Advances in Water Resources 000 (2016) 1–12

Contents lists available at ScienceDirect

Advances in Water Resources

journal homepage: www.elsevier.com/locate/advwatres

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,c,*]

[a] Laboratory of Ecohydrology, School of Architecture, Civil and Environmental Engineering, École Polytechinque Fédérale de Lausanne, Lausanne, Switzerland
[b] Dipartimento Ingegneria Civile Edile ed Ambientale, Università di Padova, Padova, Italy
[c] 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 xxx

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.

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 (E. Bertuzzo).

http://dx.doi.org/10.1016/j.advwatres.2016.10.004

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; Kalmanm 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°) 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