Tracking Epidemics with Google Flu Trends Data and a State-Space SEIR Model

Vanja Dukic, Hedibert F. Lopes and Nicholas G. Polson*

Abstract

In this paper we use Google Flu Trends data together with a sequential surveillance model based on the state-space methodology, to track the evolution of an epidemic process over time. We embed a classical mathematical epidemiology model (a susceptible-exposed-infected-recovered (SEIR) model) within the state-space framework, thereby extending the SEIR dynamics to allow changes through time. The implementation of this model is based on a particle filtering algorithm, which learns about the epidemic process sequentially through time, and provides updated estimated odds of a pandemic with each new surveillance data point. We show how our approach, in combination with sequential Bayes factors, can serve as an online diagnostic tool for influenza pandemic. We take a close look at the Google Flu Trends data describing the spread of flu in the US during 2003-2009, and in nine separate US states chosen to represent a wide range of health care and emergency system strengths and weaknesses.

Key Words: Google, Flu Trends, Google Correlate, epidemics, particle filtering, influenza, flu, SEIR, H1N1

*Vanja Dukic is an Associate Professor, Applied Mathematics, University of Colorado at Boulder (email: `Vanja.Dukic@colorado.edu`), Hedibert F. Lopes is an Associate Professor, and Nicholas G. Polson is a Professor, The University of Chicago Booth School of Business (email: `{ngp,hlopes}@chicagobooth.edu`). The authors thank the NSF EID and NIH NIGMS (U01GM087729) and NIH NIDA (R21DA027624-01) for partial support, as well as the Editor, Associate Editor, and two anonymous reviewers. Special thanks to Drs. David Bortz, Greg Dwyer and John Younger for helpful discussions. All code is available from the authors.

1 Introduction

In the spring of 2009, a novel H1N1 strain of Influenza *A* virus was detected in rural Mexico. Though not significantly more dangerous than a regular seasonal flu, this strain was met with little immunity in humans, and was able to infect almost three hundred thousand people worldwide by mid September of 2009, according to the World Health Organization (WHO). Unlike H5N1 (the avian influenza), which is slow-spreading but a more deadly strain, the fast-spreading H1N1 influenza was quickly declared a pandemic. A pandemic toll far exceeds that of a regular seasonal influenza, which usually severely sickens three to six million people, and results in between a quarter to a half million of deaths worldwide each year (Vaillant, La Ruche, Tarantola, and Barboza 2009).

Infectious disease surveillance has traditionally played a sentinel role in the public health pandemic preparedness. In the United States, the Centers for Disease Control and Prevention (CDC) serve as the main agency in charge of surveillance of "reportable" infectious diseases, such as SARS, influenza or West Nile virus. Similarly, WHO tracks infectious diseases throughout the world, including endemic diseases in the developing countries. Public health officials rely on surveillance data to estimate disease activity levels, and prepare intervention strategies. To this end, epidemic models have become an important part of public health response planning and early warning systems (Kaplan, Craft, and Wein 2002; Webby and Webster 2003; Elderd, Dukic, and Dwyer 2006; Eubank, Guclu, Kumar, Marathe, Srinivasan, Toroczkai, and Wang 2004).

1.1 Mathematical Models for Epidemics

Modern mathematical epidemiology models date back to the early twentieth century, most notably to the work by Kermack and McKendrick (1927) whose susceptible-infectious-recovered (SIR) model was used for modeling the plague (London 1665-1666, Bombay 1906) and cholera (London 1865) epidemics. The basic SIR model assumes that at any given time, a fixed population can be split into three compartments (fractions): susceptible people (those naive to the disease), infectious people (those with disease who are able to infect others), and recovered people (those who had the disease and are now immune). The total number of people in all three compartments, N, is assumed constant through time, with no births, and no deaths from causes other than the disease itself. These models assume homogeneous mixing, where each individual is equally likely to come in contact with any other.

The SIR model is an example of models commonly referred to as "compartmental models", as they describe the flow (transition) of people through different compartments (which represent the stages of disease) over time. When considering influenza, however, an immediate extension of the original SIR model is to introduce a fourth compartment corresponding to the incubation (disease latency) stage – when a person is infected with influenza but still not infectious enough to be able to

transmit it. This extension is called the "susceptible-exposed-infectious-recovered" (SEIR) model (Anderson and May 1991; Hethcote 2000), and describes the epidemic over time as follows:

$$\begin{aligned} \dot{S}_t &= -\beta S_t I_t / N \\ \dot{E}_t &= \beta S_t I_t / N - \alpha E_t \\ \dot{I}_t &= \alpha E_t - \gamma I_t \\ \dot{R}_t &= \gamma I_t. \end{aligned} \tag{1}$$

Here, the dot denotes a time derivative, and the parameters $\theta = (\beta, \alpha, \gamma)$ are related to the transition rates from one disease stage to the next. The first equation describes disease transmission resulting from contacts between susceptible and infectious people – each infectious individual transmits the pathogen to β individuals per unit time, but the new disease cases only arise if the contact is with a susceptible person (i.e. with probability S_t/N). Thus, at time t, the individuals in the class S move to the "exposed but not yet infectious" class E at the rate $\beta I_t/N$. The exposed but not yet infectious individuals move to the infectious class at the rate α per unit time, while γ is the rate (per unit time) at which infectious individuals I cease to be infectious because of recovery (or, in some cases, death). In the contact process terminology, α and γ correspond to the inverse of the average of an exponentially distributed time to onset of infectiousness and to recovery, respectively.

The model (1) is completed with the specification of initial values, S_0, E_0, I_0 and R_0: often, epidemics are modeled with an introduction of a single infectious person into a society where everyone else is susceptible, meaning that $I_0 = 1$, $S_0 = (N-1)$, $E_0 = 0$ and $R_0 = 0$. It is also possible to consider $I_0 = k$ where k is an unknown number of initially infected people, to be estimated from the data. As in the classic SIR model, SEIR model in this form assumes constant population size: $S_t + E_t + I_t + R_t = N$, for all t. Though extensions of the SIR-type models exist where the population size is allowed to vary via birth, death, and migration processes, for many fast evolving outbreaks in large populations N can be considered approximately constant, and estimated from the census statistics.

Mathematically, one can prove that the epidemic will not be able to take off if $\dot{E} + \dot{I} < 0$ for all times, or equivalently, $\beta S_0 / N\gamma < 1$. As $S_0 \approx N$ often, the quantity β/γ is commonly of interest instead, and is referred to as the basic reproductive ratio, or R_0. That quantity can be interpreted as the number of secondary infections a single infected person would cause during his or her infectious stage in an entirely susceptible population. The higher values of R_0 are associated with the faster spreading infection. When $\gamma = 1$, i.e., when there is on average 1 recovery per unit time, the value of R_0 equals the value of transmission parameter β.

Solving the system of equations (1) is done numerically, using a solver such as the one implemented in the `lsoda` function in the statistical software `R`, based on the method originally developed by Petzold (1983) and Hindmarsh (1983). An example of the solution to the deterministic SEIR

system of equations (1), for a specific value of θ, is shown in Figure 1. The solution describes S_t, E_t, I_t, and R_t trajectories over time, thus allowing the fraction of susceptible, latent, infectious, and recovered people to be determined at any point in time t. Compartmental models with various modifications (including birth and death rates for example, or migration), have proven useful in analyzing epidemics, and particularly for modeling the spread of a moderately to highly infectious diseases in a larger and well-mixed society (Anderson and May 1991; Ferguson, Keeling, Edmunds, Gant, Grenfell, Amderson, and Leach 2003; Cauchemez and Ferguson 2008; Koelle, Cobey, Grenfell, and Pascual 2006; Elderd, Dukic, and Dwyer 2006; Gani and Leach 2001).

Figure 1 about here.

1.2 State-space Models for Epidemics

The main appeal of compartmental models lies in their simplicity, well-understood behavior, and intuitive interpretation of the model parameters. Their simplicity is, however, also a limiting factor when it comes to capturing changes in the epidemic course, such as those induced by a public health intervention or a media event, variations in behavior, contact and vaccination patterns. Casting the traditional compartmental models in a state-space framework is one way to relax these assumptions and allow the models to capture changes in the dynamics over time in a flexible way. In this paper, we provide a state-space extension of the SEIR model, specifically designed to track epidemic behavior based on surveillance data.

In addition, epidemic outbreaks are almost always observed with error, making it necessary to estimate the solution of the system in (1) in the presence of statistical noise. In such situations, the true solution (the true susceptible, latent, infected, and recovered fractions), is referred to as the *hidden state* of the system. In many state-space models, estimation of the trajectory of the hidden state over time is the primary objective.

In our state-space SEIR model, one objective will be to estimate the trajectory of the hidden state vector $x_t = (S_t, E_t, I_t, R_t)$, based on a noisy time series of epidemic surveillance data y_t, (e.g. counts of the newly infected people, or some function thereof). However, along with the hidden state, we will also want to estimate the parameter vector driving the SEIR system, $\theta = (\beta, \alpha, \gamma)$ which contains the transmission, latency, and recovery parameters, and quantify the uncertainty in those parameters. Joint inference for states and parameters has been a topic of interest in the recent state-space modeling literature (Fearnhead 2002; Storvik 2002; Liu and West 2001; Kantas, Doucet, Singh, and Maciejowski 2009; Fearnhead 2008; Doucet and Johansen 2009; Lopes, Carvalho, Johannes, and Polson 2011).

2 Influenza Data

In the US, flu surveillance starts with the sentinel network of health care establishments, including individual health care professionals, clinics, diagnostic test laboratories, and public health departments, called the US Outpatient Influenza-like Illness Surveillance Network (ILINet). Some 2,400 sites in over 122 cities and 50 states are responsible for monitoring and reporting observed influenza-like cases to the CDC, who then analyze and publish consolidated reports on estimated flu activity in nine major US regions. ILINet tracks several indicators of flu activity throughout the US: hospitalizations, mortality, and outpatient visits due to "influenza-like illness" (ILI), on a weekly basis during the regular flu season (from October through mid-May). According to the CDC guidelines, ILI is defined as fever of 100 degrees F (or higher) and a cough and/or sore throat in the absence of a known cause other than influenza.

According to the CDC estimates, the average number of US ILI-related patient visits is about 16 million per year. The reported fraction of ILI-visits among all patient visits is weighted based on the population of each state, and averaged to form the overall US ILI activity, as well as the activity for ten major US regions. Estimates for the finer geographic resolution are not provided due to unevenly distributed locations and catchment areas of the ILINet members, and consequently, lower precision for ILI estimates. As with many traditional surveillance systems, the CDC reports are published with a delay of approximately two weeks, and all past postings are subject to a retroactive adjustment reflecting receipt of corrected reports from the ILINet members. More information about the CDC surveillance program and the definition of the ten regions can be found on the CDC website (`http://www.cdc.gov/flu`).

2.1 Google Flu Trends

Due to a remarkable increase in the on-line community and search engine activity over the last decade, several alternative surveillance systems have been proposed. Some are based on search engines, such as Google or Bing, and some on tracking micro-blogging content such as Twitter. Following an extensive variable selection process in collaboration with CDC, Ginsberg, Mohebbi, Patel, Brammer, Smolinski, and Brilliant (2009) were first to identify a set of search words, termed "ILI-related queries", that were most highly predictive of the CDC's ILI counts.

The Flu Trends algorithm that Google uses for prediction of ILI cases is based on a regression model that links the logit-transformed fraction of ILI visits to the logit-transformed fractions of the top search terms. The algorithm was found to track the ILI percentages well (see Figure 2), and now consistently predicts the ILI activity 1 to 2 weeks ahead of CDC publication. The results are archived every week as a part of the Google Flu Trends project (`http://www.google.org/flutrends/`). Unlike the CDC surveillance, these reports are made available instantly, and are not in general sub-

ject to future revisions. Flu Trends provides localized predictions, based on the IP address of the computer from which a search was done. IP addresses are usually tied to a specific metropolitan area, allowing for "IP surveillance" at the level of individual states as well as cities.

Figure 2 about here.

The "National Report Card on the State of Emergency Medicine" (American College of Emergency Physicians 2009) has found that the overall US emergency care system has been under a severe strain. However, as with other health-care aspects, states vary in their quality of emergency care. As a result, they vary in their pandemic preparedness, in addition to varying in their density of population and contact networks. For this reason, we will also examine individual results for nine states spanning a wide range of quality of care. We focus on two dimensions of the emergency medicine report, "public health" and "disaster preparedness", as they are directly relevant to the management of influenza epidemics. For example, one of the fields of the "public health" category is the percentage of adults 65 years of age or older who have received an influenza vaccine in the past 12 months. Similarly, "disease preparedness" measures characteristics such as the fraction of nurses and physicians registered in a state-based emergency system, presence of rapid notification systems, and regular drills for medical and emergency personnel. Perhaps not surprisingly, states which are largely rural and face challenges like workforce shortages, lack of large medical facilities, and large uninsured populations, are found to have the most difficulty with this category, and might be particularly vulnerable during a pandemic outbreak. According to the report, states that are among the best prepared are Maryland, Massachusetts, and Pennsylvania, while those that did not rank highly in the areas of "disaster preparedness" and "public health" include South Carolina, Oklahoma, Mississippi, South Dakota, Tennessee, and Arkansas. Google Flu Trends estimates for these nine states are shown in Figure 3.

In addition to individual US states and cities, Google Flu Trends has recently expanded to other countries where public health surveillance agencies provided access to training data and model validation. Countries that participate include most of Europe, Russia, Japan, Australia, New Zealand, Canada and Mexico. We also employ our method to study the influenza epidemic in New Zealand, a southern hemisphere and a relatively rural and well-off country with a good health care system, and two separated islands. We chose New Zealand as it may provide insight into the subsequent influenza season in the United States, since the southern hemisphere flu epidemics generally precede the northern hemisphere ones. The New Zealand analysis is basically similar to the US one, and the full discussion and results can be found in the supplementary material.

Figure 3 about here.

2.2 Influenza Epidemics and Pandemics in the Past

Pandemics are relatively rare, with only a handful of influenza pandemics occurring in the last hundred years. The most infamous one was the H1N1 pandemic in 1918/1919, also known as the "Spanish Flu", estimated to have caused twenty to fifty million deaths – more deaths than any pandemic since the bubonic plague (the Black Death) of the 14th century. The estimates of its basic reproductive number R_0 range from 1.8 to 3.5 in different communities (Chowell, Nishiura, and Bettencourt 2007; Chowell, Ammon, Hengartner, and Hyman 2006; Nishiura 2007; Mills, Robins, and Lipsitch 2004). The other notable influenza pandemics were the Asian Influenza (H2N2) of 1957-58 with 70,000 estimated deaths in the United States, and the Hong Kong Flu of 1968-69 (H3N2) with 34,000 estimated U.S. deaths. Both had basic reproductive numbers in the range of 1.5 to 2.2 (Vynnycky and Edmunds 2008; Gani, Hughes, Fleming, Griffin, Medlock, and Leach 2005; Longini, Halloran, Nizam, and Yang 2004). A pandemic is considered mild if its reproductive rate is below 1.5, moderate if between 1.5 and 1.8, and severe if above 1.9 (Yang, Sugimoto, Halloran, Basta, Chao, Matrajt, Potter, Kenah, and Longini 2009). On the other hand, seasonal influenza's basic reproductive number is lower, and historically estimated to range up to 1.35 (Cintrón-Arias, Castillo-Chávez, Bettencourt, Lloyd, and Banks 2009).

In the most recent H1N1 epidemic in 2009, the novel H1N1 virus' potential for a pandemic was deemed non-negligible (Fraser, *et al.*, and The WHO Rapid Pandemic Assessment Collaboration 2009). Its overall basic reproductive rate was estimated between 1.3 and 1.7 based on the first few months of data, but in some instances was found to be as high as 2.9 based on data from several city initial outbreaks (Yang, Sugimoto, Halloran, Basta, Chao, Matrajt, Potter, Kenah, and Longini 2009). In terms of the other influenza parameters, namely the latency (α) and recovery rate (γ), most estimates seem to point to the average incubation time being between three and four days, while the average infectious time is seven to eight days (Tuite, Greer, Whelan, Winter, Lee, Yan, Wu, Moghadas, Buckeridge, Pourbohloul, and Fisman 2010). People infected with the recent H1N1 virus are thought to be infectious longer however, as continued viral shedding was observed for over 10 days post infection, with nearly half of the people continuing to shed the virus on and after the seventh day of illness (Center for Infectious Disease Research & Policy 2009). Under the best fit exponential distribution, these preliminary studies would imply the mean recovery time of about 10 days.

3 State-space SEIR Models

State-space modeling (often termed *dynamic modeling*, West and Harrison (1997)) usually relies on sequential Bayes inference that facilitates sequential learning by incorporating additional information with every new surveillance data point. It can be designed to sequentially learn about

the epidemic parameters, produce near real-time estimates of the epidemic states while accounting for uncertainty in the epidemic parameters, and provide the posterior odds of a pandemic at any point in time. In this section we describe a state-space extension of the classic SEIR-type model for influenza dynamics, and introduce a sequential learning algorithm to update the posterior distributions of the hidden (dynamic) states $x_t = (S_t, E_t, I_t, R_t)'$ (the vector of susceptible, latent, infectious and recovered fractions in the population) at any time t, and the parameters guiding the disease evolution $\theta = (\beta, \alpha, \gamma)$. We also show how the algorithm can be used to provide the on-line pandemic alerts based on sequential Bayes factors.

3.1 Notation

The dynamics of influenza are described by the evolution of hidden (unobserved) states of the SEIR-type epidemics, $x_t = (S_t, E_t, I_t, R_t)'$, which depends on the unknown three-dimensional vector of epidemic parameters $\theta = (\beta, \alpha, \gamma)$ as in equation (1). A discretized version of the influenza dynamics in (1), assuming a discretization time-step of one week, can be expressed as follows:

$$\begin{aligned} S_t &= S_{t-1} - \beta S_{t-1} I_{t-1}/N \\ E_t &= (1-\alpha) E_{t-1} + \beta S_{t-1} I_{t-1}/N \\ I_t &= (1-\gamma) I_{t-1} + \alpha E_{t-1} \\ R_t &= R_{t-1} + \gamma I_{t-1}, \end{aligned} \tag{2}$$

where N is the total population size. The discretization replaces $\dot{S}_t$ in equation (1) by the weekly change in the susceptible fraction, $S_t - S_{t-1}$, and does so analogously for $\dot{E}_t$, $\dot{I}_t$ and $\dot{R}_t$.

Due to the nature of "influenza-like illness" (ILI) surveillance data, our observations will consist only of noisily observed weekly counts of ILI visits, $\tilde{I}_t$, which can be thought of as a proxy to the true fraction of infected population, I_t, in each week-long time period $(t-1, t]$. Instead of working directly with $\tilde{I}_t$ however, we will model the observed growth rate of infectious population, $y_t = (\tilde{I}_t - \tilde{I}_{t-1})/\tilde{I}_{t-1}$. This leads to the following state-space model for the growth rate:

$$y_t = g_t + \varepsilon_t^y \qquad \varepsilon_t^y \sim N(0, \sigma_y^2) \tag{3}$$

$$g_t = -\gamma + \alpha \frac{E_{t-1}}{I_{t-1}} + \varepsilon_t^g \qquad \varepsilon_t^g \sim N(0, \sigma_g^2). \tag{4}$$

We will refer to equation (3) as the "observation equation", and equation (4) as the "evolution equation" for the growth rate. The mean component of equation (4) is derived from the deterministic evolution of I_{t-1} based on the discretized SEIR model (2) above, with the true number of infections I_t related to g_t via $I_t = (1 + g_t) I_{t-1}$. With the infectious state I_t modeled directly in the growth rate evolution equation (4), the state-space SEIR model is then completed with the

evolution of the rest of the state components:

$$\begin{pmatrix} S_t \\ E_t \\ R_t \end{pmatrix} = \begin{pmatrix} S_{t-1} \\ E_{t-1} \\ R_{t-1} \end{pmatrix} + \begin{pmatrix} -\beta S_{t-1}/N & 0 \\ \beta S_{t-1}/N & -\alpha \\ \gamma & 0 \end{pmatrix} \begin{pmatrix} I_{t-1} \\ E_{t-1} \end{pmatrix}. \tag{5}$$

Given that we are now working with the growth rate which can be both positive and negative, it may be computationally convenient to assume that ε_t^y and ε_t^g are normally distributed, with means 0 and variances σ_y^2 (observation variance) and σ_g^2 (evolution variance), respectively. Before doing so, we recommend a normality check for the growth rates. In the Google dataset normality seems to be a reasonable assumption (see Figure 5 for the US growth rates). However, if normality had not seemed appropriate, a transformation of the growth rate (e.g. a log transformation) could have been employed to help achieve approximate normality.

Figure 4 about here.

The classical SEIR formulation assumes that $\sigma_g^2 = 0$. In fact, the magnitude of σ_g^2 can in essence be viewed as a measure of the underlying deterministic SEIR model fit, while the relative magnitudes of the two variances, σ_y^2 and σ_g^2, can be viewed as confidence in observations (data) and the underlying autonomous SEIR model, respectively.

While it is tempting to translate concepts and intuition from the classical compartmental models directly to their state-space counterparts, it is important to note that there are substantial differences between the two. For example, while the classical mathematical biology models produce smooth solutions for the entire disease trajectory over time, the state-space models will only yield a set of point-wise state estimates. The latter only gives an illusion of the trajectory. Also, in general, large-step discretizations and addition of weekly error pulses would not be recommended in pure non-linear compartmental models (Atkinson 1978; Cauchemez and Ferguson 2008; He, Ionides, and King 2009; King, Ionides, Pascual, and Bouma 2008); however, the state-space models are, in principle, able to compensate for the consequences of such errors via their evolution variances.

3.2 Sequential Learning Algorithm

Recently, particle filtering methods have been proposed for surveillance and early detection of epidemics (Rodeiro and Lawson 2006; Jagat, Carrat, Lajaunie, and Wackernagel 2008), though not within the context of state-space compartmental models. While powerful for rapid on-line estimation, particle filter methods can suffer from the "particle impoverishment" problem, and loss of inferential capability as the process evolves (Storvik 2002; Fearnhead 2008). Motivated by the desire for a fast on-line surveillance method, we implement a sequential learning algorithm based on a particle filter that is a hybrid of the Liu-West (2001) filter and the particle learning filter (Carvalho, Johannes, Lopes, and Polson 2010), relying on the use of sufficient statistics to help

alleviate particle impoverishment and information loss over time (Lopes, Carvalho, Johannes, and Polson 2011; Fearnhead 2002; Kantas, Doucet, Singh, and Maciejowski 2009).

The proposed sequential learning algorithm proceeds as follows. For notational convenience, we introduce Z_t, the "essential state vector" containing the hidden state vector $x_t = (I_t, E_t, I_t, R_t)'$, the vector of unknown static disease parameters $\theta = (\alpha, \beta, \gamma)$, the observation and evolution variances σ_y^2 and σ_g^2, and all (partial) sufficient statistics s_t. The sufficient statistics s_t govern sequential parameter learning via $p(\theta|s_t)$ (we will talk more about sufficient statistics in Section 3.3). The goal of the algorithm is to track the distribution of the essential state vector at each point in time t via sequential Monte Carlo – i.e., sets of M *particles,* $Z_t^{(1)}, \ldots, Z_t^{(M)}$ (denoted hereafter by $\{Z_t^{(i)}\}_{i=1}^M$). The set of particles at time t will thus need to be sampled from the posterior distribution of the essential state vector Z_t, given the observed infection growth rates up to time t, $y^t = \{y_1, y_2, ..., y_t\}$. Formally, $\{Z_t^{(i)}\}_{i=1}^M$ will need to be *i.i.d.* draws from $p(Z_t|y^t)$.

The algorithm for sampling $\{Z_t^{(i)}\}_{i=1}^M$ from $p(Z_t|y^t)$ is based on the following decomposition of the posterior distribution:

$$p(Z_{t+1}|y^{t+1}) \quad \propto \quad \int p(Z_{t+1}|Z_t, y_{t+1}) p(y_{t+1}|Z_t) \mathrm{dP}(Z_t|y^t), \tag{6}$$

which is a consequence of the following:

$$p(Z_t|y^{t+1}) \quad \propto \quad p(y_{t+1}|Z_t) p(Z_t|y^t) \tag{7}$$

$$p(Z_{t+1}|y^{t+1}) \quad = \quad \int p(Z_{t+1}|Z_t, y_{t+1}) \mathrm{dP}(Z_t|y^{t+1}). \tag{8}$$

Here, and throughout this section, $p(\cdot)$ refers to the appropriate continuous/discrete measure and

$$p(y_{t+1}|Z_t) = \int p(y_{t+1}|Z_{t+1}) p(Z_{t+1}|Z_t) dZ_{t+1} \tag{9}$$

plays the role of the predictive density of y_{t+1}.

Expressions (6)-(9) above suggest a two-step algorithm for sampling $\{Z_{t+1}^{(i)}\}_{i=1}^M$ from the posterior $p(Z_{t+1}|y^{t+1})$ at time $t+1$, given that we have stored the set of particles from the previous time t, $\{Z_t^{(i)}\}_{i=1}^M$. The first step would be to *resample* the old particles $\{Z_t^{(i)}\}_{i=1}^M$ with weights proportional to $p(y_{t+1}|Z_t^{(i)})$, and generate M resampled particles $\{Z_t^{(*)}\}_{i=1}^M$. These resampled particles can be viewed as a sample from $p(Z_t|y^{t+1})$ in (7) above. Once we have the resampled particles $\{Z_t^{(*)}\}_{i=1}^M$, we will *sample* a new set of particles $\{Z_{t+1}^{(i)}\}_{i=1}^M$ from the mixture of densities $p(Z_{t+1}|Z_t^{(*)}, y_{t+1})$, an approximation to the integral in the equation (8) above. In short, the sequential learning algorithm comprises repeating the following steps for $i = 1, \ldots, M$, at each time point:

Step 1 (*Resample*) Sample, with replacement, integers k^i from the set $\{1, \ldots, M\}$, such that $Pr(k^i = j) \propto p(y_{t+1}|Z_t^{(j)})$, for each $j = 1, \ldots, M$;

Step 2 (*Sample*) Sample $Z_{t+1}^{(i)}$ from $p(Z_{t+1}|Z_t^{(k^i)}, y_{t+1})$.

The key ingredients in the two-step algorithm are thus the posterior predictive density $p(y_{t+1}|Z_t)$, and the posterior updating rule $p(Z_{t+1}|Z_t, y_{t+1})$.

The sequential learning algorithm above can be used to produce out-of-sample forecasts, provide estimates of the sequential predictive densities and, consequently, estimates of Bayes factors. This comes from the fact that the predictive density for h periods ahead, $p(y_{t+h}|y^t)$, can be approximated by

$$p^M(y_{t+h}|y^t) = \frac{1}{M}\sum_{i=1}^{M} p(y_{t+h}|Z_t^{(i)}), \tag{10}$$

where $(Z_t)^{(i)}$ come from the current set of particles $\{Z_t^{(i)}\}_{i=1}^M$, acting as an approximation to $p(Z_t|y^t)$.

A natural further application of the above approximations is to sequential Bayes factors, which can be used to sequentially test a set of hypotheses. For example, we could sequentially compare the evidence for a seasonal epidemic ($\mathcal{M}_1$) versus evidence for a pandemic ($\mathcal{M}_2$), given all the observed data up to the week t. The approximate sequential Bayes factor is computed via:

$$BF_t^M(\mathcal{M}_1, \mathcal{M}_2) = \frac{p^M(y^t|\mathcal{M}_1)}{p^M(y^t|\mathcal{M}_2)},$$

where

$$p^M(y^t|\mathcal{M}_m) = \prod_{k=1}^{t} p^M(y_k|y^{k-1}, \mathcal{M}_m).$$

Here, $p^M(y_t|y^{t-1}, \mathcal{M}_m)$ are the one-step-ahead approximate predictive densities (based on equation 10), for $m = 1, 2$.

The two-step sequential learning algorithm presented above produces a sequence of particle sets, $\{Z_0^{(i)}\}_{i=1}^M, \ldots, \{Z_t^{(i)}\}_{i=1}^M$, which can also be used to perform on-line parameter learning for the static parameters (θ, σ_y^2 and σ_g^2). Given the current set of particles $\{Z_t^{(i)}\}_{i=1}^M$, one can simply draw, using the Metropolis-Hastings algorithm for example, a new set of $\{\theta^{(*i)}\}_{i=1}^M \sim p(\theta|s_t^{(i)}, y^t)$, which will in fact be a sample from the marginal density $p(\theta|y^t)$ (recall that sufficient statistics are a part of $\{Z_t^{(i)}\}_{i=1}^M$). Similar learning can be done for the two variance parameters, σ_y^2 and σ_g^2. These additional sampling ("learning") steps are, of course, unnecessary for posterior inference at time t, which can be performed via *Rao-Blackwellization*, but they are important in order to further replenish the particles and alleviate particle impoverishment (Lopes, Carvalho, Johannes, and Polson 2011).

The "look ahead" step in equation (7) also provides extra protection against particle degeneration in the algorithm (see Pitt and Shephard 1999; Kong, Liu, and Wong 1994), and reduces the propagation of the Monte Carlo error (Lopes, Carvalho, Johannes, and Polson 2011). To alleviate particle degeneration even further, a Liu and West (2001) kernel-shrinkage approximation can be used to reweigh and propagate ("jitter") the static parameters, and can be added to the *Sample* step. Indeed, we do so in the implementation of the sequential surveillance algorithm, as described in Section 3.3.

Although we use only one sequential learning approach, it is important to note that there are multiple other filtering variations that could be used instead, as long as they take steps to alleviate and assess particle degeneration and information loss. For recent reviews of sequential Monte Carlo methods and alternative filtering approaches, as well as issues with particle degeneration, see, amongst others, Cappé, Godsill, and Moulines (2007), Arulampalam, Maskell, Gordon, and Clapp (2002),Doucet and Johansen (2009), Ristic, Arulampalam, and Gordon (2004), Storvik (2002), Fearnhead (2008), Kantas, Doucet, Singh, and Maciejowski (2009), and Lopes and Tsay (2011). They highlight some of the recent developments over the last decade, including efficient particle smoothers, particle filters for highly dimensional dynamical systems, parameter learning, and the interconnections between MCMC and SMC methods.

Example of the sequential learning algorithm for AR(1) model. We give now an example of the sequential learning algorithm implemented for a simple AR(1) state-space model. We choose this model in part as an illustration, before moving to the full implementation of the sequential learning algorithm for the state-space SEIR model in the next subsection. The AR(1) model is also a simpler alternative which could be used to model the observed growth rate of infection, y_t, instead of the more complex SEIR model. As such, we will also treat the AR(1) state-space model as a simple benchmark model, and compare it to the state-space SEIR model performance in the Results section.

In the AR(1) plus noise model, the observed growth rate of infection, y_t, is modeled via the standard first order dynamic linear model of West and Harrison (1997), with the hidden state g_t (the growth rate at time t) evolving according to an autoregressive process of order one, i.e.:

$$\begin{aligned} y_t|g_t,\xi &\sim N(g_t,V) \\ g_t|g_{t-1},\xi &\sim N(\mu+\phi g_{t-1},W), \end{aligned}$$

where $\xi=(V,\mu,\phi,W)$, and g_0 comes from an initial distribution $N(m_0,C_0)$ with fixed values of m_0 and C_0.

When the joint prior distribution, $p(\xi)=p(V)p(\mu,\phi,W)$ with $V\sim IG(a_0,b_0)$, $W\sim IG(c_0,d_0)$ and $(\mu,\phi|W)\sim N(q_0,WQ_0)$, then the joint posterior distribution $p(\xi|y^t,g^t)\equiv p(\xi|s_t)$ is given as $p(V|s_t)p(\mu,\phi,W|s_t)$. Here, s_t is again the vector of conditional sufficient statistics for ξ given the data up to time t. More specifically, for $g^t=(g_1,\ldots,g_t)$, $x_t=(1,g_{t-1})'$ and $X^t=(x_1,\ldots,x_t)'$, we have $(\mu,\phi|W,g^t,X^t)\sim N(q_t,WQ_t)$ and $(W|g^t,X^t)\sim IG(c_t,d_t)$, where $c_t=c_{t-1}+1/2$, $\quad Q_t^{-1}=Q_{t-1}^{-1}+x_tx_t'$, $\quad Q_t^{-1}q_t=Q_{t-1}^{-1}b_{t-1}+g_tx_t$ $\quad$ and $\quad d_t=d_{t-1}+(g_t-q_t'x_t)y_t/2+(q_{t-1}-q_t)'Q_{t-1}^{-1}q_{t-1}/2$. Additionally, $(V|y^t,g^t)\sim IG(a_t,b_t)$, where $a_t=a_{t-1}+1/2$ and $b_t=b_{t-1}+(y_t-g_t)^2/2$. Therefore, $s_t=(a_t,b_t,c_t,d_t,q_t,Q_t)$.

Furthermore, $p(g_t|y^t,\xi)\equiv p(g_t|s_t^k,\xi)\sim N(m_t,C_t)$, where $s_t^k=(m_t(\xi),C_t(\xi))$ are the standard Kalman filter moments at time t. In this state-space model, the key ingredients in the sequential

learning algorithm are thus all available: $p(y_t|s^k_{t-1},\xi) = p_M(y_t;\mu+\phi m_{t-1}, V+W+\phi^2 C_{t-1})$, $p(s^k_t|s^k_{t-1},\xi)$ (a deterministic mapping) and $p(\xi|s_t)$ (above updates). In this example the essential state vector is $Z_t = (s_t, s^x_t, \xi)$ and the Step 2 (*sampling*) of the sequential learning algorithm translates into deterministic updates for s_t given (s_{t-1}, y_t, g_t) and for s^k_t given (s^k_{t-1}, ξ, y_t).

3.3 Sequential Learning Algorithm Implementation for Flu Trends Data

This subsection describes the specifics of the sequential learning algorithm implemented for Google Flu Trends surveillance. The algorithm consists of three modules - predictive density, posterior updating rule, and parameter learning. Below we describe the details of each of the three modules. We refer the reader to the Algorithm box for the detailed implementation steps for the Flu Trends Data.

Predictive density. This is Step 1 (*Resample*) of the sequential learning algorithm of Section 3.2. The tracking and learning algorithm presented in the previous section depends crucially on the predictive density $p(y_{t+1}|Z_t)$. To find this density, observe that $y_{t+1}|g_{t+1},\theta,\sigma^2_y \sim N(g_{t+1},\sigma^2_y)$, which follows from equation (3) and the fact that $\varepsilon^y_t \sim N(0,\sigma^2_y)$. Similarly, $g_{t+1}|Z_t \sim N(-\gamma+\alpha E_t/I_t, \sigma^2_g)$, based on equation (4) and the fact that $\varepsilon^g_t \sim N(0,\sigma^2_g)$. Combining these two densities, and integrating g_{t+1} out, leads to the predictive density for next growth rate observation, i.e. $(y_{t+1}|Z_t) \sim N(-\gamma+\alpha E_t/I_t, \sigma^2_y+\sigma^2_g)$. Note that this computation can be done for any step size, including those smaller than the intervals at which the observations are collected, by solving the SEIR equations numerically forward, and using the final values at the previous time-step as the initial values for the next.

Posterior updating rule. This is Step 2 (*Sample*) of the sequential learning algorithm of Section 3.2. After resampling the particles with weights proportional to the predictive distribution above, the next step is to "propagate" these particles and obtain a sample from the updated posterior at time $t+1$. The update for the hidden growth rate of infection, g_t, follows from the conditional linear state-space model, and can be done by the standard Kalman-type recursions (West and Harrison, 1997). More precisely, let the initial (time $t=0$) growth rate of infection be modeled as $g_0 \sim N(m_0, C_0)$. Then, for any time $t+1$, it follows that $(g_{t+1}|Z_t, y^{t+1}) \sim N(m_{t+1}, C_{t+1})$ with moments

$$m_{t+1} = C_{t+1}(\sigma_y^{-2} y_{t+1} + \sigma_g^{-2}(-\gamma+\alpha E_t/I_t)) \quad \text{and} \quad C^{-1}_{t+1} = \sigma_y^{-2} + \sigma_g^{-2}.$$

Then, $I_{t+1} = (1+g_{t+1})I_t$, and the other states of the SEIR model, $(S_{t+1}, E_{t+1}, R_{t+1})$ are deterministically updated via equation (5). The particle set $\{(S_{t+1}, E_{t+1}, I_{t+1}, R_{t+1})^{(i)}\}_{i=1}^M$ serves as an approximation to $p(S_{t+1}, E_{t+1}, I_{t+1}, R_{t+1}|y^{t+1})$.

Parameter learning. To carry out parameter learning, we also need to identify a set of conditional sufficient statistics for the next time $t+1$, which we denote s_{t+1}. These conditional sufficient statistics are a part of Z_{t+1}, and allow us to easily obtain new parameter samples from $p(\theta, \sigma_y^2, \sigma_g^2 | Z_{t+1})$. Note, we have implicitly assumed that given the complete state history up to time $t+1$, $x^{t+1} = (x_1, ..., x_{t+1})$, the parameters admit conditional sufficient statistics, so that $p(\theta, \sigma_y^2, \sigma_g^2 | x^{t+1}, y^{t+1}) = p(\theta, \sigma_y^2, \sigma_g^2 | s_{t+1})$, with s_{t+1} being recursively and deterministically obtained from (s_t, x_{t+1}, y_{t+1}), as follows.

Assuming an inverse gamma prior distribution for the observational variance σ_y^2 in equation (3), i.e. $\sigma_y^2 \sim IG(a_0, b_0)$, it follows $\sigma_y^2 | y^{t+1}, g^{t+1} \sim IG(a_{t+1}, b_{t+1})$, where $a_{t+1} = a_t + 1/2$ and $b_{t+1} = b_t + (y_{t+1} - g_{t+1})^2$. Then, s_{t+1} is a deterministic function of s_t, y_{t+1}^2, g_{t+1}^2 and $g_{t+1} y_{t+1}$. Similarly, a bivariate normal-inverse gamma prior for $(\gamma, \alpha, \sigma_g^2)$ leads to a bivariate normal-inverse gamma posterior with sufficient statistics, E_t/I_t, E_t^2/I_t^2 and $g_{t+1} E_t/I_t$, included in s_{t+1}. The transmission parameter β appears nonlinearly via E_t and I_t in the evolution equation and is sampled via the Liu and West (2001) filter, together with α and γ. For that reason, particle replenishing (via particle learning) is only performed for the two variances, σ_y^2 and σ_g^2.

Sequential Bayes factors. In situations where rapid decisions are needed, an estimate of the odds of pandemic might be the only quantity desired. In that case, we will be testing β_{pan} versus β_{epi}, with β_{epi} corresponding to a regular (seasonal) epidemic, and β_{pan} to a pandemic regime. Sequential computation of the Bayes factor describing the odds of a pandemic through time is then straightforward following the details in Section 3.2.

Hence, for an on-line detection of a pandemic, we can append the sequential learning algorithm with the sequential Bayes factor computation, comparing the cases where the parameter β takes one of two levels. Evidence for the high-level β indicates that the epidemic is about to become a pandemic, and evidence for the low-level β indicates a regular seasonal epidemics where the disease spreads to a relatively small fraction of the population (CDC estimates 5% to 20%) and dies out in a few months in a typical yearly cycle. Note that in the Bayes factor computation, different prior odds of a pandemic can be used: for example, they could be 1:20 (roughly corresponding to the historical frequency of flu pandemics in the past) or 1:1, which could be viewed as corresponding to a "pandemic vigilance" prior.

Sequential Learning Algorithm for state-space SEIR

Definitions:

- M = the number of particles used at each iteration (M used in the paper is 1,000,000)
- α is the latency parameter, β transmission parameter, and γ recovery parameter in SEIR
- $\psi = (\log\alpha, \log\beta, \log\gamma)$ is the log-transformation of the SEIR parameters
- m_ψ and V_ψ are the sample mean and variance of the $\psi^{(i)}$ draws ($i = 1, \ldots, M$), at each time point t, $t = 1, \ldots, T$
- η is the Liu-West shrinkage factor (η used in the paper was 0.99)
- σ_g^2 is the evolution variance, σ_y^2 is the observation variance
- ILI_t is the observed (Google Flu Trends) ILI percentage for week t

The algorithm:

1. Draw the initial particle set $\{(\beta, \alpha, \gamma, \sigma_g^2, \sigma_y^2)^{(i)}\}_{i=1}^M$ from the priors: $\beta \sim N(1.5, 0.5^2)I_{\beta>0}$, $\alpha \sim N(2, 0.5^2)I_{\alpha>0}$, $\gamma \sim N(1, 0.5^2)I_{\gamma>0}$, $\sigma_g^2 \sim IG(1.1, 0.005)$, $\sigma_y^2 \sim IG(1.1, 0.05)$ (see Section 4 of the paper)
2. Initialize the particle set for states $(S, E, I, R)^{(i)} = (1 - \mathrm{ILI}_1, 0, \mathrm{ILI}_1, 0)$, for $i = 1, \ldots, M$

Repeat the following steps for $t = 1, \ldots, T$:

1. Compute m_ψ, V_ψ
2. Compute $\tilde{\psi}^{(i)} = \eta\psi^{(i)} + (1 - \eta)m_\psi$
3. Obtain $\tilde{\alpha}^{(i)} = \exp\{\tilde{\psi}_1^{(i)}\}$, $\tilde{\beta}^{(i)} = \exp\{\tilde{\psi}_2^{(i)}\}$, and $\tilde{\gamma}^{(i)} = \exp\{\tilde{\psi}_3^{(i)}\}$
4. Compute $\mu_g^{(i)} = -\tilde{\gamma}^{(i)} + \tilde{\alpha}^{(i)} E_{t-1}^{(i)}/I_{t-1}^{(i)}$
5. Compute weights $\omega_t^{(i)} \propto p(y_t|\mu_g^{(i)}, \sigma_g^{2(i)} + \sigma_y^{2(i)})$
6. Resample $(\tilde{\psi}, \sigma_g^2, \sigma_y^2, S_{t-1}, E_{t-1}, I_{t-1}, R_{t-1})$ with weights $\omega_t^{(i)}$
7. Draw $\psi^{(i)}$ from $N(\tilde{\psi}^{(i)}, (1 - \eta)^2 V_\psi)$
8. Obtain $(\alpha^{(i)}, \beta^{(i)}, \gamma^{(i)})$ as in line 3 above
9. Obtain $\tilde{\mu}_g^{(i)} = -\gamma^{(i)} + \alpha^{(i)} E_{t-1}^{(i)}/I_{t-1}^{(i)}$
10. Sample $g_t^{(i)} \sim N(b, B)$, where $b = B(y_t/\sigma_y^{2(i)} + \tilde{\mu}_g^{(i)}/\sigma_g^{2(i)})$ and $B = 1/(1/\sigma_g^{2(i)} + 1/\sigma_y^{2(i)})$
11. Obtain
    (a) $I_t^{(i)} = I_{t-1}^{(i)}(1 + g_t^{(i)})$
    (b) $E_t^{(i)} = \beta^{(i)} I_{t-1}^{(i)} S_{t-1}^{(i)} + (1 - \alpha^{(i)})E_{t-1}^{(i)}$
    (c) $R_t^{(i)} = R_{t-1}^{(i)} + \gamma^{(i)} I_{t-1}^{(i)}$
    (d) $S_t^{(i)} = 1 - I_t^{(i)} - R_t^{(i)} - E_t^{(i)}$
12. Compute weights $\pi_t^{(i)} \propto p(y_t \mid \tilde{\mu}_g^{(i)}, \sigma_g^{2(i)} + \sigma_y^{2(i)})/\omega_t^{(i)}$
13. Resample $(\psi, \sigma_g^2, \sigma_y^2, S_t, E_t, I_t, R_t)$ with weights $\pi_t^{(i)}$
14. Sample σ_g^2 and σ_y^2 based on updated conditional sufficient statistics (according to the parameter learning paragraph in subsection 3.3)

4 Results

In this section we present the results for influenza tracking, based on the US Google Flu Trends. Individual years will be analyzed separately, with each year having a different set of epidemic parameters (latency, transmission and recovery parameters, as well as the evolution and observation variances). The population sizes in all years are assumed known, with yearly estimates provided by the Census Bureau (U. S. Census Bureau 2009). We assume that in each season the epidemics were started by an unknown number of infected individuals, estimated separately from the data.

We use the season-specific SEIR model within the state-space framework to track the epidemics. As a result, season-specific issues like cross-immunity from previous years will be partly accounted for; for example, the estimated transmission rate is expected to be lower in the years with residual immunity. While any compartmental influenza model – e.g. a model with non-constant population size (migration) or more detailed contact patterns – could be embedded into a state-space model, our goal here is not to build a more complex SEIR model, but to show how a simple SEIR model within a state-space framework can be successfully used to track the epidemic.

Given the abundance of prior information available for influenza, the hyper-parameters used were derived largely from the information based on historical epidemics and pandemics (see Section 2.2), as follows:

$$\begin{aligned}
\text{transmission parameter}: &\quad \beta \sim N(1.5, 0.5^2) I_{\beta>0} \\
\text{latency parameter}: &\quad \alpha \sim N(2, 0.5^2) I_{\alpha>0} \\
\text{recovery parameter}: &\quad \gamma \sim N(1, 0.5^2) I_{\gamma>0} \\
\text{evolution variance}: &\quad \sigma_g^2 \sim IG(1.1, 0.005) \\
\text{observation variance}: &\quad \sigma_y^2 \sim IG(1.1, 0.05).
\end{aligned}$$

Here, $I_{x>0}$ is an indicator function indicating that x is positive. The 95% ranges of the prior distributions were constructed so that they encapsulate most of the parameter estimates reported in published work. Though these priors are still somewhat informative, their influence is expected to diminish with time as more surveillance data points are incorporated into the analysis.

We show the results for two flu seasons in the US: the first season, 2003/2004, and the last season, 2008/2009. The epidemics in these two seasons had moderately more complex trajectories than those in the other four seasons. The first season, 2003/2004, shown in the first plot in Figure 2, was characterized by a notable epidemic peak in January 2004, when the number of Google-derived ILI cases increased to around 8%. The sharpness of the peak of that epidemic is somewhat at odds with the slowness of its spread early in the season. In such situations, the classic SEIR model with a time-invariant transmission rate β and no evolution variance would likely have difficulties describing the disease activity adequately. The state-space formulation of the SEIR model however

should be able to capture this sharp peak.

The 2008/2009 influenza season (the last plot in Figure 2) is the season with the most complexity in the epidemic trajectory. This season had multiple epidemic waves and multiple influenza strains merging together. The joint epidemic wave, widened by the late spring/summer H1N1 activity and the early second-wave onset of H1N1, would have presented an even greater challenge for the simple SEIR model without the state-space framework.

Although the state-space implementation is sensitive to the choice of variance parameters initially, the tracking algorithm is able to track the time progression of the 2003/2004 (Figure 5) and 2008/2009 (Figure 6) epidemics rather well. The uncertainty at each point in time is notable, and can be assessed by examining the bottom, middle and upper curves in all plots, which correspond to the lower 2.5th, median, and the upper 2.5th percentile of the posterior distribution for the hidden states and parameters as we learn more about them over time. For the 2003/2004 season, we see in Figure 5 that the transmission parameter decays over time as the epidemic subsides, while the latency and recovery parameters seem to stabilize: the latency parameter settled down around 1.45 (implying an average latency time of 4.8 days, and median latency time of 3.2 days), while the recovery parameter settled down between 0.3 and 0.4 (implying an average recovery time between 2.5 and 3 weeks – with the median recovery time between 1.6 and 2 weeks). The 95% posterior ranges at the end of the epidemic were 0.15-0.9 for the transmission parameter, 1-2 for the latency parameter, and 0.1-0.6 for the recovery parameter. The estimate of $R_0 = \beta/\gamma$, starts off between 1.5 and 2, but gradually settles down to around 1.1-1.3. This was in fact true for all seasons and regions we analyzed. The last panel in Figure 5 shows that even under 1:1 prior odds of pandemic, the Bayes factor steadily increases in favor of the regular epidemic as time progresses during the 2003/2004 season.

Figure 5 about here.

Figure 6 about here.

For the 2008/2009 season, we see in Figure 6 a similar set of findings as in the 2003/2004 season. The transmission rate of H1N1 seems slightly lower than the one for the 2003/2004 flu, while the latency parameter is approximately the same as in 2003/2004. The recovery parameter however settled down around 0.25, implying the median recovery time of 3 weeks. This is consistent with the findings that the most recent H1N1 recovery may be longer on average than the recovery from the other recent flu strains (Center for Infectious Disease Research & Policy 2009). The 95% posterior ranges at the end of the epidemic were 0.1-0.4 for the transmission parameter, 0.75-2 for the latency parameter, and 0.1-0.35 for the recovery parameter. Again, the last panel in Figure 6 shows that

even under 1:1 prior odds of pandemic, the Bayes factor steadily increased in favor of the regular epidemic as time progressed during the 2008/2009 season.

All results show that while the state-space SEIR can track the epidemic processes reasonably well, there does seem to be a fair amount of uncertainty in sequential state and parameter estimates. This is also reflected in the estimated variances, with evolution variance consistently higher than the observation variance. Note that this does not imply that the state-space SEIR model does not fit well – on the contrary, the state-space model tracks the observed data fairly well. However, the large evolution variance can be taken to indicate that the underlying autonomous SEIR model would likely not describe the epidemics trajectory adequately on its own without the state-space framework.

A notable consequence of using the state-space framework is that the updated information can result in the estimates of hidden states without the classic monotonicity constraints. In particular, the number of susceptibles can be be updated to a higher level than in the previous time period. The shown hidden states are not actual trajectories over time, as the classic SEIR forward simulation would produce, but rather a sequence of point-wise estimates of hidden states over time - as a result, they need not be monotone.

Figure 7 shows the prior sensitivity analysis under two additional priors on the transmission rate: the prior with mean of 1.4, and a slightly more "optimistic" prior with the mean of 1.1. As can be seen in both 2003/2004 and 2008/2009 seasons, the posterior means of the transmission parameter are similar under these two priors to the results under the prior mean of 1.5 shown in Figure 5 and Figure 6. The similarity is increasing, albeit slowly, with additional data, as expected. The two Bayes factors (under 1:1 prior odds of pandemic) show slight differences under the two priors, but are qualitatively the same: all still favor a regular epidemic over a pandemic.

Figure 7 about here.

In addition, Figure 8 shows the sensitivity analysis for the one-week-ahead prediction (posterior mean and 95% credible interval) for the ILI counts in 2003/2004 season (top row), and for the 2008/2009 season (bottom row). The analysis was done under 2 different priors on transmission rate: the right column corresponds to the prior mean of 1.4, and the left column to the prior mean of 1.1. As we can see, one-week-ahead prediction shows little sensitivity to the priors.

Figure 8 about here.

We also compared the performance of the state-space SEIR model with the simpler state-space AR(1) benchmark model. We only present a few of the interesting comparisons: Figure 9 shows the sequential posterior densities of the growth rate $p(g_t|y^t)$ and the infected fraction $p(I_t|y^t)$ for both state-space SEIR and AR(1) model for the 2003/2004 flu season in the US. It is immediately apparent that the AR(1) model has difficulty capturing changes in the epidemic behavior, and fails to track the epidemic trajectory closely after its peak. Similarly, Figure 10 shows the one-step-ahead prediction of AR(1) and SEIR state-space models in the 2003/2004 and 2008/2009 flu seasons. The state-space SEIR model's one-week ahead predictions seem to be closer to the actual observations, while the state-space AR(1) model's predictions are not as accurate after the peak, reflecting the inability of this simple model to capture the structure of the epidemic process well. The relative mean square error of the AR(1) model versus the state-space SEIR model is 5.09 for the 2003/2004 season, and 2.34 for the 2008/2009 season.

Figures 9 and 10 about here.

The other flu seasons for the entire US showed no evidence of strong epidemics, and we do not present them for that reason. The nine individual states chosen as widely representative of the emergency health care systems, present largely a similar story to the overall US results. Consequently, we single out only two of the more severe epidemic states, Oklahoma and South Dakota, and present the tracking algorithm results for the 2008/2009 influenza season in those two states in Figure 11.

Figure 11 about here.

The Bayes factor results are shown in the last panel of all result figures, under the 1:1 prior odds of a pandemic. A higher log-Bayes factor represents the stronger evidence for a seasonal epidemic. In all our analyses the evidence for a regular epidemic seems to be increasing steadily over the course of the epidemic, starting to level off towards the end. None of the Bayes factors supported evidence for a pandemic in the US and New Zealand. The full analysis of the New Zealand data is provided in the supplementary material.

4.1 Comparison with MCMC

Pure compartmental models (without the state-space extension) have traditionally been fitted off-line, using non-linear least-squares estimation procedures, or (as of recently) Bayesian estimation

and Markov chain Monte Carlo techniques (O'Neill and Roberts 1999; Neal and Roberts 2004; Meligkotsidou and Fearnhead 2004; Elderd, Dukic, and Dwyer 2006; Jewel, Kypraios, Neal, and Roberts 2009; Leman, Chen, and Lavine 2009). However, the lack of explicit likelihoods for these models generally results in slow estimation and lengthy Markov chain Monte Carlo (MCMC) runs. There is a large body of recent work on MCMC algorithms for dynamic models (Fearnhead 2002; Gilks and Berzuini 2001; Polson, Stroud, and Muller 2008; Fearnhead 2008), discussing some of the computational issues with MCMC in dynamic models. In spite of important improvements, the generally non-parallelizable nature of MCMC iterations for state-space model parameters may often mean long run times and possibly unassessed issues with convergence (Leman, Chen, and Lavine 2009; Meligkotsidou and Fearnhead 2004).

However, comparing a particle-filtering based algorithm with MCMC is useful in order to assess if particle collapse might have been a problem. The sequential learning algorithm proposed in this paper should perform well for dynamic models when there is a high level of conditional sufficiency for parameters of interest, which is not necessarily the case in real-life epidemics. For that reason, we take a closer look at the posterior distribution of the epidemic parameters (α, β, γ) and the two variances, and assess how the posteriors estimated via MCMC compare to those estimated via the sequential learning algorithm proposed in this paper. The results are shown in Figure 12, at the end of the 2003/2004 US flu season. There seems to be little difference between the marginal posterior densities of the three epidemic parameters and two variances. However, there was a notable difference in the length of time MCMC and sequential learning algorithm required to run: the sequential learning algorithm with 1,000,000 particles took on average less than 2 minutes on a 3.1GHz i5 processor for this season, while the MCMC with 1,000,000 iterations took approximately 15 hours on the same processor. While this does not make MCMC infeasible for on-line surveillance, the time savings with sequential learning are notable.

Figure 12 about here.

5 Conclusions

This paper presents a state-space SEIR analysis of an IP influenza surveillance dataset, the Google Flu Trends. The US Flu Trends surveillance has been found to closely track the CDC reports, and is able to precede it by one to two weeks, holding potential for developing real-time surveillance mechanisms. As a result, flexible epidemic models and fast tracking algorithms capable of near real-time estimation and prediction as new data become available, are particularly important. We present one approach to near real-time disease tracking based on the state-space methodology, compartmental modeling, and sequential Bayesian learning.

Classical compartmental models of mathematical epidemiology have been the staple of epidemic

modeling for over a century. However, the unchanging dynamical structure, present in most classical models, is often not appropriate for real life epidemics, due to seasonality (Cauchemez and Ferguson 2008), behavior changes, vaccination, quarantine, migration, or a myriad of other reasons that affect how people interact and react to a disease. The state-space approach is one of the most flexible and yet simple ways to incorporate changes in the disease dynamics through time, as it relaxes the determinism of the compartmental models through the presence of the evolution variance. Yet, compartmental models also provide simple but powerful insight into the process of disease dynamics which can be readily tied to intervention (e.g. reducing contact intensity through school closures and hygiene, shortening recovery period through antiviral drugs, etc.). The simple state-space extension of the classic SEIR model presented in this paper combines the familiar mathematical epidemiology theory with computational speed and statistical flexibility.

Although information loss and particle collapse can be a problem in our sequential learning algorithm as well as in all other particle filtering approaches, in modest-scale applications, for problems where parameters and states vary smoothly and slowly over time, any reasonable sequential Monte Carlo scheme should perform well (Kitagawa 1998). However, when sharp changes in the dynamics are present, as may happen during real-life epidemics due to media activity or public health interventions, tracking might prove challenging. As computational power increases, taking advantage of GPU and cloud computing, the serious information loss issues might be somewhat lessened through increased number of particles used in these algorithms.

The Bayesian framework utilized in the paper is able to easily provide uncertainty estimates. As a result, the method such as the one presented here can be used to guide dynamic allocation of resources and facilitate comparisons of different intervention strategies. Such comparisons can be done based on the predictive distribution of outcomes rather than just their expectations, allowing the full propagation of uncertainty in the non-linear decision problems.

Although this paper uses IP surveillance data, it is important to note that CDC surveillance plays a crucial role in the US surveillance and threat preparedness, and that the approach we present here is only one way among several designed to aid CDC in continuing their mission. Combining our approach with the CDC's scan statistic methodology would be a valuable contribution, which we hope to pursue in the future as we extend our algorithm to account for spatial structure across the US.

Finally, although CDC has done an extensive validation on Google Flu Trends, the Flu Trends algorithm has still not been validated specifically for most states, and cities. There are many ways in which states and localities differ, and search terms may be correlated within state (or even within sub-regions of states, and individual metropolitan areas). For example, the search terms found likely to be indicative of ILI for Rhode Island could differ from those used in California, especially when one allows the use of other languages. If more localized on-line surveillance is

to be put in place, Google Trends algorithms will likely need further refinement, in collaboration with local public health authorities and CDC, to capture some of these region-specific differences. Expert opinion on geographical variations and relations among search terms might be able to shed light onto this issue.

References

American College of Emergency Physicians (2009). The national report card on the state of emergency medicine.

Anderson, R. M. and R. M. May (1991). *Infectious diseases of humans: Dynamics and control*. Oxford, UK: Oxford University Press.

Arulampalam, M., S. Maskell, N. Gordon, and T. Clapp (2002). A tutorial on particle filters for on-line nonlinear/non-Gaussian Bayesian tracking. *IEEE Transactions on Signal Processing 50*, 174–188.

Atkinson, K. E. (1978). *Introduction to Numerical Analysis*. John Wiley & Sons, Inc.

Bernardo, J. M., M. J. Bayarri, J. O. Berger, A. P. Dawid, D. Heckerman, A. F. M. Smith, and M. West (Eds.) (2011). *Bayesian Statistics 9*, Oxford. Oxford University Press.

Cappé, O., S. Godsill, and E. Moulines (2007). An overview of existing methods and recent advances in sequential Monte Carlo. *IEEE Proceedings in Signal Processing 95*, 899–924.

Carvalho, C. M., M. Johannes, H. F. Lopes, and N. G. Polson (2010). Particle learning and smoothing. *Statistical Science 25*, 88–106.

Cauchemez, S. and N. M. Ferguson (2008). Likelihood-based estimation of continuous-time epidemic models from time-series data: application to measles transmission in London. *Journal of The Royal Society Interface 5*(25), 885–897.

Center for Infectious Disease Research & Policy (2009). Novel H1N1 influenza (swine flu). Technical report, Academic Health Center - University of Minnesota.

Chowell, G., C. E. Ammon, N. W. Hengartner, and J. M. Hyman (2006). Transmission dynamics of the great influenza pandemic of 1918 in Geneva, Switzerland: Assessing the effects of hypothetical interventions. *Journal of Theoretical Biology 241*, 193–204.

Chowell, G., H. Nishiura, and L. Bettencourt (2007). Comparative estimation of the reproduction number for pandemic influenza from daily case notification data. *Journal of the Royal Society Interface 4*, 155–166.

Cintrón-Arias, A., C. Castillo-Chávez, L. Bettencourt, A. Lloyd, and H. T. Banks (2009). The estimation of the effective reproductive number from disease outbreak data. *Mathematical Biosciences and Engineering 6*, 261–282.

Doucet, A. and A. Johansen (2009). *Handbook of Nonlinear Filtering*, Chapter A Tutorial on Particle Filtering and Smoothing: Fifteen years Later. Oxford: Oxford University Press.

Elderd, B., V. Dukic, and G. Dwyer (2006). Uncertainty in predictions of disease spread and public-health responses to bioterrorism and emerging diseases. *Proceedings of the National Academy of Sciences 103*, 15693–15697.

Eubank, S., H. Guclu, V. Kumar, M. Marathe, A. Srinivasan, Z. Toroczkai, and N. Wang (2004). Modelling disease outbreaks in realistic urban social networks. *Nature 429*, 180–184.

Fearnhead, P. (2002). Markov chain Monte Carlo, sufficient statistics, and particle filters. *Journal of Computational and Graphical Statistics 11*, 848–862.

Fearnhead, P. (2008). MCMC for space models. Technical report, Lancaster University.

Ferguson, N. M., M. J. Keeling, W. J. Edmunds, R. Gant, B. T. Grenfell, R. M. Amderson, and S. Leach (2003). Planning for smallpox outbreaks. *Nature 425*, 681–685.

Fraser, C., *et al.*, and The WHO Rapid Pandemic Assessment Collaboration (2009). Pandemic potential of a strain of influenza a (H1N1): Early findings. *Science 324*, 1557–1561.

Gani, R., H. Hughes, D. Fleming, T. Griffin, J. Medlock, and S. Leach (2005). Potential impact of antiviral drug use during influenza pandemic. *Emerging and Infectious Diseases 11*, 1355–1362.

Gani, R. and S. Leach (2001). Transmission potential of smallpox in contemporary populations. *Nature 414*, 748–751.

Gilks, W. and C. Berzuini (2001). Following a moving target - Monte Carlo inference for dynamic Bayesian models. *Journal of the Royal Statistical Society, Series B 63*, 127–46.

Ginsberg, J., M. Mohebbi, R. Patel, L. Brammer, M. Smolinski, and L. Brilliant (2009). Detecting influenza epidemics using search engine query data. *Nature 457*, 1012–1014.

He, D., E. L. Ionides, and A. A. King (2009). Plug-and-play inference for disease dynamics: Measles in large and small towns as a case study. *Journal of the Royal Society Interface*.

Hethcote, H. W. (2000). The mathematics of infectious diseases. *SIAM Review 42*, 599653.

Hindmarsh, A. (1983). *Scientific Computing*, Chapter ODEPACK, A Systematized Collection of ODE Solvers, pp. 55–64. Amsterdam: North-Holland.

Jagat, C., F. Carrat, C. Lajaunie, and H. Wackernagel (2008). *Geostatistics for Environmental Applications - Proceedings of the Sixth European Conference on Geostatistics for Environmental Applications*, Chapter Early Detection and Assessment of Epidemics by Particle Filtering, pp. 23–35. Amsterdam: Springer Netherlands.

Jewel, C., T. Kypraios, P. Neal, and G. Roberts (2009). Bayesian analysis for emerging infectious diseases. *Bayesian Analysis 4*, 465–496.

Kantas, N., A. Doucet, S. Singh, and J. Maciejowski (2009). An overview of sequential Monte Carlo methods for parameter estimation on general state space models. *15th IFAC Symposium on System Identification*.

Kaplan, E. H., D. L. Craft, and L. M. Wein (2002). Emergency response to a smallpox attack: The case for mass vaccination. *Proceedings of the National Academy of Sciences of the United States of America 99*, 10935–10940.

Kermack, W. and A. McKendrick (1927). Contribution to the mathematical theory of epidemics. *Proceedings of the Royal Society of London, Series A 115*, 700–721.

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

Kitagawa, G. (1998). A self-organizing state-space model. *Journal of the American Statistical Association 93*, 1203–1215.

Koelle, K., S. Cobey, B. Grenfell, and M. Pascual (2006). Epochal evolution shapes the philodynamics of interpandemic influenza a (H5N2) in humans. *Science 314*, 1898–1903.

Kong, A., J. S. Liu, and W. H. Wong (1994). Sequential imputations and Bayesian missing data problems. *Journal of the American Statistical Association 89*, 278–288.

Leman, S., Y. Chen, and M. Lavine (2009). The multiset sampler. *Journal of the American Statistical Association 104*, 1029–1041.

Liu, J. and M. West (2001). *Sequential Monte Carlo Methods in Practice*, Chapter Combined parameters and state estimation in simulation-based filtering. New York: Springer-Verlag.

Longini, I., M. Halloran, A. Nizam, and Y. Yang (2004). Containing pandemic influenza with antiviral agents. *American Journal of Epidemiology 159*, 623–633.

Lopes, H. F. and R. E. Tsay (2011). Particle filters and Bayesian inference in financial econometrics. *Journal of Forecasting 30*, 168–209.

Meligkotsidou, L. and P. Fearnhead (2004). Exact filtering for partially-observed continuous-time models. *Journal of the Royal Statistical Society, Series B 66*, 771–789.

Mills, C. E., J. M. Robins, and M. Lipsitch (2004). Transmissibility of 1918 pandemic influenza. *Nature 432*, 904–906.

Neal, P. J. and G. O. Roberts (2004). Statistical inference and model selection for the 1861 Hagelloch measles epidemic. *Biostatistics 5*, 249–261.

Nishiura, H. (2007). Time variations in the transmissibility of pandemic influenza in Prussia, Germany, from 1918-19. *Theoretical Biology and Medical Modelling*, 4–20.

O'Neill, P. and G. O. Roberts (1999). Bayesian inference for partially observed stochastic epidemics. *Journal of the Royal Statistical Society, Series A 162*, 121–129.

Petzold, L. (1983). Automatic selection of methods for solving stiff and nonstiff systems of ordinary differential equations. *SIAM Journal on Scientific and Statistical Computing 4*, 136–148.

Pitt, M. and N. Shephard (1999). Filtering via simulation: Auxiliary particle filters. *Journal of the American Statistical Association 94*, 590–599.

Polson, N., J. Stroud, and P. Muller (2008). Practical filtering with sequential parameter learning. *Journal of the Royal Statistical Society, Series B 70*, 413–428.

Ristic, B., S. Arulampalam, and N. Gordon (2004). *Beyond the Kalman filter: Particle filters for tracking applications*. Boston, MA: Artech House.

Rodeiro, C. V. and A. Lawson (2006). Online updating of space-time disease surveillance models via particle filters. *Statistical Methods in Medical Research 15*, 1–22.

Storvik, G. (2002). Particle filters in state space models with the presence of unknown static parameters. *IEEE Transactions on Signal Processing 50*, 281–289.

Tuite, A. R., A. L. Greer, M. Whelan, A.-L. Winter, B. Lee, P. Yan, J. Wu, S. Moghadas, D. Buckeridge, B. Pourbohloul, and D. N. Fisman (2010). Estimated epidemiological parameters and morbidity associated with pandemic H1N1 influenza. *Canadian Medical Association Journal 182*(2), 131–136.

U. S. Census Bureau (2009). *Annual Estimates of the Resident Population for the United States, Regions, States, and Puerto Rico: April 1, 2000 to July 1, 2009.* Population Division.

Vaillant, L., G. La Ruche, A. Tarantola, and P. Barboza (2009). Epidemiology of fatal cases associated with pandemic H1N1 influenza 2009. *Eurosurveillance*.

Vynnycky, E. and W. J. Edmunds (2008). Analyses of the 1957 (Asian) influenza pandemic in the United Kingdom and the impact of school closures. *Epidemiology & Infection 136*, 166–179.

Webby, R. J. and R. G. Webster (2003). Are we ready for pandemic influenza? *Science 302*, 1519–1522.

West, M. and J. Harrison (1997). *Bayesian Forecasting and Dynamic Models* (2nd ed.). New York: Springer-Verlag.

Yang, Y., J. Sugimoto, M. Halloran, N. Basta, D. Chao, Matrajt, G. Potter, E. Kenah, and I. Longini (2009). The transmissibility and control of pandemic influenza a (H1N1) virus. *Science Express*, 729 – 733.

FIGURES

Figure 1: An example solution to an SEIR system specified in equation (1), in a population of size 100.

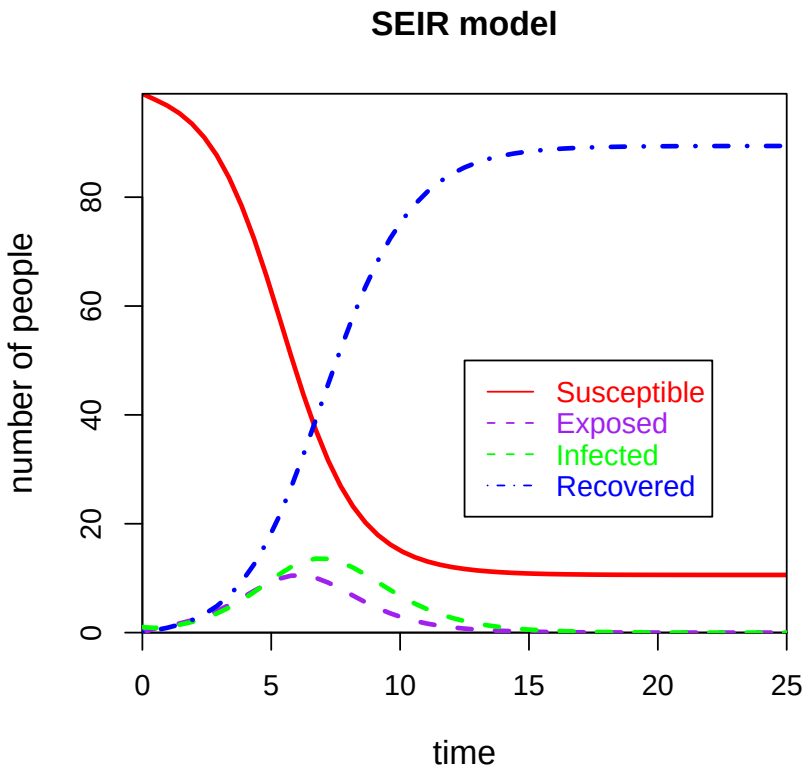

Figure 2: Google Flu Trends estimated ILI percentages (dashed line) and CDC ILI Surveillance percentages (solid line) for the United States, from June 2003 until September 2009. Separate plots correspond to separate influenza years, with each new influenza season starting in autumn, and ending in spring. Note that CDC did not used to produce ILI reports during summers before 2009, and thus no solid line appears during summer months prior to 2009.

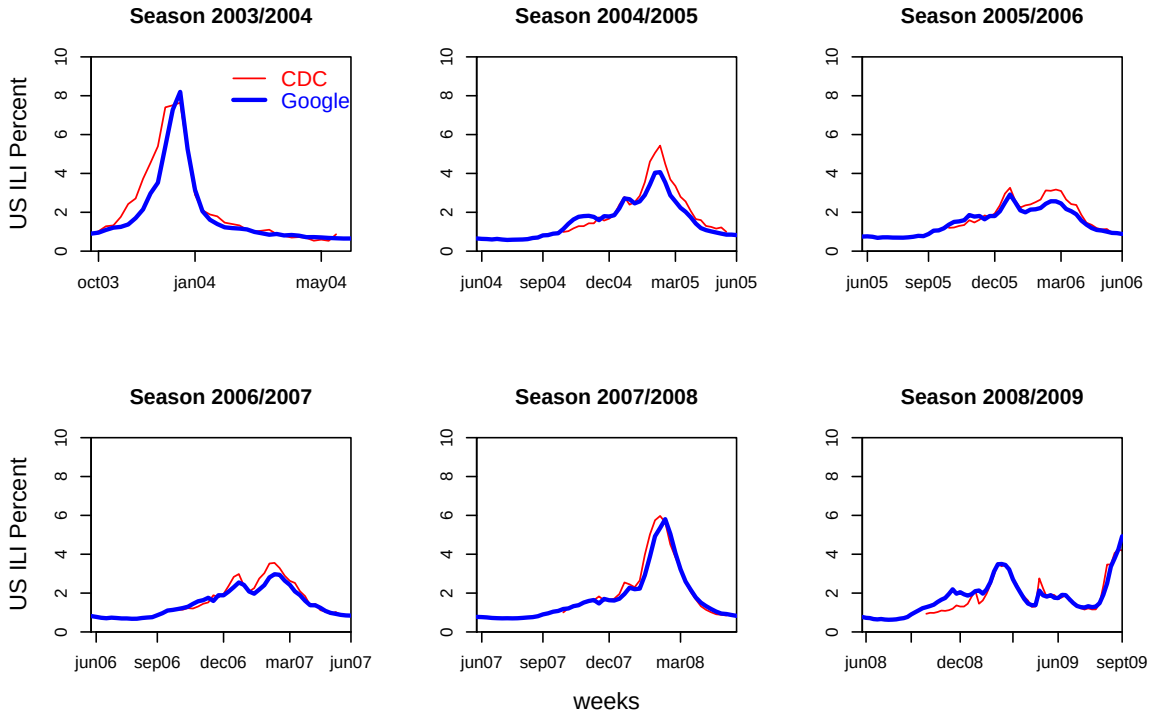

Figure 3: Google Flu Trends ILI surveillance in 9 representative states, 2003-2009. The states were chosen to span a range of health care preparedness criteria based on the results published in the American College of Emergency Physicians 2009 Report. The states that are ranked among the best in quality of health care are Maryland, Massachusetts, and Pennsylvania. The states that ranked low in the areas of "disaster preparedness", "emergency care access", and "public health" include South Carolina, Oklahoma, Mississippi, South Dakota, Tennessee, and Arkansas. Note some states' search term counts were too low to procure the Flu Trends surveillance data early on, during 2003 through 2005.

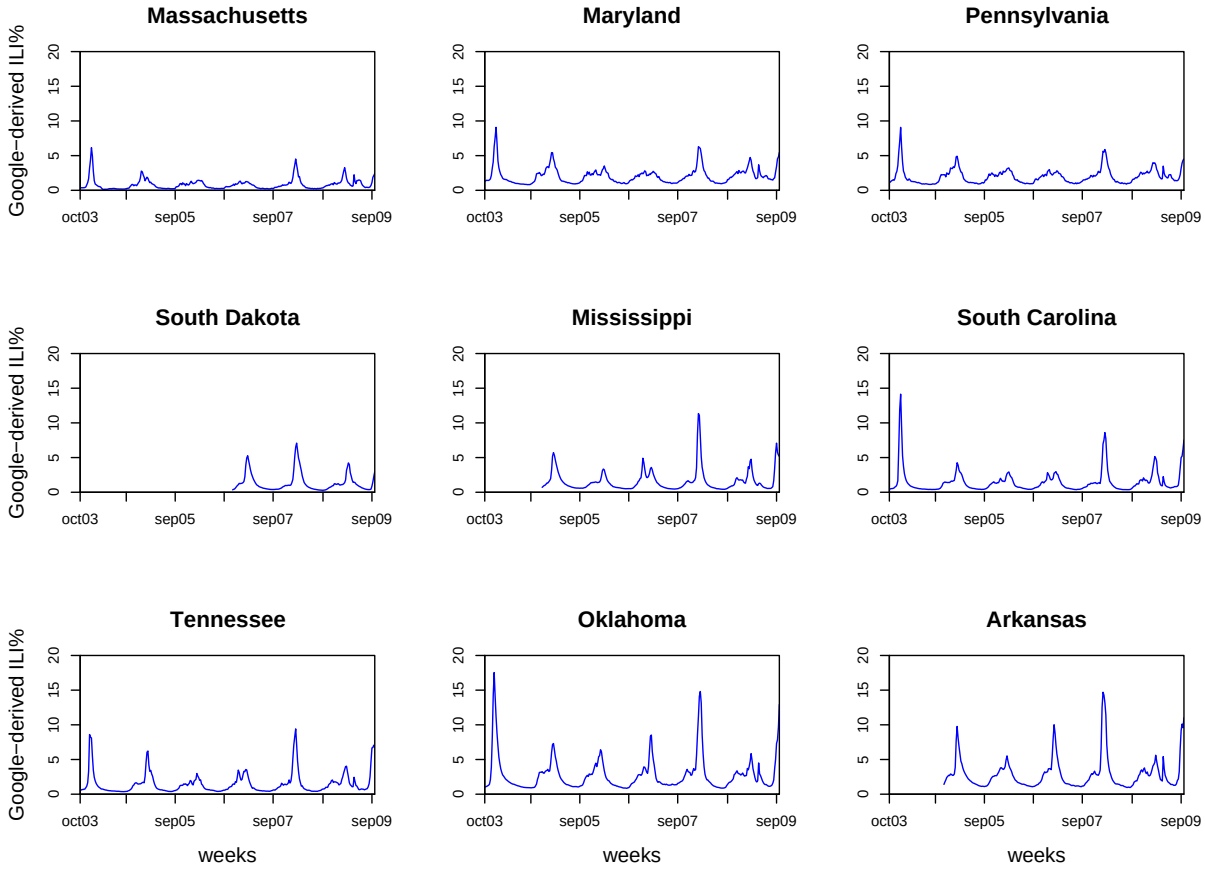

Figure 4: Normality assumption checks: The left column shows the box plots of growth rates, and the right column shows the empirical (unfilled circles) and normal CDFs (filled circles). The top row shows the 2003/2004 season, and the bottom row shows the 2008/2009 season.

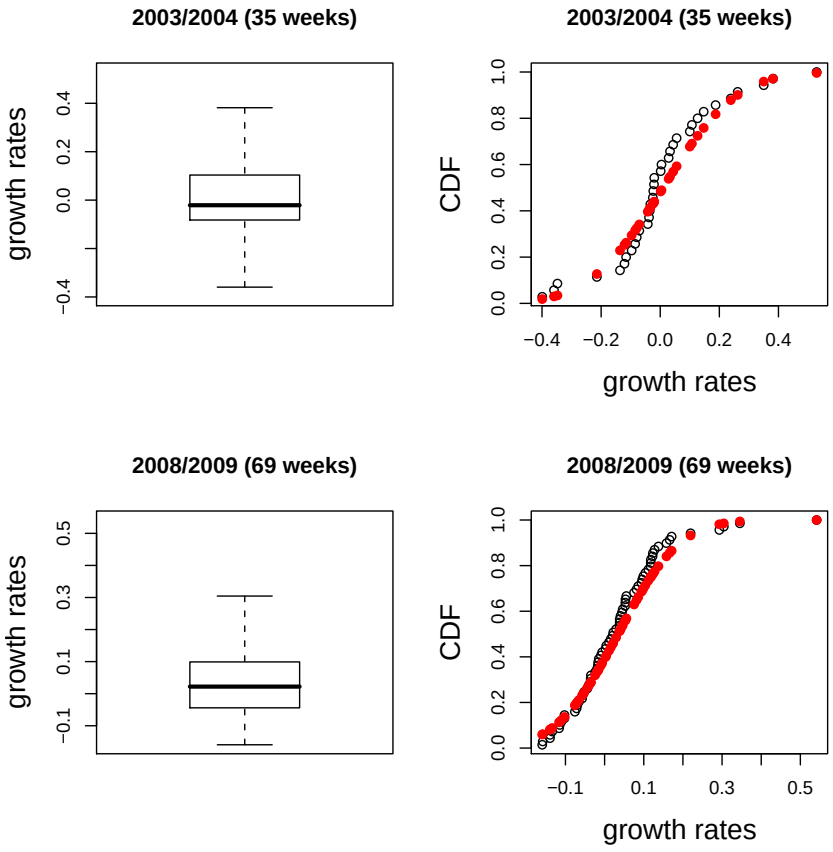

Figure 5: Flu tracking results in the US for the 2003/2004 influenza season. In the I plot (second plot in the top row), the points represent weekly Google Flu Trends values, while the lines correspond to the lower 2.5th percentile, median, and the upper 2.5th percentile of the infectious state (I_t) posterior distribution as time progresses. In the other plots, the two lines present the lower and upper 2.5th percentiles, while the points present the weekly posterior medians. The results for Bayes factors for the two competing basic reproductive ratios (1.25 vs 2.2), under 1:1 prior odds, are presented in the last panel, with higher log-Bayes factor meaning stronger evidence in favor of seasonal epidemics.

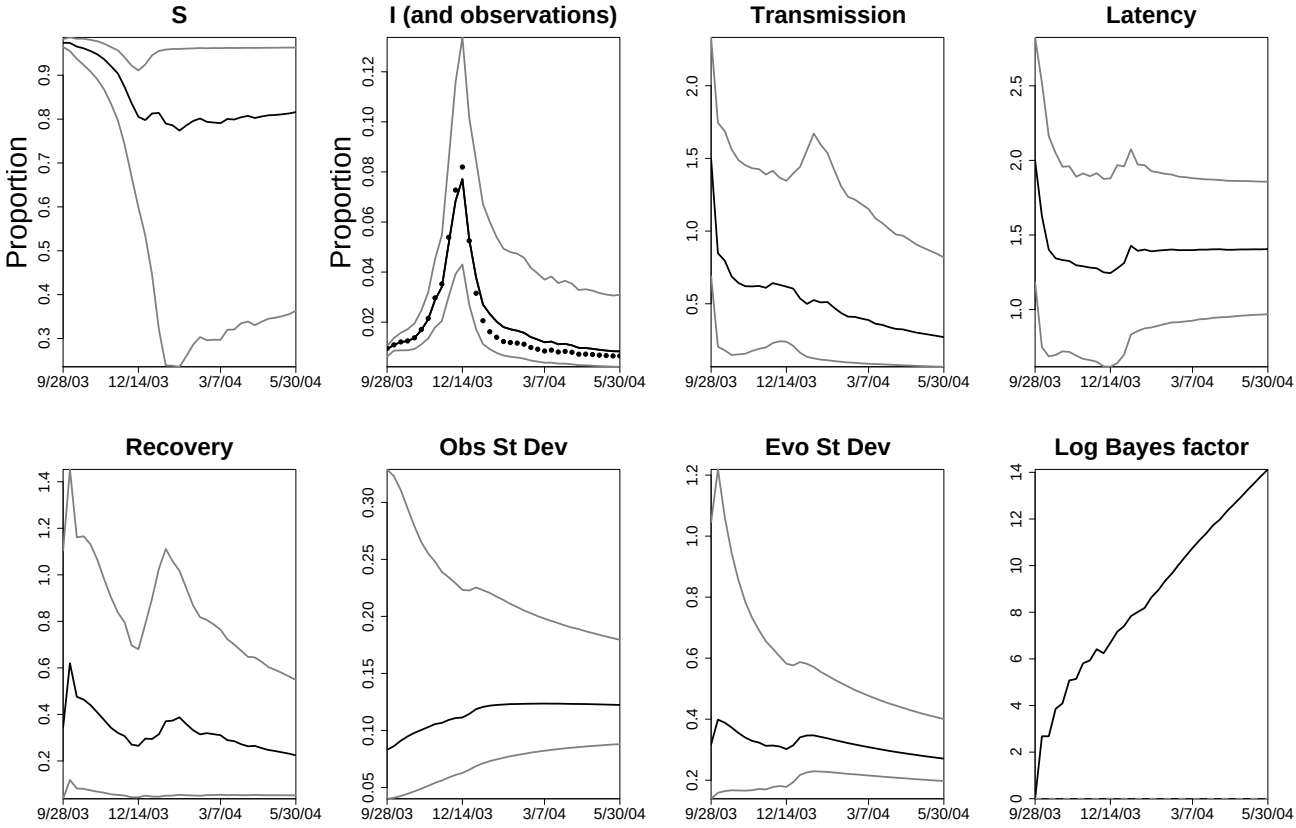

Figure 6: Flu tracking results in the US for the 2008/2009 influenza season. In the I plot, the points represent weekly Google Flu Trends values, while the lines correspond to the lower 2.5th percentile, median, and the upper 2.5th percentile of the infectious state (I_t) posterior distribution as time progresses. In the other plots, the two lines present the lower and upper 2.5th percentiles, while the points present the weekly posterior medians. The results for Bayes factors for the two competing basic reproductive ratios (1.25 vs 2.2), under 1:1 prior odds, are presented in the last panel, with higher log-Bayes factor meaning stronger evidence for seasonal epidemics.

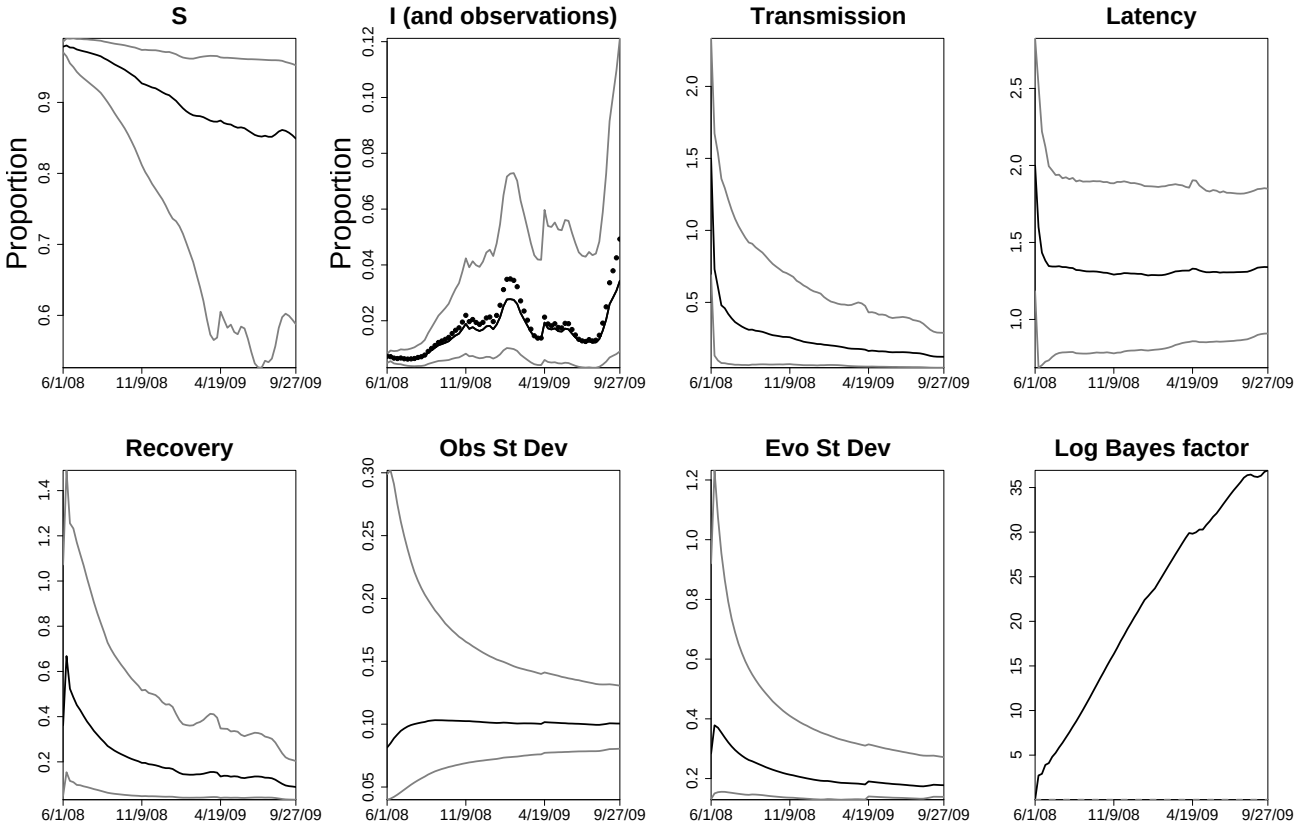

Figure 7: Sensitivity analysis under 2 additional priors on transmission rate: the gray lines correspond to a prior with the mean of 1.4, and the black lines correspond to an "optimistic" prior with the prior β mean of 1.1. The three black and gray line sets in the left column plots correspond to the upper 97.5th percentile, posterior mean, and the 2.5th percentile of the sequentially simulated marginal posteriors of the transmission parameter. The right column shows the log-Bayes factors, under 1:1 prior odds, with higher log Bayes factors indicating support for a regular epidemic. The top row shows the 2003/2004 season, and the bottom row shows the 2008/2009 season.

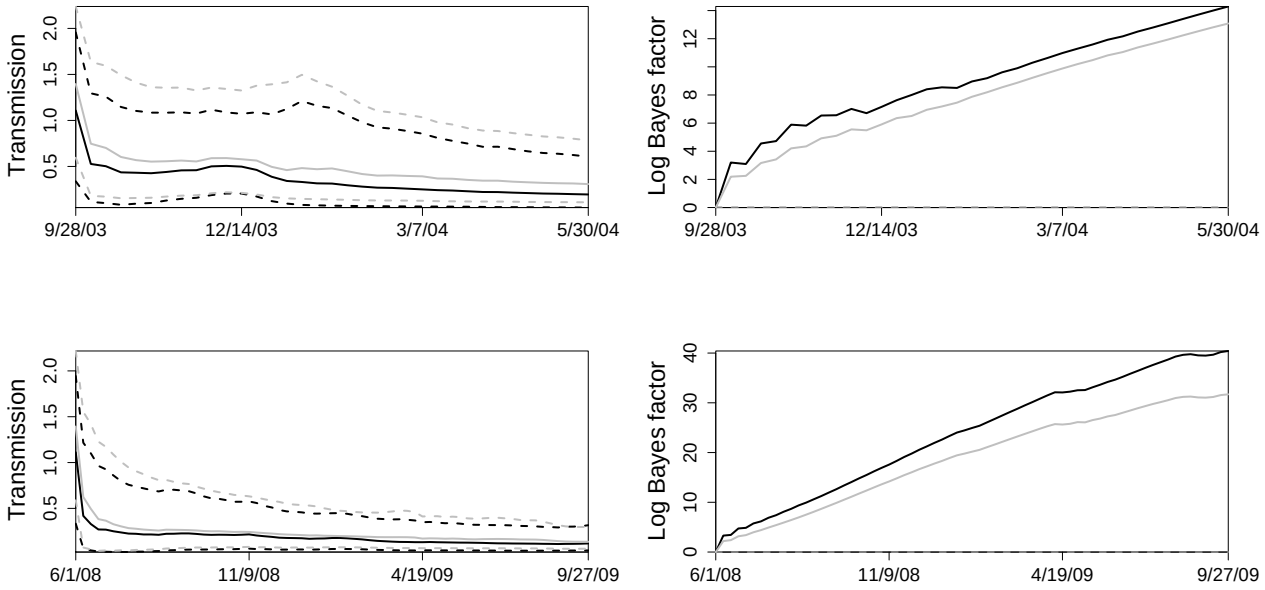

Figure 8: Sensitivity analysis for the one-week-ahead prediction under 2 different priors on transmission rate: the right column corresponds to a prior with the mean of 1.4, and the left column to an optimistic prior (with the prior β mean of 1.1). The three gray lines correspond to the upper 97.5th percentile, posterior mean, and the 2.5th percentile of the sequentially simulated predictive distributions, while the black line with points correspond to the observed data. The top row shows the 2003/2004 season, and the bottom row shows the 2008/2009 season. One-week-ahead prediction shows little sensitivity to the priors.

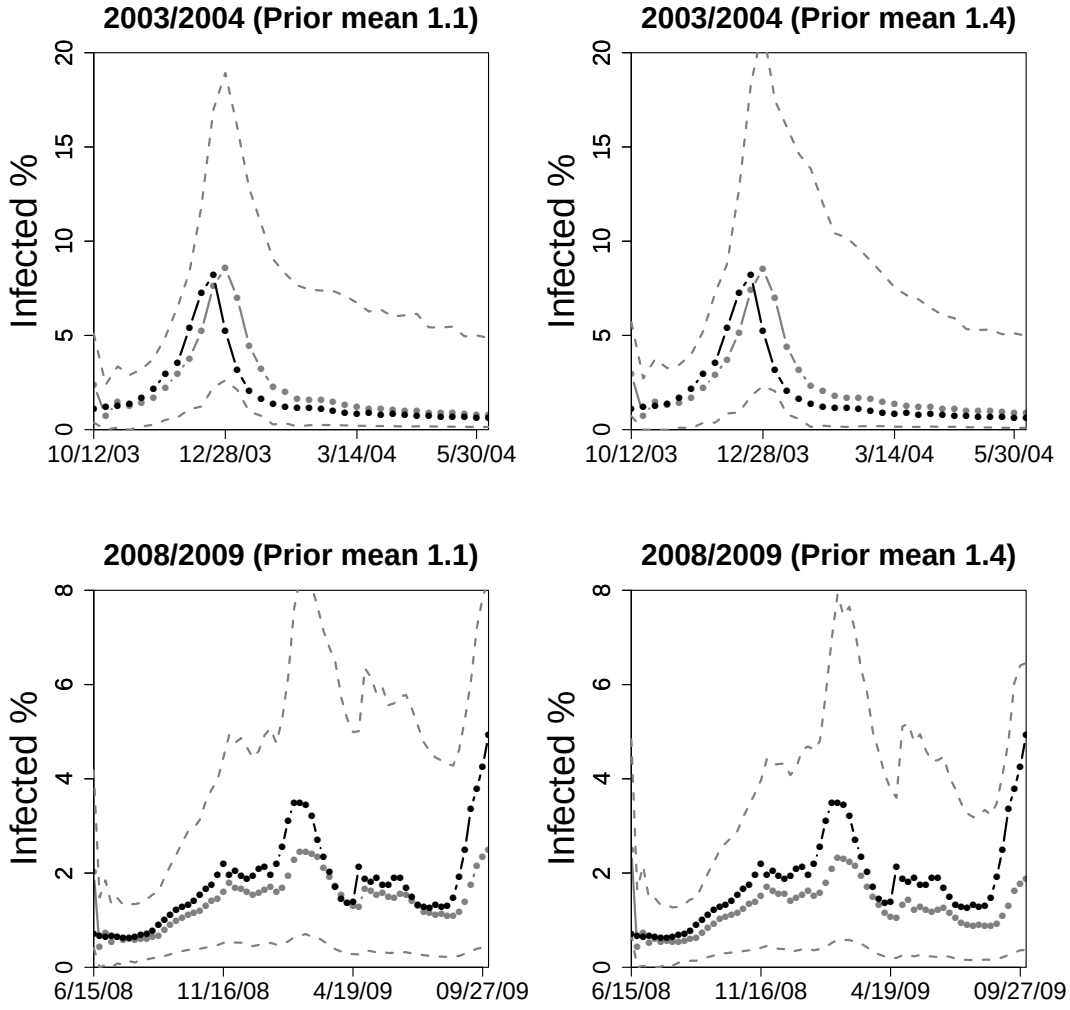

Figure 9: Sequential posterior distributions for the state-space SEIR model (left column) and the simple AR(1) benchmark model (right column) presented in Section 3, for the 2003/2004 flu season. The top row presents results for the growth rate of the infected population, and the bottom for the infected population fraction. The black circles correspond to the observations, gray squares (with gray line) are the fitted weekly values, and gray dashed lines are the 95% pointwise credible intervals. The AR(1) model is unable to capture the structure of the process as well as the state-space SEIR model.

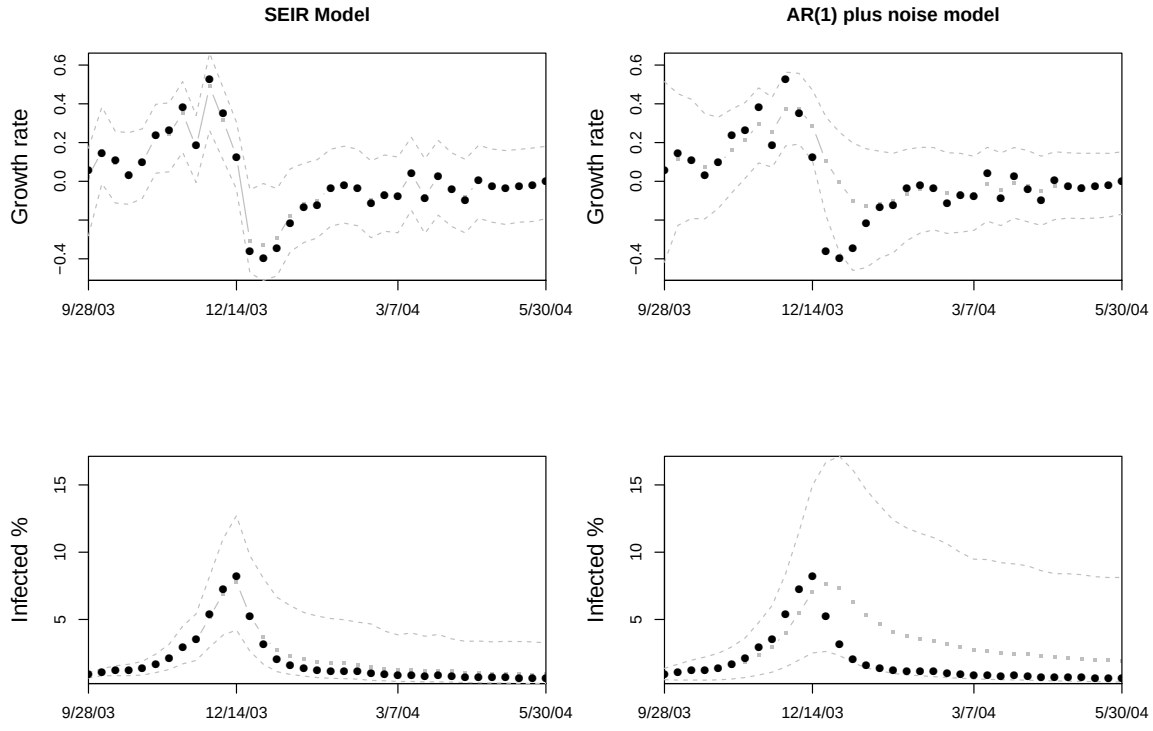

Figure 10: Comparison of the one-step ahead forecasts produced by the state-space SEIR model (left column) and the simple AR(1) benchmark model (right column) presented in Section 3. The top row presents results for the 2003/2004 flu season, and the bottom for the 2008/2009 flu season. The black squares correspond to the observations, gray circles (with gray line) are the predicted values (using data up to the previous week only), and gray dashed lines are the 95% pointwise credible intervals for the predictions. The AR(1) model predictions are not very accurate and reflect the inability of this simple model to capture the structure of the epidemic process well. The relative MSE of the AR(1) model versus the state-space SEIR model is 5.09 for the 2003/2004 season, and 2.34 for the 2008/2009 season.

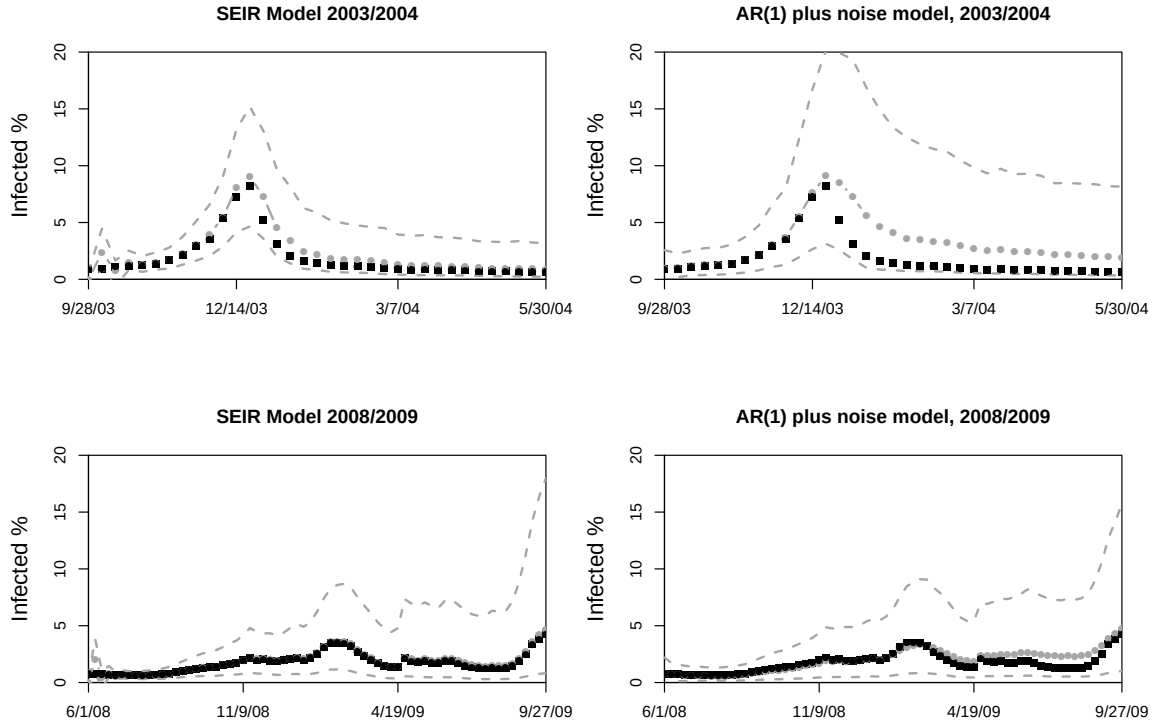

Figure 11: Flu tracking results in South Dakota (top row) and Oklahoma (bottom row) for the 2008/2009 influenza season. In the I plots (first plots in each row), the points represent weekly Google Flu Trends values, while the lines correspond to the lower 2.5th percentile, median, and the upper 2.5th percentile of the posterior distribution of I_t as time progresses. In the other plots, the two lines present the lower and upper 2.5th percentiles, while the points present the weekly posterior medians. The log-Bayes factor results for the two competing basic reproductive ratios, a mild one (1.25) and severe one (2.2), under 1:1 prior odds, are presented in the last panel. There seems to be little evidence for a pandemic.

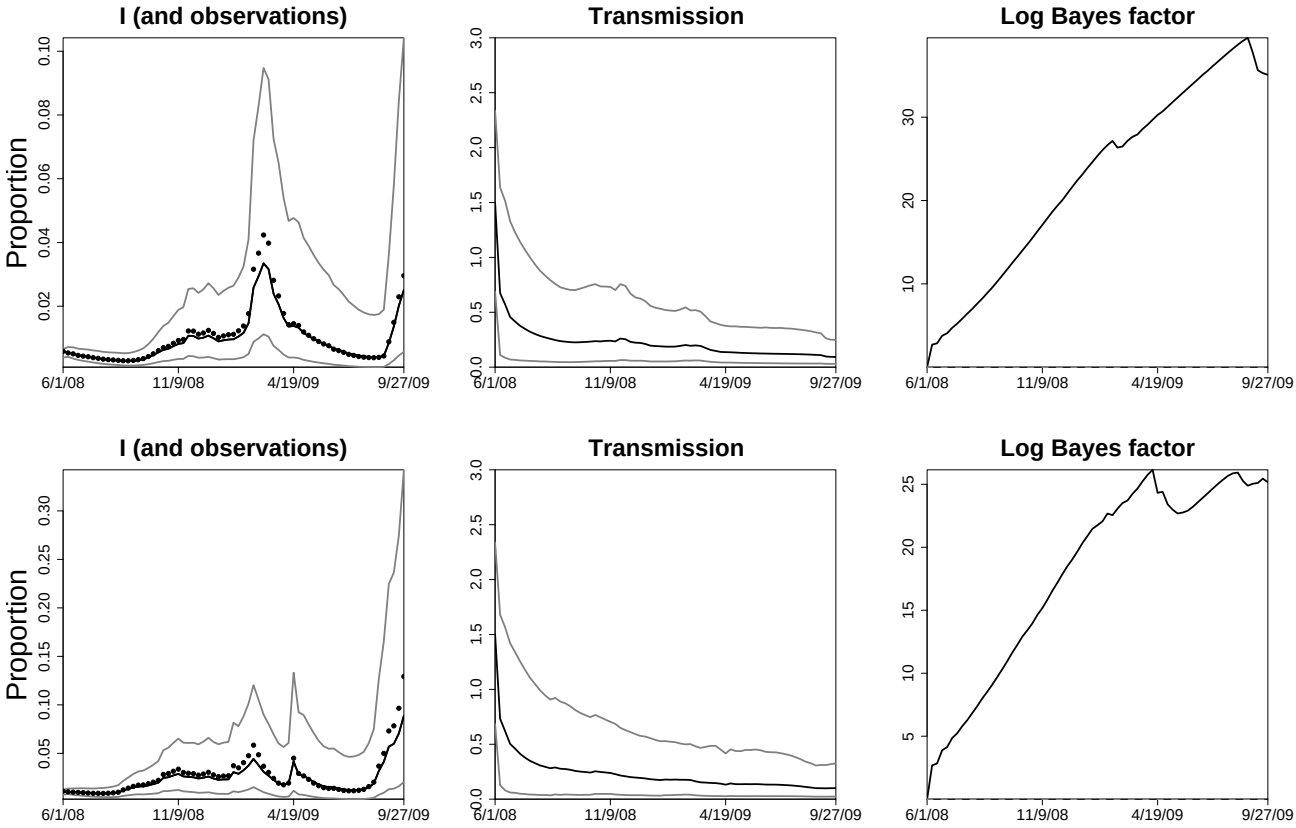

Figure 12: Comparison of posterior distributions between the sequential learning algorithm and MCMC, at the end of the 2003/2004 US flu season. Gray histograms correspond to the marginal posterior distributions obtained via MCMC (based on 1,500 samples), while the white histograms correspond to those obtained via the sequential learning algorithm ("SLA") proposed in this paper based on 1,000,000 particles.