplex satisfies a linear constraint, then every proposed point will as well. Thus, the algorithm enforces the linear constraints on the $\mathbb{P}$s automatically. Because this algorithm can ensure the identification only of local maxima, the algorithm is run several hundred times from different initial starting points to identify the global maxima.

We enforce the constraint that each probability must be between 0 and 1 by the "barrier method": If ever any probability is <0 or >1, the likelihood is taken to be negative infinity. We assume that each thana's transition is conditionally independent of each other thana's transition (conditioned on the maximum state of the nearest neighbors of the thana) and as such the likelihood of a month's transition is the product of each thanas' transition's likelihood.

For the statistical analysis, we use all 14 y of data to identify the "best" model as well as to conduct statistical tests to compare models. Because in every comparison of models that addresses a specific hypothesis, we considered nested models (i.e., one of the models can be rewritten as the other model with several of the parameters set at particular values, either equal to each other or equal to some constant), we are able to use straightforward likelihood-ratio tests. In many cases, the likelihood-ratio test statistic was extremely high and correspondingly, the associated P value was extremely low (e.g., in the case of testing the significance of the two regions, the P value was 1.09×10^{-26}).

To evaluate the model's forecasting ability we used cross-validation so that the actual observed values for a given month did not inform the model that would be used to predict that month. In particular, we fit 14 different models, leaving a different year of data out for each one. To predict any given year, we used the model fit to data exclusive of that year.

S4. Statistical Analysis

Each of the effects that have been presented in the model were statistically significant at a 5% level. For each main effect, we tested its significance in the presence of all other main effects as well as all interaction effects that did not contain it. For example, when testing the significance of ENSO, we evaluate how much its addition reduces the log-likelihood when added to a model that has all effects except both ENSO and the interaction between ENSO and space. When testing the effect of space at a large scale, because this effect is part of every interaction effect, we evaluated the increase in the log-likelihood between a model with only all other main effects and one with the transition probabilities split into core and periphery. For interaction effects, we evaluated their significance with respect to a model that had all other effects, both main and interaction. In particular, below we state, for each effect, the corresponding hypotheses. Any parameter not mentioned in any hypothesis is assumed to vary freely in both the null and the alternative models (except those of an interaction effect when testing a main effect as explained above).

Space (large scale):
  H_0: For each $(i, j) \in \{0, 1, 2\}^2$, $\mathbb{P}_{i,j,\mathcal{D}(k)} = \mathbb{P}_{i,j,\mathcal{D}(l)}$ for all thanas k, l.
  H_a: For at least one $(i, j) \in \{0, 1, 2\}^2$, $\mathbb{P}_{i,j,\mathcal{D}(k)}$ is different for thanas in different regions.
ENSO:
  H_0: $A = 0$.
  H_a: $A \neq 0$.
Seasonality:
  H_0: $\beta_{i,j} = 0$ for all i, j.
  H_a: At least one $\beta_{i,j} \neq 0$.
Space (small scale):
  H_0: $\alpha_{i,j} = 0$ for all i, j.
  H_a: At least one $\alpha_{i,j} \neq 0$.
Space (large scale) × ENSO:
  H_0: $A_{\mathcal{D}(k)} = A_{\mathcal{D}(l)}$, $h_{\mathcal{D}(k)} = h_{\mathcal{D}(l)}$ and $M_{\mathcal{D}(k)} = M_{\mathcal{D}(l)}$ for all thanas k, l.
  H_a: At least one of A, k, or M is different for thanas in different regions.
Space (large scale) × seasonality:
  H_0: For each $(i, j) \in \{0, 1, 2\}^2$, $\beta_{i,j,\mathcal{D}(k)} = \beta_{i,j,\mathcal{D}(l)}$ for all thanas k, l and $\mathrm{Se}(t, \mathcal{D}(k)) = \mathrm{Se}(t, \mathcal{D}(l))$ for all thanas k, l.
  H_a: For at least one $(i, j) \in \{0, 1, 2\}^2$, $\beta_{i,j,\mathcal{D}(k)}$ is different for thanas in different regions or for at least one $t \in \{0, 1, \ldots, 11\}$, $\mathrm{Se}(t, \mathcal{D}(k))$ is different for thanas in different regions.
Space (large scale) × space (small scale):
  H_0: For each $(i, j) \in \{0, 1, 2\}^2$, $\alpha_{i,j,\mathcal{D}(k)} = \alpha_{i,j,\mathcal{D}(l)}$ for all thanas k, l.
  H_a: For at least one $(i, j) \in \{0, 1, 2\}^2$, $\alpha_{i,j,\mathcal{D}(k)}$ is different for thanas in different regions.

The results of these hypothesis tests are summarized in Tables S1 and S2.

Below we present the 62 parameter values of our final model (with log-likelihood −3,156.6). Recall that there are 62 parameters and not 68 because in each transition matrix, due to the linear constraint, one column is determined on the basis of knowledge of the other two. Although our above statistical analysis is not based on each parameter being significant but on the effect governed by many parameters, we can see from the fitted values in the transition matrices that the base model captures the fact that the core thanas are more likely to experience high levels of cholera.

Core Thanas and Periphery Thanas.

$$\mathbb{P} = \begin{pmatrix} 0.826611405 & 0.115724566 & 0.057664029 \\ 0.567880090 & 0.325126091 & 0.106993819 \\ 0.577071169 & 0.211427120 & 0.211501711 \end{pmatrix}$$

$$\alpha = \begin{pmatrix} -0.057805277 & 0.24113984 \\ 0.073239230 & 0.05441645 \\ 0.001489096 & -0.01785350 \end{pmatrix}$$

$$\beta = \begin{pmatrix} -0.09970336 & 0.3232735 \\ -0.19332893 & 0.3378519 \\ -0.23792711 & 0.2901183 \end{pmatrix}$$

for ENSO

$$A = 0.7423721$$

$$h = 2.70762$$

$$M = 2.620335$$

and

$$\begin{array}{lll} \mathrm{Se}(1) = 0.0929788 & \mathrm{Se}(5) = 1 & \mathrm{Se}(9) = 0.9157895 \\ \mathrm{Se}(2) = 0 & \mathrm{Se}(6) = 0.7704870 & \mathrm{Se}(10) = 0.9631579 \\ \mathrm{Se}(3) = 0.3123720 & \mathrm{Se}(7) = 0.5872501 & \mathrm{Se}(11) = 0.5468458 \\ \mathrm{Se}(4) = 0.992745 & \mathrm{Se}(8) = 0.6910358 & \mathrm{Se}(12) = 0.366960. \end{array}$$

Periphery Thanas.

$$\mathbb{P} = \begin{pmatrix} 0.9087930 & 0.058051858 & 0.03315510 \\ 0.8440190 & 0.114040903 & 0.04194010 \\ 0.8498659 & 0.001530536 & 0.14860353 \end{pmatrix}$$

$$\alpha = \begin{pmatrix} -0.058722122 & 0.06863430 \\ -0.165737926 & 0.08092761 \\ -0.117189265 & 0.07731697 \end{pmatrix}$$

$$\beta = \begin{pmatrix} -0.10504627 & 0.2821137 \\ -0.09191232 & 0.1951380 \\ -0.14225600 & 0.1427373 \end{pmatrix}$$

for ENSO

$$A = 0.9964250$$

$$h = 0.06254895$$

$$M = 2.634224$$

and

$$\begin{array}{lll} \mathrm{Se}(1) = 0.3113539 & \mathrm{Se}(5) = 1 & \mathrm{Se}(9) = 0.5845409 \\ \mathrm{Se}(2) = 0 & \mathrm{Se}(6) = 0.3790798 & \mathrm{Se}(10) = 0.6583411 \\ \mathrm{Se}(3) = 0.3447757 & \mathrm{Se}(7) = 0.0978109 & \mathrm{Se}(11) = 0.6023155 \\ \mathrm{Se}(4) = 0.8887331 & \mathrm{Se}(8) = 0.2068688 & \mathrm{Se}(12) = 0.9285714. \end{array}$$

Because we also fit an alternative model to investigate the robustness of our conclusions, we can compare the resulting transition probabilities from the two different models. For the purposes of illustration, we chose to evaluate the third transition probability (the probability of moving from the current state to the high cholera state) for the two different models for the month of September. As we can see from Fig. S5, although there are some differences between the two models' fitted probabilities, there is considerable agreement both for the core and for the periphery thanas, agreement depending on the cholera level in the neighbors of the thana, and similar relationships with ENSO.

S5. Model Simulation

A nice consequence of using a Markov chain model is the ease at which simulations can be conducted. For a given month, the states of each thana can be found, as well as the maximum level of the neighbors of each thana. With the ENSO value provided as well, the transition probabilities can be computed for each thana ($p_{i,0}$, $p_{i,1}$, $p_{i,2}$) (where the current state of the thana is i). A uniform random variable is drawn and if this is less than $p_{i,0}$ we say that the thana has transitioned to state 0. If the uniform random variable is instead between $p_{i,0}$ and $p_{i,0} + p_{i,1}$, then we say the thana transitioned to state 1, and otherwise we say the thana transitioned to state 2. Repeating this for each thana gives us our 1-mo predictions. We then use these simulated data to simulate the next month, and so on. Because the ENSO values used in a given month correspond to 11 mo in the past, predictions 11 mo into the future can be obtained without using any information other than that in the initial month (in other words, we do not update the states with the "truth" throughout the 11-mo simulation). We repeated this procedure for each month to compute a set of 11-mo predictions. For the final results we averaged the predicted values for all of the thanas for a given month to compare that to the true transformed average.

Typically in matrix models, the multiple-step transition probabilities can be computed by iterative matrix multiplication. However, in our model, each thana's transition matrix depends on the states of the other thanas, so that this is impractical. Therefore, we use a Monte Carlo approach and simulate paths 10,000 times. Simulations were implemented in R (3).

S6. Forecast Probabilities

Because we are using a probabilistic model, in addition to being able to make predictions about the rate of cholera in the city, we can make predictions about the distribution of these rates. For a given month, by simulating 10,000 paths we have 10,000 11-mo predictions. The empirical distribution of these predictions estimates the true distribution of possible outcomes. In addition to examining whether the average outcome is above a certain level, we can see how many of the predictions are above this level and thus have an empirical estimate of the probability that the true outcome will cross a critical threshold. We can define an outbreak, for example, to be a month whose average cholera rate is in the top 25% of all of the data (as is done in Fig. 4 in the main text) and calculate the associated predicted probability that there will be an outbreak in any given month. Fig. S6 compares these probabilities to the actual data and shows that for months when there actually was an outbreak (indicated on the y axis as 1), the estimated probabilities are higher than in the months when there was not one (indicated on the y axis as 0).

As mentioned above, our transformation procedure removes the extreme events from the data. This procedure also causes the model to be biased away from extreme events. Fig. 4 in the main text illustrates this tendency: Although the predictions capture the interannual variability of the data, they also tend to underestimate the size of the outbreaks. Specifically, when we consider the months where we observed cases in the top 25%, we overpredict 4.5% of the time and underpredict 95.5% of the time. In the absence of such bias, we would expect that if we used our probabilities as a cutoff to indicate whether there was an outbreak, any probability >50% would indicate an outbreak and any probability <50% would indicate no outbreak (the 50% cutoff is indicated by the shaded dashed line in Fig. S6). Whereas probabilities >50% almost always correspond to outbreaks, because our model is biased low, we need to correct for this bias and find the appropriate cutoff to evaluate the risk of outbreaks. More technically, our predicted probabilities, considered as classifier scores, are not properly calibrated (4), and therefore a probability of around ≥90%, as computed for the El Niño years of 1998 and 2003 (see Fig. 4 in the main text), would actually reflect a higher risk than these actual numbers indicate. If we consider our probabilities as a measure of the likelihood of an outbreak (and not as the actual probability of an outbreak itself), then we can use decision theory to best classify a month.

Mathematically, using the predicted probability of an outbreak to decide whether there actually will be an outbreak is a binary classification problem. We are trying to place months into two groups on the basis of the predictions of the model, corresponding to "outbreak" and "no outbreak," and as such there are two different types of errors we can make. If we consider a positive event to be one where there is an outbreak, then we have the classical question of trying to define a hypothesis test that appropriately balances the false positives (months that we say will be outbreaks but are not) and the false negatives (months that we say will not be outbreaks but in fact are). To determine the appropriate location of the cutoff probability threshold to be used to classify months as predicted outbreaks or not on the basis of these predicted probabilities, we use the Kolmogorov–Smirnov test. The Kolmogorov–Smirnov test computes the empirical cumulative distribution function of the predicted probabilities for months where outbreaks occurred and compares it to the empirical cumulative distribution function of the predicted probabilities for months where there were not in fact outbreaks. We can use the location of the greatest distance between the two curves as our cutoff (5) (Fig. S7). For the predicted probability of an event in the upper 25% of all predicted months, we found that the best location of the threshold was 26.46% (the black dashed line in Fig. S6). The distance between 26.4% and 50%, where the cutoff would be if our probabilities were not biased low, is essentially the correction of that bias. We see in this case that our false negative rate and false positive rate are both ∼25%. Most importantly, if we look among all months for which no outbreak was predicted, only 10.1% (10/99) of these predictions were wrong. These errors would correspond to our model saying there would not be an outbreak, when in fact there was one. Because this error is potentially the most serious one from a public health perspective, the models appears to perform quite well with a low fraction of months misclassified.

In addition to the above cutoff probability that is essential to interpret the prediction probabilities for each month, we can also examine the particular choice of the threshold level used for computing the probabilities themselves. We specifically used for this procedure a threshold value set at 25% of the observed data. Note that this value is equal to the threshold value used to classify the empirical data into observed or nonobserved outbreaks. This is therefore a sensible choice for the predictions themselves. However, it is possible to examine systematically whether there is a better choice by relying on receiver operating curves (ROCs) (6). ROCs are are useful when the cost functions associated with the different types of errors (false positives and false negatives) are unclear (4) and they compare the sensitivity (true positive rate) and specificity (true negative rate) across all possible definitions of predicted outbreaks. We found that our results (the "scores" or probabilities assigned to the different months) are not significantly affected by the choice of a different threshold value within broad ranges between 0.2 and 0.8.

S7. Flooding

As discussed in the main text, there is a considerably different correlation between flooding and cholera rate in the early summer (June and July) between the two regions. There is, however, a large correlation for both regions between the late-monsoon cholera cases (August and September) and flooding. When the entire monsoon season (June through September) is considered as a whole throughout the entire city, we see a very high correlation between cases and flooding (r = 0.91; Fig. S8). The flooding index used here was the percentage of the country that was flooded at one point during the year. For example, in 1998, when the index was 68, this number means that at one point or another 68% of the country was flooded (although not necessarily all at the same time). This index, the only one we could find, is compiled at the end of a year and as such is inappropriate for a forecasting model. Clearly one would want to use only an index of localized flooding in advance of the target month for prediction. In Bangladesh, however, much of the flooding occurs during the monsoon season, so the observed correlations are in this case indicative of a strong effect. In particular, as we remark in the main text, higher levels of summer cholera appear to correlate with flooding (and high levels of ENSO).

S8. Spatial Heterogeneity

There are numerous levels of heterogeneity between thanas throughout the city, and these differences contribute to the observed difference between the two identified regions. No one socioeconomic index alone can be identified as the reason for the underlying difference in cholera dynamics, but there are several that correlate well with the differentiation of regions identified in the analysis. In Fig. S9, we show three such indexes collected during the 2001 census. In Fig. S9, *Left*, the density by thana is plotted, and it is clear that the core thanas are denser (the densest thana, shown in deep red, averages 131,000 people per square kilometer). The core thanas have more households that use tap water than do the periphery thanas (Fig. S9, *Center*) that rely primarily on well water. The fraction of houses in each thana that are classified as *jhupri* (small, temporary shanty houses, typically with roofs <4 feet from the ground and not able to withstand heavy weather), is shown in Fig. S9, *Right*.

1. Vrugt JA, Robinson BA (2007) Improved evolutionary optimization from genetically adaptive multimethod search. *Proc Natl Acad Sci USA* 104:708–711.
2. Nelder J, Mead R (1965) A simplex method for function minimization. *Comput J* 7: 308–313.
3. R Development Core Team (2010) *R: A Language and Environment for Statistical Computing* (R Foundation for Statistical Computing, Vienna).
4. Fawcett T (2006) An introduction to ROC analysis. *Pattern Recognit Lett* 27:861–874.
5. Rounds E (1980) A combined nonparametric approach to feature selection and binary decision tree design. *Pattern Recognit* 12:313–317.
6. Egan J (1975) *Signal Detection Theory and ROC Analysis* (Academic, New York), Vol 446.

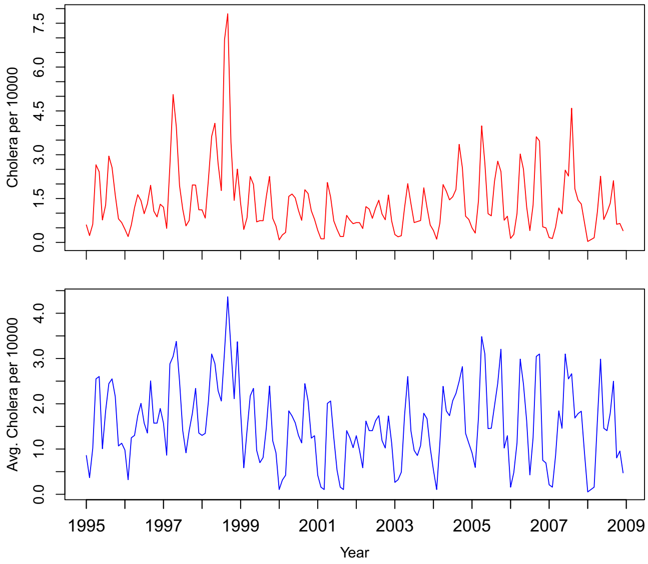


Fig. S1. Original data (red) vs. transformed data (blue).

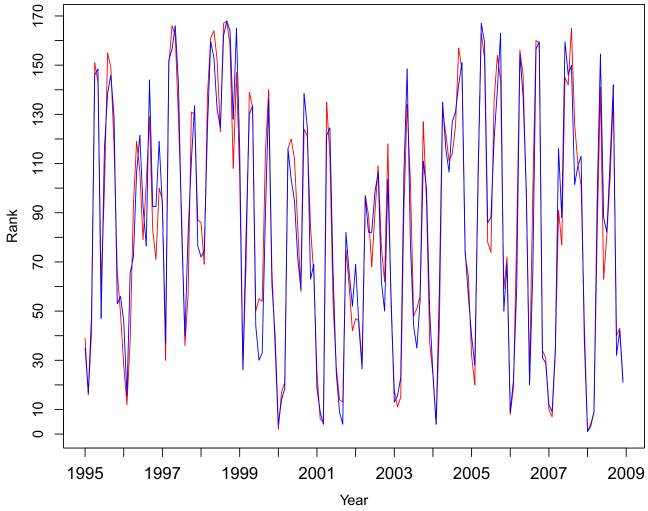


Fig. S2. Ranks by month of original data (red) and transformed data (blue).

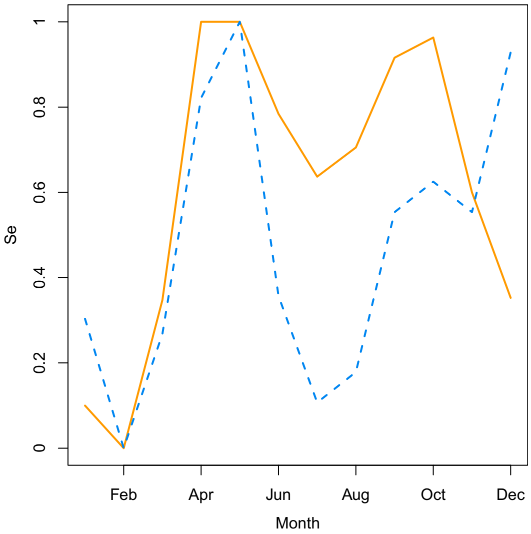


Fig. S3. Fitted Values of seasonal coefficients. The core region's Se values are indicated by the solid orange line and the peripheral region's Se values are indicated by the dashed blue line.

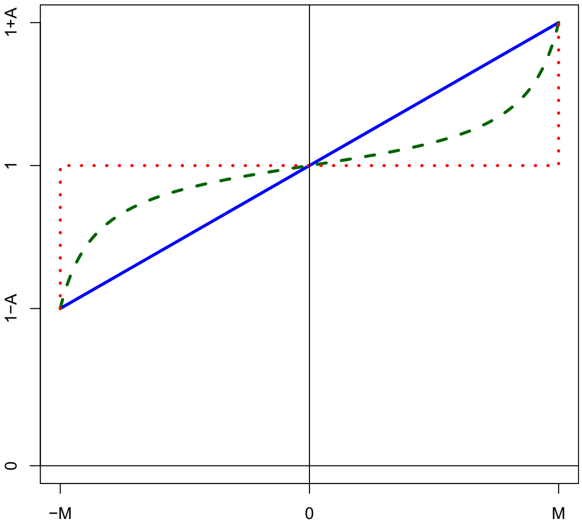


Fig. S4. Functional form of ENSO effect for three different values of *h*: $h = 0$ (solid blue line), $h = 7\pi/8$ (dashed green line), and $h = \pi$ (dotted red line).

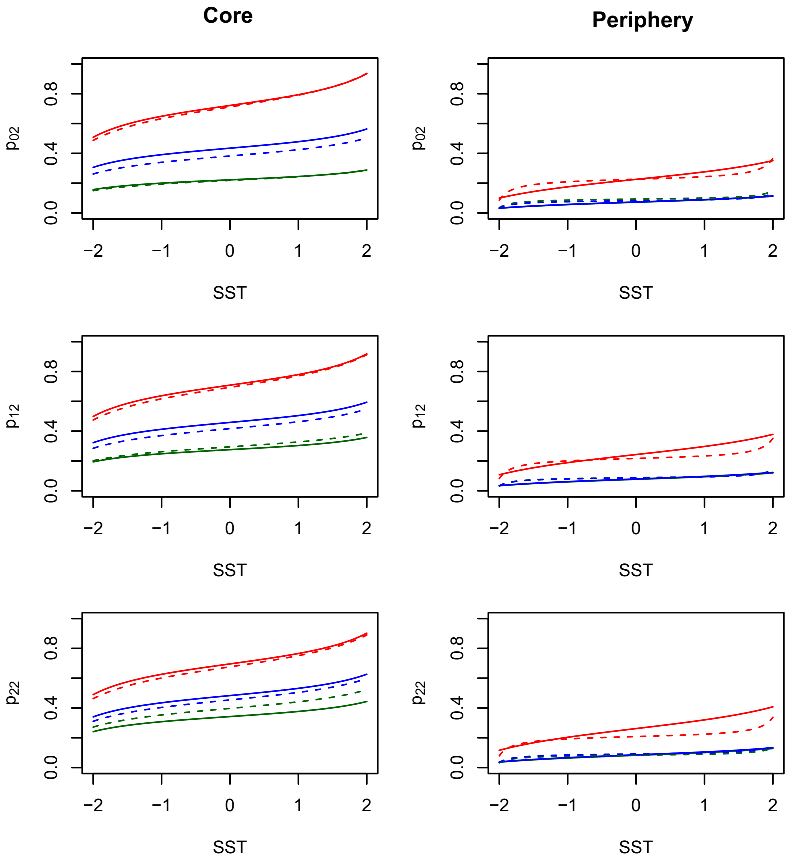


Fig. S5. Predicted $p'_{i,2,k,t}$ probabilities (transition to "high cholera" state) for both the original model (solid lines) and the alternative model (dashed lines) for the month of September for different values of ENSO. (*Left*) Core thanas; (*Right*) the probabilities for the periphery thanas. (*Top*) When the current state is "no cholera"; (*Middle*) when the current state is "low cholera"; (*Bottom*) when the current state is high cholera. Green, blue, and red lines correspond to when there is no cholera in the neighboring states, when the highest state of the neighbors is low cholera, and when at least one neighboring state has high cholera, respectively.

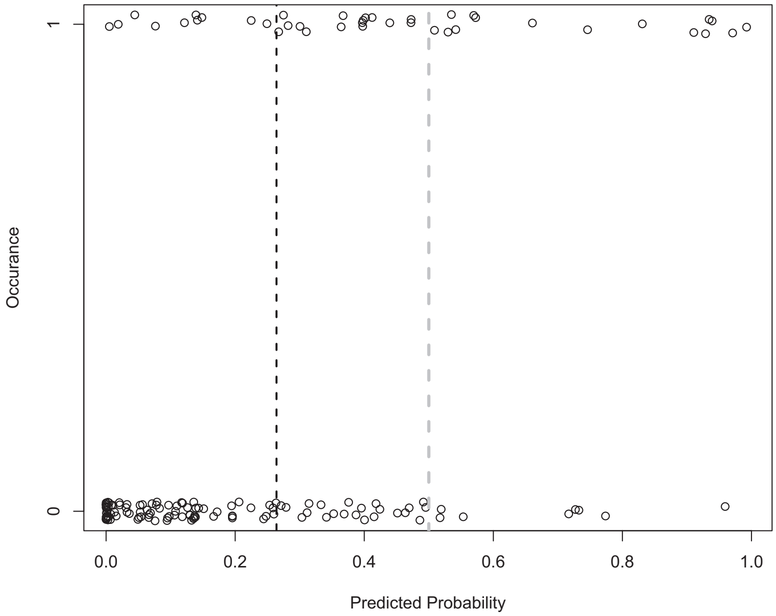


Fig. S6. Predicted probabilities of outbreaks for months when an outbreak occurred (*y* axis = 1) vs. months when an outbreak did not occur (*y* axis = 0). The shaded dashed line represents where the cutoff for classification would be if the probabilities were perfectly calibrated (i.e., at 50%). The solid dashed line is the location of the cutoff as determined through the classification analysis to calibrate the probabilities (located at 26.4%) (see text for details). We classify any month whose probability of an outbreak is to the right of this line to be one where we expect an outbreak.

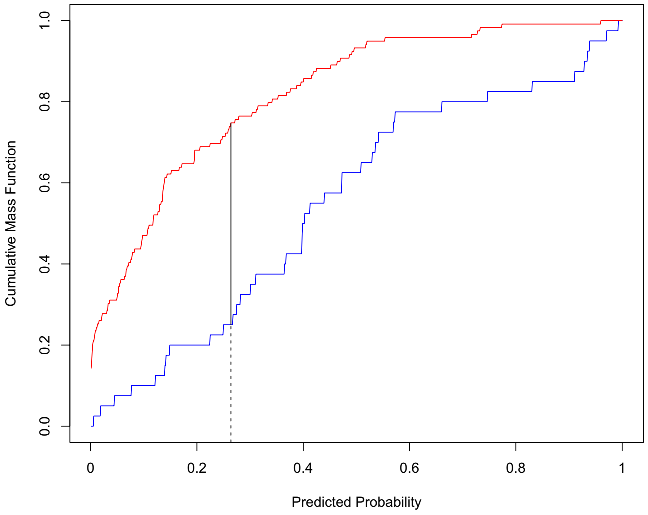


Fig. S7. Kolmogorov–Smirnov test, with cumulative mass function for predicted probabilities corresponding to months where no outbreak occurred in red, and months where an outbreak occurred in blue.

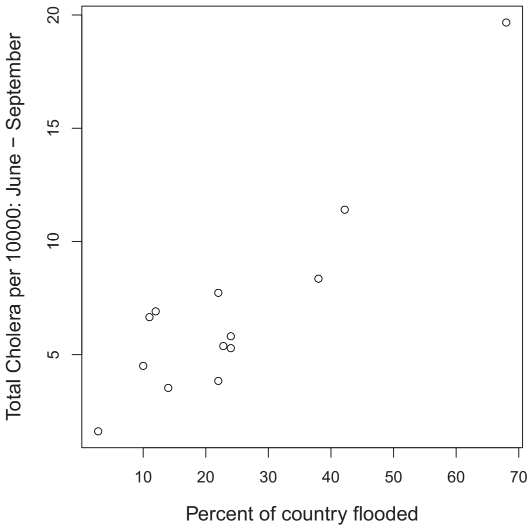


Fig. S8. Scatterplot of flooding index vs. total cases of cholera during monsoon season (per 10,000 people).

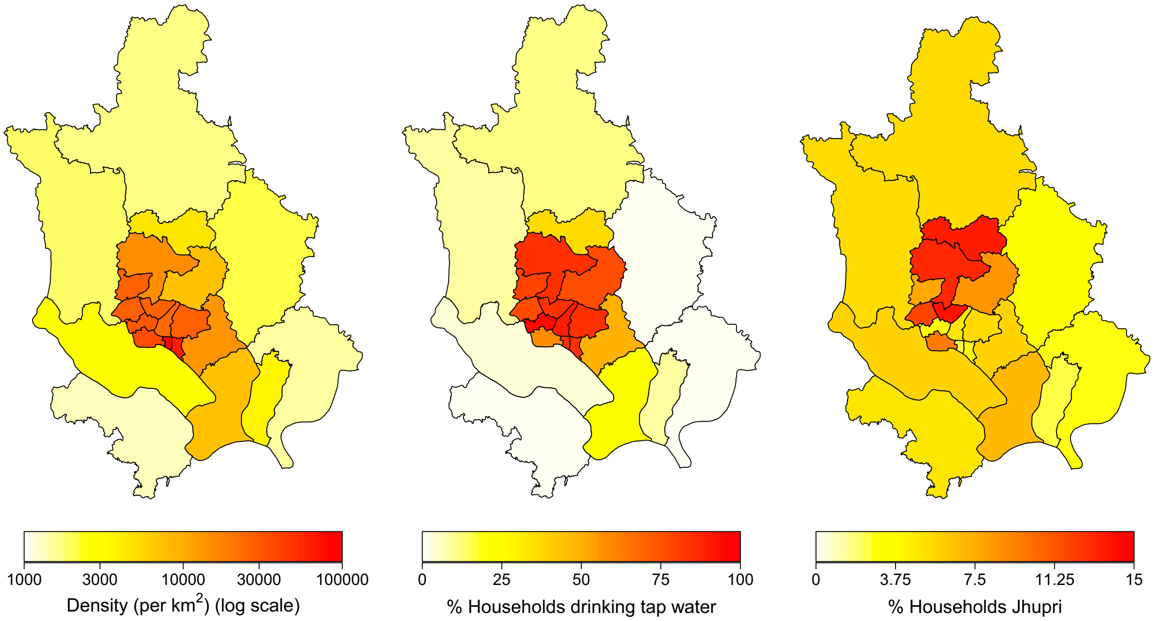


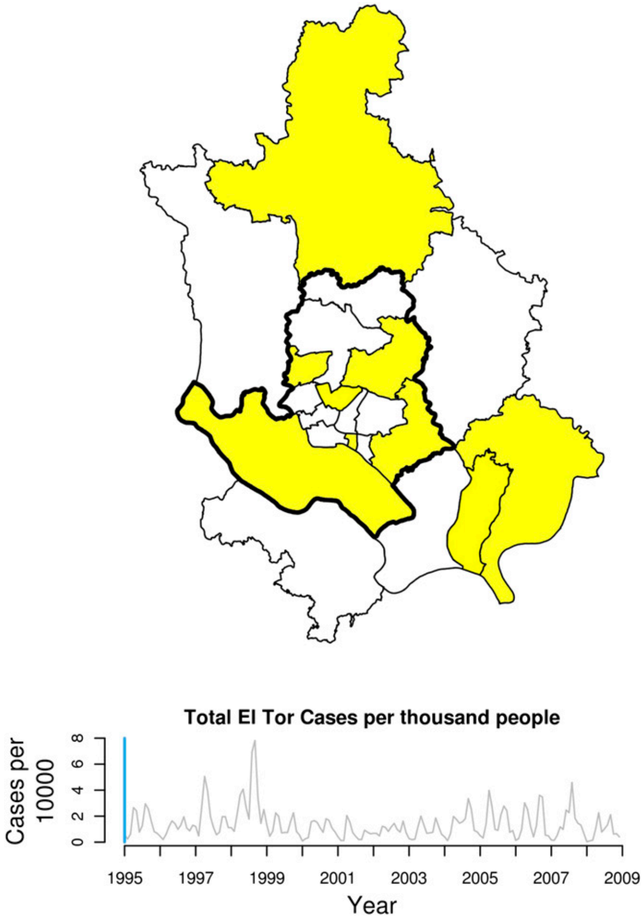
Fig. S9. Demographic maps. (*Left*) Density of thana per square kilometer (plotted in log scale). (*Center*) Percentage of households per thana whose primary drinking source is tap water. (*Right*) Percentage of households per thana whose housing is Jhupri (small temporary structure with ceiling height <4 feet).

Table S1. Summary of statistical analysis for initial parameterization

Effect	ll_0	ll_a	Δ in *ll*s	Δ df	χ^2	*P* value
Space (large scale)	−3,289.1	−3,191.7	97.4	6	194.74	2.20×10^{-39}
ENSO	−3,181.0	−3,161.5	19.5	3	39.00	1.74×10^{-8}
Seasonality	−3,315.7	−3,188.9	126.7	16	253.48	1.25×10^{-44}
Space (small scale)	−3,169.9	−3,163.2	6.8	6	13.54	0.0352
Space × ENSO	−3,161.5	−3,156.6	4.9	3	9.82	0.0202
Space × season	−3,188.9	−3,156.6	32.3	16	64.62	8.58×10^{-8}
Space × space	−3,163.2	−3,156.6	6.6	6	13.10	0.0415

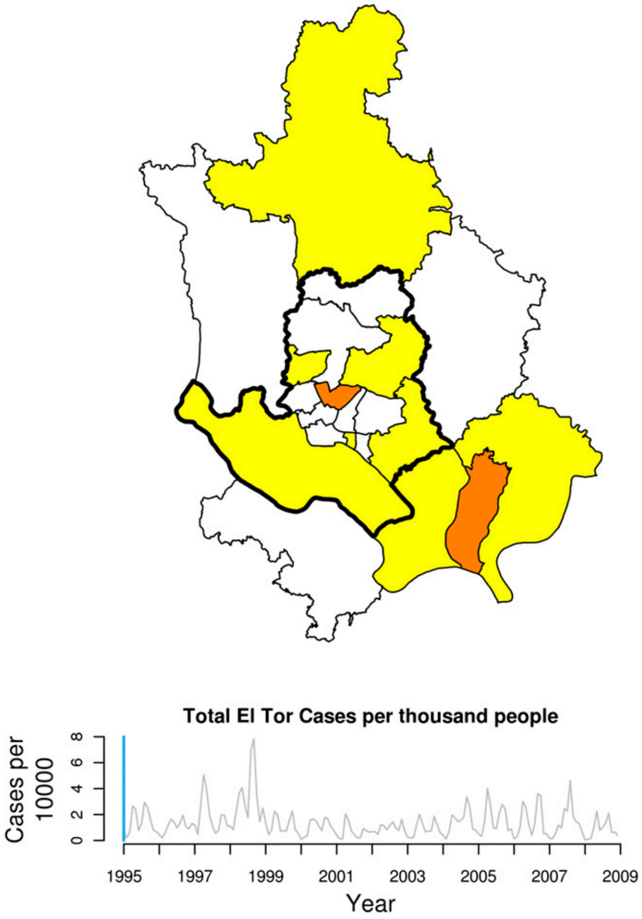
Table S2. Summary of statistical analysis for alternative parameterization

Effect	ll_0	ll_a	Δ in ll_s	Δ df	χ^2	*P* value
Space (large scale)	−3,291.9	−3,198.5	93.4	6	186.76	2.94×10^{-37}
ENSO	−3,184.0	−3,161.9	22.1	3	44.17	1.39×10^{-9}
Seasonality	−3,316.7	−3,192.7	124.0	16	248.03	1.33×10^{-43}
Space (small scale)	−3,168.4	−3,162.0	6.4	6	12.85	0.0456
Space × ENSO	−3,161.9	−3,154.7	7.1	3	14.34	0.0025
Space × season	−3,192.7	−3,154.7	38.0	16	75.91	8.99×10^{-10}
Space × space	−3,162.0	−3,154.7	7.3	6	14.60	0.0236



Movie S1. Animation of cholera rate monthly for 21 thanas from 1995 through 2008. Thick black line indicates border for core region.

Movie S1



Movie S2. Animation of transformed data monthly for 21 thanas from 1995 through 2008. White represents no cholera, yellow represents being categorized into "low cholera" bin, and orange represents being categorized into "high cholera" bin. Thick black line indicates border for core region.

Movie S2