Supporting Information

Samia et al. 10.1073/pnas.1015946108

SI Text

Data. Our wildlife reservoir data are from the PreBalkhash plague focus of southeastern Kazakhstan (74–78°E and 44–47°N). Each spring (May and June) and autumn (September and October) during the period 1949–1995, gerbil density estimates, together with flea counts, were done on a number (1–78, median = 54) of different squares in a 20 × 20-km grid (1). The consistency of these data is indeed very high, because they were obtained through a rather strict regime. Kazakhstan is the last Soviet Republic that declared independence (December 16, 1991). Moreover, Kazakhstan enjoys relatively stable political and socioeconomic developments since independence (2). Because the host densities are spatially autocorrelated over large areas, most likely through large-scale climate forcing (3), we suspect that the monitoring data from the PreBalkhash focus capture essential variation over much of Kazakhstan. By developing a series of statistical models, we find support for this view, because we find a clear correspondence between the plague/host dynamics of the PreBalkhash focus and the fluctuations in the annual number of human plague cases aggregated across Kazakhstan.

Here, we use the mean of measurements for each season, thereby eliminating smaller-scale spatial variations and measurement errors. The high degree of spatial correlation (3) means that this large scale still retains most of the temporal variation.

The human plague cases are recorded annually and aggregated spatially.

Next Generation Matrix for the Basic Reproduction Number. The basic reproduction number R_0, the expected number of new cases generated by a typical infected individual in a totally susceptible population in a demographic steady state, is defined as the dominant eigenvalue of the so-called next generation matrix for systems with a finite number of different types of infected individuals (details and algorithms are given in ref. 4). The basic reproduction number is a threshold quantity in the natural dynamic sense, meaning that a steady state of a dynamical system changes stability when the value of the quantity passes the threshold. In the rodent–flea–human system, we, in principle, have three types of infected individuals, and the next generation matrix M is a 3 × 3 matrix. The elements m_{ij} of M are defined as the expected number of new cases of type i caused by an infected individual of type j. If we number the types as 1 for fleas, 2 for rodents, and 3 for humans, we obtain (Eq. S1)

$$M = \begin{pmatrix} 0 & m_{12} & 0 \\ m_{21} & 0 & 0 \\ m_{31} & m_{32} & m_{33} \end{pmatrix}. \qquad [S1]$$

Here, $m_{12} = K_F$, the number of fleas produced by one infected rodent at carrying capacity (where we assume that all fleas on a rodent become infected during the rodent's infectious period) (5). Effectively, this number of fleas contacting a rodent is multiplied by the probability per unit of time that a rodent infects a flea and the average duration of the infectious period in the rodent. Because fleas are assumed to be attached to the rodent for a longer period, we assume, basically, that each flea becomes infected during that time (i.e., that the product of the transmission probability per unit of time and the time period of transmission is one). The element $m_{21} = (1 - e^{-aK_R})\beta_R \frac{1}{d_F}$ equals the number of rodents infected by one infected flea; it is the product of the probability that a flea successfully finds a rodent when searching for a host, the transmission rate to rodents, and the average time that the flea is infectious. We have chosen the elements m_{13} and m_{23} to be zero, signifying that humans infect neither fleas nor rodents, respectively, consistent with the view that human cases are the result of spillover from the wildlife system but that there is no spillback from humans. Furthermore, we only regard bubonic plague and neglect the pneumonic form (6). Consequently, we assume that $m_{33} = 0$, meaning that humans do not directly infect other humans. We have already assumed that fleas do not directly infect other fleas and that rodents do not directly infect other rodents.

The consequence of these reasonable assumptions is that the R_0 for the natural system does not depend on the human host. The dominant eigenvalue of M is given by (Eq. S2)

$$R_0 = \sqrt{k_{12}k_{21}} = \sqrt{\frac{K_F\beta_R(1 - e^{-aK_R})}{d_F}}. \qquad [S2]$$

This constitutes the threshold quantity for the plague outbreaks in the wildlife system. Note that, as is to be expected, the quantity depends on the rodent and flea population size, the transmission rate (from flea to rodent), and the average length of the infectious period (in fleas). It can also be derived by specifying a full compartmental transmission model, which has indeed been done by Keeling and Gilligan in ref. 7. In the text, we use the derived quantity R_{eff}, depending on time and interpreted as the effective reproduction number of the wildlife system, with the fixed carrying capacities K_F and K_R replaced by the respective actual population sizes B_t and R_t (in the text). In case of a mosquito-borne infection like malaria in humans, the traditional expression for R_0 is also the product of two terms, one number characterizing the number of mosquitoes infected by one human and the other number characterizing the number of humans infected by one mosquito [this finding was introduced by Macdonald (8), and an example is given in the work by Bailey (ref. 9, pp. 94–100)]. This finding arises in the same way from a 2 × 2 next generation matrix having mosquitoes and humans as its types. The expression from Macdonald (8) is actually the square of the R_0 from the next generation matrix, because the generations of vectors and hosts alternate so that taking the expression without the square root actually amounts to looking two generations ahead. Although the mathematical approach is the same, the difference in biology between a mosquito- and flea-based system causes the resulting expressions to be different.

It is clear from the above, however, that this threshold quantity cannot be expected to explain the second epidemiological threshold that we have statistically observed in the analysis of the human cases of plague in Kazakhstan. The elements m_{31} and m_{32} are also not likely to be very relevant, because they are based on homogeneous mixing of humans with the wildlife system.

The Threshold Quantity λ. One might imagine that, in other time periods than the 1949–1995 period under study, there was a different distribution for the contact parameter c and hence, a different distribution for λ. For example, in the period before our study period, there was likely to have been more activity of nomadically living people in the region in Kazakhstan (before more fixed agricultural communities became established, usually outside the plague-affected areas). Exposure and contacts are, therefore, likely to have been distributed differently from what we assume here, leading more frequently to a situation, perhaps, where λ is above threshold. Certainly, there were many more

human cases in the period before control. This drop is, of course, influenced (or reasonably caused) by the control measures that were taken after 1949, but there may also be a visible decline in the frequency of outbreaks in humans before that period, coinciding with the change in nomadic activity.

Statistical Modeling. ***Seasonal wildlife reservoir model.*** We test the significance of the assumption of heteroscedasticity and dependence among observations defined in the structure of the variance–covariance matrix Σ. We find that, when the spring and fall forces of infection are both in the upper regime, Σ is of the form (S3)

$$\begin{array}{c} \\ G_{t,s} \\ G_{t,f} \\ B_{t,s} \\ B_{t,f} \\ \mathcal{J}_{t,s} \\ \mathcal{J}_{t,f} \end{array} \begin{array}{c} \begin{array}{cccccc} G_{t,s} & G_{t,f} & B_{t,s} & B_{t,f} & \mathcal{J}_{t,s} & \mathcal{J}_{t,f} \end{array} \\ \begin{pmatrix} \sigma^2_{G,s} & \rho_G \delta_{G,f} \sigma^2_{G,s} & 0 & 0 & 0 & 0 \\ \rho_G \delta_{G,f} \sigma^2_{G,s} & \delta^2_{G,f} \sigma^2_{G,s} & 0 & 0 & 0 & 0 \\ 0 & 0 & \delta^2_{B,s} \sigma^2_{B,f} & 0 & \rho_{B,\mathcal{J}} \delta_{B,s} \delta_{\mathcal{J},s} \sigma^2_{B,f} & 0 \\ 0 & 0 & 0 & \sigma^2_{B,f} & 0 & \rho_{B,\mathcal{J}} \delta_{\mathcal{J},f} \sigma^2_{B,f} \\ 0 & 0 & \rho_{B,\mathcal{J}} \delta_{B,s} \delta_{\mathcal{J},s} \sigma^2_{B,f} & 0 & \delta^2_{\mathcal{J},s} \sigma^2_{B,f} & 0 \\ 0 & 0 & 0 & \rho_{B,\mathcal{J}} \delta_{\mathcal{J},f} \sigma^2_{B,f} & 0 & \delta^2_{\mathcal{J},f} \sigma^2_{B,f} \end{pmatrix} \end{array}, \quad [S3]$$

where if the infectious-flea force in the spring season is in the lower regime, the variance–covariance matrix Σ of the conditional distribution would have dimension reduced by one (i.e., reduced by the number of seasons in the lower regime). Hence, the rodent growth rate model, the flea burden model, and the change in infectious-flea force model have different variances across models and seasons. Moreover, the rodent growth rates are correlated across seasons and uncorrelated with the flea burden and change in infectious-flea force models. The flea burden is correlated with the change in infectious-flea force within each season. The flea burden and the change in infectious-flea force are not correlated across seasons.

The model is fitted using the gnls function in the nlme package of the R software (10). Fig. 1*A* summarizes the results of our analysis, where R_t^{su} is the summer rainfall at time t. The subscript $t-0.5$ refers to the previous season [i.e., if the response is in the spring (fall) of year t, then $t-0.5$ refers to the fall of last year (spring of current year)]. The subscripts $t-1$, $t-1.5$, and $t-2$ are similarly defined. Table 1 summarizes the maximum likelihood estimates of the parameters in the model along with their asymptotic SEs and asymptotic 95% confidence intervals. The plots of the normalized residuals vs. the fitted values in the rodent growth rate model, the flea burden model, and the change in infectious-flea force model reported in Fig. S2*A* do not indicate any significant heteroscedastic patterns. The normal probability plots of the normalized residuals in Fig. S2*B* do not show any significant departure from the assumption of normality. The plots of the empirical autocorrelation function, displayed in Fig. S3, indicate that the normalized residuals behave like uncorrelated noise, which was expected under the above correlation model. The observed time series, along with the fitted counterparts, are displayed in Fig. S4 for each of the three models (rodent growth rate, log flea burden, and change in infectious-flea force). Clearly, the fitted values are close to their observed counterparts in each of the three models.

We justify the use of the seasonal threshold model for the wildlife plague data by comparing the deviance of our fitted model with the deviance of a simple multivariate linear model in which the threshold is removed. The fitted model (i.e., the threshold model with seasonal delay and threshold) is compared with the simple linear multivariate model by assessing the difference in deviance when simulating from the associated smaller model based on a parametric bootstrap size of 1,000. Note that the asymptotic null χ^2 distribution for comparing changes in deviance may be invalid, because the threshold parameter is absent under the null hypothesis of no threshold effects (11). We find that the observed deviance difference between the simpler model and our fitted seasonal threshold model is 79.94 compared with the bootstrap 95% quantile of the deviance difference given by 12.28. Similarly, our fitted seasonal threshold model is compared with a model with the same delay and threshold in the spring and fall, where the delay is estimated to be 1.5. We find that the observed deviance difference between our fitted seasonal threshold model and the nonseasonal threshold model is 27.50 compared with the bootstrap 95% quantile of the deviance difference given by 11.42. Hence, our seasonal threshold wildlife reservoir model clearly provides a better fit to the data.

Human–plague model. We justify the use of the threshold model for the human plague data by comparing the deviance of our fitted model with the deviance of a simple generalized linear model in which the threshold is removed. We fitted a simple generalized linear model with a log link function to the human plague data. The resulting deviance of the simple generalized linear model without a threshold effect is 135.5. The fitted human plague model [i.e., generalized threshold model (GTM) with a deviance of 60.62] is compared with the simple generalized linear model by assessing the difference in deviance when simulating from the associated smaller model based on a parametric bootstrap size of 1,000. Note that the asymptotic null χ^2 distribution for comparing changes in deviance may be invalid, because the threshold parameter is absent under the null hypothesis of no threshold effects (11). We find that the observed deviance difference between the simpler model and our human plague GTM is 74.9 compared with the bootstrap 95% quantile of the deviance difference given by 6.91. Hence, our human plague GTM clearly provides a better fit to the human plague data.

To ensure that the threshold variable used in the fitted model, namely the lag-1 minimum of spring and fall average flea density, is a good choice, we fitted the model by considering a number of different threshold variables and assuming the same set of covariates. Table S2 reports the corresponding Akaike Information Criterion (AIC) of each model normalized to account for slight variation in the sample size because of different lag structure. The models considered in Table S2 may not be compared directly with respect to the AIC, because they are based on different number of observations. Therefore, we compare these models in Table S2 by their normalized AIC [normalized AIC (NAIC) = AIC/effective number of observations]. Note that the

fitted model in Fig. 2*A* has an NAIC of 3.64, which is the smallest value among the models considered in Table S2.

The human–plague–threshold model is fitted using a likelihood-based estimation procedure (12) with these details. The threshold parameter r and the delay parameter d in this threshold model are estimated by minimizing the NAIC. More specifically, for a fixed integer delay d between 0 and 2, the search for the threshold is done based on a grid search between the 10th and 90th percentiles of the threshold variable $\not{\theta}_{t-d}$. Then, for given estimates of the delay and threshold, the associated generalized linear submodels in each (lower and upper) regime are fitted using the glm function in R (13). Using this estimation approach, the optimal delay d and its corresponding threshold r, for which the NAIC is smallest, are estimated to be 1 y and 148.71, respectively. The maximum likelihood estimates of the remaining parameters in the human–plague model are summarized in Table 1.

Model diagnostics are used to check the adequacy of the fitted model. In particular, the autocorrelation function (ACF) plot of the standardized residuals, reported in Fig. S5, suggests that the standardized residuals are serially uncorrelated. The standardized residuals are, indeed, independent over time based on the nonparametric method of runs test, with a P value of 0.64. The good agreement between observed and fitted values (on the original scale) of human plague cases in Fig. 2*E* attests to the adequacy of the model.

Forecasting Using the Full Plague Ecoepidemiological Model. We use the fitted multivariate wildlife reservoir model, described in Fig. 1*A*, to forecast the rodent growth rate, the flea burden, and the infectious-flea force. To illustrate the empirical forecasting performance of this multivariate model, we fitted the model using the seasonal observations corresponding to the years 1949–1990 and reserved the seasonal observations corresponding to 1991–1995 for forecasting evaluation. In addition, the forecasting procedure is carried up to year 2003. A parametric bootstrap procedure is used to compute the forecasts and their prediction limits based on 1,500 iterations. Each point forecast is the sample average, and the lower and upper 95% prediction limits correspond to the 2.5 and 97.5 percentiles, respectively. The forecasts, along with their 95% interval forecasts displayed in Fig. S6, show a strong seasonal pattern and closely track the observed data.

We then use the fitted human–plague model, described in Fig. 2*A*, to forecast the number of human plague cases. The biannual point forecasts of the lag-1 infectious-flea force, the lag-1 flea burden, the lag-1 flea density, and the lag-1 rodent density, corresponding to the years 1991–2003, are used in the forecasting procedure of the number of annual human plague cases. A parametric bootstrap procedure is used to compute the forecasts and their prediction limits based on 3,000 bootstrap replications. Each point forecast is the sample average, and the lower and upper 95% prediction limits correspond to the 2.5 and 97.5 percentiles, respectively. The forecasts, along with their 95% interval forecasts displayed in Fig. 2*E*, closely track the observed data.

1. Frigessi A, et al. (2005) Bayesian population dynamics of interacting species: Great gerbils and fleas in Kazakhstan. *Biometrics* 61:230–238.
2. United Nations (2010) *Table B of the 2010 Human Development Report by the UN.* Available at http://www.undp.kz/en/articles/1/75.jsp. Accessed January 7, 2011.
3. Kausrud KL, et al. (2007) Climatically-driven synchrony of gerbil populations allows large-scale plague outbreaks. *Proc R Soc Lond B Biol Sci* 274:1963–1969.
4. Diekmann O, Heesterbeek JAP, Roberts MG (2010) The construction of next-generation matrices for compartmental epidemic models. *J R Soc Interface* 7:873–885.
5. Eisen RJ, et al. (2006) Early-phase transmission of *Yersinia pestis* by unblocked fleas as a mechanism explaining rapidly spreading plague epizootics. *Proc Natl Acad Sci USA* 103:15380–15385.
6. Gage KL, Kosoy MY (2005) Natural history of plague: Perspectives from more than a century of research. *Annu Rev Entomol* 50:505–528.
7. Keeling MJ, Gilligan CA (2000) Bubonic plague: A metapopulation model of a zoonosis. *Proc Biol Sci* 267:2219–2230.
8. MacDonald G (1952) The analysis of equilibrium in malaria. *Trop Dis Bull* 49:813–829.
9. Bailey NTJ (1982) *The Biomathematics of Malaria* (Charles Griffin, London), p xii.
10. Pinheiro JC, Bates DM (2000) *Mixed-Effects Models in S and S-PLUS* (Springer, New York), p xvi.
11. Chan KS (1990) Testing for threshold autoregression. *Ann Stat* 18:1886–1894.
12. Samia NI, Chan KS (2011) Maximum likelihood estimation of a generalized threshold stochastic regression model. *Biometrika* 98:433–448.
13. Venables WN, Ripley BD (2002) *Modern Applied Statistics with S* (Springer New York), 4th Ed, p xi.

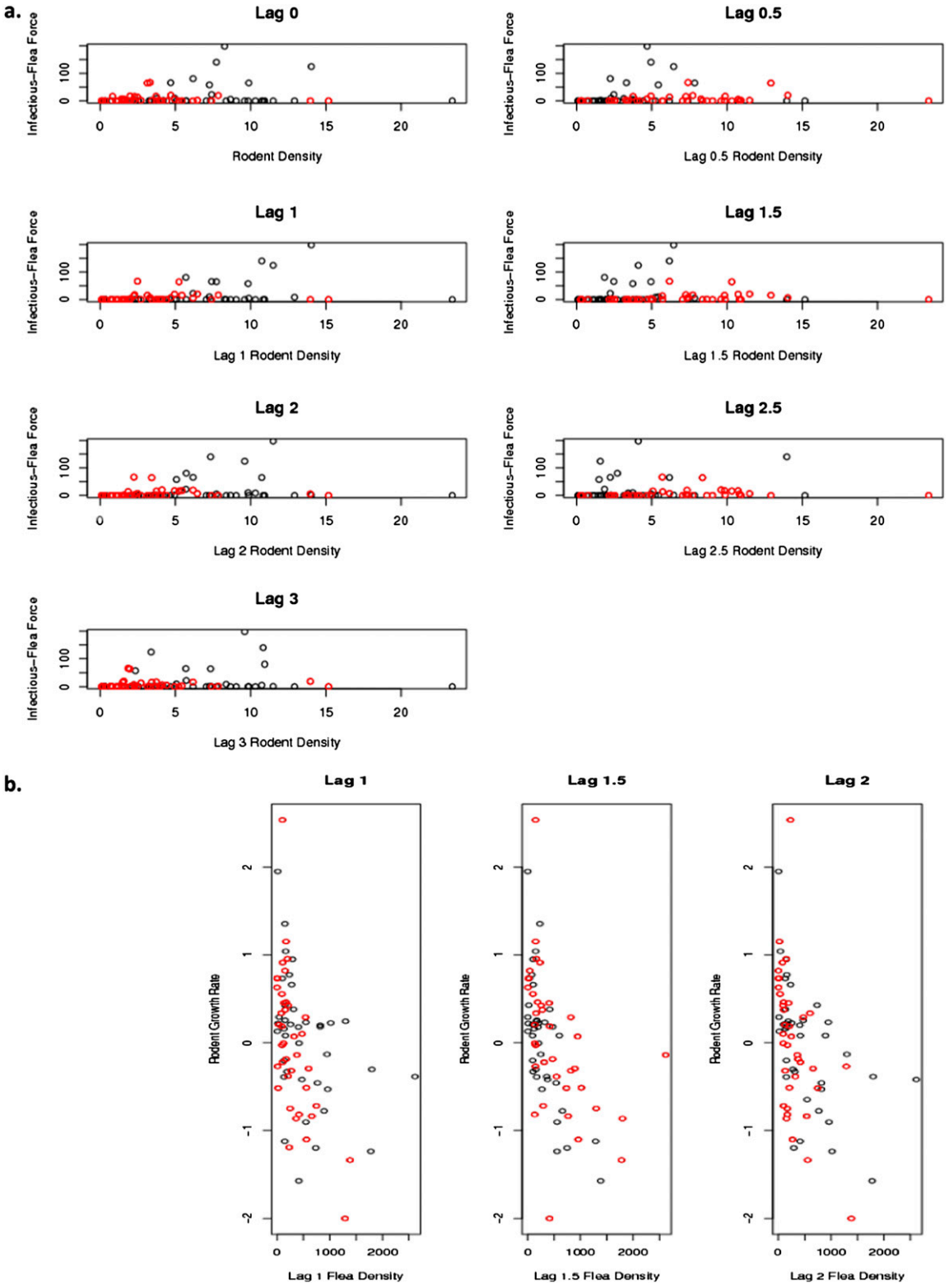

Fig. S1. (*A*) Plots of the infectious-flea force vs. the rodent density at different lags (lag = 0, 0.5, 1.0, 1.5, 2.0, 2.5, and 3.0). The red circles correspond to the spring season, and the black circles correspond to the fall season. (*B*) Plots of the rodent growth rate vs. the flea density at different lags (lag = 1.0, 1.5, and 2.0). The red circles correspond to the spring season, and the black circles correspond to the fall season.

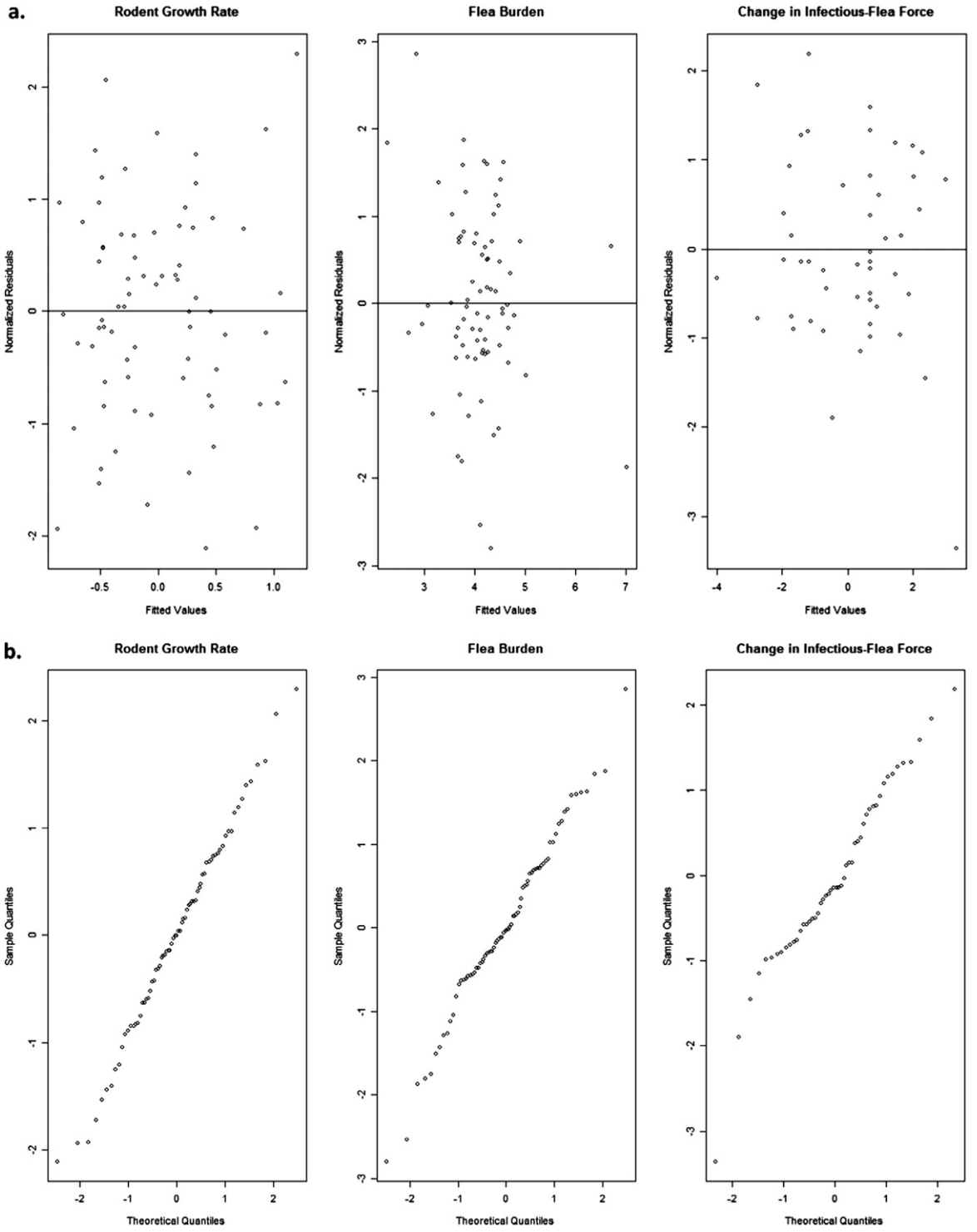

Fig. S2. (*A*) Plots of the normalized residuals vs. the fitted values for the rodent growth rate, the flea burden, and the change in infectious-flea force models. (*B*) Normal probability plots of the normalized residuals for the rodent growth rate, the flea burden, and the change in infectious-flea force models.

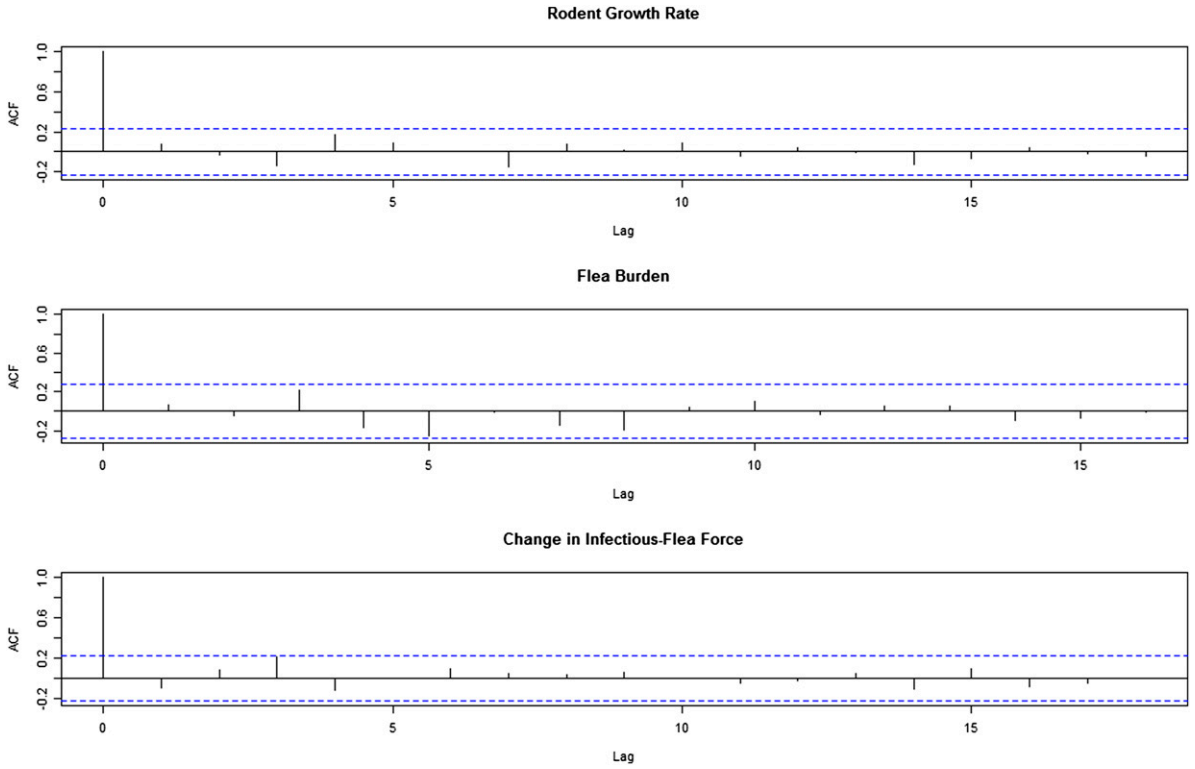

Fig. S3. Plots of the empirical autocorrelation function for the rodent growth rate, the flea burden, and the change in infectious-flea force models.

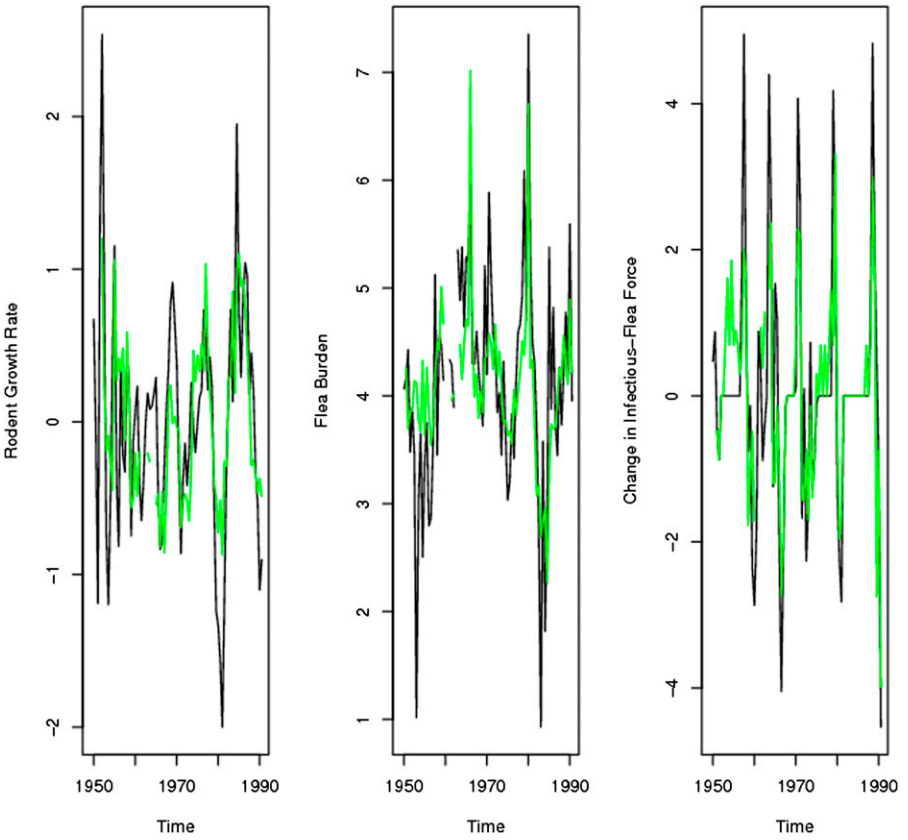

Fig. S4. Time series plots for the rodent growth rate (*Left*), the flea burden (*Center*), and the change in infectious-flea force (*Right*) models. The time series indicated in black are observed; the time series in green correspond to the fitted values.

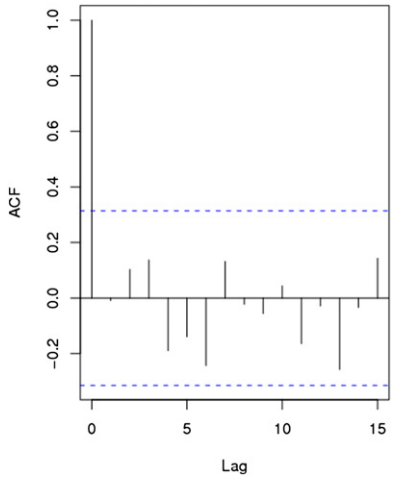

Fig. S5. ACF plot of the standardized residuals for the human plague model.

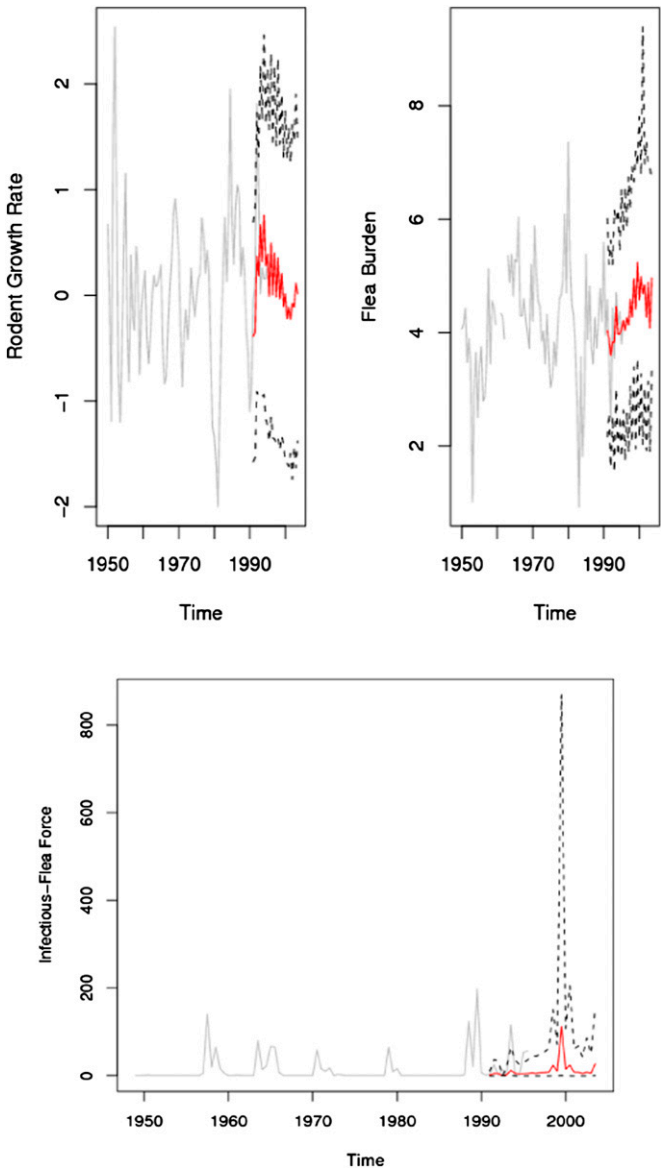

Fig. S6. Out of sample point and interval forecasts for the rodent growth rate (*Upper Left*), the (log) flea burden (*Upper Right*), and the infectious-flea force (*Lower*). The forecast origin is the fall season of 1990. The solid black line shows the actual observations, the solid red line shows the point forecasts, and the dashed black lines show the 95% interval forecasts.

Table S1. Notations and variables used in the full ecoepidemiological model summarized in Figs. 1*A* and 2*A*

Variable	Description	Definition
Seasonal reservoir model		
Sub- and superscripts *s*, *f*	Refer to spring and fall seasons, respectively	
$t - d$	Time $t - d$ (e.g., $t - 0.5$ refers to the previous season and $t - 1$ refers to the previous year)	
R_t	Rodent density at time t	
F_t	Flea density at time t	
G_t	Gerbil population yearly growth rate at time t	$G_t = \log(R_t/R_{t-1})$
B_t	Flea burden on the log scale at time t	$B_t = \log(F_t/R_t)$
I_t	Infectious-flea force at time t (computed as the product of flea density and prevalence of plague in rodents averaged across sites)	If $R_{t-l} < z$, I_t is a degenerate random variable such that $I_t = 0$, where l is estimated to be 1.5 y (2 y) if t is spring (fall), and z is estimated to be 3.25 (4.94) if t is spring (fall)
J_t	Change in infectious-flea force at time t	$\mathcal{J}_t = \log(1 + I_t) - \log(1 + I_{t-1})$
R_t^{su}	Summer precipitation at time t	Average amount of rainfall over June, July, and August
Yearly human plague model		
t	Year t	
$\mathcal{D}_t$	Rodent density in year t	Seasonal flea density averaged across all sites and then maximized across seasons
$\mathcal{F}_t$	Flea density in year t	Seasonal flea density averaged across all sites and then minimized across seasons (Table S2)
H_t	Number of infectious human cases in year t	
C_t	Flea burden in year t	Ratio of the flea density divided by the rodent density averaged across all sites and then maximized across seasons
$\mathcal{I}_t$	Infectious-flea force in year t	Product of flea density and prevalence of plague in rodents averaged across all sites and then maximized across seasons
T_t^{sp}	Spring temperature in year t	Average monthly temperature over March and April
T_t^{su}	Summer temperature in year t	Average monthly temperature over June, July, and August
R_t^{fa}	Fall precipitation in year t	Average amount of precipitation over September, October, and November

Table S2. Normalized AIC (NAIC) of models fitted with various threshold values

Threshold variable	Threshold lag	NAIC	Effective sample size
Minimum of spring and fall average flea density	0	4.44	38
Minimum of spring and fall average flea density	1	3.64	39
Minimum of spring and fall average flea density	2	4.69	37
Maximum of spring and fall average flea density	0	4.71	38
Maximum of spring and fall average flea density	1	4.17	39
Maximum of spring and fall average flea density	2	5.08	37
Mean of spring and fall average flea density	0	4.71	38
Mean of spring and fall average flea density	1	4.17	39
Mean of spring and fall average flea density	2	4.87	37