



# Real-time projections of cholera outbreaks through data assimilation and rainfall forecasting



Damiano Pasetto<sup>a</sup>, Flavio Finger<sup>a</sup>, Andrea Rinaldo<sup>a,b</sup>, Enrico Bertuzzo<sup>a,c,\*</sup>

<sup>a</sup> Laboratory of Ecohydrology, School of Architecture, Civil and Environmental Engineering, École Polytechnique Fédérale de Lausanne, Lausanne, Switzerland

<sup>b</sup> Dipartimento Ingegneria Civile Edile ed Ambientale, Università di Padova, Padova, Italy

<sup>c</sup> Department of Environmental Sciences, Informatics and Statistics, University Cà Foscari Venice, Venezia Mestre, Italy

## ARTICLE INFO

### Article history:

Received 11 May 2016

Revised 31 August 2016

Accepted 6 October 2016

Available online 8 October 2016

### Keywords:

Epidemiological model

Data assimilation

Cholera

Rainfall forecast

Climate forecast system

## ABSTRACT

Although treatment for cholera is well-known and cheap, outbreaks in epidemic regions still exact high death tolls mostly due to the unpreparedness of health care infrastructures to face unforeseen emergencies. In this context, mathematical models for the prediction of the evolution of an ongoing outbreak are of paramount importance. Here, we test a real-time forecasting framework that readily integrates new information as soon as available and periodically issues an updated forecast. The spread of cholera is modeled by a spatially-explicit scheme that accounts for the dynamics of susceptible, infected and recovered individuals hosted in different local communities connected through hydrologic and human mobility networks. The framework presents two major innovations for cholera modeling: the use of a data assimilation technique, specifically an ensemble Kalman filter, to update both state variables and parameters based on the observations, and the use of rainfall forecasts to force the model. The exercise of simulating the state of the system and the predictive capabilities of the novel tools, set at the initial phase of the 2010 Haitian cholera outbreak using only information that was available at that time, serves as a benchmark. Our results suggest that the assimilation procedure with the sequential update of the parameters outperforms calibration schemes based on Markov chain Monte Carlo. Moreover, in a forecasting mode the model usefully predicts the spatial incidence of cholera at least one month ahead. The performance decreases for longer time horizons yet allowing sufficient time to plan for deployment of medical supplies and staff, and to evaluate alternative strategies of emergency management.

© 2016 The Authors. Published by Elsevier Ltd.

This is an open access article under the CC BY-NC-ND license.

(<http://creativecommons.org/licenses/by-nc-nd/4.0/>)

## 1. Introduction

Cholera is a diarrheal disease caused by the ingestion of water or food contaminated by the bacterium *Vibrio cholerae*, the causative agent of the disease. Although treatment is cheap and well-known (chiefly rehydration therapy), cholera is still one of the leading causes of death in developing countries (Mathers et al., 2008). In regions where the disease is endemic (e.g., Bangladesh) the case fatality rate is relatively low (around 0.1%, see e.g., Ryan et al., 2000) because health-care staff and infrastructures are prepared and thus symptomatic cases are readily reported and treated. On the contrary, epidemic regions that are scourged by irregular and severe cholera outbreaks usually exhibit higher mortality, mostly due to the unpreparedness of health care facilities. In addition, severe cholera outbreaks in epidemic regions, where

the number of infections is boosted by a relatively low level of population immunity, can locally exceed the allocated treatment capacity (e.g., number of beds in treatment facilities, number of oral rehydration therapy units available). A revealing example is the cholera epidemic that struck Haiti in October 2010, 10 months after a catastrophic earthquake that destroyed an already faltering civil and sanitary infrastructure, and is still lingering as of May 2016. The epidemic has totalled almost 800,000 reported cases and 9200 deaths with an overall case fatality rate of 1.15%, which was even higher (around 2%) during the first months (Barzilay et al., 2013) (data available on-line at <http://mspp.gouv.ht>). Thus, modeling tools which can possibly predict the evolution of an ongoing outbreak in time for interventions are of paramount importance to guide health care officials in allocating staff and resources and evaluating alternative control strategies.

The quasi-real time release of the epidemiological data during the Haitian cholera outbreak prompted many research teams to develop epidemiological models of the outbreak in an

\* Corresponding author.

E-mail address: [enrico.bertuzzo@unive.it](mailto:enrico.bertuzzo@unive.it) (E. Bertuzzo).

effort to provide meaningful insights to guide the emergency management (Abrams et al., 2013; Andrews and Basu, 2011; Bertuzzo et al., 2014, 2011; Chao et al., 2011; Eisenberg et al., 2013; Gatto et al., 2012; Kirpich et al., 2015; Mari et al., 2015; Mukandavire et al., 2013; Righetto et al., 2013; Rinaldo et al., 2012; Tuite et al., 2011). Some of these studies (Abrams et al., 2013; Andrews and Basu, 2011; Bertuzzo et al., 2014, 2011; Righetto et al., 2013; Rinaldo et al., 2012) attempted to actually forecast the evolution of the unfolding outbreak by calibrating a model on the data available at a certain moment in time and projecting the simulations into the future. Early attempts show contrasting results (for a complete reassessment see Rinaldo et al., 2012). The ability to predict under different modeling assumptions has later been analyzed in details (Mari et al., 2015), showing that, when data is scarce, spatially-explicit models (e.g., Bertuzzo et al., 2011) clearly outperform models that do not account for the spatial coupling among individual local models (e.g., Andrews and Basu, 2011). The revamping of the outbreak in conjunction with the rainy season in spring 2011 revealed empirically that, at least in the Haitian context, intense rainfall enhances cholera transmission and therefore has to be taken into account for future model developments and predictions (Rinaldo et al., 2012). This consideration further complicates modelers' task because it implies that in order to predict cholera incidence one must also predict precipitation intensity in space and time. So far, this issue has been tackled by producing realistic rainfall scenarios using stochastic models of rainfall generation (Righetto et al., 2013) or by bootstrapping of past observed rainfall fields (Bertuzzo et al., 2014; Rinaldo et al., 2012).

All the previous examples represent isolated attempts to forecast cholera dynamics, each based on different assumptions to accommodate relevant processes and recalibration on the available data. Here, we aim at proving the feasibility of a real-time forecasting framework during emergencies that: i) flexibly adapts to account for the dominant processes driving the outbreak, ii) readily integrates new information available, and iii) periodically issues an updated forecast for a predefined time horizon. We therefore set ourselves at the initial phase of the Haitian cholera outbreak and produce weekly bulletins forecasting the spatio-temporal distribution of new cases for the first two years of the epidemic using only information that was available at that time.

The first major innovation of this study with respect to previous efforts is the use of a data assimilation (DA) framework to integrate new epidemiological data as soon as they become available and to update the model forecast without recalibrating. DA has long been used in weather forecasting (Navon, 2009; Rabier, 2005), where numerical models require frequent re-initializations to track the real dynamics and to avoid the rapid divergence of the numerical solution. This procedure is typically performed by the data assimilation cycle (Thompson, 1961), the sequential repetition of a forecast step and its correction in the analysis (or update) step using the newly available system observations. Forecast and analysis steps are naturally formulated in a Bayesian framework by the so-called filtering problem (Jazwinski, 1970), which seeks the posterior probability distribution of the system state, given all the observations in a time window of interest, and takes into account the model uncertainties and the observation errors. While the well-known Kalman–Bucy filter (Kalman, 1960; Kalman and Bucy, 1961) solves the filtering problem in the simple case of linear models with additive and Gaussian errors, an analytical solution in the presence of nonlinearities does not exist and several alternative filters have been proposed in literature (see e.g., Arulampalam et al., 2002). The ensemble Kalman filter (EnKF), developed by Evensen (Evensen, 1994, 2003) for nonlinear applications in the context of ocean modeling, is one of the most popular DA techniques and consists in an ensemble approximation of the Kalman filter. Although optimal only for Gaussian distributions

of state variables, EnKF typically delivers satisfactory performances using a small number of model realizations also for non-Gaussian models (Zhou et al., 2006), a feature that favored its application in different fields including atmospheric sciences (e.g., Houtekamer et al., 2005) and hydrology (e.g., Camporese et al., 2009; ElSheikh et al., 2012; Pasetto et al., 2012). Another appealing feature of EnKF is the possibility to infer model parameters at each assimilation step by the augmented state technique (Evensen, 2009; Pasetto et al., 2015). In this manner, the filter corrects the probability distribution of the parameters during the simulation, reducing the model bias and tracking the parameter evolution in time. Lately, DA frameworks have also been applied to forecast epidemics, in particular for seasonal and pandemic influenza (Chretien et al., 2014; Shaman and Karspeck, 2012; Shaman et al., 2013; Yang et al., 2015b, 2015a), HIV/AIDS (Cazelles and Chau, 1997; Wu and Tan, 2000), the Ebola outbreak in Sierra Leone (Yang et al., 2015c), and the cattle disease *Theileria orientalis* (Jewell and Brown, 2015).

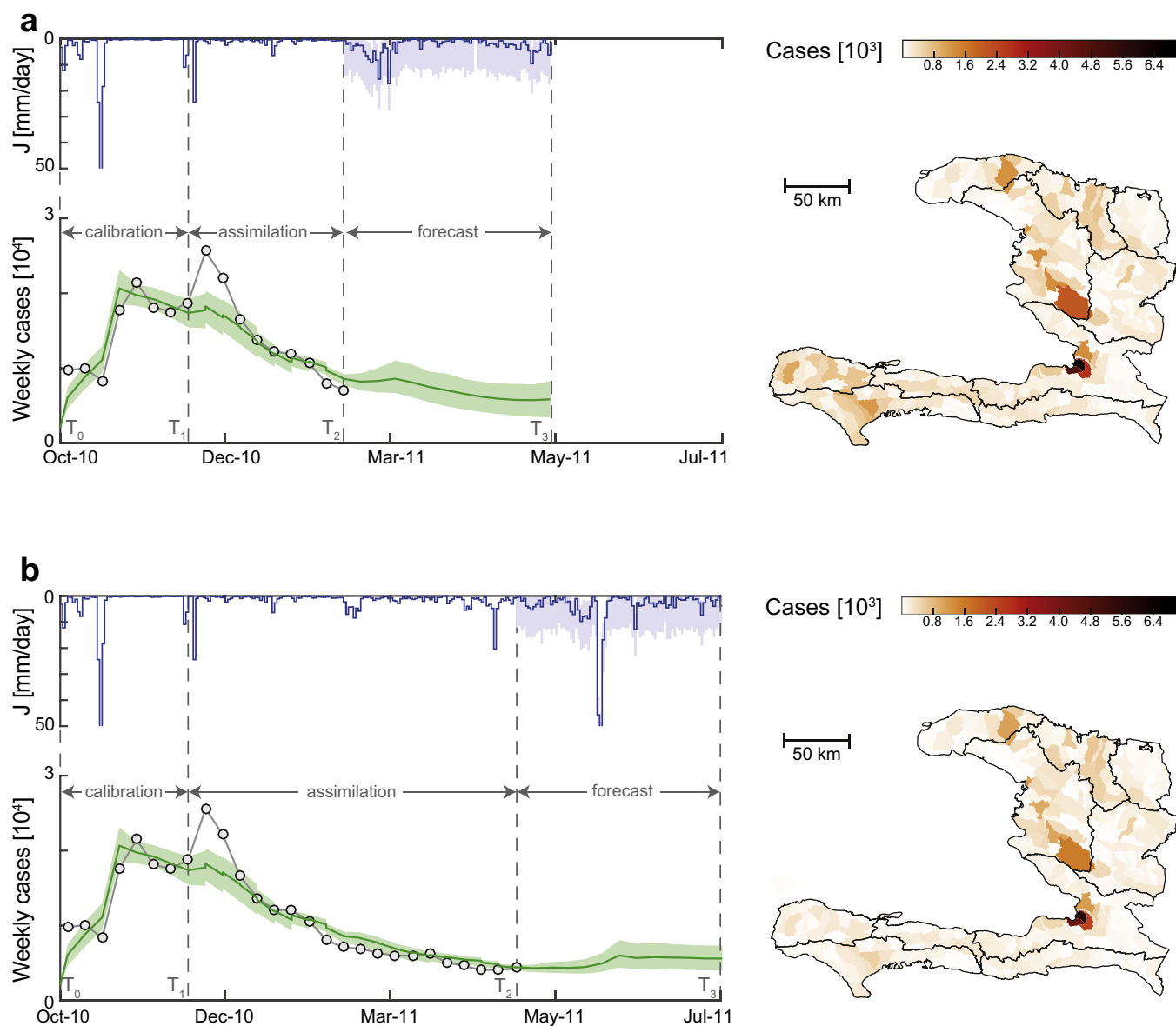
The second main novelty of our approach is the direct use of rainfall forecasts as predicted by the Climate Forecast System (CFS) (Saha et al., 2014) of the National Centers for Environmental Prediction (NCEP). CFS models the interaction between oceans, land, and atmosphere at a global scale assimilating remotely acquired variables. Operational climate forecasts are produced daily at different spatial scales (down to  $0.5^\circ$ ) and temporal intervals (up to six months of forecast with a frequency of six hours). An appealing feature of such datasets is their long forecast horizon, which allows epidemiological modelers to analyze the long-term impact of hydrologic drivers on the course of an outbreak. Moreover, CFS forecasts are freely available at the global scale, thus providing precipitation data and forecasts also over developing countries where waterborne diseases are likely but meteorological data are typically scarce.

## 2. Conceptual framework

In this section we present the conceptual framework for the operational forecast of a cholera outbreak. The individual components of the framework, namely the epidemiological model, the calibration and the DA schemes and the rainfall forecast are described in details in Section 3.

We assume that there must be a time-lag between the onset of an outbreak and the moment when the epidemic forecasts are fully operational. First, a certain amount of time is necessary for healthcare authorities to identify and declare a cholera outbreak. Second, if not already in place, a surveillance system that centralizes epidemiological data must be implemented. The duration of this lag crucially depends on the preparedness of the healthcare infrastructures. In the case of Haiti, the whole process took about one month (Barzilay et al., 2013). From the modeling perspective, data regarding population distribution, climatic and hydrological variables must be collected and suitably processed. In the following we term  $T_0$  the onset of the epidemic and  $T_1$  the moment when forecasts begin to be issued.

The first set of data pertaining the onset of the outbreak is used to calibrate the model through a Markov chain Monte Carlo (MCMC, see Section 3.2) scheme, in order to obtain a preliminary estimation of the posterior parameter distribution. In this case study, the first seven weeks of epidemiological data are used for calibration, thus  $T_0$  = October 20, 2010 and  $T_1$  = December 12, 2010 (see Fig. 1). The posterior parameter distribution computed employing MCMC is used to initialize the DA framework and start the operational forecast. Specifically,  $N$  parameter sets are sampled from the posterior distribution, along with the corresponding simulations. This set of trajectories, periodically updated through DA, is kept throughout the whole forecasting period. After the calibration period  $[T_0, T_1]$ , the epidemiological forecasts are issued



**Fig. 1.** Example of two cholera forecasts for the Haitian outbreak computed at  $T_2 =$  February 12, 2011 (a) and  $T_2 =$  April 23, 2011 (b). Green lines and light-green areas represent, respectively, the expected value and the 90% confidence interval of the total weekly cases estimated by the model. Grey circles show recorded cases. Blue lines represent the average daily rainfall used as forcing in the cholera model, which is measured from  $T_0$  to  $T_2$  and forecast from  $T_2$  to  $T_3$ . Light blue areas show the 90% uncertainty associated with the forecast rainfall. Maps show the expected value of the forecast cumulative cases on each model sub-unit. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

weekly, at every assimilation of the newly reported cases. The main steps of the proposed real-time forecast framework are detailed below. At the end of an epidemiological week:

- Rainfall data measured during the previous week are collected;
- Each of the  $N$  system trajectories is advanced, forced by the measured rainfall, by one week such as to arrive at the present time;
- The newly available epidemiological data is assimilated by means of an ensemble Kalman filter (EnKF, Section 3.2). Therefore the model state variables (i.e., infected and recovered individuals, and bacterial concentrations in the water reservoir that conceptualizes to various degrees infection exposures (Codeço, 2001) at each model node) of each of the  $N$  trajectories are updated. The EnKF is applied with the augmented state, thus the  $N$  parameter sets are also updated;
- The rainfall forecast issued on the same day is retrieved.  $N$  different time series of rainfall are generated by adding a random error to the forecast.
- Each of the  $N$  system trajectories is associated to one rainfall time-series and projected into the future for the prescribed time horizon;
- The forecast epidemiological variables, such as the total number of cases and the spatial distribution of the cholera incidence are published.

Fig. 1 depicts two examples of forecasts for the Haitian outbreak where  $T_2$  indicates the time when the bulletin is issued and  $T_3$  the end of the forecast horizon. In this example  $T_3 - T_2$  is equal to three months. The model trajectories computed using the MCMC posterior distribution of the model parameters fit the data collected during the calibration period [ $T_0, T_1$ ], but are not suitable for the subsequent time step. The sequential assimilation of the data collected during [ $T_1, T_2$ ] corrects the model trajectories and

parameter values toward the real epidemic dynamics. The model uncertainty gradually increases during the forecast period  $[T_2, T_3]$  due to the uncertainty in the forecast rainfall, which is a driver of the model.

### 3. Material and methods

#### 3.1. Epidemiological model

The cholera epidemiological model used herein derives directly from the one developed in Bertuzzo et al. (2014), which, in turn, builds on previous spatially-explicit epidemiological models (Bertuzzo et al., 2008, 2010; Mari et al., 2012a, 2012b). The model subdivides the total population into  $n$  human communities spatially distributed within a domain of  $n$  nodes connected by both human mobility and hydrological networks. Let  $S_i(t)$ ,  $I_i(t)$  and  $R_i(t)$  denote the local abundances of susceptible, symptomatic infected and recovered individuals at time  $t$  in each node  $i$  of the network, and let  $B_i(t)$  be the environmental concentration of *V. cholerae* in  $i$ . Cholera transmission dynamics can be described by the following set of coupled differential equations:

$$\frac{dS_i}{dt} = \mu(H_i - S_i) - F_i(t)S_i + \rho R_i \tag{1}$$

$$\frac{dI_i}{dt} = \sigma F_i(t)S_i - (\gamma + \mu + \alpha)I_i \tag{2}$$

$$\frac{dR_i}{dt} = (1 - \sigma)F_i(t)S_i + \gamma I_i - (\rho + \mu)R_i \tag{3}$$

$$\begin{aligned} \frac{dB_i}{dt} = & -\mu_B B_i + \frac{p}{W_i} [1 + \phi J_i(t)] \left( (1 - m)I_i + m \sum_{j=1}^n Q_{ij} I_j \right) \\ & - l \left( B_i - \sum_{j=1}^n P_{ji} \frac{W_j}{W_i} B_j \right), \end{aligned} \tag{4}$$

where each node population  $H_i$  is assumed to be at demographic equilibrium. Under this assumption, Eq. (1) is equivalent to  $S_i(t) = H_i - I_i(t) - R_i(t)$ , which ensures the conservation of the population during the numerical simulation while reducing the model dimension. The force of infection  $F_i(t)$ , which represents the rate at which susceptible individuals become infected due to contact with contaminated water, is expressed as:

$$F_i(t) = \beta \left[ (1 - m) \frac{B_i}{K + B_i} + m \sum_{j=1}^n Q_{ij} \frac{B_j}{K + B_j} \right]. \tag{5}$$

The parameter  $\beta$  represents the maximum exposure rate. The model assumes that  $\beta$  is constant, but the framework allows the EnKF to potentially change its value in time. The fraction  $B_i/(K + B_i)$  is the probability of becoming infected due to exposure to a concentration  $B_i$  of *V. cholerae*,  $K$  being the half-saturation constant (Codeço, 2001). Because of human mobility, a susceptible individual residing at node  $i$  can, while travelling, be exposed to pathogens in the destination community  $j$ . This is modeled assuming that the force of infection in a given node depends on the local concentration  $B_i$  for a fraction  $(1 - m)$  of the susceptible hosts and on the concentration  $B_j$  of the remote communities for the remaining fraction  $m$ . The parameter  $m$  represents the community-level probability that individuals travel outside their node. The concentrations  $B_j$  are weighted according to the probabilities  $Q_{ij}$  that an individual living in node  $i$  reaches  $j$  as a destination. We apply a gravity approach (Erlander and Stewart, 1990) to model human mobility. Accordingly, connection probabilities are defined as

$$Q_{ij} = \frac{H_j e^{-d_{ij}/D}}{\sum_{k \neq i}^n H_k e^{-d_{ik}/D}}, \tag{6}$$

where the attractiveness of node  $j$  depends on its population size  $H_j$ , while the deterrence factor is assumed to be dependent on the distance  $d_{ij}$  between the two communities via an exponential kernel (with shape factor  $D$ ). A fraction  $\sigma$  of the infected individuals develops symptoms, thus entering class  $I_i$ . The remaining fraction  $(1 - \sigma)$  does not develop symptoms and therefore does not contribute to disease transmission and enters directly the recovered compartment. Symptomatic infected individuals recover at a rate  $\gamma$ , or die due to cholera or other causes at rates  $\alpha$  or  $\mu$ , respectively. Recovered individuals lose their immunity and return to the susceptible compartment at a rate  $\rho$  or die at a rate  $\mu$ . A fraction  $m$  of the symptomatic infected individuals are assumed to move among the nodes according to the human mobility model, and thus contribute to the environmental concentration of *V. cholerae* at a rate  $p/W_i$ , where  $p$  is the rate at which bacteria excreted by an infected individual reach and contaminate the local water reservoir of volume  $W_i$  (assumed to be proportional to population size, i.e.,  $W_i = cH_i$  as in Rinaldo et al. (2012)). *V. cholerae* are assumed to decay in the environment at a rate  $\mu_B$ . Bacteria undergo hydrologic dispersal at a rate  $l$ : pathogens travel from node  $i$  to  $j$  with probability  $P_{ij}$ , which is assumed to be one if node  $j$  is the downstream nearest neighborhood of  $i$ , and zero otherwise. In order to express the worsening of sanitation conditions caused by rainfall-induced runoff, which causes additional pathogen loads to enter the water reservoir due to effects such as overflow of pit latrines and washout of open-air defecation sites (Gaudart et al., 2013), the contamination rate  $p$  is increased by the rainfall intensity  $J_i(t)$  via a coefficient  $\phi$  (Righetto et al., 2013; Rinaldo et al., 2012). By introducing the dimensionless bacterial concentration  $B_i^* = B_i/K$ , it is possible to group three model parameters into a single ratio  $\theta = p/(cK)$  (Bertuzzo et al., 2008).

The estimation of weekly cholera cases (the quantity usually reported in epidemiological records) from the model output requires to compute

$$C_i(t_k) = \sigma \int_{t_{k-1}}^{t_k} F_i S_i dt, \tag{7}$$

where  $t_k$  marks the end of the  $k$ th week.

The time-integration of Eqs. (1–4) is performed through the Runge-Kutta (4,5) method, as described in Shampine and Reichelt (1997).

#### 3.2. Parameter estimation

At the end of the  $k$ th epidemiological week, the model describes the state of the epidemic by the system vector  $\mathbf{x}_k \in \mathbb{R}^{4n}$ ,  $\mathbf{x}_k = \{(I_{i,k}, R_{i,k}, B_{i,k}, C_{i,k})\}$ , with  $i = 1, \dots, n$ , where  $n$  is the number of nodes. The solution of (1–4) is driven by the daily rainfall over each node,  $\mathbf{J} = (J_1, \dots, J_n)$ , and the model epidemiological parameters. While model parameters  $\mu$ ,  $\gamma$ , and  $\alpha$  can reasonably be estimated from demographic and epidemiological literature (see, e.g., Bertuzzo et al., 2014), the remaining model parameters are typically unknown and require to be inferred by calibration. In this case, the epidemiological data used for calibration are the observed weekly cases, in the following indicated with  $\mathbf{y}_k \in \mathbb{R}^d$ , where  $d$  is the number of measurements points at time  $t_k$  (in the Haitian case study  $d=10$  is the number of departments). The relationship  $\mathbf{y}_k = \mathbf{H}\mathbf{x}_k + \boldsymbol{\xi}_k$  links the observations to the model state variables, where  $\mathbf{H} \in \mathbb{R}^{d \times 4n}$  transfers the modeled weekly cases  $C$  from the node level to the observation points, and the vector  $\boldsymbol{\xi}_k \in \mathbb{R}^d$  represents the measurement error. The error components  $\xi_{i,k}$ ,  $i = 1, \dots, d$ , are modeled as independent Gaussian random variables with zero mean and standard deviation  $\sigma_\xi$ . In the following we call  $\boldsymbol{\vartheta}$  the set of the unknown model parameters,  $\boldsymbol{\vartheta} = (\beta, \psi, m, D, \rho, \sigma, \mu_B, \theta, l)$ . We consider two methods for the Bayesian estimation of model parameters as described in the following sections.

### 3.2.1. Markov chain Monte Carlo

We use the Differential Evolution Adaptive Metropolis (DREAM<sub>ZS</sub>) (Vrugt et al., 2009) implementation of the MCMC algorithm. Given the prior probability density function (pdf) of the parameters and the collection of observations in the temporal window of interest, e.g.,  $t_0, \dots, t_k$ , DREAM<sub>ZS</sub> samples the desired number of parameter realizations from the posterior distribution using multiple MCMC chains that run in parallel and that jointly contribute to the computation of the proposal parameter samples. This technique has already been effectively applied to this epidemiological model (Bertuzzo et al., 2014; Mari et al., 2015). However, the calibration with DREAM<sub>ZS</sub> over long time windows might result in overfitted posterior distributions which in most cases are not realistic and are the consequence of model bias and errors possibly due to temporal changes in the parameters. Moreover, in an operational scenario, the calibration should be repeated each time new epidemiological data becomes available, with high computational cost and a reduced capability to promptly forecast the epidemic.

### 3.2.2. Data assimilation

The second method we propose consists in inferring the distribution of the cholera model parameters in a dynamical way using DA. The main idea is that the parameter distribution can change in time and the DA scheme can sequentially track them using the collected data. The parameter update is performed in the analysis steps, which correct both the state variables and the parameter pdfs in the direction of the new observations, seeking to compute the filtering (or analysis) pdf  $p(\mathbf{x}_k, \boldsymbol{\vartheta}_k | \mathbf{y}_1, \dots, \mathbf{y}_k)$ . Using the Bayes formula, the filtering pdf rewrites in the product of the forecast pdf, i.e., the system state pdf predicted by the evolution of the model from  $t_{k-1}$  to  $t_k$ , and the likelihood function  $\mathcal{L}(\mathbf{y}_k | \mathbf{x}_k, \boldsymbol{\vartheta}_k)$  (see, e.g., Jazwinski, 1970).

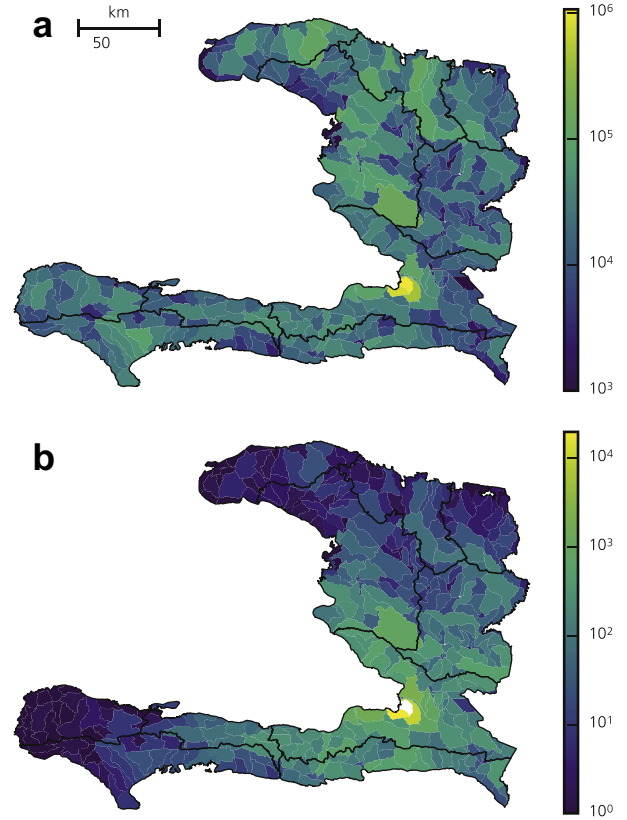
The recursion of forecast and analysis pdfs have an analytical solution only for linear and Gaussian models. Here we use an EnKF (Evensen, 2009), a method that approximates the forecast pdf with empirical distribution of several model solutions, a technique frequently adopted when dealing with nonlinear state-space model, such as the one defined in (1–4). Using the augmented-state technique, the filter is initialized with an ensemble of  $N$  random samples from the initial distribution of the state and parameter vectors,  $\{\mathbf{x}_0^{a,j}, \boldsymbol{\vartheta}_0^{a,j}\}_{j=1}^N \sim p(\mathbf{x}_0, \boldsymbol{\vartheta}_0)$ . The forecast pdf at an assimilation time  $t_k$  is approximated by the numerical solutions associated to the different realizations  $\{\mathbf{x}_k^{f,j}\}_{j=1}^N$ ,

$$\mathbf{x}_k^{f,j} = \mathcal{F}(\mathbf{x}_{k-1}^{a,j}, \mathbf{J}(t), \boldsymbol{\vartheta}_{k-1}^{a,j}, t_{k-1}, t_k), \quad (8)$$

where  $\mathcal{F}$  is the nonlinear operator solving (2)–(7) and the superscripts  $a$  and  $f$  indicate analysis and forecast, respectively. Note that the parameters are constant during the forecast, i.e.,  $\boldsymbol{\vartheta}_k^{f,j} = \boldsymbol{\vartheta}_{k-1}^{a,j}$ . In scenarios with uncertain rainfall conditions (e.g., when forecasting the future rainfall), it is convenient to model  $\mathbf{J}(t)$  as a random variable. In these cases, different samples of the precipitation,  $\mathbf{J}^j(t)$ , can be used in (8) for different realizations. In the analysis step of EnKF, both the state vector and the parameters are updated using the state augmentation:

$$\begin{pmatrix} \mathbf{x}_k^{a,j} \\ \boldsymbol{\vartheta}_k^{a,j} \end{pmatrix} = \begin{pmatrix} \mathbf{x}_k^{f,j} \\ \boldsymbol{\vartheta}_k^{f,j} \end{pmatrix} + \mathbf{K}_k^f (\mathbf{y}_k^j - \mathbf{y}_k^{f,j}) \quad (9)$$

where  $\mathbf{y}_k^{f,j} = \mathbf{H}\mathbf{x}_k^{f,j}$ . The vector  $\mathbf{y}_k^j$  represents the random perturbations of the observed measurements  $\mathbf{y}_k$ ,  $\mathbf{y}_k^j = \mathbf{y}_k + \boldsymbol{\xi}_k^j$ , which are introduced to correctly estimate the variance of the updated variables (e.g., Evensen, 2003).  $\mathbf{K}_k^f$  is an empirical approximation of the Kalman filter, where the correlations between forecast and obser-



**Fig. 2.** Model setup for the Haitian epidemic. (a) Population associated to each of the 365 watersheds estimated from the remotely sensed dataset of population distribution (LandScan algorithm of the Oak Ridge National Laboratory, <http://www.ornl.gov/landscan>). (b) Estimates of daily human mobility from the Haitian capital, Port-au-Prince, to the other watersheds, computed through the gravity model (6), using  $m = 0.05$  and  $D = 31.0$  km.

vations are computed through the ensemble (for more details see, e.g., Evensen, 2003).

A possible drawback of EnKF is the so-called filter inbreeding: the rapid convergence of the parameter distribution toward one value, with the consequent underestimation of the model uncertainty. Here, we use an adaptive inflation of the covariance error used in the computation of the Kalman gain (e.g., Anderson, 2007; Li et al., 2009) to reduce the inbreeding effect. The idea is to repeat the update step by gradually increasing the measurement error variance until the parameter variances  $\sigma_{\vartheta_k^a}$  are higher than a desired tolerance. At the  $i$ -th repetition of the update, we set the measurement error variance equal to  $c_1^i \sigma_\xi$ , with  $c_1 > 1$ , and the update is accepted if  $\sigma_{\vartheta_k^a} > c_2 \sigma_{\vartheta_k^f}$  for each parameter, with  $0 < c_2 < 1$ . This condition controls the decrease of the parameter variances during the simulation and, thus, of the probability space explored by the ensemble. The proposed approach is justified in our application by the high uncertainty associated with the epidemiological data, whose error variance is largely unknown.

### 3.3. Haitian model setup

Our model setup is equivalent to a previous application to the Haitian epidemic (Bertuzzo et al., 2014). The computational domain of the model has been derived by subdividing the Haitian territory into 365 watersheds, each of them hosting a human community whose size is determined using remotely acquired data (Fig. 2a). This allowed for the identification of the hydrological network, defining a unique downstream node (or, in coastal areas, the

sea) for each watershed and leading to the connectivity matrix  $\mathbf{P}$ . Distances  $d_{ij}$  among communities (Fig. 2b) have been extracted from the road network provided by the OpenStreetMap contributors (available on-line at [www.openstreetmap.org](http://www.openstreetmap.org)).

Two rainfall datasets are used to drive the cholera model, distinguishing between rainfall measurements and forecast. Daily rainfall measurements for each watershed have been computed starting from data collected by the NASA-JAXA's Tropical Rainfall Measuring Mission (TRMM\_3B42 precipitation estimates, resolution: 0.25°, see <http://trmm.gsfc.nasa.gov/> for details). Precipitation fields are first re-sampled at the resolution of the digital terrain model used to derive the watersheds (0.00083°) through linear interpolation and then averaged over the watershed area to obtain a representative value for each node. We assume these rainfall measurements to be error free. Under this assumption, the model uncertainty during simulations before the assimilation is completely determined by the probability distribution of the parameter, which accounts also for possible bias in the TRMM precipitation estimates.

The daily rainfall forecasts are obtained from the CFS climate reforecast from the beginning of the outbreak to March 31, 2011, and from the CFS operational climate forecast from the April 1, 2011 to present (data available on-line at <https://www.ncdc.noaa.gov>). CFS operational forecasts are computed daily starting at four different times (00, 06, 12, 18 UTC). For each of the four starting points the climatic data are forecast every six hours for about six months with a spatial resolution of 0.938° in longitude and 0.246° in latitude (about 104.3 × 27.46 km over Haiti). CFS climate reforecasts have the same spatial and temporal resolution but are available only every five days. We computed the daily forecast rainfall considering the CFS forecasts starting at 00 UTC. For each forecast day, the rainfall is averaged over the four forecast hours and then downscaled to the watershed scale as described for TRMM. To take into account the uncertainties introduced with the CFS forecasts and their possible bias, we perturb the rainfall time-series used in the cholera forecast with an additive error. The empirical error distribution is computed from the comparison between the CFS forecast rainfall and TRMM estimates from October 2010 to December 2013 and is found to have a temporal correlation of two days.

The initial conditions for infected people in each watershed,  $I_i(0)$ , are set according to the number of reported cases detailed in Piarroux et al. (2011) as of October 20, 2010 ( $t=0$ ). Specifically,  $I_i(0) = 1000$  in the watershed hosting Mirebalais, the commune where the first case of cholera was reported. Additional 1100 cases were distributed, according to population size, in the three watersheds downstream of Mirabelais along the Artibonite river that host the seven communes that was simultaneously struck by the outbreak on October 20. The initial number of recovered and the value of bacteria concentration are assumed to be in equilibrium with the infected cases, that is  $R_i(0) = \frac{1-\sigma}{\sigma} I_i(0)$  and  $B_i^*(0) = \theta I_i(0) / (H_i \mu_B)$ . The remaining fraction of the population is assumed to be susceptible because of the lack of any pre-existing immunity.

The DREAM algorithm is run with three chains, assuming a uniform prior distribution of the parameters and reflecting parameter boundaries. Concerning the EnKF setup, preliminary sensitivity analyses on the ensemble size  $N$  and on the tuning parameters  $c_1$  and  $c_2$  show that stable results are obtained with  $N = 1000$ ,  $c_1 = 2$ , and  $c_2 = 0.8$ . Reflecting parameter boundaries are enforced after the update to constrain the parameters within the prior boundaries. The condition  $S_i = H_i - I_i - R_i > 0$ , for  $i = 1, \dots, n$ , is checked for each realization of the ensemble, and is required to accept the updated state variables. The state variables of the realizations that do not satisfy this condition are not updated.

## 4. Results

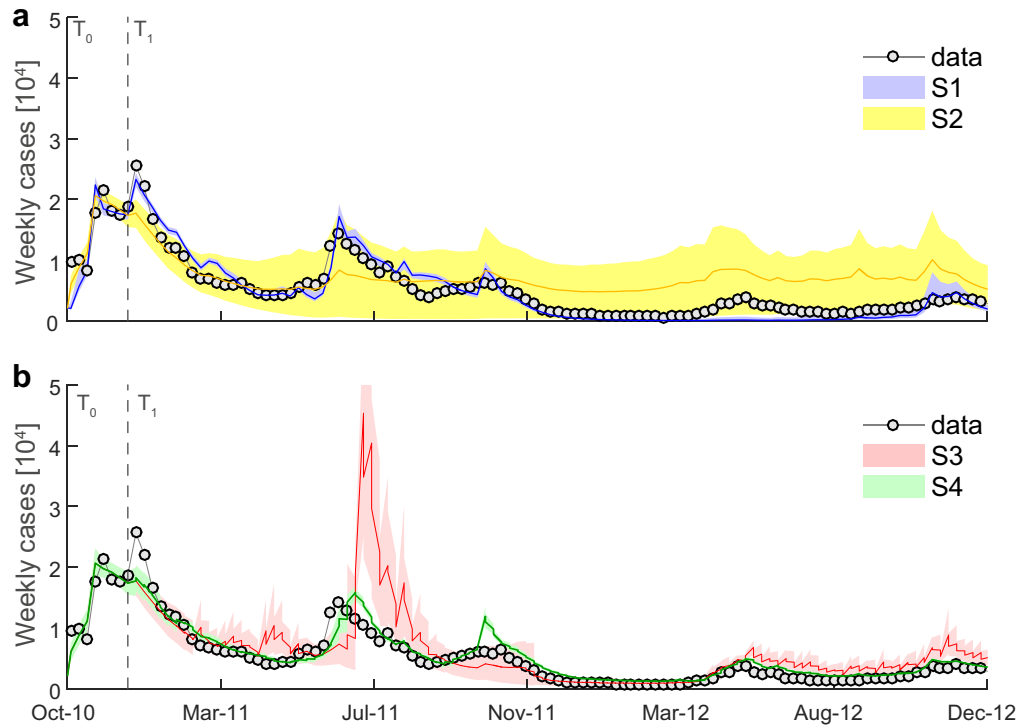
### 4.1. Assimilation analysis

To demonstrate the reliability of the proposed methodology, we consider four different scenarios (S1, S2, S3, and S4) for calibration and data assimilation of the model. The performance of the scenarios is assessed based on the ability to reproduce the first two years of the Haitian epidemic, from  $T_0 =$  October 20, 2010 to  $T_F =$  December 31, 2012.

- **S1:** The model is calibrated using DREAM on the complete set of data collected from  $T_0$  to  $T_F$ ;  $N = 1000$  random samples of the posterior distribution of the parameters are then used to assess the model response during  $[T_0, T_F]$ .
- **S2:** The posterior distribution of the parameters is computed using DREAM and considering only the data collected from  $T_0$  to  $T_1 =$  December 12, 2010. During this time window, the leading driver of the outbreak changed from hydrologic transport to human mobility, as detailed in Gaudart et al. (2013). We thus argue that the calibration window is long enough to sample different epidemiological dynamics and achieve a reasonable preliminary estimate of the parameters. The data collected during  $[T_1, T_F]$  are not assimilated during the simulation of the epidemics.
- **S3:** The parameters are calibrated as in S2; in the time interval  $[T_1, T_F]$  the model state variables are updated weekly using the EnKF procedure, without changing the associated parameters;
- **S4:** As in S3, but performing the EnKF update on the augmented state, correcting weekly both state variables and parameters. S4 corresponds to the methodology proposed in Section 2.

Scenario S1 computes the posterior parameter distribution that better retrieves the collected data without the use of a DA procedure. While this scenario is not feasible for operational forecasts, its comparison with S4 is useful to assess the performance of the proposed methodology in simulating the epidemics. Scenario S3 allows the assessment of the impact of the EnKF procedure in correcting the model trajectories. Finally, the comparison between S3 and S4 assesses the effect of the dynamical update of the parameters.

The model responses associated to each scenario are illustrated in terms of weekly cases for the whole country (Fig. 3) and disaggregated for each department (Figs. 4 and 5). Table 1 reports the ranges of the posterior parameter distribution computed in scenarios S1 and S2, while the dynamical update of the parameters in S4 is depicted in Fig. 6. These results show that the posterior parameter distribution obtained in scenario S1 is able to retrieve most of the epidemic curve. However, parameter uncertainty is underestimated and, as a consequence, the model does not reproduce all the epidemic peaks, e.g., the one occurring between April and August 2012 (Fig. 3a). The opposite situation is obtained in S2, where the posterior parameter distribution computed during the first weeks of the outbreak is too wide. In this scenario the model response is highly uncertain and suitable to assess the epidemic dynamics only for few months after the end of the calibration period (Fig. 3a). The weekly assimilation of the newly available data with EnKF improves model results, as highlighted in scenario S3 (Fig. 3b). During each week the erroneous parameters are driving the model far from the real state of the system, but the update steps correct the epidemic trajectories toward the measurements and reduce the model uncertainty. The combined update of the state variables and model parameters introduced in S4 reduces the errors during the forecast step, allowing the model to accurately follow the epidemic curve for the two years considered, improving the results obtained in S1 (Fig. 3).



**Fig. 3.** Comparison between the total reported weekly cases during the Haitian cholera epidemic and those estimated by the model in scenarios S1, S2 (a) and scenarios S3, S4 (b). For each scenario, the results are obtained with  $N = 1000$  model runs associated to random samples of the posterior distribution of the parameters. Lines and shaded areas represent the ensemble mean and the 90% confidence interval, respectively.

**Table 1**

Parameter values of the Haitian cholera model with the associated units as well as upper and lower boundaries. The 50th (5th–95th) percentiles of the posterior distributions computed in scenarios S1 and S2 are indicated.

Par.	Units	Prior	S1	S2
$\mu$	day <sup>-1</sup>	$4.5 \times 10^{-5}$		
$\gamma$	day <sup>-1</sup>	0.20		
$\alpha$	day <sup>-1</sup>	0.004		
$\beta$	day <sup>-1</sup>	0.01–10	0.25 (0.11–0.34)	4.80 (0.68–9.42)
$m$	–	0–1	0.054 (0.040–0.065)	0.043 (0.021–0.088)
$D$	km	1–300	245.5 (199.0–277.4)	31.0 (23.5–38.4)
$\rho$	day <sup>-1</sup>	0.0005–0.02	0.0013 (0.0010–0.0014)	0.0069 (0.0012–0.0137)
$\sigma$	–	0.03–0.20	0.0827 (0.0771–0.0937)	0.0333 (0.0303–0.0408)
$\mu_B$	day <sup>-1</sup>	0.01–1	0.15 (0.0104–0.26)	0.0147 (0.0102–0.054)
$\theta$	day <sup>-1</sup>	0.01–10	3.60 (3.43–4.01)	0.63 (0.21–6.05)
$l$	day <sup>-1</sup>	0.01–1	0.22 (0.18–0.29)	0.37 (0.21–0.80)
$\phi$	day/mm	0.01–2	0.111 (0.084–0.145)	0.022 (0.012–0.041)

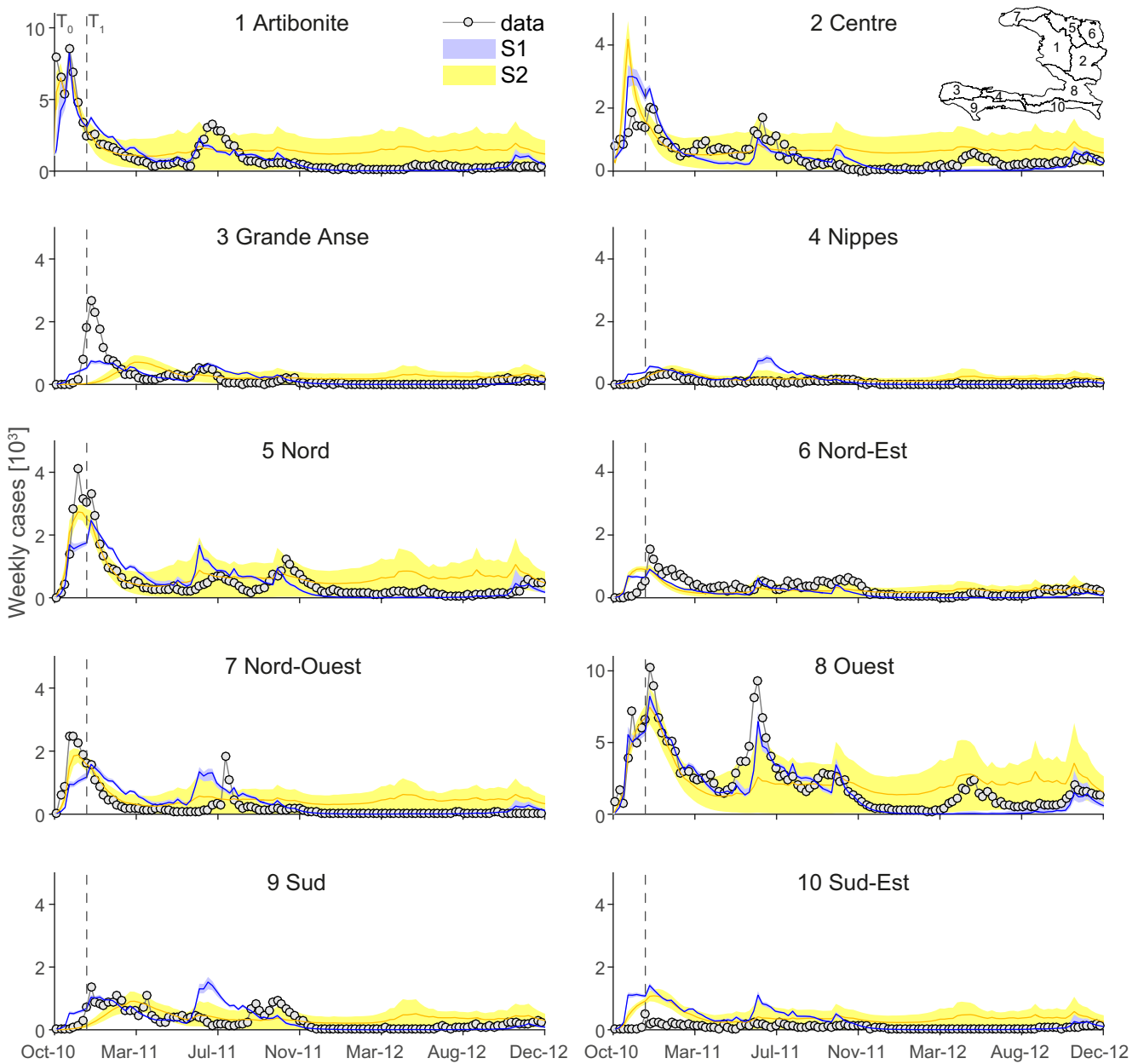
The spatial nature of the cholera model allows us to compare the four scenarios with the epidemiological data at the department level. Fig. 4 presents the results for S1 and S2, while Fig. 5 refers to S3 and S4. The limitations of scenarios S2 and S3 evinced in the aggregated results (Fig. 3) are here repeated in most of the departments. The uncertainty associated to scenario S2 is too wide and the mean number of modeled cases overestimates the reported cases in every department during 2012. The analysis of the performance of S3 in the different departments (Fig. 5) allows the identification of a possible drawback of data assimilation technics. In order to track the epidemic peak occurring in May/July 2011 mainly in the Ouest department, the EnKF overestimates, due to an erroneous representation of the spatial cross-correlation at that moment, the number of cases in all the other departments. This results in an overall overestimation of this peak at the country scale (Fig. 3). The assimilation with the augmented state (scenario S4, Fig. 5) effectively limits such drawback. Under scenarios S1 and S4 the cholera model well retrieves the epidemic curve in the departments of Artibonite, Nord, and Ouest, which are characterized

by large numbers of reported cases (Figs. 4 and 5). Both scenario poorly perform in the departments of Nippes and Sud-Est, where the small number of reported cases is constantly overestimated. The main advantage of S4 over S1 is the retrieval of the epidemic peak occurred during May/June 2012 in the Ouest and Centre departments (Fig. 5).

To quantitatively compare model performances, we compute the root mean square error (RMSE) between the modeled,  $y^f$ , and the total reported weekly cases,  $y$ , for each ensemble realization  $j$ :

$$\epsilon_k^j = \sqrt{\frac{\sum_{i=1}^d (y_{i,k} - y_{i,k}^{f,j})^2}{d}} \quad (10)$$

where  $k$  is the epidemiological week, and  $d$  is the number of the measurement points in space, here corresponding to the number the Haitian departments. Table 2 reports the 50th (5th–95th) percentiles of  $\epsilon_k^j$  averaged over the time-points  $k$ . The average errors associated with S4 are smaller than those in the other scenarios.



**Fig. 4.** Comparison between the weekly cases reported at the ten Haitian departments during the cholera epidemic and those estimated by the model in scenarios S1 and S2. Symbols as in Fig. 3. Inset shows the map of the ten Haitian departments. Notice that the dataset provided by Ministère de la Santé Publique et de la Population (<http://mspp.gouv.ht>) lists cases for the capital Port-au-Prince separately from its department, i.e. Ouest. However, due to the difficulties in determining where the cases reported at Port-au-Prince were actually coming from, these two time series have been aggregated for calibration.

**Table 2**

Scenario characteristics and associated temporal mean of the 50th (5th–95th) ensemble percentiles of the RMSE  $\epsilon^j = \langle \epsilon_k^j \rangle_k$  (i.e.  $\epsilon_k^j$  averaged over the time-points  $k$ ) described in (10).

Scenario	Calibration period	Data assimilation	$\epsilon^j (\times 10^3)$
S1	$[T_0, T_F]$	–	0.41 (0.35–0.44)
S2	$[T_0, T_1]$	–	0.64 (0.31–1.17)
S3	$[T_0, T_1]$	States	0.49 (0.32–0.79)
S4	$[T_0, T_1]$	States and parameters	0.33 (0.30–0.38)

4.2. Forecast analysis

Having demonstrated the reliability of the proposed methodology in reproducing cholera dynamics in the past, here we present the results of real-time forecasts and the associated errors. Fig. 7

presents the number of weekly cases forecast by the framework described in Section 2 one, two, and three months in advance, i.e., the ensemble state variables at the beginning of each forecast and the associated model parameters are computed according to scenario S4. The one-month forecasts retrieve many features of the temporal dynamic of the epidemic, but slightly delay the peak of infection that occurred during June 2011 (Fig. 7a). This is probably due to erroneous model parameters that require to be updated to reproduce that particular period of the epidemic. This statement is corroborated by the parameter dynamics depicted in Fig. 6, where it is evident that several parameters ( $\mu, \beta, \phi$ ) are subject to strong changes during such months. The delay in forecasting the cholera peaks seems to be mitigated after July 2011, showing the general reliability of the one-month cholera predictions (Fig. 7a). The two-month forecasts are subject to similar problems, but with



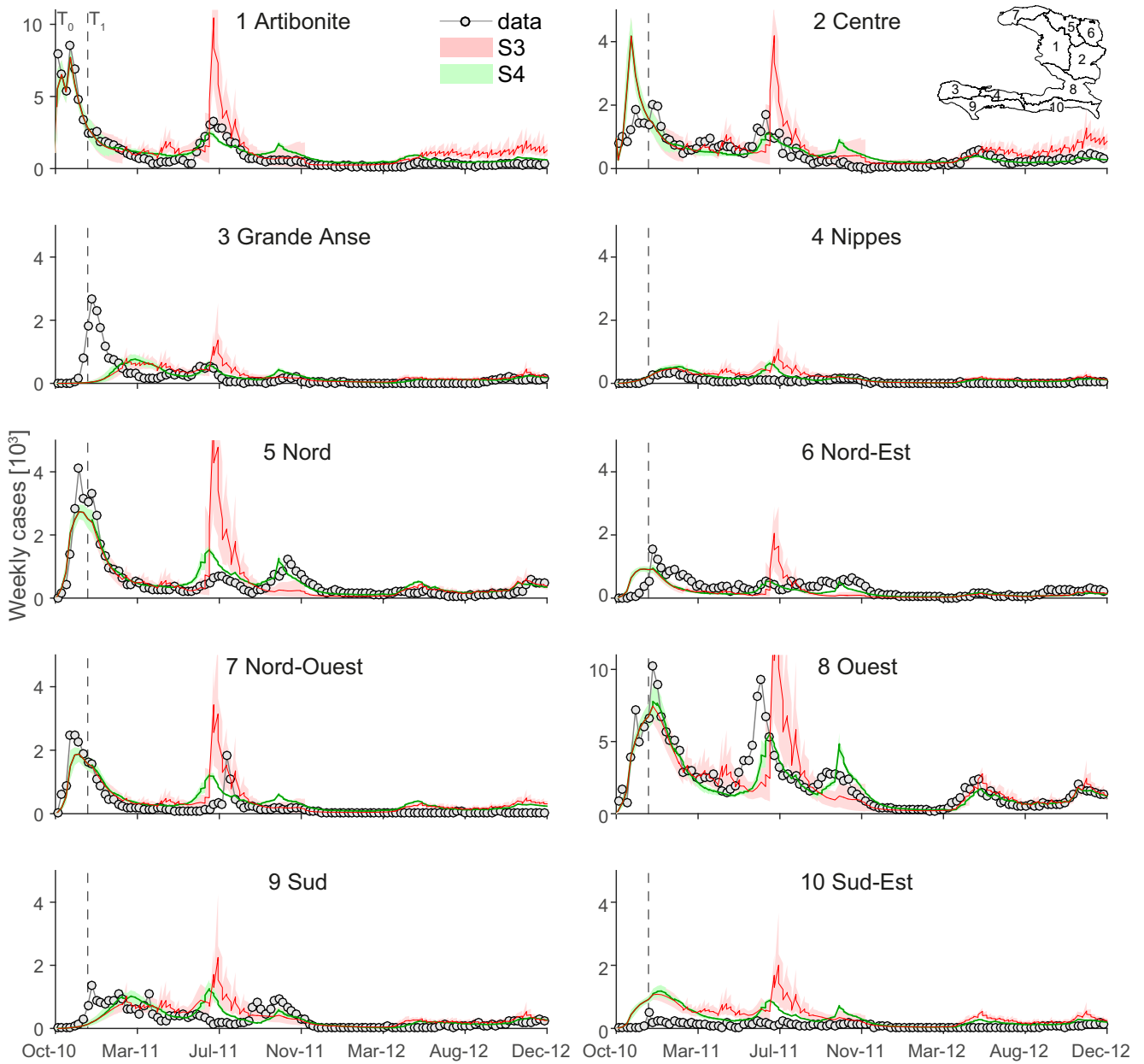


Fig. 5. Comparison between the weekly cases reported at the ten Haitian departments during the cholera epidemic and those estimated by the model in scenarios S3 and S4. Symbols as in Fig. 4.

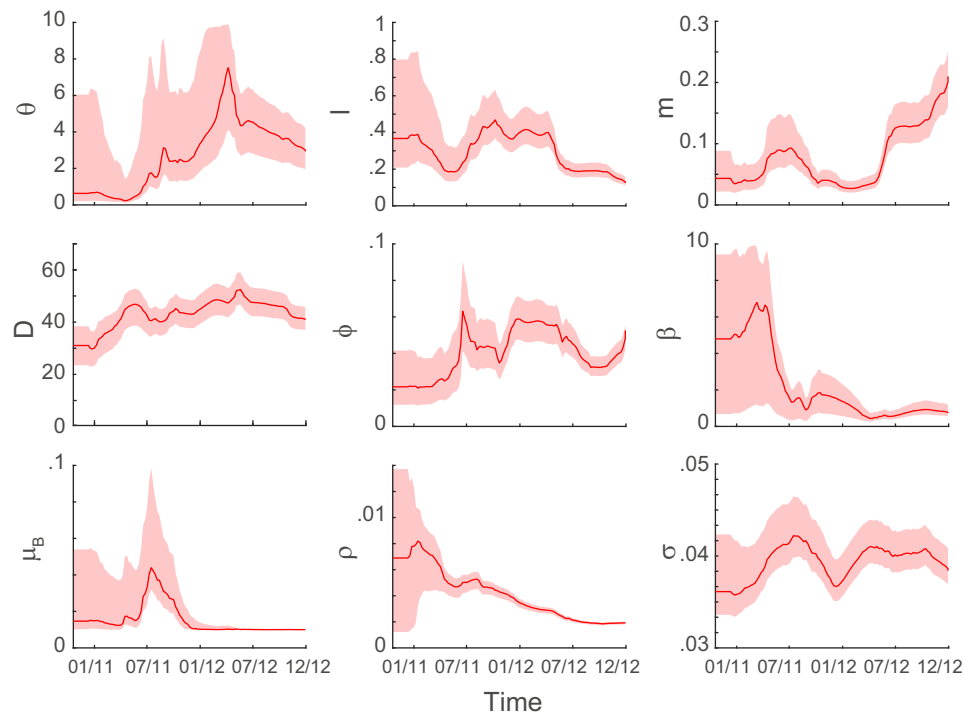
a stronger discrepancy between model and reported cholera peaks (Fig. 7b). The errors highly increase in the three-month forecasts, where the model response do not reproduce the data (Fig. 7c).

To assess if the update of the parameter distribution is effectively improving the forecast, we compared the results of S4 with the forecast obtained using scenarios S1, where the parameters are computed using DREAM on the whole time series, and S3, where only the state variables are updated by the EnKF. Note that S1 is used only for comparison purposes, not as a possible alternative candidate. Indeed S1 cannot be used in a real forecast scenario, as the parameters are estimated using the whole epidemiological dataset, information not available at the time of forecast. For each forecast issued and for each model realization  $j$ , we compute the following spatio-temporal RMSE:

$$\eta_k^j = \sqrt{\frac{\sum_{i=1}^d \sum_{l=1}^{n_l} (y_{i,k+l} - y_{i,k+l}^{f,j})^2}{d n_l}} \quad (11)$$

where  $k$  is the epidemiological week when the forecast is computed,  $d$  the number of measurement points in space,  $n_l$  is the number of weeks in the forecast, and  $y_{i,k+l}^{f,j}$  is the number of forecasted cases for the  $j$ th realization, at the  $i$ th measurement point,  $l$  weeks after  $k$ . Table 3 reports the 50th (5th–95th) percentiles of  $\eta_k^j$  averaged over the time-points  $k$  for  $n_l = 4, 8$  and 12 weeks. The table shows that the errors associated with the proposed methodology S4 are slightly smaller than the errors obtained by the DREAM calibration S1, while the forecast obtained without the parameter update (S3) show much higher errors. These results clearly highlight that the update of the parameter distribution in the EnKF assimilation is crucial in order to reduce the forecast errors.

To assess the performance of the CFS estimated rainfall on the cholera forecasts, we consider a final scenario where cholera forecasts are performed with the observed rainfall. The associated errors, reported in the last row of Table 3, are comparable with those of scenario 4 obtained using the forecast rainfall.



**Fig. 6.** Empirical distribution of parameters computed by the proposed methodology (scenario S4) with an ensemble size  $N = 1000$ . The thick lines and the extremes of the shaded areas represent the 50th and 5th–95th percentiles of the empirical distribution.

**Table 3**

Forecast scenarios characteristics and associated temporal mean of the 50th (5th–95th) ensemble percentiles of the RMSE  $\eta^j = \langle \eta_k^j \rangle_k$  (i.e.  $\eta_k^j$  averaged over the time-points  $k$ ) described in (11).

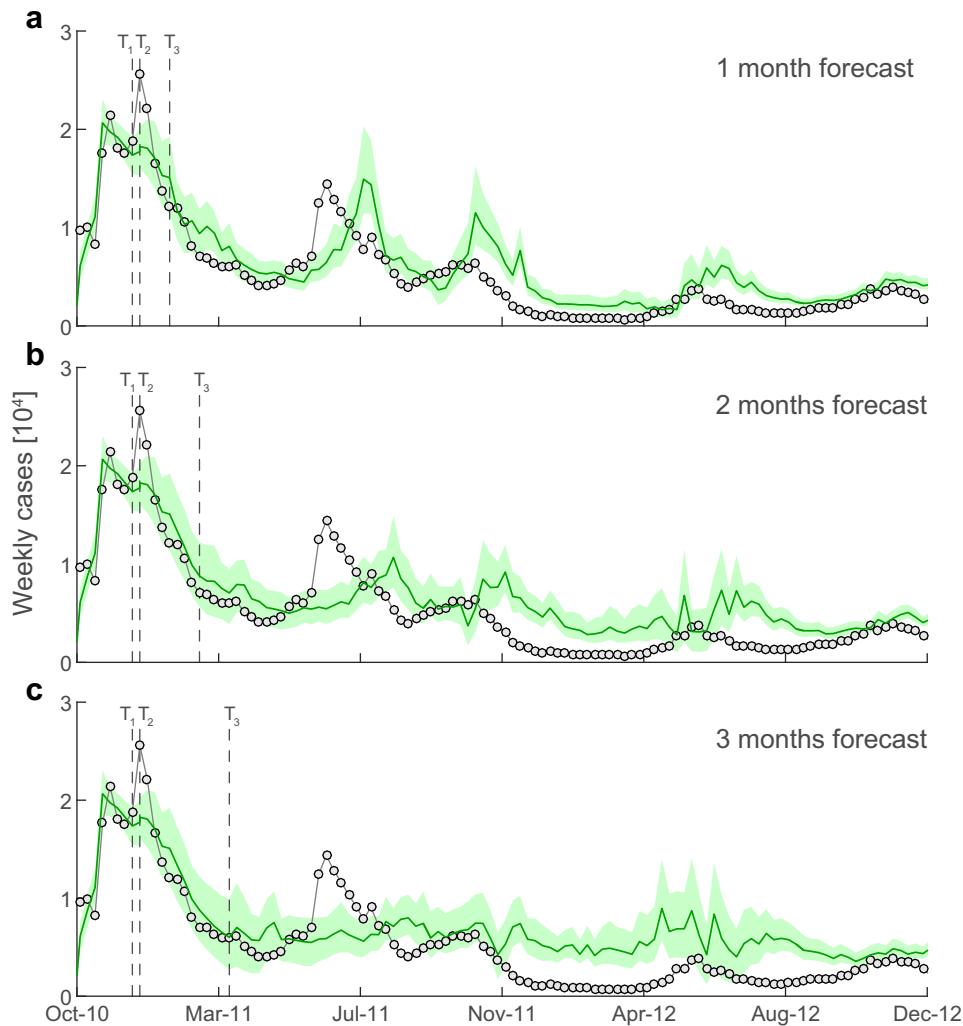
Scenario	Rainfall model	$\eta^j (\times 10^3)$		
		At one month	At two months	At three months
S2	CFS forecasts	0.88 (0.61–1.63)	1.40 (0.93–2.62)	1.81 (1.20–3.14)
S3	CFS forecasts	1.08 (0.70–1.72)	1.54 (1.02–2.66)	2.04 (1.32–3.70)
S4	CFS forecasts	0.74 (0.62–0.99)	1.23 (1.01–1.67)	1.69 (1.38–2.28)
S4	Measured	0.76 (0.63–1.02)	1.27 (1.03–1.73)	1.79 (1.42–2.45)

## 5. Discussion and conclusions

We presented an innovative methodology for the real-time forecasting of cholera epidemics, which employs a spatially-explicit numerical model driven by CFS rainfall forecasts. The model is calibrated on the epidemiological data collected at the beginning of the epidemic using a MCMC scheme, while EnKF is used for the on-line update of the state variable and parameter distributions during the simulation. Forecast spatial distribution of cholera incidence is reliable for at least one month, while errors expectedly increase for longer time horizons. However, one month lead time seems like a worthwhile goal of public health predictions anywhere, in the writers' view, to assess needs of medical supplies and staff/infrastructure.

A possible alternative to the use of DA schemes is to recalibrate the model parameters every week, using the whole dataset from the beginning of the epidemic to the latest datapoint available, and to use the parameter posterior distribution to project the future evolution of the outbreak. We argue that a DA scheme offers several advantages with respect to a repeated calibration scheme. The first advantage is rather technical and is related to computational time. An update of state variables and parameters in our DA scheme takes less than a second on a standard desktop machine. On the other hand, calibration requires on the order of  $10^5$  model runs (around 1 day of computing time spread over 12 cores). The

computing time argument might not be conclusive, in particular if the forecast bulletins are foreseen to be issued with a weekly frequency; however, it should not be completely discarded as the framework is supposed to be implemented during emergencies. In terms of forecasting performance, the DA scheme updates the state of the system based on the latest observations; therefore the forecast starts from a state which is close to the observed one, a feature that is not necessarily satisfied by a standard calibration scheme. Finally, the main advantage of using a DA scheme with the state augmentation technique (Evensen, 2009) is that it can track the possible time-evolution of parameters and thus detect possible directional changes. Indeed, some of the model parameters may change during an outbreak. In particular, exposure to cholera (here represented by the rate  $\beta$ ) reportedly decreases as interventions unfold (e.g., distribution of safe water and information campaigns) and the population awareness of cholera transmission risk factors increases (de Rochars et al., 2011). The time evolution of the parameter  $\beta$  (Fig. 6) seems to suggest such a trend. However, a trend in the time evolution of a parameter might not necessarily reflect changes in the actual processes but rather be a byproduct of the challenge of identifying such parameter as the dominant drivers of the epidemic change. This is likely to be the case for the rate of loss of acquired immunity,  $\rho$ . Indeed, the process of immunity loss can affect the outbreak dynamics only when previously infected people replenish the susceptible population. Therefore the onset of



**Fig. 7.** Weekly cholera cases forecast one (a), two (b), and three (c) months ahead using scenario S4. The mean and the 90% confidence interval of the model trajectories are reported from  $T_0$  to  $T_3$ , as in Fig. 1. From  $T_3$  to the end of the simulation only the cases at the end of the forecast period are reported.

the outbreak is almost insensible to the rate  $\rho$ , which can be identified only at later stages (Fig. 6). The same issue could also apply to the fraction of mobile people  $m$ , which is crucial at the beginning of the outbreak but becomes difficult to estimate once the epidemic has spread over the whole country.

Our results indicate that the EnKF sequential assimilation improves the spatio-temporal reproduction of the epidemics with respect to the classical model calibration on the whole dataset, with the consequential reduction of the forecast error on the reported cases (Table 3). Discrepancies between model forecasts and observations can be attributed to three main sources: model structural errors, parameter uncertainty and rainfall forecast uncertainty. The comparison between the projections obtained using forecast and observed rainfall shows that rainfall projections are relatively good in this context and that most of the uncertainty comes from the epidemiological dynamics rather than external forcing. This result might also be due to the fact that high frequency components of rainfall are filtered out by the epidemiological dynamics and only seasonal components, which are arguably well-captured by the forecasts, matter.

The feasibility of a real-time forecasting system for cholera outbreak critically depends on the immediate implementation of a surveillance system and the release of the relevant epidemiological data. In the case of the outbreak in Haiti, data aggregated

over the ten departments (Fig. 2) were made readily available. This dataset was processed starting from higher resolution data (Barzilay et al., 2013), which, however, were not publicly released. Spatially-distributed data is indeed crucial for the calibration of spatially-explicit models like the one considered in this study, as they might allow to consider heterogeneous parameters over the infected area. In this scenario, owing to a much larger number of parameters, classical Bayesian methods for the calibration of the model might face major difficulties. On the contrary, several studies (e.g., ELSheikh et al., 2012; Pasetto et al., 2015) show the effectiveness of sequential approaches, such as EnKF, to retrieve the spatial distributions of model parameters, demonstrating the importance of considering DA procedures for epidemiological projections.

#### Acknowledgments

EB, FF and AR acknowledge the support from the Swiss National Science Foundation (SNF/FNS) project “Dynamics and controls of large-scale cholera outbreaks” (DYCHO CR2312 138104).

#### References

- Abrams, J.Y., Copeland, J.R., Tauxe, R.V., Date, K.A., Belay, E.D., Mody, R.K., Mintz, E.D., 2013. Real-time modelling used for outbreak management during a cholera epidemic, Haiti, 2010–2011. *Epidemiol. Infect.* 141 (6), 1276–1285.

- Anderson, J.L., 2007. An adaptive covariance inflation error correction algorithm for ensemble filters. *Tellus A* 59 (2), 210–224. <https://doi.org/10.3402/tellusa.v59i2.14925>.
- Andrews, J.R., Basu, S., 2011. Transmission dynamics and control of cholera in Haiti: An epidemic model. *Lancet* 377, 1248–1252.
- Arulampalam, M.S., Maskell, S., Gordon, N., Clapp, T., 2002. A tutorial on particle filters for online nonlinear/non-Gaussian Bayesian tracking. *IEEE T. Signal Process.* 50 (2), 174–188. <https://doi.org/10.1109/78.978374>.
- Barzilay, E.J., Schaad, N., Magloire, R., Mung, K.S., Boncy, J., Dahourou, G.A., Mintz, E.D., Steenland, M.W., Vertefeuille, J.F., Tappero, J.W., 2013. Cholera surveillance during the Haiti epidemic - the first 2 years. *N. Engl. J. Med.* 368 (7), 599–609.
- Bertuzzo, E., Azaele, S., Maritan, A., Gatto, M., Rodriguez-Iturbe, I., Rinaldo, A., 2008. On the space-time evolution of a cholera epidemic. *Water Res. Res.* 44, W01424.
- Bertuzzo, E., Casagrandi, R., Gatto, M., Rodriguez-Iturbe, I., Rinaldo, A., 2010. On spatially explicit models of cholera epidemics. *J. R. Soc. Interface* 7, 321–333.
- Bertuzzo, E., Finger, F., Mari, L., Gatto, M., Rinaldo, A., 2014. On the probability of extinction of the Haiti cholera epidemic. *Stochastic Environ. Res. Risk Assess.* 28. <https://doi.org/10.1007/s00477-014-0906-3>.
- Bertuzzo, E., Mari, L., Righetto, L., Gatto, M., Casagrandi, R., Rodriguez-Iturbe, I., Rinaldo, A., 2011. Prediction of the spatial evolution and effects of control measures for the unfolding Haiti cholera outbreak. *Geophys. Res. Lett.* 38 (L06403).
- Camporese, M., Paniconi, C., Putti, M., Salandini, P., 2009. Ensemble Kalman filter data assimilation for a process-based catchment scale model of surface and subsurface flow. *Water Resources Research* 45, W10421. <https://doi.org/10.1029/2008WR007031>.
- Cazelles, B., Chau, N.P., 1997. Using the Kalman filter and dynamic models to assess the changing HIV/AIDS epidemic. *Math. Biosci.* 140 (2), 131–154. [https://doi.org/10.1016/S0025-5564\(96\)00155-1](https://doi.org/10.1016/S0025-5564(96)00155-1).
- Chao, D.L., Halloran, M.E., Longini Jr., I.M., 2011. Vaccination strategies for epidemic cholera in Haiti with implications for the developing world. *Proc. Nat. Acad. Sci. U. S. A.* 108 (17), 7081–7085.
- Chretien, J.-P., George, D., Shaman, J., Chitale, R.A., McKenzie, F.E., 2014. Influenza forecasting in human populations: a scoping review. *Plos ONE* 9 (4), e94130. <https://doi.org/10.1371/journal.pone.0094130>.
- Codeço, C., 2001. Endemic and epidemic dynamics of cholera: The role of the aquatic reservoir. *BMC Infect. Dis.* 1 (1).
- Eisenberg, M.C., Kujbida, G., Tuite, A.R., Fisman, D.N., Tien, J.H., 2013. Examining rainfall and cholera dynamics in Haiti using statistical and dynamic modeling approaches. *Epidemics* 5 (4), 197–207.
- ELSheikh, A.H., Pain, C.C., Fang, F., Gomes, J.L.M.A., Navon, I.M., 2012. Parameter estimation of subsurface flow models using iterative regularized ensemble Kalman filter. *Stochastic Environ. Res. Risk Assess.* 27 (4), 877–897. <https://doi.org/10.1007/s00477-012-0613-x>.
- Erlander, S., Stewart, N.F., 1990. *The Gravity Model in Transportation Analysis - Theory and Extensions*. VSP Books, Zeist, The Netherlands.
- Evensen, G., 1994. Sequential data assimilation with a nonlinear quasi-geostrophic model using Monte Carlo methods to forecast error statistics. *J. Geophys. Res.-Oceans* 99 (C5), 10143–10162. <https://doi.org/10.1029/94JC00572>.
- Evensen, G., 2003. The ensemble Kalman filter: theoretical formulation and practical implementation. *Ocean Dyn.* 53 (4), 343–367. <https://doi.org/10.1007/s10236-003-0036-9>.
- Evensen, G., 2009. The ensemble Kalman filter for combined state and parameter estimation. *IEEE CONTR. Syst. Mag.* 29 (3), 83–104. <https://doi.org/10.1109/MCS.2009.932223>.
- Gatto, M., Mari, L., Bertuzzo, E., Casagrandi, R., Righetto, L., Rodriguez-Iturbe, I., Rinaldo, A., 2012. Generalized reproduction numbers and the prediction of patterns in waterborne disease. *Proc. Nat. Acad. Sci. USA* 48, 19703–19708.
- Gaudart, J., Rebaudet, S., Barraix, R., Boncy, J., Faucher, B., Piarroux, M., Magloire, R., Thimothe, G., Piarroux, R., 2013. Spatio-temporal dynamics of cholera during the first year of the epidemic in Haiti. *PLoS Negl. Trop. Dis.* 7 (4), e2145.
- Houtekamer, P.L., Mitchell, H.L., Pellerin, G., Buehner, M., Charron, M., Spacek, L., Hansen, B., 2005. Atmospheric data assimilation with an ensemble Kalman filter: results with real observations. *Mon. Weather Rev.* 133 (3), 604–620. <https://doi.org/10.1175/MWR-2864.1>.
- Jazwinski, A.H., 1970. *Stochastic Processes and Filtering Theory*. Academic Press, New York.
- Jewell, C.P., Brown, R.G., 2015. Bayesian data assimilation provides rapid decision support for vector-borne diseases. *J. R. Soc. Interf.* 12 (108). <https://doi.org/10.1098/rsif.2015.0367>.
- Kalman, R.E., 1960. A new approach to linear filtering and prediction problems. *ASME J. Basic Eng.* 82 (1), 35–45. <https://doi.org/10.1115/1.3662552>.
- Kalman, R.E., Bucy, R.S., 1961. New results in linear filtering and prediction theory. *ASME J. Basic Eng.* 83 (1), 95–108. <https://doi.org/10.1115/1.3658902>.
- Kirpich, A., Weppelmann, T.A., Yang, Y., Ali, A., Morris Jr., J.G., Longini, I.M., 2015. Cholera transmission in Ouest department of Haiti: dynamic modeling and the future of the epidemic. *PLoS Negl. Trop. Dis.* 9 (10), e0004153. <https://doi.org/10.1371/journal.pntd.0004153>.
- Li, H., Kalnay, E., Miyoshi, T., 2009. Simultaneous estimation of covariance inflation and observation errors within an ensemble Kalman filter. *Q. J. R. Meteor. Soc.* 135 (639), 523–533. <https://doi.org/10.1002/qj.371>.
- Mari, L., Bertuzzo, E., Finger, F., Casagrandi, R., Gatto, M., Rinaldo, A., 2015. On the predictive ability of mechanistic models for the Haitian cholera epidemic. *J. R. Soc. Interf.* 12 (104), 20140840. <https://doi.org/10.1098/rsif.2014.0840>.
- Mari, L., Bertuzzo, E., Righetto, L., Casagrandi, R., Gatto, M., Rodriguez-Iturbe, I., Rinaldo, A., 2012a. On the role of human mobility in the spread of cholera epidemics: Towards an epidemiological movement ecology. *Ecology* 5, 531–540.
- Mari, L., Bertuzzo, E., Righetto, L., Casagrandi, R., Gatto, M., Rodriguez-Iturbe, I., Rinaldo, A., 2012b. Modelling cholera epidemics: The role of waterways, human mobility and sanitation. *J. R. Soc. Interf.* 9, 376–388.
- Mathers, C., Fat, D. M., Boerma, J. T., 2008. *The global burden of disease: 2004 update*. World Health Organization (Geneva).
- Mukandavire, Z., Smith, D.L., Morris Jr., J.G., 2013. Cholera in Haiti: reproductive numbers and vaccination coverage estimates. *Nat. Sci. Rep.* 3, 1–8 article number: 997.
- Navon, I.M., 2009. Data assimilation for numerical weather prediction: a review. In: Park, S., Xu, L. (Eds.), *Data Assimilation for Atmospheric, Oceanic and Hydrologic Applications*. Springer Berlin Heidelberg, Berlin, Heidelberg, pp. 21–65. [https://doi.org/10.1007/978-3-540-71056-1\\_2](https://doi.org/10.1007/978-3-540-71056-1_2).
- Pasetto, D., Camporese, M., Putti, M., 2012. Ensemble Kalman filter versus particle filter for a physically-based coupled surface-subsurface model. *Adv. Water Res.* 47 (1), 1–13. <https://doi.org/10.1016/j.advwatres.2012.06.009>.
- Pasetto, D., Niu, G.-Y., Pangle, L., Paniconi, C., Putti, M., Troch, P.A., 2015. Impact of sensor failure on the observability of flow dynamics at the biosphere 2 LEO hillslopes. *Adv. Water Res.* 86, 327–339. <https://doi.org/10.1016/j.advwatres.2015.04.014>. Part B
- Piarroux, R., Barraix, R., Faucher, B., Haus, R., Piarroux, M., Gaudart, J., Magloire, R., Raoult, D., 2011. Understanding the cholera epidemic, Haiti. *Emerging Infect. Dis.* 17, 1161–1168.
- Rabier, F., 2005. Overview of global data assimilation developments in numerical weather-prediction centres. *Q. J. R. Meteorol. Soc.* 131 (613), 3215–3233. <https://doi.org/10.1256/qj.05.129>.
- Righetto, L., Bertuzzo, E., Mari, L., Schild, E., Casagrandi, R., Gatto, M., Rodriguez-Iturbe, I., Rinaldo, A., 2013. Rainfall mediations in the spreading of epidemic cholera. *Adv. Water Res.* 60, 34–46.
- Rinaldo, A., Bertuzzo, E., Mari, L., Righetto, L., Blokesch, M., Gatto, M., Casagrandi, R., Murray, M., Veseneckh, S.M., Rodriguez-Iturbe, I., 2012. Reassessment of the 2010–2011 Haiti cholera outbreak and rainfall-driven multiseason projections. *Proc. Nat. Acad. Sci. U. S. A.* 109 (17), 6602–6607.
- de Rochars, V.E.M.B., Tipret, J., Patrick, M., Jacobson, L., Barbour, K.E., Berendes, D., Bensyl, D., Frazier, C., Domercant, J.W., Archer, R., Roels, T., Tappero, J.W., Handzel, T., 2011. Knowledge, attitudes, and practices related to treatment and prevention of cholera, Haiti, 2010. *Emerging Infect. Dis.* 17 (11), 2158–2161.
- Ryan, E.T., Dhar, U., Khan, W.A., Salam, M.A., Faruque, A.S.G., Fuchs, G.J., Calderwood, S.B., Bennis, M.L., 2000. Mortality, morbidity, and microbiology of endemic cholera among hospitalized patients in dhaka, bangladesh. *Am. J. Trop. Med. Hygiene* 63 (1–2), 12–20.
- Saha, S., Moorthi, S., Wu, X., Wang, J., et al., 2014. The ncep climate forecast system version 2. *J. Climate* 27 (6), 2185–2208. <https://doi.org/10.1175/JCLI-D-12-00823.1>.
- Shaman, J., Karspeck, A., 2012. Forecasting seasonal outbreaks of influenza. *Proc. Nat. Acad. Sci. U. S. A.* 109 (50), 20425–20430. <https://doi.org/10.1073/pnas.1208772109>.
- Shaman, J., Karspeck, A., Yang, W., Tamerius, J., Lipsitch, M., 2013. Real-time influenza forecasts during the 2012–2013 season. *Nat. Commun.* 4. <https://doi.org/10.1038/ncomms3837>.
- Shampine, L.F., Reichelt, M.W., 1997. *The MATLAB ODE suite*. *SIAM J. Sci. Comput.* 18 (50).
- Thompson, P.D., 1961. A dynamical method of analyzing meteorological data. *Tellus* 13 (3), 334–349. <https://doi.org/10.1111/j.2153-3490.1961.tb00094.x>.
- Tuite, A.L., Tien, J., Eisenberg, M., Earn, D.J.D., Ma, J., Fisman, D.N., 2011. Cholera epidemic in Haiti, 2010: Using a transmission model to explain spatial spread of disease and identify optimal control interventions. *Ann. Int. Med.* 154, 593–601.
- Vrugt, J.A., Braak, C.J.F., Diks, C.G.H., Robinson, B.A., Hyman, J.M., Higdon, D., 2009. Accelerating Markov chain Monte Carlo simulation by differential evolution with self-adaptive randomized subspace sampling. *Int. J. Nonlin. Sci. Numer. Simul.* 10, 271–288.
- Wu, H., Tan, W.-Y., 2000. Modelling the HIV epidemic: A state-space approach. *Math. Comput. Modell.* 32 (1–2), 197–215. [https://doi.org/10.1016/S0895-7177\(00\)00129-1](https://doi.org/10.1016/S0895-7177(00)00129-1).
- Yang, W., Cowling, B.J., Lau, E.H.Y., Shaman, J., 2015b. Forecasting influenza epidemics in hong kong. *PLoS Computat. Biol.* 11 (7). <https://doi.org/10.1371/journal.pcbi.1004383>.
- Yang, W., Lipsitch, M., Shaman, J., 2015a. Inference of seasonal and pandemic influenza transmission dynamics. *Proc. Nat. Acad. Sci. U. S. A.* 112 (9), 2723–2728. <https://doi.org/10.1073/pnas.1415012112>.
- Yang, W., Zhang, W., Kargbo, D., Yang, R., Chen, Y., Chen, Z., Kamara, A., Kargbo, B., Kandula, S., Karspeck, A., Liu, C., Shaman, J., 2015c. Transmission network of the 2014–2015 ebola epidemic in sierra leone. *J. R. Soc. Interf.* 12 (112). <https://doi.org/10.1098/rsif.2015.0536>.
- Zhou, Y., McLaughlin, D., Entekhabi, D., 2006. Assessing the performance of the ensemble Kalman filter for land surface data assimilation. *Mon. Weather Rev.* 134 (8), 2128–2142. <https://doi.org/10.1175/MWR3153.1>.