Cholera in the Lake Kivu region (DRC): integrating remote sensing and spatially-explicit epidemiological modeling

Flavio Finger,[1] Allyn Knox,[1] Enrico Bertuzzo,[1] Lorenzo Mari,[1,2] Didier Bompangue,[3,4] Marino Gatto,[2] Ignacio Rodriguez-Iturbe,[5] and Andrea Rinaldo[1,6]

Abstract. Mathematical models of cholera dynamics can not only help in identifying environmental drivers and processes that influence disease transmission, but may also represent valuable tools for the prediction of the epidemiological patterns in time and space as well as for the allocation of health care resources. Cholera outbreaks have been reported in the Democratic Republic of the Congo since the 1970s. They have been ravaging the shore of Lake Kivu in the east of the country repeatedly during the last decades. Here we employ a spatially explicit, inhomogeneous Markov chain model to describe cholera incidence in eight health zones on the shore of the lake. Remotely sensed datasets of chlorophyll *a* concentration in the lake, precipitation and indices of global climate anomalies are used as environmental drivers in addition to baseline seasonality. The effect of human mobility is also modelled mechanistically. We test several models on a multi-year dataset of reported cholera cases. The best fourteen models, accounting for different environmental drivers, and selected using the Akaike information criterion, are formally compared via proper cross-validation. Among these, the one accounting for seasonality, El Niño Southern Oscillation, precipitation and human mobility outperforms the others in cross-validation. Some drivers (such as human mobility and rainfall) are retained only by a few models, possibly indicating that the mechanisms through which they influence cholera dynamics in the area will have to be investigated further.

1. Introduction

The risk, loss and social disruption brought in by cholera outbreaks can hardly be overestimated and the global relevance of preventive assessments and controls of cholera spreading is manifest. The recent epidemics in Haiti, the Congo river basin, Cuba, Sierra Leone and the Sahel region [*Luque Fernández et al.*, 2009; *Kelvin*, 2011; *Bompangue et al.*, 2011; *Al-Tawfiq and Memish*, 2012; *Gaudart et al.*, 2013] witness the ongoing, widespread inadequacy of reliable drinking water supply and sanitation infrastructure all over the developing world. As a result, cholera remains a major cause of morbidity and mortality in developing countries even to date, despite all public health policies and humanitarian efforts deployed worldwide. As an example, according to the World Health Organization, as much as 85% increase in the number of reported cholera cases has been observed globally in 2011 relative to 2010, with 58 countries involved and a total of 589,854 yearly cases leading to an overall case fatality rate of 1.3% [*World Health Organization*, 2012].

To promote reliable and timely preventive assessments and controls of cholera spreading, and to evaluate emergency management alternatives, two main modeling approaches have been followed. One approach consists of predictive empirical models relying on environmental drivers which possibly influence the ecology of *Vibrio cholerae* [*Bouma and Pascual*, 2001; *Pascual et al.*, 2002; *Lipp et al.*, 2002; *Ruiz-Moreno et al.*, 2007; *Matsuda et al.*, 2008], often using remotely acquired information [*Lobitz et al.*, 2000; *de Magny et al.*, 2008; *Ford et al.*, 2009; *Akanda et al.*, 2009; *Jutla et al.*, 2010, 2013a, b]. Such methods, suited in particular to regions where cholera is endemic but applied to predict other infectious disease outbreaks as well [*Ford et al.*, 2009], have been shown to relate significant changes in remotely-acquired optical signatures to interannual and annual cyclic patterns of infections [*de Magny et al.*, 2008; *Emch et al.*, 2008; *Matsuda et al.*, 2008; *Jutla et al.*, 2013b]. For cholera, such signatures often consist of chlorophyll *a*, sea surface temperature (SST), sea surface height, precipitation, air temperature, and/or their anomalies (i.e. deviations from interannual averages).

Predictive empirical models have contributed new epidemiological perspectives; however, their use for the understanding, prediction and control of waterborne disease outbreaks is challenged by the fact that infection patterns are influenced by spatial structure and temporal asynchrony. Another type of approach relies on mechanistic models of disease spread, whether deterministic or stochastic, metacommunity- or individual-based. Such mathematical models, either spatially implicit [*Codeço*, 2001; *Koelle et al.*, 2005; *Hartley et al.*, 2006; *Riley*, 2007; *King et al.*, 2008; *Grad et al.*, 2012] or explicit [*Bertuzzo et al.*, 2008, 2010, 2011; *Tuite et al.*, 2011; *Chao et al.*, 2011; *Bertuzzo et al.*, 2012; *Mari et al.*, 2012a, b; *Rinaldo et al.*,

[1]Laboratory of Ecohydrology, Faculté de l'Environnement Naturel, Architectural et Construit, École Polytechnique Fédérale de Lausanne, 1015 Lausanne, Switzerland
[2]Dipartimento di Elettronica, Informazione e Bioingegneria, Politecnico di Milano, Milano, Italy
[3]Laboratoire Chrono-environnement, UMR6249, CNRS, University of Franche-Comte, Place Leclerc 25030 Besancon, France
[4]Laboratory of Microbiology, Faculty of Medicine, University of Kinshasa, BP: 834, Kinshasa, Democratic Republic of Congo
[5]Department of Civil and Environmental Engineering, Princeton University, Princeton NJ 65332, USA
[6]Dipartimento di Ingegneria Civile, Edile ed Ambientale, Università di Padova, 35131 Padova, Italy

0043-1397 /14/$9.00

2012; *Righetto et al.*, 2013], can lend key insights into the course of an ongoing epidemic, provide predictive frameworks and potentially aid real-time emergency management in allocating health care resources, also by anticipating the impact of alternative interventions. A selection of recent applied cholera models and their main characteristics is presented in Table 1.

A spatially-explicit approach, made possible by the now widespread access to data mapping of hydrological drivers, transportation and sanitation infrastructure and population distribution, can furthermore address the spatiotemporal evolution of disease propagation as well as the precise conditions under which a waterborne disease epidemic grows or dies out. Significantly, it has recently been demonstrated formally [*Gatto et al.*, 2012, 2013] that spatially implicit stability conditions (based on local reproduction numbers) are neither necessary nor sufficient for outbreaks to occur owing to spatial effects. In particular, spatial phenomena are bound to become fundamental when local settlements are connected by networks of primary (environmental to human) and secondary (human to human) infection mechanisms. Networked connectivity models, describing the interplay between hydrology, epidemiology, and social behavior sustaining human mobility, thus prove to be key tools for prediction and emergency management of waterborne infections.

Climate variables that have been related to cholera can be categorized into global and local, where global variables may mediate local ones [*de Magny et al.*, 2006]. Global scale climate phenomena, such as the El Niño Southern Oscillation (ENSO), have been reported to influence cholera dynamics mostly in endemic regions such as the Bay of Bengal [*Colwell*, 1996; *Pascual et al.*, 2000; *Lipp et al.*, 2002; *Pascual et al.*, 2002; *Rodò et al.*, 2002; *Koelle et al.*, 2005; *Pascual et al.*, 2008], but also in African countries [*de Magny et al.*, 2006; *Olago et al.*, 2007; *Bompangue et al.*, 2011]. However, the causative mechanistic links between ENSO and disease dynamics remains controversial to date. Some authors point out correlations between cholera incidence and suitably delayed phyto- and zooplankton abundances in coastal waters, which in turn are driven by local climate anomalies, mediated by ENSO [*de Magny et al.*, 2008; *Ford et al.*, 2009; *Jutla et al.*, 2010; *Bompangue et al.*, 2011; *Mishra et al.*, 2011; *Jutla et al.*, 2013a]. It is in fact known that *V. cholerae* may attach to zooplankton [*Colwell*, 1996]. This hypothesis thus places great importance on the role of aquatic environmental reservoirs in maintaining the disease. Others argue that the local climate anomalies caused by ENSO may as well influence disease dynamics via other pathways, such as droughts, water salinity, or human behaviour and population dynamics [*Pascual et al.*, 2002; *Rodò et al.*, 2002; *Rebaudet et al.*, 2013].

While the role of local climatic conditions, rainfall, air temperature and sea surface temperature (SST) in particular, on patterns of cholera transmission has long been studied especially in empirical frameworks [*Lipp et al.*, 2002; *Koelle et al.*, 2005; *Altizer et al.*, 2006; *de Magny et al.*, 2008], spatially explicit mechanistic models of cholera epidemics have incorporated hydroclimatological drivers only more recently, most notably in models used to study the course of the Haitian epidemic, starting from the very first months after its outbreak in late 2010 and following disease resurgence (May 2011) in connection with unusually intense tropical rains [*Rinaldo et al.*, 2012; *Eisenberg et al.*, 2013; *Gaudart et al.*, 2013; *Righetto et al.*, 2013]. Possible mechanisms of enhanced cholera spread due to heavy rains include increased bacterial concentration in drinking water due to failure of sanitation systems, washout of open-air defecation sites [*Rinaldo et al.*, 2012; *Gaudart et al.*, 2013], or the seasonal modification of human water sources and human behaviour [*Gaudart et al.*, 2013], possibly including enhanced exposure owing to crowding effects.

Here we use a semi-mechanistic, spatially-explicit modeling framework to describe cholera dynamics around Lake Kivu, Democratic Republic of the Congo (DRC). Our approach builds on the multidimensional inhomogeneous Markov chain (MDIMC) method proposed by *Reiner et al.* [2012]. This method requires the discretization of the variable to be modeled (i.e. cholera incidence) into a finite number of states, and applies a semi-mechanistic description of the transitions between discrete dynamical states. Transition probabilities vary in time as they account for environmental drivers (estimated through remotely sensed and objectively manipulated datasets) and human mobility patterns. The case study at hand refers to the regions adjacent to Lake Kivu (eastern DRC). Routinely collected surveillance data have been used to construct epidemic curves of cholera cases and map the spatio-temporal progress of the disease [*Bompangue et al.*, 2009]. Datasets of precipitation, chlorophyll *a* concentration in Lake Kivu as well as indices of global climate phenomena are used as model input, together with a mechanistic description of human mobility among the health zones adjacent to the lake. Specifically, the MDIMC model is fed with all possible combinations of environmental drivers, with variable lags. The performances of different model settings are compared using formal model selection techniques in order to draw conclusions about the relative importance of environmental drivers for the proliferation of cholera in the study area. Furthermore, cross-validation is applied to assess the possibility of predictive modelling of epidemiological dynamics based on environmental data.

2. Case Study

2.1. Spatial Setting

Lake Kivu is situated in eastern DRC on the border with Rwanda (Figure 1). In this study we concentrate on eight health zones (or their aggregations) located on the Congolese shore, which include the two major cities of Goma and Bukavu, respectively at the northern and southern ends of the lake. Areas further from the lake are not considered due to their low number of cholera cases and limited population. The total population size of the study area is of about 1.8 millions.

2.2. Pathogen Transport

The Lake Kivu catchment consists of numerous small subcatchments along steep slopes leading down to the lake and a northern region characterized by porous volcanic soils that allow for little (to no) surface run-off. Therefore we assume that the hydrological transport of the pathogen (sensu *Bertuzzo et al.* [2010] and *Rinaldo et al.* [2012]) is negligible at the regional scale. However, rainfall can facilitate local pathways of transmission and/or amplify contamination through failure of inappropriate sanitation systems. The health zones in our model are connected through human mobility fluxes, simulated here by a gravity model. Given the stark difference in sociopolitical stability between the eastern DRC and neighboring Rwanda, and the low number of cases reported in Rwanda during the study period, we assume fluxes between the two countries to be negligible.

2.3. Climate

Local climate in the study area is characterized by a rainy season from October to May [*Plisnier et al.*, 2000; *Bompangue et al.*, 2009], which is interrupted by a short dry period early in the year. The annual precipitation corresponds to around 1 200 mm. Monthly average temperatures are fairly constant, close to 20 °C throughout the year (available

from: *http://en.climate-data.org/location/1074/*). *Plisnier et al.* [2000] reported highly complex and spatially differentiated effects of ENSO on local climate, such as a positive correlation of ENSO with rainfall, air pressure and temperature in the area. According to *Stager et al.* [2007] ENSO also influences lake levels in the African Great Lakes Region. *Marchant et al.* [2007] found that the the Indian Ocean Dipole (IOD), a cyclic climate phenomenon independent of ENSO, influences the rainfall in East Africa too. This is in accordance with *Becker et al.* [2010], who revealed an influence of ENSO and IOD on the total water storage in the area. The exact mechanisms leading to all the above so-called teleconnections are yet to be determined [*Plisnier et al.*, 2000; *Marchant et al.*, 2007].

2.4. Data

2.4.1. Cholera

Weekly cholera incidence data (2004–2011) were made available through the work of *Bompangue et al.* [2009]. Briefly, data were collected from registries at each Cholera Treatment Center, aggregated weekly and reported to the Ministry of Health officials of each health zone, where they were preserved in electronic or paper format [*Piarroux and Bompangue*, 2007; *Bompangue et al.*, 2008; *Bompangue*, 2009; *Piarroux et al.*, 2009; *Bompangue et al.*, 2009, 2011, 2012]. For the purpose of this study we aggregated the data to obtain monthly numbers of cases for the eight lakeside health zones described above (Figure 2).

The time scale of this study has been chosen to be monthly because of the level of noise and the number of missing values in both reported cholera cases and remotely sensed Chlorophyll *a* concentrations. Note for instance that in order to get a high-quality time series of Chlorophyll *a* data, every time step must contain a certain number of cloudless days, which is sometimes difficult to enforce during the rainy season in the study region.

2.4.2. Demography

A remotely sensed dataset of the estimated 2010 population distribution (available from `http://www.worldpop.org.uk`) was used to approximate the population of each health zone, serving as a base to compute monthly cholera incidence (reported cases divided by population abundance of each health zone).

2.4.3. Plankton

The optimization of remotely-sensed plankton biomass estimates for Lake Kivu, described in *Knox et al.* [submitted], enabled the selection of the plankton biomass proxy best-suited for this study. Here, we use a chlorophyll *a* database generated with the OC3 bio-optical algorithm and a coastal atmospheric correction model with 90% relative humidity, spanning the years 2002–2012 (Figure 3). Daily data were spatially averaged across the entire lake, and monthly averages were created by weighting each day by the number of data yielding pixels.

2.4.4. Precipitation

Daily precipitation fields were obtained from a remotely sensed dataset by the National Aeronautics and Space Agency (NASA) [*Huffman et al.*, 2010]. The resolution of the dataset is 0.25 degrees of latitude and longitude. Precipitation estimates were then projected to each health zone and aggregated monthly (Figure 3). The projection was done by assigning the corresponding precipitation value to each cell in a rasterized version of the health zones delimitation and subsequently taking the mean over each health zone.

2.4.5. ENSO and IOD

In order to account for possible relations between global climate anomalies and the dynamics of the disease in the study region, as reported by *Bompangue et al.* [2011], we included two additional climatic drivers in our study (Figure 3). SST anomaly from the Niño 3.4 region made available by the National Oceanic and Atmospheric Administration (NOAA, available online at `http://www.cpc.ncep.noaa.gov/data/indices/sstoi.indices`), was used as index for ENSO. For IOD we used the so called Dipole Mode Index (DMI) [*Saji et al.*, 1999], a measure of the SST gradient between two regions in the Indian Ocean (available online at `http://www.jamstec.go.jp/frcgc/research/d1/iod/DATA/dmi.monthly.ascii`).

In order to quantify correlations between the different environmental drivers (precipitation, chlorophyll *a*, ENSO and IOD), we computed their cross-correlation functions. Precipitation and chlorophyll *a* show low significant correlations ($r \approx 0.2$) between lags 0 and 2 months, whereas DMI and ENSO are weakly correlated at lags around 1 month ($r \approx 0.2$) as well as anti-correlated at higher lags. All other combinations do not show significant correlations ($p > 0.05$).

3. Model

3.1. Inhomogeneous Markov Chain Model for Endemic Cholera

The theoretical framework adopted here builds on a previous semi-deterministic modeling approach (MDIMC) for endemic cholera developed by *Reiner et al.* [2012] that is based on finite-state Markov chain modeling. This approach requires cholera incidence data to be categorized into discrete states. It assigns a probability to the transitions between epidemic states in a given spatial setting and with a defined time step. The Markov chain model can be made inhomogeneous by allowing transition probabilities to depend on temporal and spatial environmental drivers, namely precipitation, chlorophyll *a* concentration in the lake, ENSO and IOD, as well as on spatial interactions induced by human mobility.

Following *Reiner et al.* [2012] we categorize cholera incidence into three discrete states, namely low (state 1), mild (state 2) and high (state 3). We select the monthly incidence thresholds for the definition of the three discrete states such that half of the records fall in the low incidence category and the remaining data are equally partitioned into the remaining two categories. Therefore the thresholds correspond to the 50^{th} (0.0133% incidence) and the 75^{th} (0.0415%) percentiles of the monthly incidence data, respectively.

Figure 2 shows cholera incidence in the eight health zones considered, and the category each data point belongs to. Low (state 1), mild (state 2) and high (state 3) incidence are denoted by blue, green and red color respectively. The categorization of the data into finite classes causes the merging of the events in the tail of the incidence distribution with less severe ones. Although some information is therefore discarded, the model focuses on levels of variation that are relevant to public health.

We first define a baseline homogeneous Markov chain model that does not account for environmental drivers and spatial interactions. Let $X_{k,t} = 1, 2, 3$ be the state of health zone k at time t. According to this model, the generic transition $(X_{k,t} = i) \to (X_{k,t+1} = j)$ occurs, at any time t, with probability $p_{i,j}$. The baseline model is therefore completely defined by the following transition probability matrix $\mathcal{P}$:

$$\mathcal{P} = \begin{bmatrix} p_{1,1} & (1 - p_{1,1} - p_{1,3}) & p_{1,3} \\ p_{2,1} & (1 - p_{2,1} - p_{2,3}) & p_{2,3} \\ p_{3,1} & (1 - p_{3,1} - p_{3,3}) & p_{3,3} \end{bmatrix} \quad (1)$$

where the probabilities of transition to state 2 are expressed so as to enforce $\mathcal{P}$ to be a stochastic matrix (i.e. row-wise sums equal to 1).

We further assume that environmental drivers and spatial interactions among neighboring human communities can modify baseline transition probabilities (1). In particular, we hypothesize that these external drivers can modify, through a multiplicative factor, the probabilities of transition from low/mild cholera incidence to a worse state ($p_{i,j},\ i<j$), as well as the probability of remaining in the highest incidence state ($p_{3,3}$). The resulting transition probabilities $p'_{i,j,k,t}$ are thus site- and time-specific, as both environmental drivers and cholera incidence patterns vary in space and time. We adopt the following formulation:

$$\begin{aligned} p'_{i,j,k,t} =& p_{i,j}\,(1+f^{rain}_{k,t})(1+f^{chl}_{t})(1+f^{ENSO}_{k,t}) \\ & (1+f^{IOD}_{k,t})(1+f^{mob}_{k,t})(1+f^{sea}_{t}) \\ & \text{for } \ i<j \ \text{ or } i=j=3\,. \end{aligned} \tag{2}$$

The remaining probabilities are adjusted, proportionally to their baseline values, to ensure that matrix $\mathcal{P}'_{k,t}=(p'_{i,j,k,t})$ is stochastic, i.e.:

$$\begin{aligned} p'_{1,1,k,t} &= 1-p'_{1,2,k,t}-p'_{1,3,k,t} \\ p'_{i,1,k,t} &= (1-p'_{i,3,k,t})\,\frac{p_{i,1}}{p_{i,1}+p_{i,2}} \quad \text{for } i=2,3 \\ p'_{i,2,k,t} &= (1-p'_{i,3,k,t})\,\frac{p_{i,2}}{p_{i,1}+p_{i,2}} \quad \text{for } i=2,3\,. \end{aligned} \tag{3}$$

The term $f^{rain}_{k,t}$ is assumed to be linearly dependent on the actual precipitation intensity $f^{rain}_{k,t}=\alpha J_k(t)$, where $J_k(t)$ is the mean normalized monthly precipitation of health zone k during month t. Precipitation data have been normalized to span the range $[0,1]$. Therefore, to enforce $p'_{i,j,k,t}>0$ we impose the constraint $\alpha \geq -1$. As an example, if $\alpha>0$, rainfall enhances cholera transmission and therefore all the probabilities to make a transition to a higher incidence state (or to stay at the highest) increase. Consequently, all the other transition probabilities decrease. Analogously, the potential effect of chlorophyll a concentration on cholera transmission is modelled as $f^{chl}_{t}=\beta C\ (t-t_C)$ ($\beta\geq -1$), where $C(t)$ is the mean normalized (i.e. rescaled in the range $[0,1]$) monthly chlorophyll a concentration of month t. The lag t_C is introduced to possibly account for a delay between the dynamics of phyto- and/or zoo-plankton and favorable conditions for bacteria survival in the lake. To account for the possible enhancing effect of climatic drivers on disease dynamics we model the terms f^{ENSO}_{t} and f^{IOD}_{t} equivalently to $f^{chl}_{k,t}$, i.e. $f^{ENSO}_{t}=\phi ENSO(t-t_{ENSO})$ and $f^{ENSO}_{t}=\psi IOD(t-t_{IOD})$, where $ENSO(t)$ is the normalized SST anomaly in the Niño 3.4 region during month t, $IOD(t)$ is the normalized DMI (see section 2.4.5) during month t, t_{ENSO} and t_{DMI} are time lags and $\phi\geq -1$ as well as $\psi\geq -1$ are proportionality constants.

Cholera transmission in a health zone can also be enhanced by the mobility of people toward health zones with ongoing outbreaks. This potential effect is accounted for in equation (2) by the term $f^{mob}_{k,t}$, which reads

$$f^{mob}_{k,t}=\gamma\ \sum_{z\neq k} Q_{kz}X^{\nu}_{z,t}\,,$$

where Q_{kz} is the probability that a traveller from zone k visits zone z, and γ and ν are two positive parameters. We model human mobility through a gravity model [*Erlander and Stewart*, 1990]. Accordingly, connection probabilities are defined as

$$Q_{kz}=\frac{H_z e^{-d_{kz}/D}}{\sum_{n\neq k} H_n e^{-d_{kn}/D}}\,,$$

where the attractiveness factor of zone z depends on its population size H_z, while the deterrence factor is assumed to be dependent on the distance d_{kz} between the two communities and represented by an exponential kernel (with shape factor D). Distances between health zones are measured along the road network.

Finally the term $f^{sea}_{k,t}$ in equation (2) accounts for the seasonality possibly induced by drivers other than those explicitly considered above. Baseline seasonality is modelled through a simple sinusoidal function:

$$f^{sea}_{t}=\delta\left(1+\sin\left(2\pi\frac{t-t_s}{12}\right)\right)\,,$$

where $\delta\geq -1$ and t_s is the lag of seasonality.

3.2. Model Calibration and Validation

We consider all the $2^6=64$ model combinations obtained by accounting for or neglecting the effects of rainfall, chlorophyll a, human mobility, ENSO, IOD and baseline seasonality. Models are fitted by maximizing their likelihood. In the most complex setting we must optimize the values of 17 parameters, including the set of lags (if applicable) up to 6 months that produces the best fit to data. Under the Markovian assumption of the model, the transition from one month to the next is independent of all other transitions. Therefore, likelihood can be defined as the product of the probabilities of the transitions actually observed for each month. We use the simplex search algorithm proposed by *Nelder and Mead* [1965] to maximize the log-likelihood. We enforce the constraint that each transition probability must be between 0 and 1 by a barrier method, i.e. we set likelihood to 0 whenever a transition probability falls outside these limits [*Reiner et al.*, 2012]. Because the Nelder-Mead method can only ensure the identification of local stationary points of the considered objective function, the optimization algorithm is run 100 times with different initial starting points to better approximate the global maximum of the likelihood function. The best model is then selected out of all candidate model combinations through the Akaike information criterion (AIC) which evaluates model performance and discounts for complexity.

The Markovian nature of the model also allows to easily implement a simulation algorithm. Let us consider a generic health zone k with discretized cholera incidence i at time t, i.e. $X_{k,t}=i$. Knowing the epidemic state of the other health zones and the magnitude of the environmental forcing at the same time t, it is possible to compute the transition probabilities $p'_{i,j,k,t}$ for $j=1,2,3$ through equations 2 and 3. A random variable U, uniformly distributed in the $[0,1]$ interval, is drawn to determine which transition occurs. If $U<p'_{i,1,k,t}$, the considered health zone transitions to the low incidence state in the next month, i.e. $X_{k,t+1}=1$. Otherwise, if $U<p'_{i,1,k,t}+p'_{i,2,k,t}$, a transition to a mild cholera state occurs, i.e. $X_{k,t+1}=2$. In the remaining case the transition is to a high incidence state, i.e. $X_{k,t+1}=3$. Repeating this procedure for all the health zones gives a 1-month time step simulation. The simulated state can be used to advance the chain for another time step and so on to simulate the model for any number of time steps.

To evaluate the predictive ability of the different models, we perform a validation analysis. Specifically, we perform leave-one-out cross-validation, i.e. we remove one month of data for all the health zones and recalibrate the model being tested using the remaining data. Starting from the state of the system observed in the month before the one removed, we simulate the model for one time step, using the newly calibrated parameter set, and compare model prediction to the removed data. The accuracy of the different models in validation is estimated by computing the likelihood of the

observed state. To that end, we infer the probability distribution of the predicted state performing 10,000 independent simulations. This procedure is then sequentially repeated removing, once at a time, all the monthly data points available. We also perform a validation analysis removing 2 and 3 contiguous months of data at a time. In this case the chain is advanced for 2 and 3 time steps, respectively. Performance is evaluated through the likelihood of the state observed in the latest month removed. Note that likelihood values evaluated at different lags or between validation and calibration runs cannot be compared because of different numbers of data points.

4. Results

Table 2 shows the results of the calibration procedure described above. The 20 best combinations of model components are shown, ranked according to their AIC score. All 64 possible combinations are shown in Supplementary Table 1. The best ranked model accounts for the effects of SST anomalies and seasonality only (parameters are shown in Table 3). However, models number 2 to 14 have an AIC score close to that of the top-ranked candidate ($\Delta\text{AIC} < 4$) and thus cannot be safely discarded [*Burnham and Anderson*, 2002]. All these models include seasonality. To test the significance of the individual components of models 1 to 14 (alternative hypothesis) against the model including seasonality only (number 8, null hypothesis) we employ a likelihood ratio test. Improvements in likelihood for models 1 and 2 are significant, and so are the effects of ENSO alone, as well as ENSO combined with precipitation ($p < 0.05$). Improvements in likelihood for models 4 to 7 and 9 to 14 are not significant at $p = 0.05$.

Model validation is performed using the 14 models retained in model selection. Table 4 shows log-likelihood values obtained by applying cross-validation at lags of one, two and three months. Model 13 has the highest likelihood values for all lags. Figure 4 shows the validation of model 13 at respective lags of one, two and three months. Note the decreasing accuracy of the median as predictor of cholera incidence as well as the higher uncertainty of the simulations as the lag increases. A less formal but more intuitive measure of model accuracy is the fraction of times in which the model correctly predicts the observed cholera incidence state. If we assume the mode of the distribution over 10'000 runs as the best predictor, model 13 predicts 68% of the state correctly at lag one month, 62% at lag two months and 59% at lag three months.

In addition to the results reported above, we also tested the effect of adding the water surface temperature of Lake Kivu [*MacCallum and Merchant*, 2012; *Thiery et al.*, 2014a, b] as a further explanatory variable. No significant improvements were found (result not shown for brevity).

5. Discussion

In this work we have applied a MDIMC-based approach to model cholera dynamics in eight health zones in the Lake Kivu region (DRC). The semi-deterministic framework chosen allows for a mechanistic description of processes such as human mobility or the enhancing effect of rainfall on disease transmission, as well as for an explicit treatment of space. Its discrete nature allows to characterize spatiotemporal cholera dynamics robustly, even if the reported case-data available present high uncertainties because of over- and under-reporting and missing records. This robustness is especially important in endemic regions such as the eastern DRC, where incidence is generally lower than in epidemic settings. Conversely, classical SIR-type models (like, e.g., the one applied in *Rinaldo et al.* [2012] to describe the Haiti cholera epidemic) do heavily rely on detailed epidemiological datasets for parameter estimation and are thus very difficult to apply to the current case at a fine spatial resolution because of the low signal to noise ratio.

Several models were retained during model selection. All of them account for seasonality, which is thus found to be an important factor to explain endemic cholera transmission in the study area. The model that performed best according to AIC accounts for the effect of ENSO in addition to seasonality. During validation, though, a more complex model, including also the effects of mobility and precipitation, proved to perform best. This might indicate that higher complexity in this case does not lead to overfitting but to improved predictive abilities.

Global climate anomalies (ENSO and IOD) seem to be the most important environmental factors, as they appear in all but two of the best performing models. Because of the known interactions between these anomalies and local climate [*Plisnier et al.*, 2000; *Stager et al.*, 2007; *Marchant et al.*, 2007] this is not surprising. The effect of precipitation alone did not prove significant compared to a model with seasonality only. However, its combination with ENSO and IOD is significant, presumably because of rainfall deteriorating sanitary conditions.

Chlorophyll *a* concentration in Lake Kivu as a driver of cholera dynamics was retained only by 5 of the models selected in calibration. Apart from the interaction between *V. cholerae* and plankton, possibly being more complex than presumed here, significant correlations might be clouded by other environmental factors such as precipitation, or by the fact that the estimates of chlorophyll *a* concentrations used here are of insufficient accuracy and/or spatial detail. The subtleties involved in the remote sensing of chlorophyll *a* concentrations in lakes indeed deserve further investigations [*Knox et al.*, submitted]. However, we maintain that the search for remotely sensed proxies for parameters of mechanistic epidemiological models is an important field of study towards a new concept of mathematical epidemiology.

The effect of human mobility has been retained only by one of the models selected in calibration, which may be an artifact of our decision to limit the study to the lakeside region of Lake Kivu, with only two distant population centers separated by mostly rural areas. Indeed, the most important mobility patterns identified in the eastern DRC are between the lakeside regions and non-lakeside regions [*Bompangue et al.*, 2009] and are thus not accounted for in this study. In addition, since the description of mobility in the model includes three parameters, it is penalized in model selection relative to other model components. Note however that the best performing model in validation accounts for human mobility as well, which thus proves to be important for epidemiological projections.

All models including seasonality have shown to perform better than models not including it. This can be interpreted as a clear indication that other environmental and/or social factors, which have not been explicitly accounted for, might play a major role in the dynamics of the disease in the study area. In particular, the influence of population movements due to war, civil unrest or seasonal migrations [*Bompangue et al.*, 2009], particularly from and to other endemic areas in the country, would merit further investigation.

We have been able to show that cholera incidence in the region is influenced by global (ENSO, IOD) and local (rainfall) climatic variables. Thus it seems clear that climatic and environmental conditions play an important role in the disease dynamics. Our results do not support the hypothesis of phyto- and zooplankton being a major factor for persistence and proliferation of the disease in this area, as indicated by the lack of any significant effect of chlorophyll *a*. Alternative explanations exist, such as the influence of climate

on human behaviour and metapopulation dynamics [*Pascual et al.*, 2002; *Rodò et al.*, 2002; *Bompangue et al.*, 2011; *Rebaudet et al.*, 2013].

Note that, although the framework applied here allows for prediction of epidemic state a few months in advance, our results did not identify specific environmental drivers with long lead times. Thus, in order to use the model for the purpose of longer-term predictions, projections of environmental drivers have to be made first.

Our results provide further evidence that different geographic and social contexts call for different dominant infection mechanisms, and hence for proxies and modeling approaches shifting on a case-dependent basis. As an example, chlorophyll *a* acts as a suitable proxy in Bangladesh, whereas rainfall is the most important environmental driver in Haiti. The proposed modeling framework is flexible and capable of selecting the dominant infection mechanisms. Thus it can easily be exported to the study of other regions.

Acknowledgments. The authors would like to thank Mercedes Pascual and two anonymous reviewers for their valuable comments. Moreover we would like to thank Wim Thiery (KU Leuven) and his colleagues for the lake water surface temperature data they provided.

E.B., L.M. and A.R. acknowledge the support provided by the European Research Council (ERC) advanced grant program through the project "River networks as ecological corridors for species, populations and waterborne disease" (RINEC 227612). F.F., E.B., L.M. and A.R. acknowledge the support from the Swiss National Science Foundation (SNF/FNS) project "Dynamics and controls of large-scale cholera outbreaks" (DYCHO CR23I2_138104). Data are available upon request.

References

Akanda, A. S., S. Jutla, and S. Islam (2009), Dual peak cholera transmission in Bengal Delta: a hydroclimatological explanation, *Geophysical Research Letters*, *36*, L19,401.

Al-Tawfiq, J., and Z. Memish (2012), The Hajj: updated health hazards and current recommendations for 2012, *Eurosurveillance*, *17*, 6–10.

Altizer, S., A. Dobson, P. Hosseini, P. Hudson, M. Pascual, and P. Rohani (2006), Seasonality and the dynamics of infectious diseases, *Ecology Letters*, *9*, 467–484.

Andrews, J. R., and S. Basu (2011), The transmission dynamics and control of cholera in Haiti: An epidemic model, *Lancet*, *377*, 1248–1255.

Becker, M., W. LLovel, A. Cazenave, A. Güntner, and J.-F. Crétaux (2010), Recent hydrological behavior of the east african great lakes region inferred from GRACE, satellite altimetry and rainfall observations, *Comptes Rendus Geoscience*, *342*(3), 223–233.

Bertuzzo, E., M. Gatto, A. Maritan, S. Azaele, I. Rodriguez-Iturbe, and A. Rinaldo (2008), On the space-time evolution of a cholera epidemic, *Water Resources Research*, *44*, W01,424.

Bertuzzo, E., R. Casagrandi, M. Gatto, I. Rodriguez-Iturbe, and A. Rinaldo (2010), On spatially explicit models of cholera epidemics, *Journal of the Royal Society Interface*, *7*, 321–333.

Bertuzzo, E., L. Mari, L. Righetto, M. Gatto, R. Casagrandi, M. Blokesch, I. Rodriguez-Iturbe, and A. Rinaldo (2011), Prediction of the spatial evolution and effects of control measures for the unfolding Haiti cholera outbreak, *Geophysical Research Letters*, *38*, L06,403.

Bertuzzo, E., L. Mari, L. Righetto, M. Gatto, R. Casagrandi, I. Rodriguez-Iturbe, and A. Rinaldo (2012), Hydroclimatology of dual-peak cholera epidemics: inferences from a spatially explicit model, *Geophysical Research Letters*, *39*, L05,403.

Bompangue, D. (2009), Dynamique des epidemies de cholera dans la region des grands lacs africains: cas de la Republique Democratique du Congo, Ph.D. thesis, University of Franche-Compte.

Bompangue, D., P. Giraudoux, P. Handschumacher, M. Piarroux, B. Sudre, M. Ekwanzala, I. Kebela, and R. Piarroux (2008), Lakes as sources of cholera outbreaks, Democratic Republic of Congo, *Emerging Infectious Diseases*, *14*, 798–800.

Bompangue, D., P. Giraudoux, M. Piarroux, G. Mutombo, R. Shamavu, B. Sudre, A. Mutombo, V. Mondonge, and R. Piarroux (2009), Cholera epidemics, war and disasters around Goma and Lake Kivu: An eight-year study, *PLoS Neglected Tropical Diseases*, *3*, e436.

Bompangue, D., P. Giraudoux, P. D. Plisnier, A. M. Tinda, M. Piarroux, B. Sudre, S. Horion, J. M. Tamfum, B. K. Ilunga, and R. Piarroux (2011), Dynamics of cholera outbreaks in Great Lakes region of Africa, 1978–2008, *Emerging Infectious Diseases*, *17*, 2026–2036.

Bompangue, D., S. M. Vesenbeckh, P. Giraudoux, M. Castro, J. J. Muyembe, B. K. Ilunga, and M. Murray (2012), Cholera ante portas – the re-emergence of cholera in Kinshasa after a ten-year hiatus, *PLoS Currents Disasters*, *4*.

Bouma, M. J., and M. Pascual (2001), Seasonal and interannual cycles of endemic cholera in Bengal 1891–1940 in relation to climate and geography, *Hydrobiologia*, *460*, 147–156.

Burnham, K. P., and D. R. Anderson (2002), *Model selection and multimodel inference: a practical information-theoretic approach*, Springer-Verlag, New York, NY.

Chao, D. L., M. E. Halloran, and I. M. Longini Jr. (2011), Vaccination strategies for epidemic cholera in Haiti with implications for the developing world, *Proceedings of the National Academy of Sciences USA*, *108*, 7081–7085.

Codeço, C. T. (2001), Endemic and epidemic dynamics of cholera: the role of the aquatic reservoir, *BMC Infectious Diseases*, *1*, 1.

Colwell, R. R. (1996), Global climate and infectious disease: the cholera paradigm, *Science*, *274*(5295), 2025–2031.

de Magny, G. C., B. Cazelles, and J.-F. Guégan (2006), Cholera threat to humans in ghana is influenced by both global and regional climatic variability, *EcoHealth*, *3*(4), 223–231.

de Magny, G. C., R. Murtugudde, M. R. P. Sapiano, A. N. C. W. Brown, A. J. Busalacchi, M. Yunus, M. andG. B. Nair, A. I. Gil, C. F. Lanata, J. Calkins, B. M. K. Rajendran, M. K. Bhattacharya, A. Huq, R. B. Sack, and R. R. Colwell (2008), Environmental signatures associated with cholera epidemics, *Proceedings of the National Academy of Sciences USA*, *105*, 17,676–17,681.

Eisenberg, M. C., G. Kujbida, A. R. Tuite, D. N. Fisman, and J. H. Tien (2013), Examining rainfall and cholera dynamics in Haiti using statistical and dynamic modeling approaches, *Epidemics*, *5*, 197–207.

Emch, M., C. Feldacker, M. Yunus, P. K. Streatfield, V. Dinh Thiem, D. G. Canh, and M. Ali (2008), Local environmental predictors of cholera in Bangladesh and Vietnam, *American Journal of Tropical Medicine and Hygiene*, *78*, 823–832.

Erlander, S., and N. F. Stewart (1990), *The Gravity Model in Transportation Analysis – Theory and Extensions*, VSP Books, Zeist, The Netherlands.

Ford, T. E., R. R. Colwell, J. B. Rose, S. S. Morse, D. J. Rogers, and T. L. Yates (2009), Using satellite images of environmental changes to predict infectious disease outbreaks, *Emerging Infectious Diseases*, *15*, 1341–1346.

Gatto, M., L. Mari, E. Bertuzzo, R. Casagrandi, L. Righetto, I. Rodriguez-Iturbe, and A. Rinaldo (2012), Generalized reproduction numbers and the prediction of patterns in waterborne disease, *Proceedings of the National Academy of Sciences USA*, *48*, 19,703–19,708.

Gatto, M., L. Mari, E. Bertuzzo, R. Casagrandi, L. Righetto, I. Rodriguez-Iturbe, and A. Rinaldo (2013), Spatially explicit conditions for waterborne pathogen invasion, *American Naturalist*, *182*, 328–346.

Gaudart, J., S. Rebaudet, R. Barrais, J. Boncy, B. Faucher, M. Piarroux, R. Magloire, G. Thimothe, and R. Piarroux (2013), Spatio-temporal dynamics of cholera during the first year of the epidemic in Haiti, *PLoS Neglected Tropical Diseases*, *7*, e2145.

Grad, Y. H., J. C. Miller, and M. Lipsitch (2012), Cholera modeling: Challenges to quantitative analysis and predicting the impact of interventions, *Epidemiology*, *23*, 523–530.

Hartley, D. M., J. G. Morris, and D. L. Smith (2006), Hyperinfectivity: A critical element in the ability of *V. cholerae* to cause epidemics?, *PLoS Medicine*, *39*, 3289–3294.

Huffman, G. J., R. F. Adler, D. T. Bolvin, and E. J. Nelkin (2010), The TRMM multi-satellite precipitation analysis (TMPA), in *Satellite Rainfall Applications for Surface Hydrology*, edited by M. Gebremichael and F. Hossain, pp. 3–22, Springer Netherlands.

Islam, M. S., M. A. Y. Sharker, S. Rheman, S. Hossain, Z. H. Mahmud, M. S. Islam, A. M. K. Uddin, M. Yunus, M. S. Osman, R. Ernst, I. Rector, C. P. Larson, S. P. Luby, H. P. Endtz, and A. Cravioto (2009), Effects of local climate variability on transmission dynamics of cholera in Matlab, Bangladesh, *Transactions of the Royal Society of Tropical Medicine and Hygiene*, *103*, 1165–1170.

Jutla, A., A. S. Akanda, A. Huq, A. Syed, G. Faruque, R. R. Colwell, and S. Islam (2013a), A water marker monitored by satellites to predict seasonal endemic cholera, *Remote Sensing Letters*, *4*, 822–831.

Jutla, A. S., A. S. Akanda, and S. Islam (2010), Tracking cholera in coastal regions using satellite observations, *Journal of the American Water Resources Association*, *46*, 651–662.

Jutla, A. S., A. S. Akanda, and S. Islam (2013b), A framework for predicting endemic cholera using satellite derived environmental determinants, *Environmental Modeling and Software*, *47*, 148–158.

Kelvin, A. A. (2011), Cholera outbreak in the Republic of Congo, the Democratic Republic of Congo, and cholera worldwide, *Journal of Infection in Developing Countries*, *5*, 137–143.

King, A. A., E. L. Ionides, M. Pascual, and M. J. Bouma (2008), Inapparent infections and cholera dynamics, *Nature*, *454*, 877–880.

Koelle, K., X. Rodó, M. Pascual, M. Yunus, and G. Mostafa (2005), Refractory periods and climate forcing in cholera dynamics, *Nature*, *436*, 696–700.

Lipp, E. K., A. Huq, and R. R. Colwell (2002), Effects of global climate on infectious disease: the cholera model, *Clinical Microbiology Reviews*, *15*, 757–770.

Lobitz, B., L. Beck, A. Huq, B. Wood, G. Fuchs, A. S. G. Faruque, and R. R. Colwell (2000), Climate and infectious disease: Use of remote sensing for detection of *Vibrio cholerae* by indirect measurement, *Proceedings of the National Academy of Sciences USA*, *97*, 1438–1443.

Luque Fernández, M. A., A. Bauernfeind, J. D. Jiménez, C. L. Gil, N. El Omeiri, and D. H. Guibert (2009), Influence of temperature and rainfall on the evolution of cholera epidemics in Lusaka, Zambia, 2003–2006: analysis of a time series, *Transactions of the Royal Society of Tropical Medicine and Hygiene*, *103*, 137–143.

MacCallum, S. N., and C. J. Merchant (2012), Surface water temperature observations of large lakes by optimal estimation, *Canadian Journal of Remote Sensing*, *38*, 25–45.

Marchant, R., C. Mumbi, S. Behera, and T. Yamagata (2007), The indian ocean dipole – the unsung driver of climatic variability in east africa, *African Journal of Ecology*, *45*(1), 4–16.

Mari, L., E. Bertuzzo, L. Righetto, R. Casagrandi, M. Gatto, I. Rodriguez-Iturbe, and A. Rinaldo (2012a), Modelling cholera epidemics: the role of waterways, human mobility and sanitation, *Journal of the Royal Society Interface*, *9*, 376–388.

Mari, L., E. Bertuzzo, L. Righetto, R. Casagrandi, M. Gatto, I. Rodriguez-Iturbe, and A. Rinaldo (2012b), On the role of human mobility in the spread of cholera epidemics: towards an epidemiological movement ecology, *Ecohydrology*, *5*, 531–540.

Matsuda, F., S. Ishimura, Y. Wagatsuma, T. Higashi, T. Hayashi, A. S. G. Faruque, D. A. Sack, and M. Nishibuchi (2008), Prediction of epidemic cholera due to *Vibrio cholerae* o1 in children younger than 10 years using climate data in Bangladesh, *Epidemiology and Infection*, *136*, 73–79.

Mishra, A., N. Taneja, and M. Sharma (2011), Environmental and epidemiological surveillance of *Vibrio cholerae* in a cholera-endemic region in India with freshwater environs, *Journal Of Applied Microbiology*, *112*, 225–237.

Mukandavire, Z., S. Liao, J. Wang, H. Gaff, D. L. Smith, and J. G. Morris Jr (2011), Estimating the reproductive numbers for the 2008–2009 cholera outbreaks in Zimbabwe, *Proceedings of the National Academy of Sciences USA*, *108*, 8767–8772.

Mukandavire, Z., D. L. Smith, and J. G. Morris Jr (2013), Cholera in Haiti: reproductive numbers and vaccination coverage estimates, *Scientific Reports*, *3*, 997.

Nelder, J. A., and R. Mead (1965), A simplex method for function minimization, *Computer Journal*, *7*, 308–313.

Olago, D., M. Marchall, and S. O. Wandiga (2007), Climatic, socio-economic, and health factors affecting human vulnerability to cholera in the Lake Victoria basin, East Africa, *Ambio*, *36*, 350–358.

Pascual, M., X. Rodó, S. P. Ellner, R. R. Colwell, and M. J. Bouma (2000), Cholera dynamics and El Niño Southern Oscillation, *Science*, *289*, 1766–1769.

Pascual, M., M. J. Bouma, and A. P. Dobson (2002), Cholera and climate: revisiting the quantitative evidence, *Microbes and Infection*, *4*, 237–245.

Pascual, M., L. F. Chaves, B. Cash, X. Rodó, and M. Yunus (2008), Predicting endemic cholera: the role of climate variability and disease dynamics, *Climate Research*, *36*, 131–140.

Piarroux, R., and D. Bompangue (2007), *Needs for an integrative approach of epidemics: the example of cholera*, encyclopedia of infectious diseases: modern methodologies, John Wiley And Sons, Inc.

Piarroux, R., D. Bompangue, P. Y. Oger, F. Haaser, A. Boinet, and T. Vandevelde (2009), From research to field action: Example of the fight against cholera in the Democratic Republic of Congo, *Field Actions Scientific Reports*, *1*, 1–9.

Plisnier, P. D., S. Serneels, and E. F. Lambin (2000), Impact of ENSO on east african ecosystems: a multivariate analysis based on climate and remote sensing data, *Global Ecology and Biogeography*, *9*(6), 481–497.

Rebaudet, S., B. Sudre, B. Faucher, and R. Piarroux (2013), Cholera in coastal africa: A systematic review of its heterogeneous environmental determinants, *Journal of Infectious Diseases*, *208*(suppl 1), S98–S106.

Reiner, R. C., A. A. King, M. Emch, M. Yunus, A. S. G. Faruque, and M. Pascual (2012), Highly localized sensitivity to climate forcing drives endemic cholera in a megacity, *Proceedings of the National Academy of Sciences USA*, *109*, 2033–2036.

Reyburn, R., D. R. Kim, M. Emch, A. Khatib, L. von Seidlein, and M. Ali (2011), Climate variability and the outbreaks of cholera in Zanzibar, East Africa: A time series analysis, *American Journal of Tropical Medicine and Hygiene*, *84*, 862–869.

Righetto, L., E. Bertuzzo, L. Mari, E. Schild, R. Casagrandi, M. Gatto, I. Rodriguez-Iturbe, and A. Rinaldo (2013), Rainfall mediations in the spreading of epidemic cholera, *Advances in Water Resources*, *60*, 34–46.

Riley, S. (2007), Large-scale spatial-transmission models of infectious disease, *Science*, *316*, 1298–1301.

Rinaldo, A., E. Bertuzzo, L. Mari, L. Righetto, M. Blokesch, M. Gatto, R. Casagrandi, M. Murray, S. Vesenbeckh, and I. Rodriguez-Iturbe (2012), Reassessment of the 2010–2011 Haiti cholera outbreak and rainfall-driven multiseason projections, *Proceedings of the National Academy of Sciences USA*, *109*, 6602–6607.

Rodò, X., M. Pascual, G. Fuchs, and A. S. G. Faruque (2002), ENSO and cholera: A nonstationary link related to climate change?, *Proceedings of the National Academy of Sciences*, *99*(20), 12,901–12,906.

Ruiz-Moreno, D., M. Pascual, M. Bouma, A. Dobson, and B. Cash (2007), Cholera seasonality in Madras: dual role for rainfall in endemic and epidemic regions, *EcoHealth*, *4*, 52–62.

Saji, N. H., B. N. Goswami, P. N. Vinayachandran, and T. Yamagata (1999), A dipole mode in the tropical indian ocean, *Nature*, *401*, 360–363.

Sardar, T., S. Mukhopadhyay, A. R. Bhowmick, and J. Chattopadhyay (2013), An optimal cost effectiveness study on Zimbabwe cholera seasonal data from 2008–2011, *PLoS ONE*, *8*, e81,231.

Stager, J. C., A. Ruzmaikin, D. Conway, P. Verburg, and P. J. Mason (2007), Sunspots, el niño, and the levels of lake victoria, east africa, *Journal of Geophysical Research: Atmospheres*, *112*(D15), D15,106.

Thiery, W., A. Martynov, F. Darchambeau, J.-P. Descy, P.-D. Plisnier, L. Sushama, and N. P. M. van Lipzig (2014a), Understanding the performance of the FLake model over two african great lakes, *Geosci. Model Dev.*, *7*, 317–337.

Thiery, W., V. M. Stepanenko, X. Fang, K. D. Jöhnk, Z. Li, A. Martynov, M. Perroud, Z. M. Subin, F. Darchambeau, D. Mironov, and N. P. M. Van Lipzig (2014b), LakeMIP kivu: evaluating the representation of a large, deep tropical lake by a set of one-dimensional lake models, *Tellus A*, *66*, 21,390.

Tuite, A. R., J. Tien, M. Eisenberg, D. J. Earn, J. Ma, and D. N. Fisman (2011), Cholera epidemic in Haiti, 2010: Using a transmission model to explain spatial spread of disease and identify optimal control interventions, *Annals of Internal Medicine*, *154*, 593–601.

World Health Organization (2012), Cholera, 2011, *Weekly epidemiological record*, *87*, 289–304.

Corresponding author: A. Rinaldo, Laboratory of Ecohydrology (EPFL/ENAC/IIE/ECHO), GR C1 512, Station 2, Lausanne CH-1015 (andrea.rinaldo@epfl.ch)

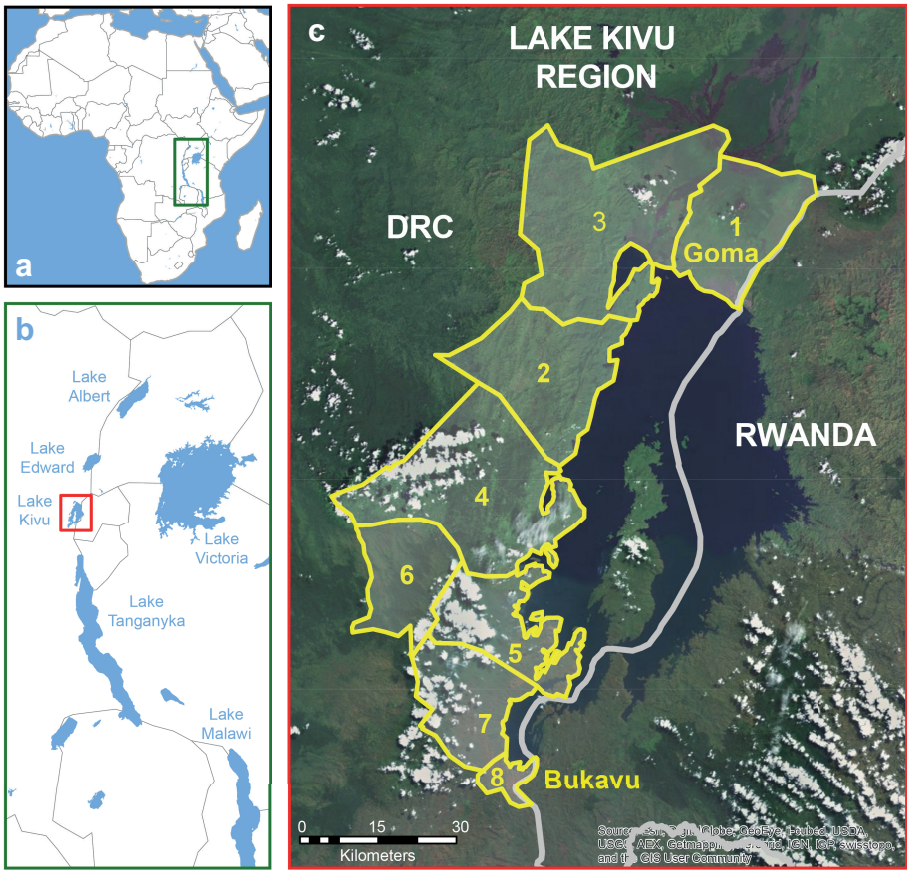

Figure 1. Study area. Lake Kivu is one of the African Great Lakes (a), located to the north of Lake Tanganyika and west of Lake Victoria (b), at the border between the DRC and Rwanda (c). The eight lakeside health zones included in this study (1: Goma, 2: Minova, 3: Kirotshe, 4: Kalehe, 5: Katana, 6: Miti Murhesa, 7: Kabare, 8: Bukavu) as well as Goma and Bukavu, the two main cities on its shore, are shown in panel (c).

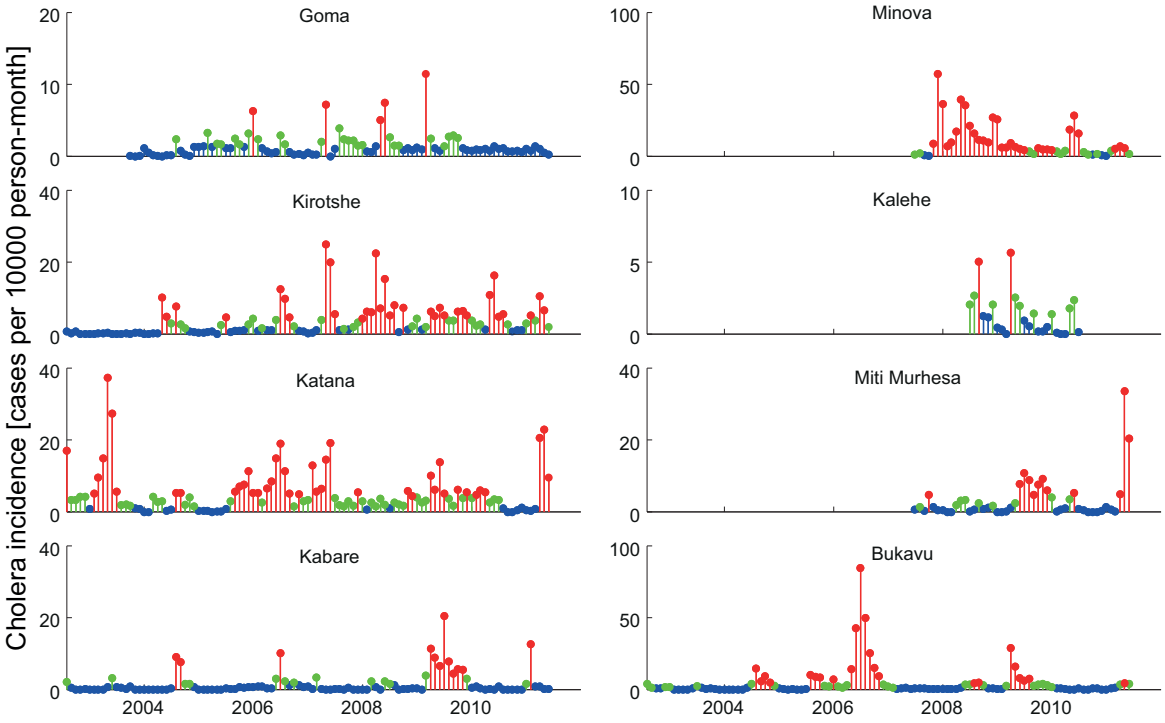

Figure 2. Cholera incidence data for the eight health zones considered. Low (state 1), mild (state 2) and high incidence (state 3) are denoted by respectively blue, green and red color. Note that in some health zones data are available only during parts of the study period.

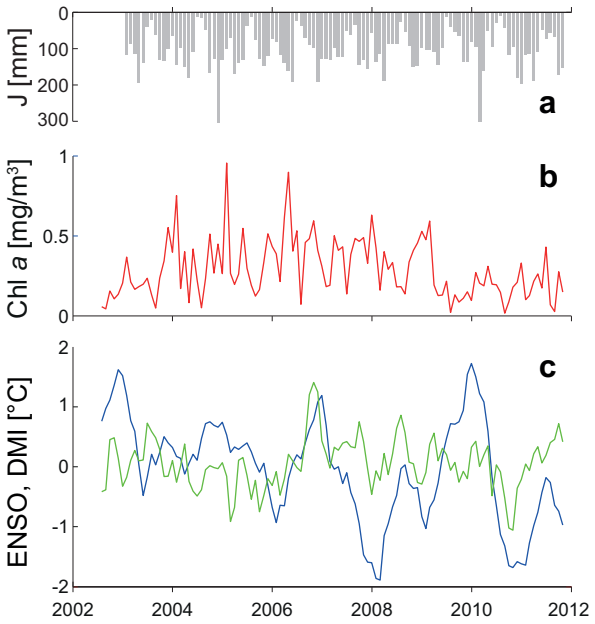

Figure 3. Environmental drivers. (a) Total monthly precipitation averaged over all considered health zones. (b) Remotely sensed, spatially averaged chlorophyll *a* concentration in Lake Kivu. (c) Global climate anomalies. ENSO SST anomaly in the Niño 3.4 region (blue) and IOD Dipole Mode Index (green). Note that the chlorophyll *a* and climate anomalies start 6 months before cholera incidence data in order to allow for lags in the model.

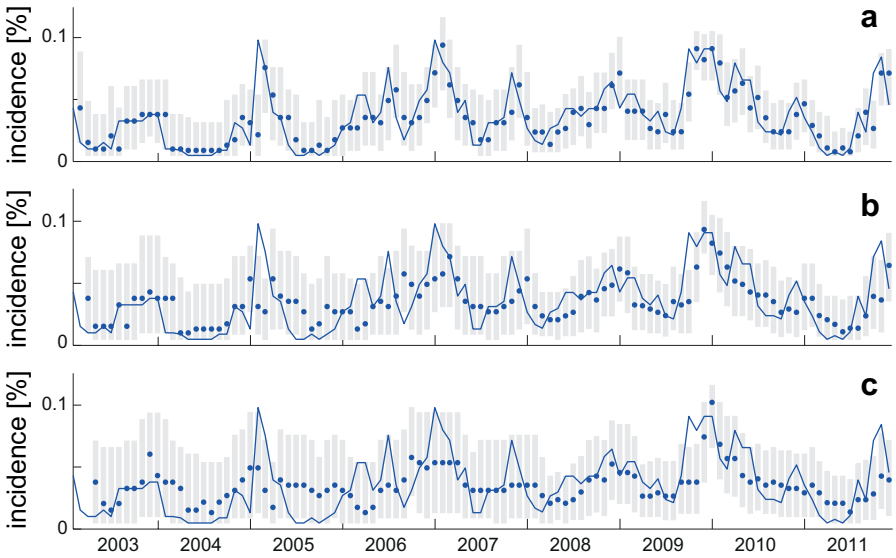

Figure 4. Simulations of model 13, which accounts for mobility, precipitation and SST anomaly in addition to seasonality, at one (a), two (b) and three (c) months lags using the cross-validation procedure described in section 3.2. The blue line shows the average over the states of the eight health zones, where each state is represented by its mean incidence. The blue dots show the median of the simulated values and the grey bars the 5^{th}–95^{th} percentile range over 10,000 simulations.

Table 1. Summary of recently published applied cholera models.

	Region	Empirical	Mechanistic	Implicit[b]	Explicit[b]	Data-driven[c]	Simulated[c]	None[c]	Epidemic[d]	Endemic[d]
Lobitz et al. [2000]	BoB[a]	*		*		*				*
Pascual et al. [2000]	BoB[a]	*		*			*			*
Koelle et al. [2005]	BoB[a]		*	*			*			*
de Magny et al. [2008]	BoB[a]	*		*		*				*
Bertuzzo et al. [2008]	S. Africa		*		*		*		*	
Luque Fernández et al. [2009]	Zambia	*		*		*				*
King et al. [2008]	BoB[a]		*	*			*			*
Matsuda et al. [2008]	BoB[a]	*		*		*				*
Pascual et al. [2008]	BoB[a]		*	*			*			*
Akanda et al. [2009]	BoB[a]	*		*		*				*
Islam et al. [2009]	BoB[a]	*		*		*				*
Andrews and Basu [2011]	Haiti		*	*				*	*	
Bertuzzo et al. [2011]	Haiti		*		*			*	*	
Chao et al. [2011]	Haiti		*		*			*	*	
Mukandavire et al. [2011]	Zimbabwe		*	*				*	*	
Reyburn et al. [2011]	Zanzibar	*		*		*				*
Tuite et al. [2011]	Haiti		*		*			*	*	
Rinaldo et al. [2012]	Haiti		*		*	*			*	
Eisenberg et al. [2013]	Haiti	*	*		*			*	*	
Jutla et al. [2013a]	BoB[a]	*		*		*				*
Jutla et al. [2013b]	BoB[a]	*		*		*				*
Mukandavire et al. [2013]	Haiti		*	*				*	*	
Reiner et al. [2012]	BoB[a]	*	*	*		*	*			*
Gatto et al. [2012]	Haiti / S. Africa		*		*	*			*	
Mari et al. [2012a]	S. Africa		*		*			*	*	
Righetto et al. [2013]	Haiti		*		*	*			*	
Sardar et al. [2013]	Zimbawe		*	*			*		*	

[a] Bay of Bengal

[b] Attributes refer to the treatment of space by the models. Explicit treatment of space means that the model incorporates terms for the spatial spread of disease vehiculed by human mobility, water-ways, etc.

[c] Attributes describing the treatment of environmental forcings. Simulated refers to cyclicity based on seasonal and interannual (ENSO) forcings.

[d] The disease is considered endemic if outbreaks appear to occur every year and if the outbreak is not apparently the result of recent introduction.

Table 2. Results of the fitting procedure ordered by increasing AIC score (first 20 lines).[a]

Model	Seasonality	Mobility	Precipitation	Chlorophyll *a*	ENSO	IOD	Degrees of freedom[b]	log(Likelihood)	AIC	ΔAIC
1	+	-	-	-	+	-	11	−470.02	962.03	0
2	+	-	+	-	+	-	12	−469.22	962.45	0.4
3	+	-	-	-	+	+	13	−468.89	963.78	1.6
4	+	-	-	+	+	-	13	−468.95	963.89	1.9
5	+	-	-	-	-	+	11	−471.03	964.05	2.0
6	+	-	+	-	+	+	14	−468.07	964.15	2.1
7	+	-	-	+	-	+	13	−469.26	964.52	2.5
8	+	-	-	-	-	-	9	−473.27	964.53	2.5
9	+	-	+	+	+	-	14	−468.37	964.74	2.7
10	+	-	-	+	+	+	15	−467.38	964.76	2.7
11	+	-	+	-	-	+	12	−470.6	965.2	3.2
12	+	-	+	-	-	-	10	−472.84	965.67	3.6
13	+	+	+	-	+	-	15	−467.84	965.68	3.6
14	+	-	+	+	+	+	16	−466.9	965.79	3.8
15	+	-	+	+	-	+	14	−469.16	966.31	4.3
16	+	+	+	-	-	-	13	−470.27	966.53	4.5
17	+	-	-	+	-	-	11	−472.36	966.72	4.7
18	+	+	-	-	+	-	14	−469.75	967.51	5.5
19	+	+	-	-	-	-	12	−471.85	967.71	5.7
20	+	-	+	+	-	-	12	−471.95	967.89	5.9

[a] See Supplementary Table 1 for all 64 lines.
[b] Number of parameters plus one (residual variance)

Table 3. Parameter sets corresponding to the best ranked models in calibration and cross-validation (respectively models 1 and 13 in Table 2).

	Model 1	Model 13
$p_{1,1}$	0.8841	0.9203
$p_{2,1}$	0.3822	0.4021
$p_{3,1}$	0.1151	0.1394
$p_{1,3}$	0.02575	0.01879
$p_{2,3}$	0.1366	0.09161
$p_{3,3}$	0.3736	0.2446
δ	0.4029	0.3702
t_s	6.358	5.755
γ		0.2697
D		30.53
ν		0.7578
α		0.5598
ϕ	0.5014	0.3642
t_{ENSO}	0	0

Table 4. Log-likelihood values obtained during validation by comparing one, two or three months in advance simulations with reported incidence.

Lag (months)	1	2	3
Model[a] 1	−482.35	−559.28	−583.37
Model 2	−482.79	−558.34	−581.98
Model 3	−476.72	−552.91	−579.46
Model 4	−472.56	−549.97	−573.83
Model 5	−472.78	−548.69	−575.05
Model 6	−470.53	−546.38	−571.45
Model 7	−479.03	−556.25	−583.36
Model 8	−483.40	−563.09	−591.90
Model 9	−477.77	−554.54	−578.10
Model 10	−473.55	−549.05	−575.05
Model 11	−475.61	−552.38	−579.58
Model 12	−485.75	−564.61	−592.76
Model 13	−469.67	−543.63	−569.04
Model 14	−471.28	−547.75	−571.63

[a] Model numbering corresponds to the rank obtained according to Table 2.