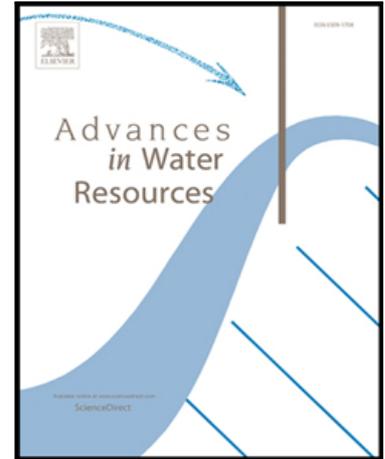


Accepted Manuscript

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

Damiano Pasetto, Flavio Finger, Andrea Rinaldo, Enrico Bertuzzo

PII: S0309-1708(16)30521-8
DOI: [10.1016/j.advwatres.2016.10.004](https://doi.org/10.1016/j.advwatres.2016.10.004)
Reference: ADWR 2706



To appear in: *Advances in Water Resources*

Received date: 11 May 2016
Revised date: 31 August 2016
Accepted date: 6 October 2016

Please cite this article as: Damiano Pasetto, Flavio Finger, Andrea Rinaldo, Enrico Bertuzzo, Real-time projections of cholera outbreaks through data assimilation and rainfall forecasting, *Advances in Water Resources* (2016), doi: [10.1016/j.advwatres.2016.10.004](https://doi.org/10.1016/j.advwatres.2016.10.004)

This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

Highlights

- Rainfall-driven cholera epidemic forecasts through a spatially-explicit model
- Precipitation estimates of a climate forecast system drive future epidemic forecast
- Sequential assimilation of reported infected cases improves the forecast accuracy

ACCEPTED MANUSCRIPT

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

Damiano Pasetto^a, Flavio Finger^a, Andrea Rinaldo^{a,b}, Enrico Bertuzzo^{a,*}

^a*Laboratory of Ecohydrology, School of Architecture, Civil and Environmental Engineering,
Ecole Polytechnique Fédérale de Lausanne, Lausanne, Switzerland*

^b*Dipartimento ICEA, Università di Padova, Padova, Italy*

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

*Corresponding author

Email address: enrico.bertuzzo@epfl.ch (Enrico Bertuzzo)

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 for 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.

Keywords: Epidemiological model, Data assimilation, Cholera, Rainfall forecast, Climate Forecast System

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 [1]. In regions where the disease is endemic (e.g., Bangladesh) the case fatality rate is relatively low (around 0.1%, see e.g., [2]) 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 9,200 deaths with an overall case fatality rate

19 of 1.15%, which was even higher (around 2%) during the first months [3] (data
20 available on-line at <http://mspp.gouv.ht>). Thus, modeling tools which can pos-
21 sibly predict the evolution of an ongoing outbreak in time for interventions are
22 of paramount importance to guide health care officials in allocating staff and re-
23 sources and evaluating alternative control strategies.

24 The quasi-real time release of the epidemiological data during the Haitian
25 cholera outbreak prompted many research teams to develop epidemiological models
26 of the outbreak in an effort to provide meaningful insights to guide the emergency
27 management [4–16]. Some of these studies [4, 5, 8, 10, 13, 14] attempted to ac-
28 tually forecast the evolution of the unfolding outbreak by calibrating a model on
29 the data available at a certain moment in time and projecting the simulations into
30 the future. Early attempts show contrasting results (for a complete reassessment
31 see [8]). The ability to predict under different modeling assumptions has later
32 been analyzed in detail [16], showing that, when data is scarce, spatially-explicit
33 models [e.g., 4] clearly outperform models that do not account for the spatial cou-
34 pling among individual local models [e.g., 5]. The revamping of the outbreak in
35 conjunction with the rainy season in spring 2011 revealed empirically that, at least
36 in the Haitian context, intense rainfall enhances cholera transmission and therefore
37 has to be taken into account for future model developments and predictions [8].
38 This consideration further complicates modelers' task because it implies that in
39 order to predict cholera incidence one must also predict precipitation intensity in
40 space and time. So far, this issue has been tackled by producing realistic rainfall
41 scenarios using stochastic models of rainfall generation [13] or by bootstrapping of
42 past observed rainfall fields [8, 14].

43 All the previous examples represent isolated attempts to forecast cholera dy-
44 namics, each based on different assumptions to accommodate relevant processes
45 and recalibration on the available data. Here, we aim at proving the feasibility of

46 a real-time forecasting framework during emergencies that: i) flexibly adapts to
47 account for the dominant processes driving the outbreak, ii) readily integrates new
48 information available, and iii) periodically issues an updated forecast for a prede-
49 fined time horizon. We therefore set ourselves at the initial phase of the Haitian
50 cholera outbreak and produce weekly bulletins forecasting the spatio-temporal dis-
51 tribution of new cases for the first two years of the epidemics using only information
52 that was available at that time.

53 The first major innovation of this study with respect to previous efforts is the
54 use of a data assimilation (DA) framework to integrate new epidemiological data
55 as soon as they become available and to update the model forecast without recal-
56 ibrating. DA has long been used in weather forecasting [17, 18], where numerical
57 models require frequent re-initializations to track the real dynamics and to avoid
58 the rapid divergence of the numerical solution. This procedure is typically per-
59 formed by the data assimilation cycle [19], the sequential repetition of a forecast
60 step and its correction in the analysis (or update) step using the newly available
61 system observations. Forecast and analysis steps are naturally formulated in a
62 Bayesian framework by the so-called filtering problem [20], which seeks the pos-
63 terior probability distribution of the system state, given all the observations in a
64 time window of interest, and takes into account the model uncertainties and the
65 observation errors. While the well-known Kalman-Bucy filter [21, 22] solves the
66 filtering problem in the simple case of linear models with additive and Gaussian
67 errors, an analytical solution in the presence of nonlinearities does not exist and
68 several alternative filters have been proposed in literature [see e.g., 23]. The ensem-
69 ble Kalman filter (EnKF), developed by *Evensen* [24, 25] for nonlinear applications
70 in the context of ocean modeling, is one of the most popular DA techniques and
71 consists in an ensemble approximation of the Kalman filter. Although optimal only
72 for Gaussian distributions of state variables, EnKF typically delivers satisfactory

73 performances using a small number of model realizations also for non-Gaussian
74 models [26], a feature that favored its application in different fields including at-
75 mospheric sciences [e.g., 27] and hydrology [e.g., 28–30]. Another appealing feature
76 of EnKF is the possibility to infer model parameters at each assimilation step by
77 the augmented state technique [31, 32]. In this manner, the filter corrects the prob-
78 ability distribution of the parameters during the simulation, reducing the model
79 bias and tracking the parameter evolution in time. Lately, DA frameworks have
80 also been applied to forecast epidemics, in particular for seasonal and pandemic
81 influenza [33–37], HIV/AIDS [38, 39], the Ebola outbreak in Sierra Leone [40], and
82 the cattle disease *Theileria orientalis* [41].

83 The second main novelty of our approach is the direct use of rainfall forecasts
84 as predicted by the Climate Forecast System (CFS) [42] of the National Centers for
85 Environmental Prediction (NCEP). CFS models the interaction between oceans,
86 land, and atmosphere at a global scale assimilating remotely acquired variables.
87 Operational climate forecasts are produced daily at different spatial scales (down
88 to 0.5°) and temporal intervals (up to six months of forecast with a frequency of
89 six hours). An appealing feature of such datasets is their long forecast horizon,
90 which allows epidemiological modelers to analyze the long-term impact of hydro-
91 logic drivers on the course of an outbreak. Moreover, CFS forecasts are freely
92 available at the global scale, thus providing precipitation data and forecasts also
93 over developing countries where waterborne diseases are likely but meteorological
94 data are typically scarce.

95 2. Conceptual framework

96 Here, we present the conceptual framework for the operational forecast of a
97 cholera outbreak. The individual components of the framework, namely the epi-
98 demiological model, the calibration and DA schemes and the rainfall forecast are

99 described in details in the Section 3.

100 We assume that there must be a time-lag between the onset of an outbreak
101 and the moment when the epidemic forecasts are fully operational. First, a cer-
102 tain amount of time is necessary for healthcare authorities to identify and declare
103 a cholera outbreak. Second, if not already in place, a surveillance system that
104 centralizes epidemiological data must be implemented. The duration of this lag
105 crucially depends on the preparedness of the healthcare infrastructures. In the case
106 of Haiti, the whole process took about one month [3]. From the modeling perspec-
107 tive, data regarding population distribution, climatic and hydrological variables
108 must be collected and suitably processed. In the following we term T_0 the onset
109 of the epidemic and T_1 the moment when forecasts begin to be issued.

110 The first set of data pertaining the onset of the outbreak is used to calibrate the
111 model through a Markov Chain Monte Carlo (MCMC, see Section 3.2) scheme, in
112 order to obtain a preliminary estimation of the posterior parameter distribution.
113 In this case study, the first seven weeks of epidemiological data are used for cali-
114 bration, thus $T_0 =$ October 20, 2010 and $T_1 =$ December 12, 2010 (see Fig. 1). The
115 posterior parameter distribution computed employing MCMC is used to initialize
116 the DA framework and start the operational forecast. Specifically, N parameter
117 sets are sampled from the posterior distribution, along with the corresponding
118 simulations. This set of trajectories, periodically updated through DA, is kept
119 throughout the whole forecasting period. After the calibration period $[T_0, T_1]$, the
120 epidemiological forecasts are issued weekly, at every assimilation of the newly re-
121 ported cases. The main steps of the proposed real-time forecast framework are
122 detailed below. At the end of an epidemiological week:

- 123
- Rainfall data measured during the previous week are collected;
- 124
- Each of the N system trajectories is advanced, forced by the measured rain-
125 fall, by one week such as to arrive at the present time;

- 126 • The newly available epidemiological data is assimilated by means of an en-
127 semble Kalman filter (EnKF, Section 3.2). Therefore the model state vari-
128 ables (i.e., infected and recovered individuals, and bacterial concentrations
129 in the water reservoir that conceptualizes to various degrees infection expo-
130 sures [43] at each model node) of each of the N trajectories are updated.
131 The EnKF is applied with the augmented state, thus the N parameter sets
132 are also updated;
- 133 • The rainfall forecast issued on the same day is retrieved. N different time
134 series of rainfall are generated by adding a random error to the forecast.
- 135 • Each of the N system trajectories is associated to one rainfall time-series
136 and projected into the future for the prescribed time horizon;
- 137 • The forecast epidemiological variables, such as the total number of cases and
138 the spatial distribution of the cholera incidence are published.

139 Fig. 1 depicts two examples of forecasts for the Haitian outbreak where T_2 in-
140 dicates the time when the bulletin is issued and T_3 the end of the forecast horizon.
141 In this example $T_3 - T_2$ is equal to three months. The model trajectories computed
142 using the MCMC posterior distribution of the model parameters fit the data col-
143 lected during the calibration period $[T_0, T_1]$, but are not suitable for the subsequent
144 time step. The sequential assimilation of the data collected during $[T_1, T_2]$ corrects
145 the model trajectories and parameter values toward the real epidemic dynamics.
146 The model uncertainty gradually increases during the forecast period $[T_2, T_3]$ due
147 to the uncertainty in the forecast rainfall, which is a driver of the model.

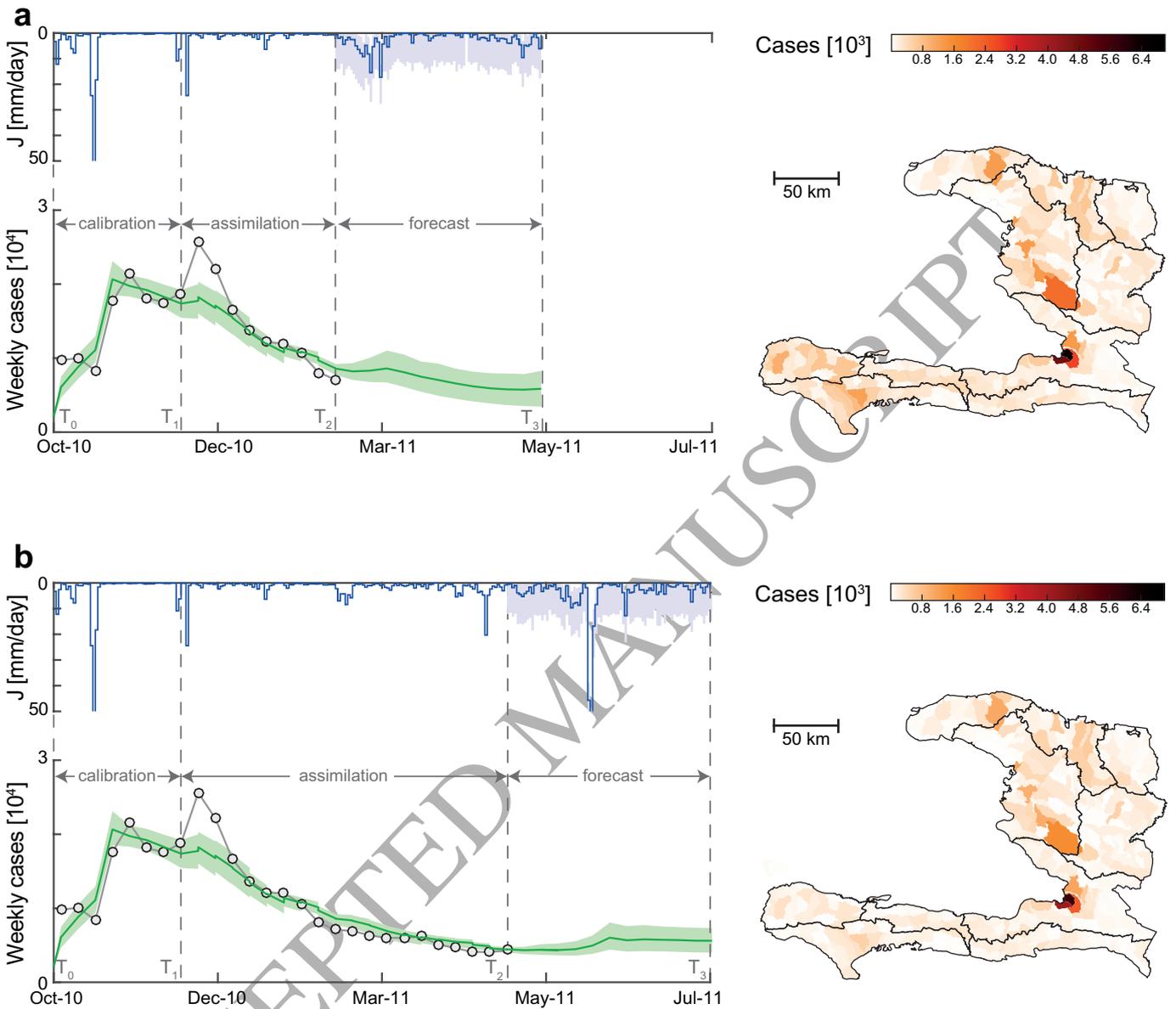


Figure 1: Example of two cholera forecasts for the Haitian outbreak computed at $T_2 =$ February 12, 2011 (a) and $T_2 =$ April 23, 2011 (b). Green lines and light-green areas represent, respectively, the expected value and the 90% confidence interval of the total weekly cases estimated by the model. 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 to the forecast rainfall. Maps show the expected value of the forecast cumulative cases on each model sub-unit.

148 **3. Material and methods**

149 *3.1. Epidemiological model*

150 The cholera epidemiological model used herein derives directly from the one
 151 developed in [14], which, in turn, builds on previous spatially-explicit epidemio-
 152 logical models [44–47]. The model subdivides the total population into n human
 153 communities spatially distributed within a domain of n nodes connected by both
 154 human mobility and hydrological networks. Let $S_i(t)$, $I_i(t)$ and $R_i(t)$ denote the
 155 local abundances of susceptible, symptomatic infected and recovered individuals
 156 at time t in each node i of the network, and let $B_i(t)$ be the environmental con-
 157 centration of *V. cholerae* in i . Cholera transmission dynamics can be described by
 158 the following set of coupled differential equations:

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

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

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

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

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

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

165 The parameter β represents the maximum exposure rate. The model assumes that
 166 β is constant, but the framework allows the EnKF to potentially change its value
 167 in time. The fraction $B_i/(K + B_i)$ is the probability of becoming infected due
 168 to the exposure to a concentration B_i of *V. cholerae*, K being the half-saturation
 169 constant [43]. Because of human mobility, a susceptible individual residing at node
 170 i can, while travelling, be exposed to pathogens in the destination community j .
 171 This is modeled assuming that the force of infection in a given node depends
 172 on the local concentration B_i for a fraction $(1 - m)$ of the susceptible hosts and
 173 on the concentration B_j of the remote communities for the remaining fraction
 174 m . The parameter m represents the community-level probability that individuals
 175 travel outside their node. The concentrations B_j are weighted according to the
 176 probabilities Q_{ij} that an individual living in node i reaches j as a destination. We
 177 apply a gravity approach [48] to model human mobility. Accordingly, connection
 178 probabilities are defined as

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

179 where the attractiveness of node j depends on its population size H_j , while the
 180 deterrence factor is assumed to be dependent on the distance d_{ij} between the two
 181 communities via an exponential kernel (with shape factor D). A fraction σ of
 182 the infected individuals develops symptoms, thus entering class I_i . The remaining
 183 fraction $(1 - \sigma)$ does not develop symptoms and therefore does not contribute to the
 184 disease transmission and enters directly the recovered compartment. Symptomatic
 185 infected individuals recover at a rate γ , or die due to cholera or other causes at
 186 rates α or μ , respectively. Recovered individuals lose their immunity and return
 187 to the susceptible compartment at a rate ρ or die at a rate μ . A fraction m
 188 of the symptomatic infected individuals are assumed to move among the nodes
 189 according to the human mobility model, and thus contribute to the environmental
 190 concentration of *V. cholerae* at a rate p/W_i , where p is the rate at which bacteria

191 excreted by an infected individual reach and contaminate the local water reservoir
 192 of volume W_i (assumed to be proportional to population size, i.e., $W_i = cH_i$
 193 as in [8]). *V. cholerae* are assumed to decay in the environment at a rate μ_B .
 194 Bacteria undergo hydrologic dispersal at a rate l : pathogens travel from node i
 195 to j with probability P_{ij} , which is assumed to be one if node j is the downstream
 196 nearest neighborhood i , and zero otherwise. In order to express the worsening of
 197 sanitation conditions caused by rainfall-induced runoff, which causes additional
 198 pathogen loads to enter the water reservoir due to effects such as overflow of pit
 199 latrines and washout of open-air defecation sites [49], the contamination rate p is
 200 increased by the rainfall intensity $J_i(t)$ via a coefficient ϕ [8, 13]. By introducing
 201 the dimensionless bacterial concentrations $B_i^* = B_i/K$, it is possible to group three
 202 model parameters into a single ratio $\theta = p/(cK)$ [44].

203 The estimation of weekly cholera cases (the quantity usually reported in epi-
 204 demiological records) from the model output requires to compute

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

205 where t_k marks the end of the k -th week.

206 The time-integration of equations (1-4) is performed through the Runge-Kutta (4,5)
 207 method, as described in [50].

208 3.2. Parameter estimation

209 At the end of the k -th epidemiological week, the model describes the state of the
 210 epidemic by the system vector $\mathbf{x}_k \in \mathbb{R}^{4n}$, $\mathbf{x}_k = \{(I_{i,k}, R_{i,k}, B_{i,k}, C_{i,k}), \text{ with } i = 1, \dots, n\}$,
 211 where n is the number of nodes. The solution of (1-4) is driven by the daily rain-
 212 fall over each node, $\mathbf{J} = (J_1, \dots, J_n)$, and the model epidemiological parameters.
 213 While model parameters μ , γ , and α can reasonably be estimated from demo-
 214 graphic and epidemiological literature [see, e.g., 14], the remaining model param-
 215 eters are typically unknown and require to be inferred by calibration. In this case,

216 the epidemiological data used for calibration are the observed weekly cases, in the
 217 following indicated with $\mathbf{y}_k \in \mathbb{R}^d$, where d is the number of measurements points
 218 at time t_k (in the Haitian case study $d=10$ is the number of departments). The
 219 relationship $\mathbf{y}_k = \mathbf{H}\mathbf{x}_k + \boldsymbol{\xi}_k$ links the observations to the model state variables,
 220 where $\mathbf{H} \in \mathbb{R}^{d \times n}$ transfers the modeled weekly cases C from the node level to the
 221 observation points, and the vector $\boldsymbol{\xi}_k \in \mathbb{R}^d$ represents the measurement error. The
 222 error components $\xi_{i,k}$, $i = 1, \dots, d$, are modeled as independent Gaussian random
 223 variables with zero mean and standard deviation σ_ξ . In the following we call $\boldsymbol{\vartheta}$ the
 224 set of the unknown model parameters, $\boldsymbol{\vartheta} = (\beta, \psi, m, D, \rho, \sigma, \mu_B, \theta, l)$. We consider
 225 two methods for the Bayesian estimation of model parameters as described in the
 226 following sections.

227 3.2.1. Markov chain Monte Carlo

228 We use the Differential Evolution Adaptive Metropolis
 229 (DREAM_{ZS}) [51] implementation of the MCMC algorithm. Given the prior proba-
 230 bility density function (pdf) of the parameters and the collection of observations in
 231 the temporal window of interest, e.g., t_0, \dots, t_k , DREAM_{ZS} samples the desired
 232 number of parameter realizations from the posterior distribution using multiple
 233 MCMC chains that run in parallel and that jointly contribute to the computa-
 234 tion of the proposal parameter samples. This technique has already been effec-
 235 tively applied to this epidemiological model [14, 16]. However, the calibration
 236 with DREAM_{ZS} over long time windows might result in overfitted posterior dis-
 237 tributions which in most cases are not realistic and are the consequence of model
 238 bias and errors possibly due to temporal changes in the parameters. Moreover,
 239 in an operational scenario, the calibration should be repeated each time new epi-
 240 demiological data becomes available, with high computational cost and a reduced
 241 capability to promptly forecast the epidemic.

242 *3.2.2. Data assimilation*

243 The second method we propose consists in inferring the distribution of the
 244 cholera model parameters in a dynamical way using DA. The main idea is that the
 245 parameter distribution can change in time and the DA scheme can sequentially
 246 track them using the collected data. The parameter update is performed in the
 247 analysis steps, which correct both the state variables and the parameter pdfs in
 248 the direction of the new observations, seeking to compute the filtering (or analysis)
 249 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
 250 product of the forecast pdf, i.e., the system state pdf predicted by the evolution of
 251 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., [20]).

252 The recursion of forecast and analysis pdfs have an analytical solution only
 253 for linear and Gaussian models. Here we use an EnKF [31], a method that ap-
 254 proximates the forecast pdf with empirical distribution of several model solutions,
 255 a technique frequently adopted when dealing with nonlinear state-space model,
 256 such as the one defined in (1-4). Using the augmented-state technique, the filter
 257 is initialized with an ensemble of N random samples from the initial distribution
 258 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
 259 at an assimilation time t_k is approximated by the numerical solutions associated
 260 to the different realizations $\{\mathbf{x}_k^{f,j}\}_{j=1}^N$,

$$261 \quad \mathbf{x}_k^{f,j} = \mathcal{F}(\mathbf{x}_{k-1}^{a,j}, \mathbf{J}(t), \boldsymbol{\vartheta}_{k-1}^{a,j}, t_{k-1}, t_k), \quad (8)$$

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

268 state vector and the parameters are updated using the state augmentation:

$$269 \quad \begin{pmatrix} \mathbf{x}_k^{a,j} \\ \boldsymbol{\vartheta}_k^{a,j} \end{pmatrix} = \begin{pmatrix} \mathbf{x}_k^{f,j} \\ \boldsymbol{\vartheta}_k^{f,j} \end{pmatrix} + \mathbf{K}_k^f (\mathbf{y}_k^j - \mathbf{y}_k^{f,j}) \quad (9)$$

270 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 ob-
 271 served measurements \mathbf{y}_k , $\mathbf{y}_k^j = \mathbf{y}_k + \boldsymbol{\xi}_k^j$, which are introduced to correctly estimate
 272 the variance of the updated variables [e.g., 25]. K_k^f is an empirical approximation
 273 of the Kalman filter, where the correlations between forecast and observations are
 274 computed through the ensemble (for more details see, e.g., [25]).

275 A possible drawback of EnKF is the so-called filter inbreeding: the rapid
 276 convergence of the parameter distribution toward one value, with the consequent
 277 underestimation of the model uncertainty. Here, we use an adaptive inflation of
 278 the covariance error used in the computation of the Kalman gain [e.g., 52, 53] to
 279 reduce the inbreeding effect. The idea is to repeat the update step by gradually
 280 increasing the measurement error variance until the parameter variances $\sigma_{\vartheta_k^a}$ are
 281 higher than a desired tolerance. At the i -th repetition of the update, we set the
 282 measurement error variance equal to $c_1^i \sigma_\xi$, with $c_1 > 1$, and the update is accepted
 283 if $\sigma_{\vartheta_k^a} > c_2 \sigma_{\vartheta_k^f}$ for each parameter, with $0 < c_2 < 1$. This condition controls
 284 the decrease of the parameter variances during the simulation and, thus, of the
 285 probability space explored by the ensemble. The proposed approach is justified in
 286 our application by the high uncertainty associated with the epidemiological data,
 287 whose error variance is largely unknown.

288 3.3. Haitian model setup

289 Our model setup is equivalent to a previous application to the Haitian epi-
 290 demic [14]. The computational domain of the model has been derived by subdivi-
 291 ding the Haitian territory into 365 watersheds, each of them hosting a human
 292 community whose size is determined using remotely acquired data (Fig. 2a). This

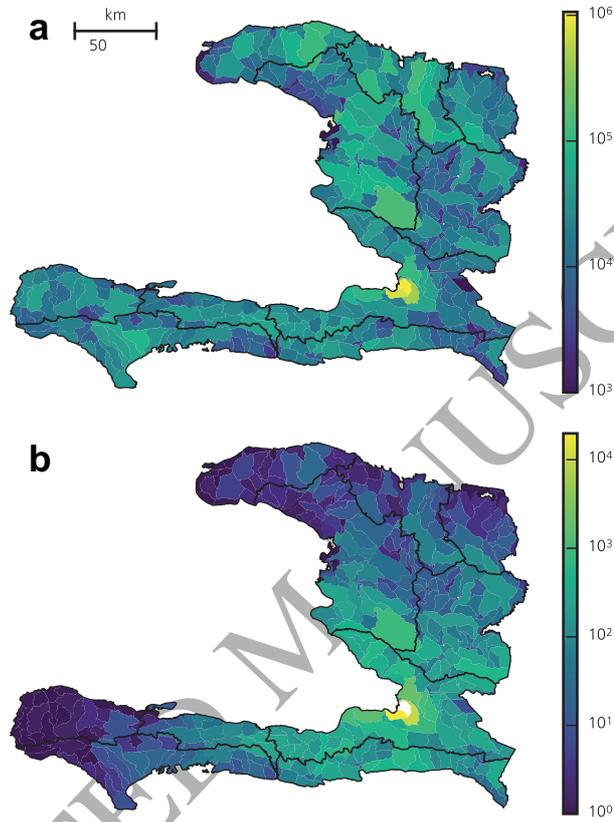


Figure 2: Model setup for the Haitian epidemic. (a) Population associated to each of the 365 watersheds estimated from the remotely sensed dataset of population distribution (LandScan algorithm of the Oak Ridge National Laboratory, <http://www.ornl.gov/landscan>). (b) Estimates of daily human mobility 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.

293 allowed for the identification of the hydrological network, defining a unique down-
294 stream node (or, in coastal areas, the sea) for each watershed and leading to the
295 connectivity matrix \mathbf{P} . Distances d_{ij} among communities (Fig. 2b) have been
296 extracted from the road network provided by the OpenStreetMap contributors
297 (available on-line at www.openstreetmap.org).

298 Two rainfall datasets are used to drive the cholera model, distinguishing be-
299 tween rainfall measurements and forecast. Daily rainfall measurement for each
300 watershed have been computed starting from data collected by the NASA-JAXA's
301 Tropical Rainfall Measuring Mission (TRMM_3B42 precipitation estimates, resolu-
302 tion: 0.25 degrees, see <http://trmm.gsfc.nasa.gov/> for details). Precipitation
303 fields are first re-sampled at the resolution of the digital terrain model used to
304 derived the watersheds (0.00083 degrees) through linear interpolation and then av-
305 eraged over the watershed area to obtain a representative value for each node. We
306 assume these rainfall measurements to be error free. Under this assumption, the
307 model uncertainty during simulations before the assimilation is completely deter-
308 mined by the probability distribution of the parameter, which accounts also for
309 possible bias in the TRMM precipitation estimates.

310 The daily rainfall forecasts are obtained from the CFS climate reforecast from
311 the beginning of the outbreak to March 31, 2011, and from the CFS opera-
312 tional climate forecast from the April 1, 2011 to present (data available on-line
313 at <https://www.ncdc.noaa.gov>). CFS operational forecasts are computed daily
314 starting at four different times (00, 06, 12, 18 UTC). For each of the four starting
315 points the climatic data are forecast every six hours for about six months with a
316 spatial resolution of 0.938 degrees in longitude and 0.246 degrees in latitude (about
317 104.3×27.46 km over Haiti). CFS climate reforecasts have the same spatial and
318 temporal resolution but are available only every five days. We computed the daily
319 forecast rainfall considering the CFS forecasts starting at 00 UTC. For each fore-

320 cast day, the rainfall is averaged over the four forecast hours and then downscaled
 321 to the watershed scale as described for TRMM. To take into account the uncer-
 322 tainties introduced with the CFS forecasts and their possible bias, we perturb the
 323 rainfalls used in the cholera forecasts with an additive error. The empirical error
 324 distribution is computed from the comparison between the CFS forecast rainfall
 325 and TRMM estimates from October 2010 to December 2013 and is found to have
 326 a temporal correlation of two days.

327 The initial conditions for infected people in each watershed, $I_i(0)$, are set ac-
 328 cording to the number of reported cases detailed in [54] as of October 20, 2010
 329 ($t=0$). Specifically, $I_i(0) = 1,000$ in the watershed hosting Mirebalais, the com-
 330 mune where the first case of cholera was reported. Additional 1,100 cases were
 331 distributed, according to population size, in the three watersheds downstream
 332 of Mirabelais along the Artibonite river that host the seven communes that was
 333 simultaneously struck by the outbreak on October 20. The initial number of re-
 334 covered and the value of bacteria concentration are assumed to be in equilibrium
 335 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
 336 remaining fraction of the population is assumed to be susceptible because of the
 337 lack of any pre-existing immunity.

338 The DREAM algorithm is run with three chains, assuming a uniform prior
 339 distribution of the parameters and reflecting parameter boundaries. Concerning
 340 the EnKF setup, preliminary sensitivity analyses on the ensemble size N and
 341 on the tuning parameters c_1 and c_2 show that stable results are obtained with
 342 $N = 1000$, $c_1=2$, and $c_2=0.8$. Reflecting parameter boundaries are enforced after
 343 the update to constrain the parameters within the prior boundaries. The condition
 344 $S_i = H_i - I_i - R_i > 0$, for $i = 1, \dots, n$, is checked for each realization of the
 345 ensemble, and is required to accept the updated state variables. The state variables
 346 of the realizations that do not satisfy this condition are not updated.

347 4. Results

348 4.1. Assimilation analysis

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

- 354 • **S1:** The model is calibrated using DREAM on the complete set of data col-
355 lected from T_0 to T_F ; $N = 1000$ random samples of the posterior distribution
356 of the parameters are then used to assess the model response during $[T_0, T_F]$.
- 357 • **S2:** The posterior distribution of the parameters is computed using DREAM
358 and considering only the data collected from T_0 to $T_1 =$ December 12, 2010.
359 During this time window, the leading driver of the outbreak changed from
360 hydrologic transport to human mobility, as detailed in [49]. We thus argue
361 that the calibration window is long enough to sample different epidemio-
362 logical dynamics and achieve a reasonable preliminary estimate of the pa-
363 rameters. The data collected during $[T_1, T_F]$ are not assimilated during the
364 simulation of the epidemics.
- 365 • **S3:** The parameters are calibrated as in S2; in the time interval $[T_1, T_F]$
366 the model state variables are updated weekly using the EnKF procedure,
367 without changing the associated parameters;
- 368 • **S4:** As in S3, but performing the EnKF update on the augmented state,
369 correcting weekly both state variables and parameters. S4 corresponds to
370 the methodology proposed in Section 2.

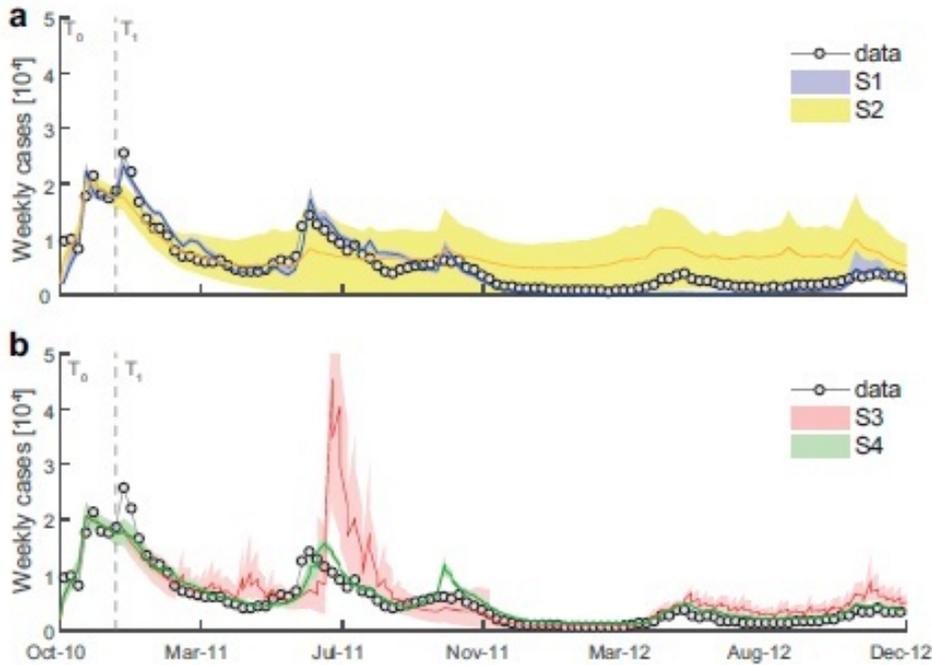


Figure 3: Comparison between the total reported weekly cases during the Haitian cholera epidemic and those estimated by the model in scenarios S1, S2 (a) and scenarios S3, S4 (b). For each scenario, the results are obtained with $N = 1000$ model runs associated to random samples of the posterior distribution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

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

378 The model responses associated to each scenario are illustrated in terms of
 379 weekly cases for the whole country (Fig. 3) and disaggregated for each depart-

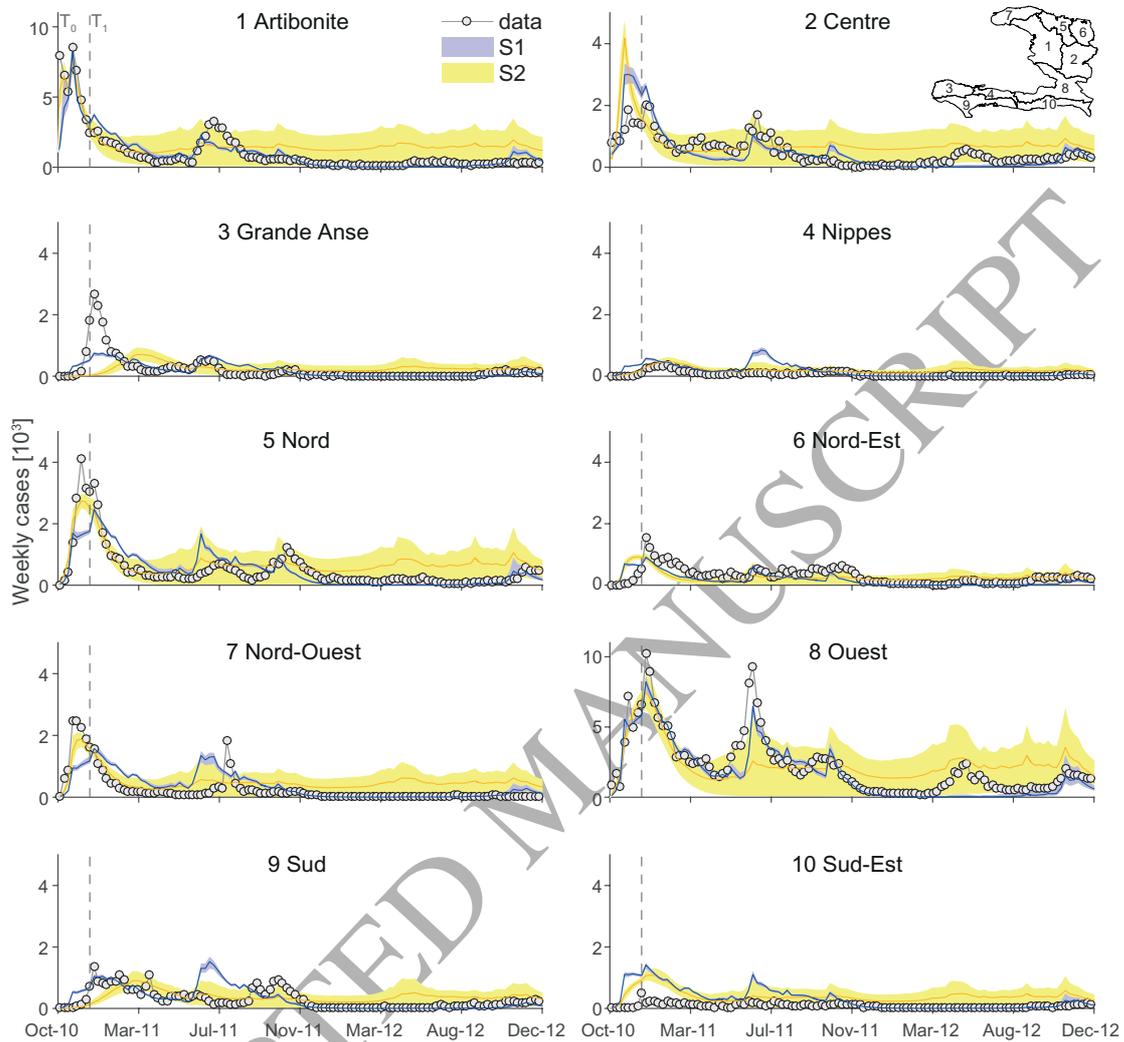


Figure 4: Comparison between the weekly cases reported at the ten Haitian departments during the cholera epidemic and those estimated by the model in scenarios S1 and S2. Symbols as in Fig. 3. Inset shows the map of the ten Haitian departments. Notice that the dataset provided by Ministère de la Santé 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.

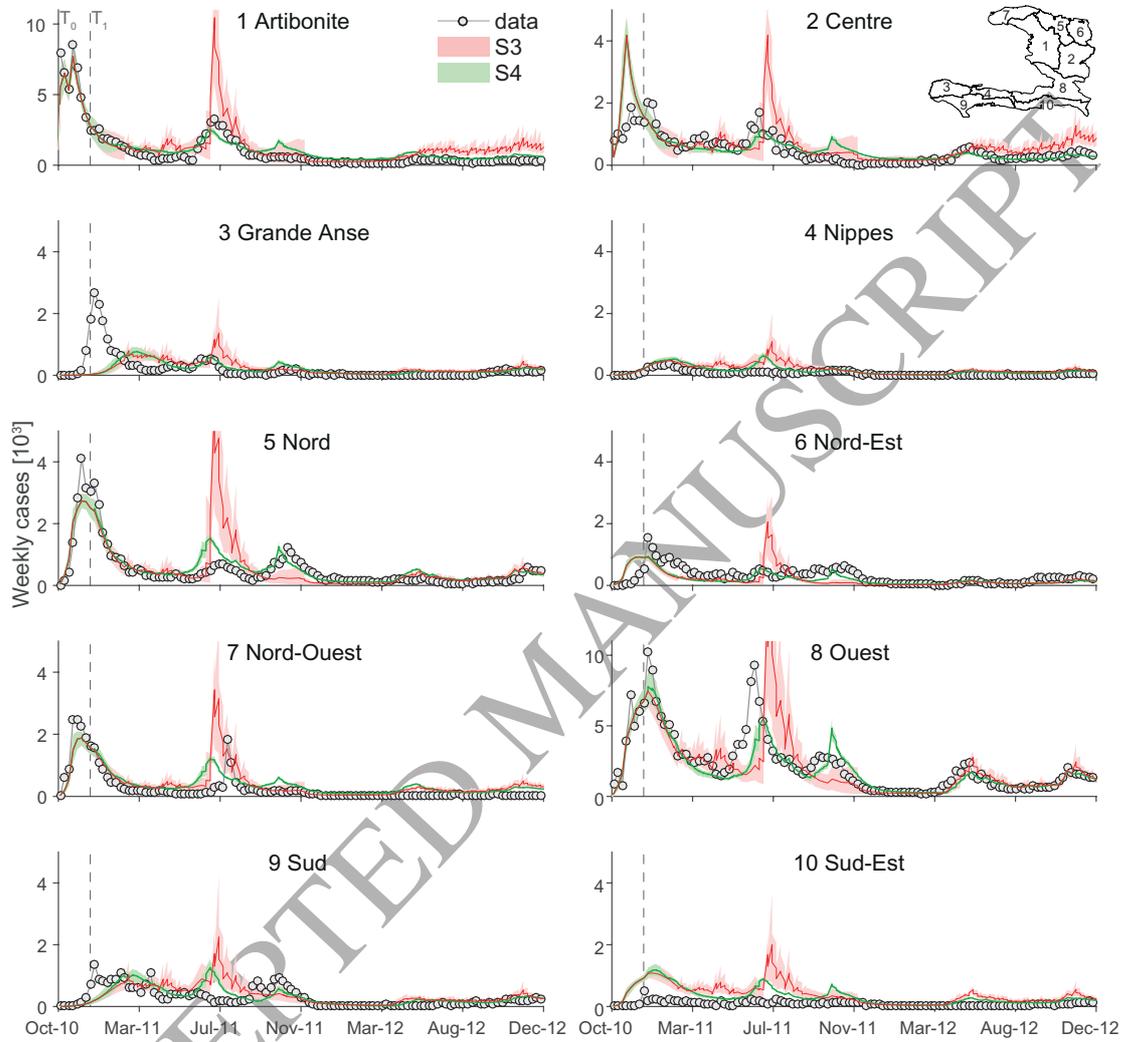


Figure 5: Comparison between the weekly cases reported at the ten Haitian departments during the cholera epidemic and those estimated by the model in scenarios S3 and S4. Symbols as in Fig. 4.

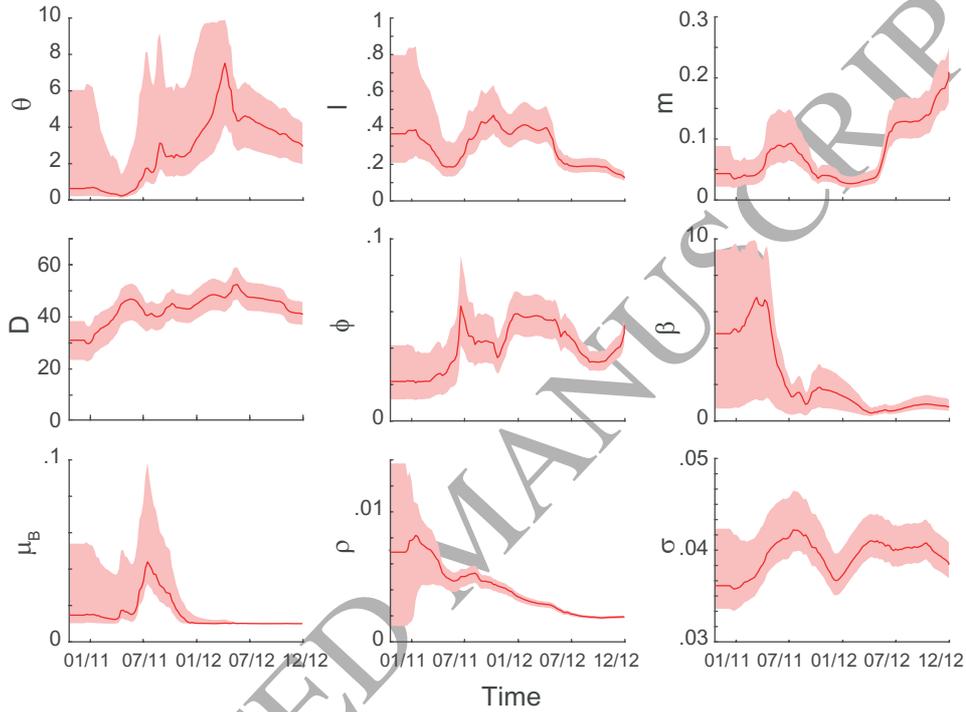


Figure 6: Empirical distribution of parameters computed by the proposed methodology (scenario S4) with an ensemble size $N = 1000$. The thick lines and the extremes of the shaded areas represent the 50th and 5th-95th percentiles of the empirical distribution.

Table 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
μ	day ⁻¹	4.5·10 ⁻⁵		
γ	day ⁻¹	0.20		
α	day ⁻¹	0.004		
β	day ⁻¹	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)
ρ	day ⁻¹	0.0005 - 0.02	0.0013 (0.0010 - 0.0014)	0.0069 (0.0012 - 0.0137)
σ	–	0.03 - 0.20	0.0827 (0.0771 - 0.0937)	0.0333 (0.0303 - 0.0408)
μ_B	day ⁻¹	0.01 - 1	0.15 (0.0104 - 0.26)	0.0147 (0.0102 - 0.054)
θ	day ⁻¹	0.01 - 10	3.60 (3.43 - 4.01)	0.63 (0.21 - 6.05)
l	day ⁻¹	0.01 - 1	0.22 (0.18 - 0.29)	0.37 (0.21 - 0.80)
ϕ	day/mm	0.01 - 2	0.111 (0.084 - 0.145)	0.022 (0.012 - 0.041)

380 ment (Fig. 4 and Fig. 5). Table 1 reports the ranges of the posterior parameter
381 distribution computed in scenarios S1 and S2, while the dynamical update of the
382 parameters in S4 is depicted in Fig. 6. These results show that the posterior pa-
383 rameter distribution obtained in scenario S1 is able to retrieve most of the epidemic
384 curve. However, parameter uncertainty is underestimated and, as a consequence,
385 the model does not reproduce all the epidemic peaks, e.g., the one occurring be-
386 tween April and August 2012 (Fig. 3a). The opposite situation is obtained in S2,
387 where the posterior parameter distribution computed during the first weeks of the
388 outbreak is too wide. In this scenario the model response is highly uncertain and
389 suitable to assess the epidemic dynamics only for few months after the end of the
390 calibration period (Fig. 3a). The weekly assimilation of the newly available data
391 with EnKF improves model results, as highlighted in scenario S3 (Fig. 3b). During
392 each week the erroneous parameters are driving the model far from the real state

393 of the system, but the update steps correct the epidemic trajectories toward the
394 measurements and reduce the model uncertainty. The combined update of the
395 state variables and model parameters introduced in S4 reduces the errors during
396 the forecast step, allowing the model to accurately follow the epidemic curve for
397 the two years considered, improving the results obtained in S1 (Fig. 3).

398 The spatial nature of the cholera model allows us to compare the four scenarios
399 with the epidemiological data at the department level. Fig. 4 presents the results
400 for S1 and S2, while Fig. 5 refers to S3 and S4. The limitations of scenarios S2
401 and S3 evinced in the aggregated results (Fig. 3) are here repeated in most of the
402 departments. The uncertainty associated to scenario S2 is too wide and the mean
403 number of modeled cases overestimates the reported cases in every department
404 during 2012. The analysis of the performance of S3 in the different departments
405 (Fig. 5) allows the identification of a possible drawback of data assimilation tech-
406 nics. In order to track the epidemic peak occurring in May/July 2011 mainly in
407 the Ouest department, the EnKF overestimates, due to an erroneous representa-
408 tion of the spatial cross-correlation at that moment, the number of cases in all the
409 other departments. This results in an overall overestimation of this peak at the
410 country scale (Fig. 3). The assimilation with the augmented state (scenario S4,
411 Fig. 5) effectively limits such drawback. Under scenarios S1 and S4 the cholera
412 model well retrieves the epidemic curve in the departments of Artibonite, Nord,
413 and Ouest, which are characterized by large numbers of reported cases (Fig. 4
414 and Fig. 5). Both scenarios poorly perform in the departments of Nippes and Sud-
415 Est, where the small number of reported cases is constantly overestimated. The
416 main advantage of S4 over S1 is the retrieval of the epidemic peak occurred during
417 May/June 2012 in the Ouest and Centre departments (Fig. 5).

418 To quantitatively compare model performances, we compute the root mean
419 square error (RMSE) between the modeled, y^f , and the total reported weekly

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. ϵ_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)

420 cases, y , for each ensemble realization j :

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

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

425 4.2. Forecast analysis

426 Having demonstrated the reliability of the proposed methodology in reproduc-
 427 ing cholera dynamics in the past, here we present the results of real-time forecasts
 428 and the associated errors. Fig. 7 presents the number of weekly cases forecast by
 429 the framework described in Section 2 one, two, and three months in advance, i.e.,
 430 the ensemble state variables at the beginning of each forecast and the associated
 431 model parameters are computed according to scenario S4. The one-month forecasts
 432 retrieve many features of the temporal dynamic of the epidemic, but slightly delay
 433 the peak of infection that occurred during June 2011 (Fig. 7a). This is probably
 434 due to erroneous model parameters that require to be updated to reproduce that
 435 particular period of the epidemic. This statement is corroborated by the param-
 436 eter dynamics depicted in Fig. 6, where it is evident that several parameters (μ , β ,

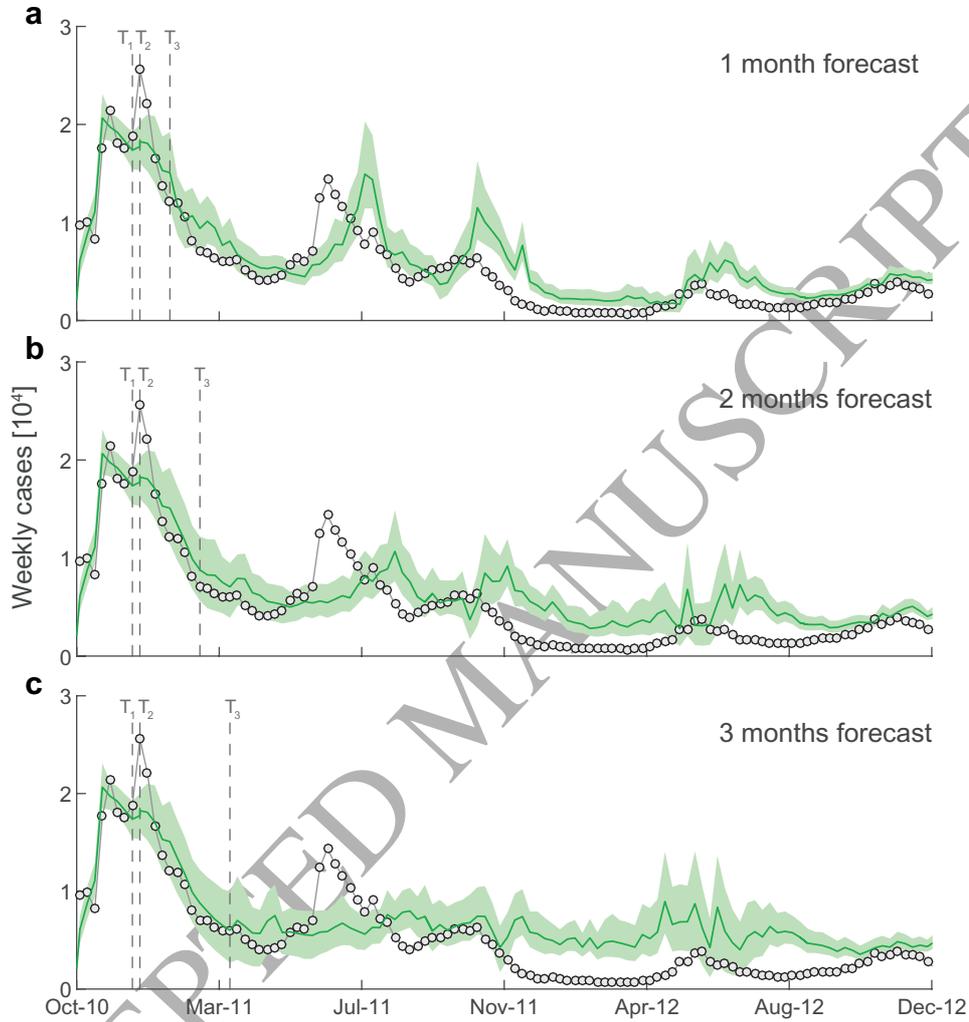


Figure 7: Weekly cholera cases forecast one (a), two (b), and three (c) months ahead using scenario S4. The mean and the 90% confidence interval of the model trajectories are reported from T_0 to T_3 , as in Fig. 1. From T_3 to the end of the simulation only the cases at the end of the forecast period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

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. η_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)

437 ϕ) are subject to strong changes during such months. The delay in forecasting the
 438 cholera peaks seems to be mitigated after July 2011, showing the general reliabil-
 439 ity of the one-month cholera predictions (Fig. 7a). The two-month forecasts are
 440 subject to similar problems, but with a stronger discrepancy between model and
 441 reported cholera peaks (Fig. 7b). The errors highly increase in the three-month
 442 forecasts, where the model response do not reproduce the data (Fig. 7c).

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

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

452 where k is the epidemiological week when the forecast is computed, d the number

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

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

466 5. Discussion and conclusions

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

476 A possible alternative to the use of DA schemes is to recalibrate the model pa-
 477 rameters every week, using the whole dataset from the beginning of the epidemic
 478 to the latest datapoint available, and to use the parameter posterior distribution

479 to project the future evolution of the outbreak. We argue that a DA scheme of-
480 fers several advantages with respect to a repeated calibration scheme. The first
481 advantage is rather technical and is related to computational time. An update of
482 state variables and parameters in our DA scheme takes less than a second on a
483 standard desktop machine. On the other hand, calibration requires on the order
484 of 10^5 model runs (around 1 day of computing time spread over 12 cores). The
485 computing time argument might not be conclusive, in particular if the forecast
486 bulletins are foreseen to be issued with a weekly frequency; however, it should not
487 be completely discarded as the framework is supposed to be implemented during
488 emergencies. In terms of forecasting performance, the DA scheme updates the
489 state of the system based on the latest observations; therefore the forecast starts
490 from a state which is close to the observed one, a feature that is not necessarily
491 satisfied by a standard calibration scheme. Finally, the main advantage of using a
492 DA scheme with the the state augmentation technique [31] is that it can track the
493 possible time-evolution of parameters and thus detect possible directional changes.
494 Indeed, some of the model parameters may change during an outbreak. In partic-
495 ular, exposure to cholera (here represented by the rate β) reportedly decreases as
496 interventions unfold (e.g., distribution of safe water and information campaigns)
497 and the population awareness of cholera transmission risk factors increases [55].
498 The time evolution of the parameter β (Fig. 6) seems to suggest such a trend.
499 However, a trend in the time evolution of a parameter might not necessarily re-
500 flect changes in the actual processes but rather be a byproduct of the challenge of
501 identifying such parameter as the dominant drivers of the epidemic change. This
502 is likely to be the case for the rate of loss of acquired immunity, ρ . Indeed, the
503 process of immunity loss can affect the outbreak dynamics only when previously
504 infected people replenish the susceptible population. Therefore the onset of the
505 outbreak is almost insensible to the rate ρ , which can be identified only at later

506 stages (Fig. 6). The same issue could also apply to the fraction of mobile peo-
507 ple m , which is crucial at the beginning of the outbreak but becomes difficult to
508 estimate once the epidemic has spread over the whole country.

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

521 The feasibility of a real-time forecasting system for cholera outbreak critically
522 depends on the immediate implementation of a surveillance system and the release
523 of the relevant epidemiological data. In the case of the outbreak in Haiti, data
524 aggregated over the ten departments (Fig. 2) were made readily available. This
525 dataset was processed starting from higher resolution data [3], which, however,
526 were not publicly released. Spatially-distributed data is indeed crucial for the
527 calibration of spatially-explicit models like the one considered in this study, as they
528 might allow to consider heterogeneous parameters over the infected area. In this
529 scenario, owing to a much larger number of parameters, classical Bayesian methods
530 for the calibration of the model might face major difficulties. On the contrary,
531 several studies [e.g., 29, 32] show the effectiveness of sequential approaches, such
532 as EnKF, to retrieve the spatial distributions of model parameters, demonstrating

533 the importance of considering DA procedures for epidemiological projections.

534 **Acknowledgments**

535 EB, FF and AR acknowledge the support from the Swiss National Science
536 Foundation (SNF/FNS) project “Dynamics and controls of large-scale cholera out-
537 breaks” (DYCHO CR23I2 138104).

ACCEPTED MANUSCRIPT

538 **References**

- 539 [1] C. Mathers, D.M. Fat, and J.T. Boerma. *The global burden of disease: 2004*
540 *update*. World Health Organization (Geneva), 2008.
- 541 [2] E.T. Ryan, U. Dhar, W.A. Khan, M. Abdus Salam, A.S.G. Faruque, G.J.
542 Fuchs, S.B. Calderwood, and M.L. Bennish. Mortality, morbidity, and micro-
543 biology of endemic cholera among hospitalized patients in Dhaka, Bangladesh.
544 *American Journal of Tropical Medicine and Hygiene*, 63(1-2):12–20, 2000.
- 545 [3] E.J. Barzilay, N. Schaad, R. Magloire, K.S. Mung, J. Boncy, G.A. Dahourou,
546 E.D. Mintz, M.W. Steenland, J. F. Vertefeuille, and J.W. Tappero. Cholera
547 surveillance during the Haiti epidemic - The first 2 years. *New England Jour-*
548 *nal of Medicine*, 368(7):599–609, 2013.
- 549 [4] E. Bertuzzo, L. Mari, L. Righetto, M. Gatto, R. Casagrandi, I. Rodriguez-
550 Iturbe, and A. Rinaldo. Prediction of the spatial evolution and effects of
551 control measures for the unfolding Haiti cholera outbreak. *Geophysical Re-*
552 *search Letters*, 38:L06403, 2011.
- 553 [5] J.R. Andrews and S. Basu. Transmission dynamics and control of cholera in
554 Haiti: an epidemic model. *Lancet*, 377:1248–1252, 2011.
- 555 [6] A.L. Tuite, J. Tien, M. Eisenberg, D.J.D. Earn, J. Ma, and D.N. Fisman.
556 Cholera epidemic in Haiti, 2010: using a transmission model to explain spatial
557 spread of disease and identify optimal control interventions. *Annals of Internal*
558 *Medicine*, 154:593–601, 2011.
- 559 [7] D.L. Chao, M.E. Halloran, and I.M. Longini Jr. Vaccination strategies for
560 epidemic cholera in Haiti with implications for the developing world. *Proceed-*
561 *ings of the National Academy of Sciences of the United States of America*,
562 108(17):7081–7085, 2011.

- 563 [8] A. Rinaldo, E. Bertuzzo, L. Mari, L. Righetto, M. Blokesch, M. Gatto,
564 R. Casagrandi, M. Murray, S.M. Vesenbeckh, and I. Rodriguez-Iturbe. Re-
565 assessment of the 2010-2011 Haiti cholera outbreak and rainfall-driven mul-
566 tiseason projections. *Proceedings of the National Academy of Sciences of the*
567 *United States of America*, 109(17):6602–6607, 2012.
- 568 [9] M. Gatto, L. Mari, E. Bertuzzo, R. Casagrandi, L. Righetto, I. Rodriguez-
569 Iturbe, and A. Rinaldo. Generalized reproduction numbers and the prediction
570 of patterns in waterborne disease. *Proceedings of the National Academy of*
571 *Sciences USA*, 48:19703–19708, 2012.
- 572 [10] J.Y. Abrams, J.R. Copeland, R.V. Tauxe, K.A. Date, E.D. Belay, R.K. Mody,
573 and E.D. Mintz. Real-time modelling used for outbreak management during
574 a cholera epidemic, Haiti, 2010-2011. *Epidemiology and Infection*, 141(6):
575 1276–1285, 2013.
- 576 [11] M.C. Eisenberg, G. Kujbida, A.R. Tuite, D.N. Fisman, and J.H. Tien. Ex-
577 amining rainfall and cholera dynamics in Haiti using statistical and dynamic
578 modeling approaches. *Epidemics*, 5(4):197–207, 2013.
- 579 [12] Z. Mukandavire, D.L. Smith, and J.G. Morris Jr. Cholera in Haiti: Reproduc-
580 tive numbers and vaccination coverage estimates. *Nature Scientific Reports*,
581 3, 2013.
- 582 [13] L. Righetto, E. Bertuzzo, L. Mari, E. Schild, R. Casagrandi, M. Gatto,
583 I. Rodriguez-Iturbe, and A. Rinaldo. Rainfall mediations in the spreading
584 of epidemic cholera. *Advances in Water Resources*, 60:34–46, 2013.
- 585 [14] E. Bertuzzo, F. Finger, L. Mari, M. Gatto, and A. Rinaldo. On the probability
586 of extinction of the Haiti cholera epidemic. *Stochastic Environmental Research*
587 *and Risk Assessment*, 28, 2014. doi: 10.1007/s00477-014-0906-3.

- 588 [15] A. Kirpich, T.A. Weppelmann, Y. Yang, A. Ali, J.G. Morris Jr., and I.M.
589 Longini. Cholera transmission in ouest department of Haiti: Dynamic model-
590 ing and the future of the epidemic. *PLoS Neglected Tropical Diseases*, 9(10):
591 e0004153, 2015. doi: 10.1371/journal.pntd.0004153.
- 592 [16] L. Mari, E. Bertuzzo, F. Finger, R. Casagrandi, M. Gatto, and A. Rinaldo. On
593 the predictive ability of mechanistic models for the Haitian cholera epidemic.
594 *Journal of the Royal Society Interface*, 12(104):20140840, 2015. doi: 10.1098/
595 rsif.2014.0840.
- 596 [17] F. Rabier. Overview of global data assimilation developments in numerical
597 weather-prediction centres. *Quarterly Journal of the Royal Meteorological*
598 *Society*, 131(613):3215–3233, 2005. doi: 10.1256/qj.05.129.
- 599 [18] I.M. Navon. Data assimilation for numerical weather prediction: A review.
600 In S.K. Park and L. Xu, editors, *Data Assimilation for Atmospheric, Oceanic*
601 *and Hydrologic Applications*, pages 21–65. Springer Berlin Heidelberg, Berlin,
602 Heidelberg, 2009. doi: 10.1007/978-3-540-71056-1_2.
- 603 [19] P.D. Thompson. A dynamical method of analyzing meteorological data. *Tel-*
604 *lus*, 13(3):334–349, 1961. doi: 10.1111/j.2153-3490.1961.tb00094.x.
- 605 [20] A.H. Jazwinski. *Stochastic Processes and Filtering Theory*. Academic Press,
606 New York, 1970.
- 607 [21] R.E. Kalman. A new approach to linear filtering and prediction problems.
608 *ASME. J. Basic Eng.*, 82(1):35–45, 1960. doi: 10.1115/1.3662552.
- 609 [22] R.E. Kalman and R.S. Bucy. New results in linear filtering and prediction
610 theory. *ASME. J. Basic Eng.*, 83(1):95–108, 1961. doi: 10.1115/1.3658902.

- 611 [23] M.S. Arulampalam, S. Maskell, N. Gordon, and T. Clapp. A tutorial on
612 particle filters for online nonlinear/non-Gaussian Bayesian tracking. *IEEE T.*
613 *Signal Proces.*, 50(2):174–188, 2002. doi: 10.1109/78.978374.
- 614 [24] G. Evensen. Sequential data assimilation with a nonlinear quasi-geostrophic
615 model using Monte Carlo methods to forecast error statistics. *J. Geophys.*
616 *Res.-Oceansi*, 99(C5):10143–10162, 1994. doi: 10.1029/94JC00572.
- 617 [25] G. Evensen. The ensemble Kalman filter: theoretical formulation and prac-
618 tical implementation. *Ocean Dyn.*, 53(4):343–367, 2003. doi: 10.1007/
619 s10236-003-0036-9.
- 620 [26] Y. Zhou, D. McLaughlin, and D. Entekhabi. Assessing the performance of the
621 ensemble Kalman filter for land surface data assimilation. *Monthly Weather*
622 *Review*, 134(8):2128–2142, 2006. doi: 10.1175/MWR3153.1.
- 623 [27] P.L. Houtekamer, H.L. Mitchell, G. Pellerin, M. Buehner, M. Charron,
624 L. Spacek, and B. Hansen. Atmospheric data assimilation with an ensem-
625 ble Kalman filter: Results with real observations. *Monthly Weather Review*,
626 133(3):604–620, 2005. doi: 10.1175/MWR-2864.1.
- 627 [28] M. Camporese, C. Paniconi, M. Putti, and P. Salandin. Ensemble Kalman
628 filter data assimilation for a process-based catchment scale model of surface
629 and subsurface flow. *Water Resources Research*, 45:W10421, 2009. doi: 10.
630 1029/2008WR007031.
- 631 [29] A.H. ELSheikh, C.C. Pain, F. Fang, J.L.M.A. Gomes, and I. M. Navon. Pa-
632 rameter estimation of subsurface flow models using iterative regularized en-
633 semble Kalman filter. *Stochastic Environmental Research and Risk Assess-*
634 *ment*, 27(4):877–897, 2012. doi: 10.1007/s00477-012-0613-x.

- 635 [30] D. Pasetto, M. Camporese, and M. Putti. Ensemble Kalman filter versus par-
636 ticle filter for a physically-based coupled surface-subsurface model. *Advances*
637 *in Water Resources*, 47(1):1–13, 2012. doi: 10.1016/j.advwatres.2012.06.009.
- 638 [31] G. Evensen. The ensemble Kalman filter for combined state and parameter
639 estimation. *IEEE CONTR. Syst. Mag.*, 29(3):83–104, 2009. doi: 10.1109/
640 MCS.2009.932223.
- 641 [32] D. Pasetto, G.-Y. Niu, L. Pangle, C. Paniconi, M. Putti, and P.A. Troch.
642 Impact of sensor failure on the observability of flow dynamics at the biosphere
643 2 LEO hillslopes. *Advances in Water Resources*, 86, Part B:327–339, 2015.
644 doi: 10.1016/j.advwatres.2015.04.014.
- 645 [33] J.-P. Chretien, D. George, J. Shaman, R.A. Chitale, and F.E. McKenzie.
646 Influenza forecasting in human populations: A scoping review. *Plos ONE*, 9
647 (4):e94130, 2014. doi: 10.1371/journal.pone.0094130.
- 648 [34] J. Shaman, A. Karspeck, W. Yang, J. Tamerius, and M. Lipsitch. Real-time
649 influenza forecasts during the 2012-2013 season. *Nature Communications*, 4,
650 2013. doi: 10.1038/ncomms3837.
- 651 [35] J. Shaman and A. Karspeck. Forecasting seasonal outbreaks of influenza. *Pro-*
652 *ceedings of the National Academy of Sciences of the United States of America*,
653 109(50):20425–20430, 2012. doi: 10.1073/pnas.1208772109.
- 654 [36] W. Yang, M. Lipsitch, and J. Shaman. Inference of seasonal and pandemic
655 influenza transmission dynamics. *Proceedings of the National Academy of*
656 *Sciences of the United States of America*, 112(9):2723–2728, 2015. doi: 10.
657 1073/pnas.1415012112.
- 658 [37] W. Yang, B.J. Cowling, E.H.Y. Lau, and J. Shaman. Forecasting influenza

- 659 epidemics in hong kong. *PLoS Computational Biology*, 11(7), 2015. doi:
660 10.1371/journal.pcbi.1004383.
- 661 [38] B. Cazelles and N.P. Chau. Using the Kalman filter and dynamic models to
662 assess the changing HIV/AIDS epidemic. *Mathematical Biosciences*, 140(2):
663 131–154, 1997. doi: 10.1016/S0025-5564(96)00155-1.
- 664 [39] H. Wu and W.-Y. Tan. Modelling the HIV epidemic: A state-space approach.
665 *Mathematical and Computer Modelling*, 32(1–2):197–215, 2000. doi: 10.1016/
666 S0895-7177(00)00129-1.
- 667 [40] W. Yang, W. Zhang, D. Kargbo, R. Yang, Y. Chen, Z. Chen, A. Kamara,
668 B. Kargbo, S. Kandula, A. Karspeck, C. Liu, and J. Shaman. Transmission
669 network of the 2014-2015 ebola epidemic in sierra leone. *Journal of the Royal*
670 *Society Interface*, 12(112), 2015. doi: 10.1098/rsif.2015.0536.
- 671 [41] C.P. Jewell and R.G. Brown. Bayesian data assimilation provides rapid deci-
672 sion support for vector-borne diseases. *Journal of the Royal Society Interface*,
673 12(108), 2015. doi: 10.1098/rsif.2015.0367.
- 674 [42] S. Saha, S. Moorthi, X. Wu, and J. Wang et al. The ncep climate forecast
675 system version 2. *Journal of Climate*, 27(6):2185–2208, 2014. doi: 10.1175/
676 JCLI-D-12-00823.1.
- 677 [43] C. Codeço. Endemic and epidemic dynamics of cholera: the role of the aquatic
678 reservoir. *BMC Infectious Diseases*, 1(1), 2001.
- 679 [44] E. Bertuzzo, S. Azaele, A. Maritan, M. Gatto, I. Rodriguez-Iturbe, and A. Ri-
680 naldo. On the space-time evolution of a cholera epidemic. *Water Resources*
681 *Research*, 44:W01424, 2008.

- 682 [45] E. Bertuzzo, R. Casagrandi, M. Gatto, I. Rodriguez-Iturbe, and A. Rinaldo.
683 On spatially explicit models of cholera epidemics. *Journal of the Royal Society*
684 *Interface*, 7:321–333, 2010.
- 685 [46] L. Mari, E. Bertuzzo, L. Righetto, R. Casagrandi, M. Gatto, I. Rodriguez-
686 Iturbe, and A. Rinaldo. On the role of human mobility in the spread of cholera
687 epidemics: towards an epidemiological movement ecology. *Ecohydrology*, 5:
688 531–540, 2012.
- 689 [47] L. Mari, E. Bertuzzo, L. Righetto, R. Casagrandi, M. Gatto, I. Rodriguez-
690 Iturbe, and A. Rinaldo. Modelling cholera epidemics: the role of waterways,
691 human mobility and sanitation. *Journal of the Royal Society Interface*, 9:
692 376–388, 2012.
- 693 [48] S. Erlander and N.F. Stewart. *The Gravity Model in Transportation Analysis*
694 *– Theory and Extensions*. VSP Books, Zeist, The Netherlands, 1990.
- 695 [49] J. Gaudart, S. Rebaudet, R. Barraï, J. Boncy, B. Faucher, M. Piarroux,
696 R. Magloire, G. Thimothe, and R. Piarroux. Spatio-temporal Dynamics of
697 cholera during the first year of the epidemic in Haiti. *PLoS Neglected Tropical*
698 *Diseases*, 7(4):e2145, 2013.
- 699 [50] L.F. Shampine and M.W. Reichelt. The MATLAB ODE Suite. *SIAM Journal*
700 *on Scientific Computing*, 18(50), 1997.
- 701 [51] J.A. Vrugt, C.J.F. ter Braak, C.G.H. Diks, B.A. Robinson, J.M. Hyman, and
702 D. Higdon. Accelerating Markov Chain Monte Carlo simulation by Differential
703 Evolution with self-adaptive randomized subspace sampling. *International*
704 *Journal of Nonlinear Sciences and Numerical Simulation*, 10:271–288, 2009.
- 705 [52] J.L. Anderson. An adaptive covariance inflation error correction algorithm for

- 706 ensemble filters. *Tellus A*, 59(2):210–224, 2007. doi: 10.3402/tellusa.v59i2.
707 14925.
- 708 [53] H. Li, E. Kalnay, and T. Miyoshi. Simultaneous estimation of covariance
709 inflation and observation errors within an ensemble Kalman filter. *Quarterly*
710 *Journal of the Royal Meteorological Society*, 135(639):523–533, 2009. doi:
711 10.1002/qj.371.
- 712 [54] R. Piarroux, R. Barraix, B. Faucher, R. Haus, M. Piarroux, J. Gaudart,
713 R. Magloire, and D. Raoult. Understanding the cholera epidemic, Haiti.
714 *Emerging Infectious Diseases*, 17:1161–1168, 2011.
- 715 [55] V.E.M.B. de Rochars, J. Tipret, M. Patrick, L. Jacobson, K.E. Barbour,
716 D. Berendes, D. Bensyl, C. Frazier, J.W. Domercant, R. Archer, T. Roels,
717 J.W. Tappero, and T. Handzel. Knowledge, attitudes, and practices related
718 to treatment and prevention of cholera, Haiti, 2010. *Emerging Infectious*
719 *Diseases*, 17(11):2158–2161, 2011.