Supplementary Material
Climate-driven endemic cholera is modulated by human mobility in a megacity

Javier Perez-Saez, Aaron A. King, Andrea Rinaldo, Mohammad Yunus, Abu S. G. Faruque and Mercedes Pascual

1 Parameter profiling

We chose to perform the profiling of the baseline transmission probabilities to high cholera states of the MDIMC model, which correspond to parameters $p_{i,3}$. Results shown in Figure S1 illustrate the higher intrinsic risk of transitioning to high cholera states in the inner thanas (group I) than in the outer.

2 SARIMA baseline modeling

A seasonal autoregressive integrated moving average model (SARIMA) was used as a benchmark for evaluating the predictive capacity of the MDIMC framework. The seasonal component of the model was included in the form of seasonal forcing covariates composed of a trigonometric series of the form $\{\sin(2\pi\frac{\tau}{12}), \cos(2\pi\frac{\tau}{12})\}$, where $1/\tau$ is the period in fraction of years. The optimal combination of forcing covariates was computed using the compensated Akaike Information Criterion (AICc). The best performing seasonal forcing was found by sequentially selecting the best ARIMA structure in terms of AICc using each combination of included seasonal components form. This was done in R using the automatic ARIMA fitting function *auto.arima*. The effect of ENSO was also taken into account by adding it as a covariate to the ARIMA model. The overall SARIMA model structure to model the time series of the mean city-wide attack rates Y_t therefore reads

$$\mathrm{Y}_t = \mathrm{X}_t^T \boldsymbol{\beta}^{seas} + \mathrm{ENSO}_{t-t_{lag}} \beta^{ENSO} + \epsilon_t, \qquad \epsilon_t \sim \mathrm{ARMA}(n,p), \tag{1}$$

where $\mathrm{X}_{\boldsymbol{t}} = \left[\sin(2\pi\frac{\tau_1}{12}t), \cos(2\pi\frac{\tau_1}{12}t), \cdots, \sin(2\pi\frac{\tau_n}{12}t), \cos(2\pi\frac{\tau_n}{12}t)\right]$ is the vector of seasonal forcings for a given vector of n periods $\boldsymbol{\tau} = [\tau_1, \cdots, \tau_n]$, $\boldsymbol{\beta}^{seas}$ is the vector of seasonal regression coefficients, $\mathrm{ENSO}_{t-t_{lag}}$ is the sea-surface temperature anomaly at lag $t_{lag} = 10$mo, β^{ENSO} the regression coefficient for the ENSO effect, and ϵ_t are the regression residuals which are assumed to follow an ARMA(n,p) structure. The residuals of the overall best-fitting model were tested using the Ljung-Box statistic corrected for the AR and MA components. The best fitting SARIMA model had $\boldsymbol{\tau} = [1, 2, 3]$, with regression residuals following an ARIMA(1,2) structure (details reported in Table S1, residual diagnostics in Figure S2).

The predictive performance of the SARIMA model was evaluated using the same methodology as for the MDIMC model (see main text). For each month 10'000 simulations were performed with a 11 month lead. The prediction output was then used to compute the optimal threshold for outbreak classification and constructing the contingency matrix and related statistics. The output of the SARIMA simulations are given in Figure S3. The SARIMA model had an overall accuracy of 68.9% (lower than the 76.4% of the MDIMC model, and false negative (37.5%) and positive (24.8%) rates respectively higher and similar to the MDIMC model.

3 Directionality of cholera transmission

The patterns of thana-to-than risk contribution illustrated in Figure S4, Figure S5 and Supplementary Video 1 illustrate the directionality of cholera transmission in the city of Dhaka.
The regression plots of the risk contribution on the probability of an inter-thana trip show the differential role of connectivity depending on the direction of human mobility (Figure S6 and Table S2). Interestingly the largest slope, meaning the largest increase of cholera risk per increase in trip probability, is for the inner-to-outer direction (p-values < 0.05).

4 Sensitivity analysis of differential urbanization rates

To study the potential effect of differential rates of urbanization in the periphery in comparison with the core of the city, we performed a sensitivity analysis of model outputs in terms of city-wide predicted cholera attack rates to variations in the inner-to-outer mobility fluxes.
The sensitivity analysis was done by altering directly the mobility matrix Q extracted from the 2011 population distribution estimates (Q_{2011}). More specifically, the analysis consisted of: 1) modifying the outer-to-inner movement probabilities between -20% and +20% of the 2011 estimates parametrized by parameter $\epsilon \in [0.8, 1.2]$, 2) enforcing the row-sums to 1 by modifying the intra-thana and outer-to-outer movement probabilities by assuming the residual (may it be deficit or excess if $\epsilon > 1$ or $\epsilon < 1$ respectively) movement probability of thana k, $\Delta Q_{k,.}(\epsilon)$, is partition between intra-thana and outer-to-outer depending on parameter $\lambda \in [0, 1]$ (where $\lambda = 0$ means that all of $\Delta Q_{k,.}(\epsilon)$ is assigned to outer-to-outer mobility), and 3) applying a linear trend in time for the mobility matrix $Q(t)$ to match the 2011 mobility matrix at the end of the simulation period. The resulting perturbed time varying matrix $Q(\epsilon, \lambda, t)$ thus depends on two parameters which are used to explore a range of altered mobility scenarios. Model sensitivity to the input mobility matrix $Q(t)$ was measured by running 5000 simulations of the full MDIMC model (using the best-fitting parameter set) with the mobility matrices resulting from each combination of parameters $[\epsilon, \lambda]$, and comparing the mean simulated city-wide cholera attack rates to the original simulation results using the fixed mobility matrix Q_{2011}.

Results show that the average difference in the simulations between the constant matrix Q_{2011} and alternative matrices $Q(\epsilon, \lambda, t)$ is consistently lower than 0.1% (Figure S7), which suggests that omitting the differential rate of urbanization in the two parts of the city does not undermine the results.

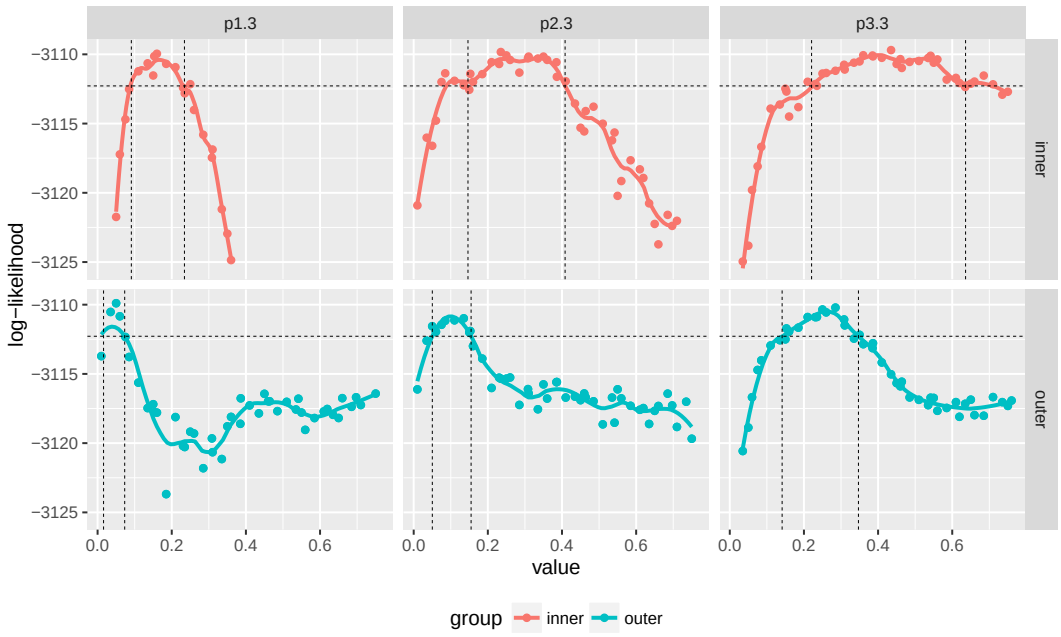

Figure S1: Parameter profiling of the transition probability to high choler states. Profiling was perfomerd over a sequence of 30 parameter values between 10^{-3} and 0.75 (higher parameter values did not satisfy model constraints). The log-likelihood profiles are subdivided by group and by parameter. A polynomial regression with a span of 0.3 was used to determine the 95% confidence intervals for each parameter (vertical dashed lines). The results for $p_{1,3}$ of the inner group for which the log-likelihood was lower than -3125 ($p_{1,3} > 0.4$) have been omitted.

AR(1)	MA(1)	MA(2)	$\beta^{seas}_{\sin,\tau=1}$	$\beta^{seas}_{\cos,\tau=1}$	$\beta^{seas}_{\sin,\tau=2}$	$\beta^{seas}_{\cos,\tau=2}$	$\beta^{seas}_{\sin,\tau=3}$	$\beta^{seas}_{\cos,\tau=3}$	β^{ENSO}
−0.75	1.43	0.61	−0.29	−0.35	−0.53	−0.26	0.30	0.21	3.54
(0.12)	(0.11)	(0.09)	(0.12)	(0.12)	(0.11)	(0.11)	(0.09)	(0.09)	(0.83)

Table S1: Parameter values of best-fitting ARIMA model with seasonal forcing and ENSO effect. The standard errors of the parameter values are given in parenthesis. The best fitting model has $\boldsymbol{\tau} = \{1, 2, 3\}$, and corresponding AICc= 371.76. The second best-fitting seasonal covariate model had $\boldsymbol{\tau} = \{2, 3\}$ with AICc= 379.80.

	inner-inner	inner-outer	outer-inner	outer-outer
fluxes	1′050.43***	15′793.03***	886.56***	6′758.63***
	(28.32)	(801.46)	(55.55)	(650.55)
Adj. R^2	0.88	0.80	0.72	0.72
Num. obs.	182	98	98	42
RMSE	23.07	107.02	7.50	173.95

$^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$

Table S2: Linear model fits between connectivity and risk contribution by direction. RMSE: root mean-squared error.

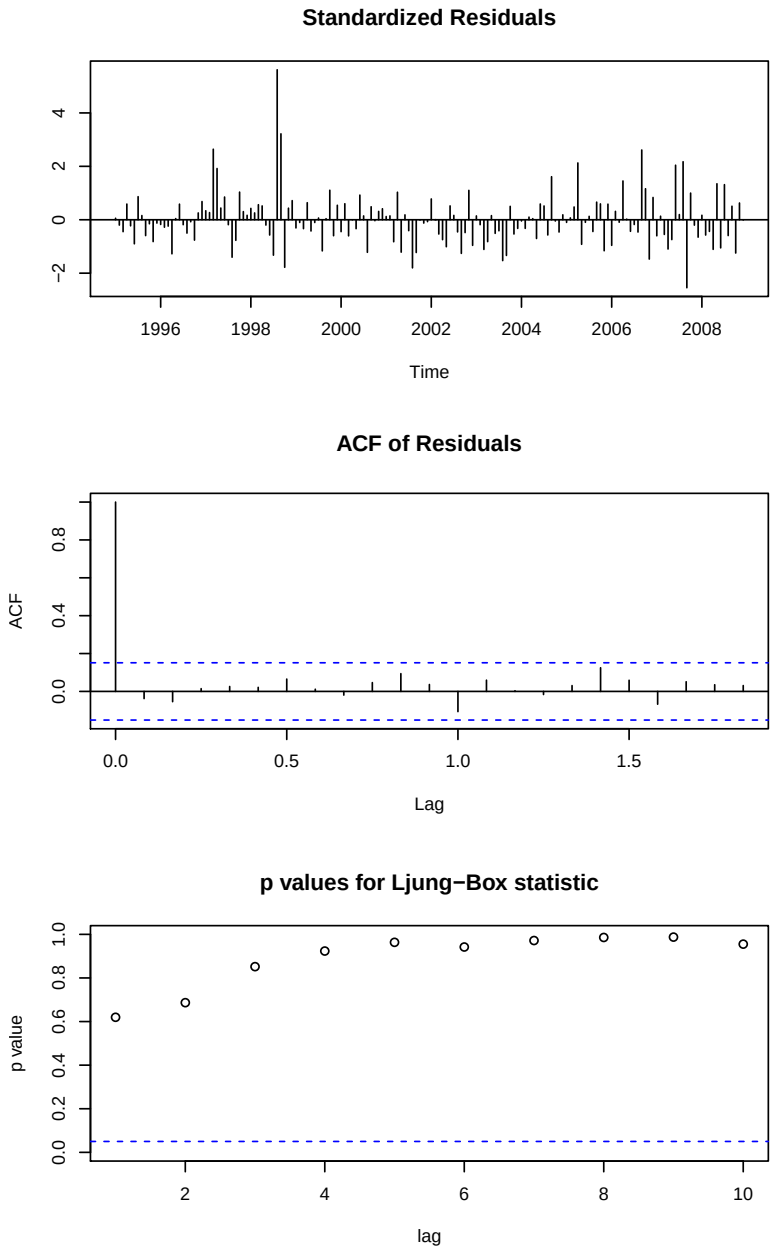

Figure S2: Diagnostics on best performing SARIMA model residuals. Both the ACF and the Ljung-Box statistics suggest that the residuals are uncorrelated.

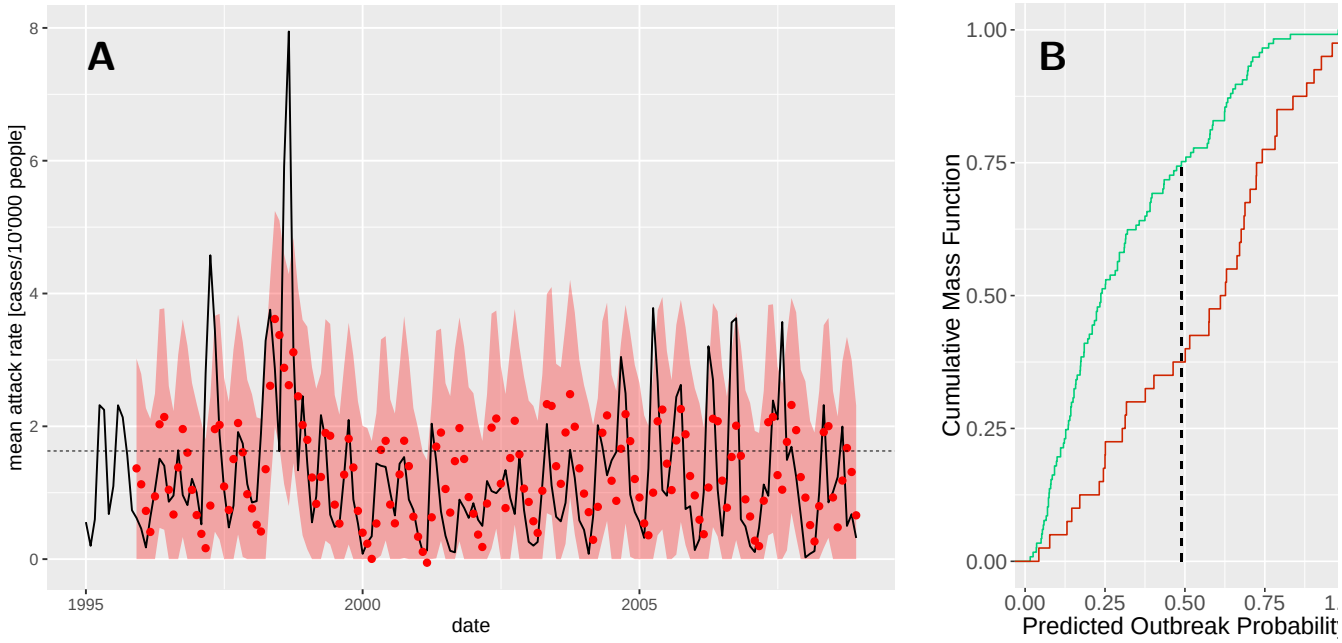

Figure S3: SARIMA model cholera outbreak prediction performance. (A) Eleven-month predictions of cholera incidence. The means (red dots) and 95% confidence intervals (red shadings) of the simulations are given together with the observed average city-wide cholera state (solid line). (B) Same as in Fig. 6 of the main text, here the outbreak probability threshold is 0.489.

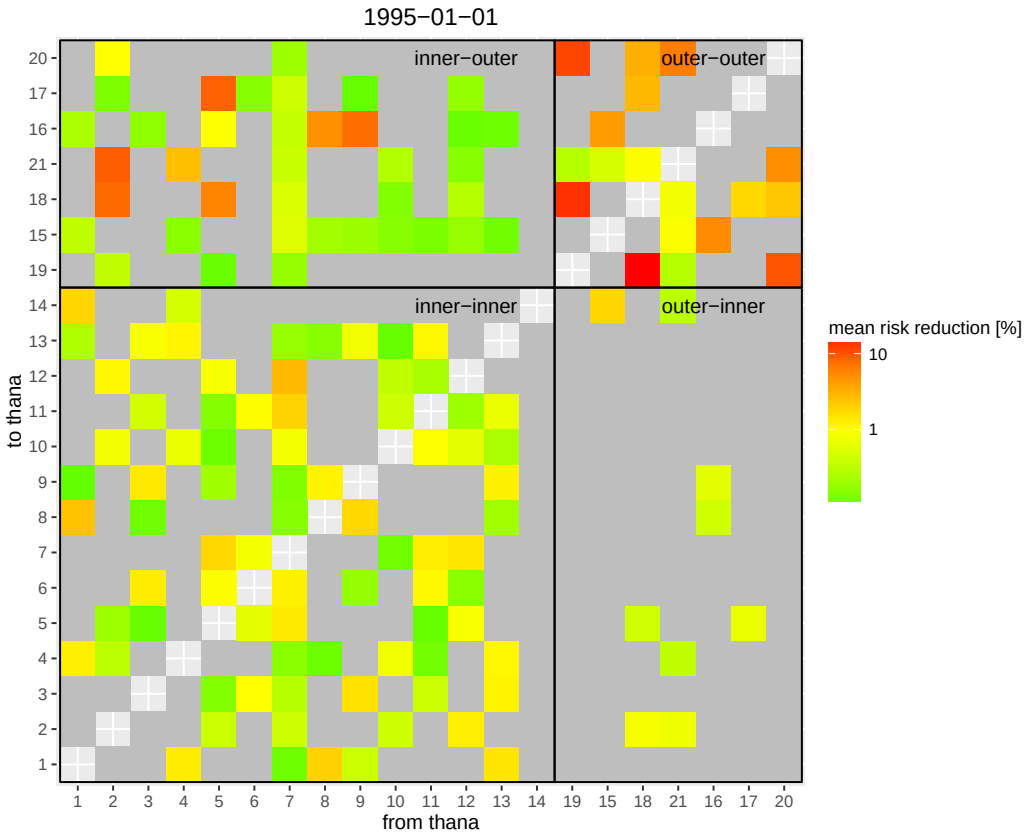

Figure S4: Thana-to-thana mean risk reduction animation frame. The mean risk reduction over 1'000 state transition simulations of the month of January 1995 is given by the color of the cells in a logarithmic scale (gray cells indicate risk contribution lower than 0.1%), with transmission directionality delineated by black outlines. The Supplementary Video 1 is composed of a series of frames for each available date in the timeseries.

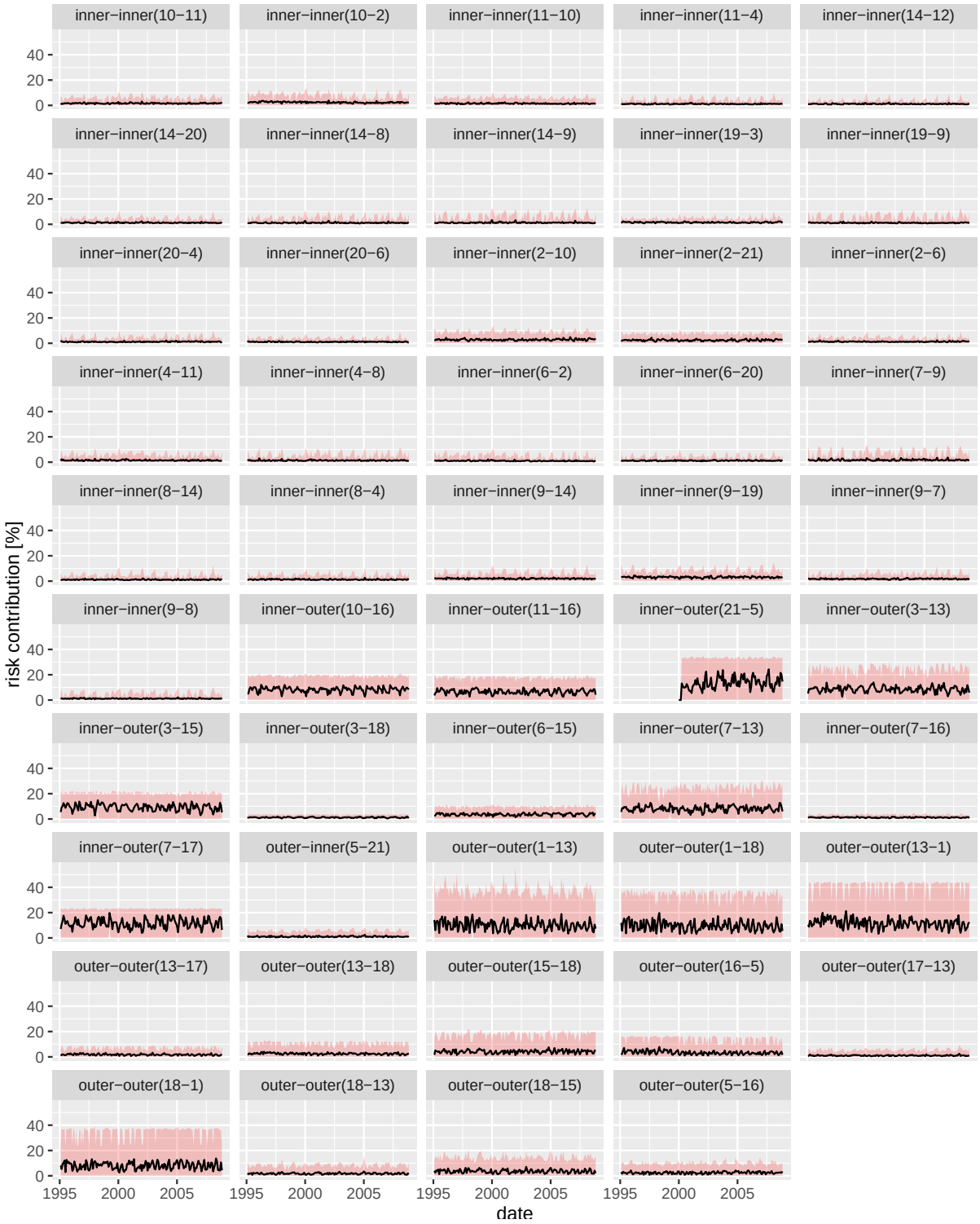

Figure S5: Thana-to-thana risk contribution. The risk contribution was computed over 1'000 simulations of the state transitions in terms of the mean (black lines) and the 95% inter-quantile of the simulated values (red envelop). Results are only shown for thana-to-thana risk contribution for which the mean risk contribution throughout the study period is larger than 1%. The facet labels indicate the group-to group risk transmission direction and the IDs of the thanas as given in Fig. 1 of the main text (in parenthesis).

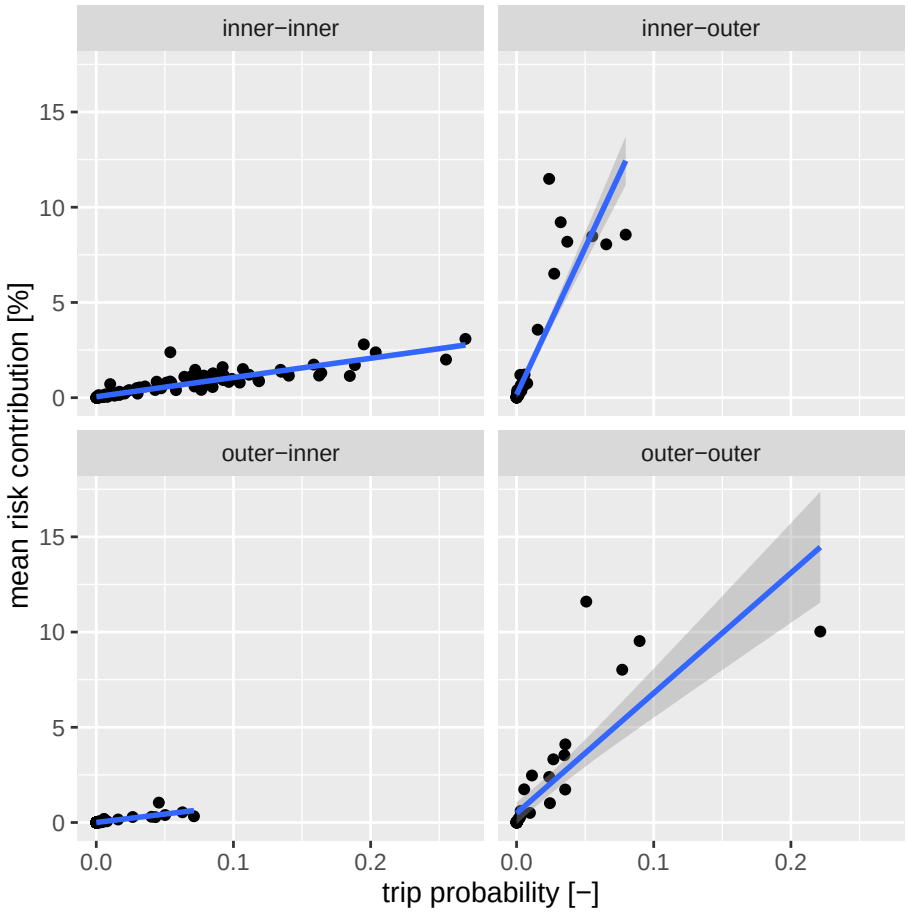

Figure S6: Correlation between connectivity through human mobility and cholera risk transfer. The mean risk contribution corresponds to the average across 1'000 simulations and over the whole study period, partitioned by origin and destination group. The regression lines (blue) and 95% CI prediction (gray shading) are given by group-to-group direction. The significantly different slopes between groups illustrate the different roles of human mobility on the directionality of cholera risk spread in the megacity. The parameters of the regression coefficients and their significance are given in Table S2.

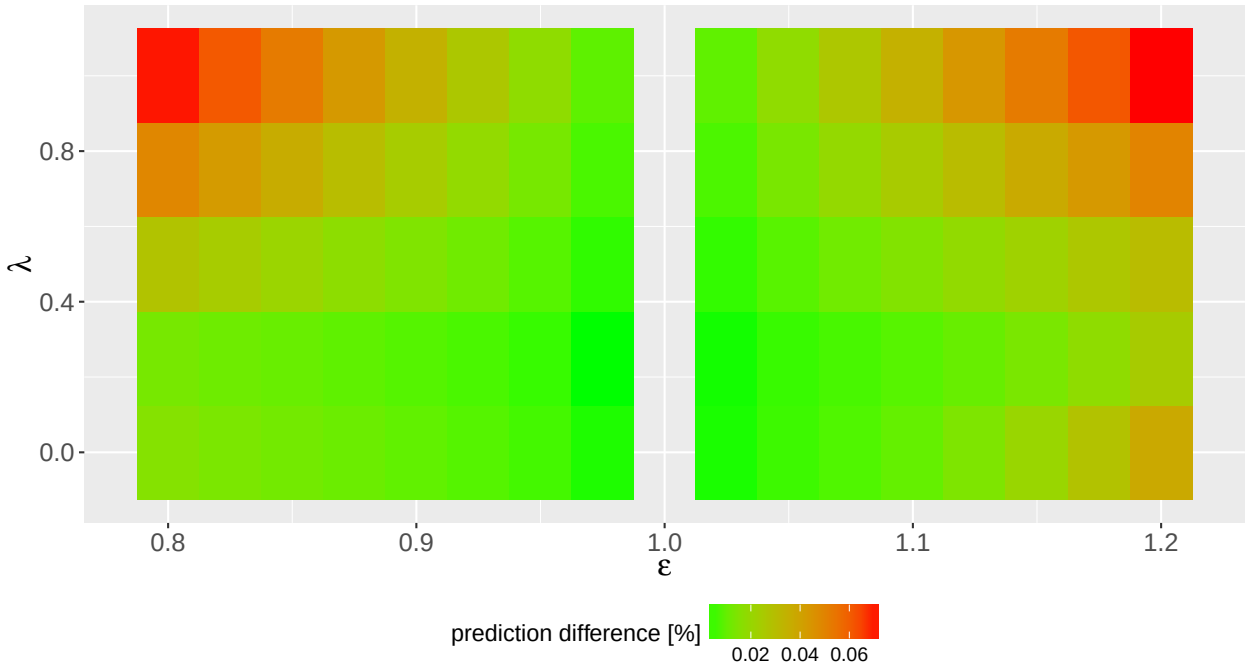

Figure S7: Sensitivity analysis of differential urbanization rates. The outer-to-inner fluxes of estimated mobility matrix were modified by a factor of $\epsilon \in [0.8, 1.2]$ and a fraction $\lambda[1-\lambda]$ of the residual (or excess) mobility re-assigned to the intra-thana[outer-to-outer] mobility. A linear trend was imposed for the mobility matrix to match the 2011 estimate at the end of the study period. The percentage difference between the average of 5000 simulated city-wide attack rates using the modified and constant mobility matrices was used a measure of model sensitivity to the hypothesis of differential urbanization rates between the two parts of the city.