Particle filtering in compartmental projection models

A Thesis Submitted to the
College of Graduate and Postdoctoral Studies
in Partial Fulfillment of the Requirements
for the degree of Master of Science
in the Department of Computer Science
University of Saskatchewan
Saskatoon

By
Anahita Safarishahrbijari

Permission to Use

In presenting this thesis in partial fulfillment of the requirements for a Postgraduate degree from the University of Saskatchewan, I agree that the Libraries of this University may make it freely available for inspection. I further agree that permission for copying of this thesis in any manner, in whole or in part, for scholarly purposes may be granted by the professor or professors who supervised my thesis work or, in their absence, by the Head of the Department or the Dean of the College in which my thesis work was done. It is understood that any copying or publication or use of this thesis or parts thereof for financial gain shall not be allowed without my written permission. It is also understood that due recognition shall be given to me and to the University of Saskatchewan in any scholarly use which may be made of any material in my thesis.

Requests for permission to copy or to make other use of material in this thesis in whole or part should be addressed to:

Head of the Department of Computer Science
176 Thorvaldson Building
110 Science Place
University of Saskatchewan
Saskatoon, Saskatchewan
Canada
S7N 5C9

Or

Dean
College of Graduate and Postdoctoral Studies
University of Saskatchewan
116 Thorvaldson Building
110 Science Place
Saskatoon, Saskatchewan
Canada
S7N 5C9

Abstract

Simulation models are important tools for real-time forecasting of pandemics. Models help health decision makers examine interventions and secure strong guidance when anticipating outbreak evolution. However, models usually diverge from the real observations. Stochastics involved in pandemic systems, such as changes in human contact patterns play a substantial role in disease transmissions and are not usually captured in traditional dynamic models. In addition, models of emerging diseases face the challenge of limited epidemiological knowledge about the natural history of disease. Even when the information about natural history is available – for example for endemic seasonal diseases – transmission models are often simplified and are involved with omissions. Availability of data streams can provide a view of early days of a pandemic, but fail to predict how the pandemic will evolve. Recent developments of computational statistics algorithms such as Sequential Monte Carlo and Markov Chain Monte Carlo, provide the possibility of creating models based on historical data as well as re-grounding models based on ongoing data observations. The objective of this thesis is to combine particle filtering – a Sequential Monte Carlo algorithm – with system dynamics models of pandemics. We developed particle filtering models that can recurrently be re-grounded as new observations become available. To this end, we also examined the effectiveness of this arrangement which is subject to specifics of the configuration (e.g., frequency of data sampling). While clinically-diagnosed cases are valuable incoming data stream during an outbreak, new generation of geo-spatially specific data sources, such as search volumes can work as a complementary data resource to clinical data. As another contribution, we used particle filtering in a model which can be re-grounded based on both clinical and search volume data. Our results indicate that the particle filtering in combination with compartmental models provides accurate projection systems for the estimation of model states and also model parameters (particularly compared to traditional calibration methodologies and in the context of emerging communicable diseases). The results also suggest that more frequent sampling from clinical data improves predictive accuracy outstandingly. The results also present that assumptions to make regarding the parameters associated with the particle filtering itself and changes in contact rate were robust across adequacy of empirical data since the beginning of the outbreak and inter-observation interval. The results also support the use of data from Google search API along with clinical data.

Acknowledgements

I would like to express my gratitude to my supervisor, Prof. Nathaniel Osgood, who has inspired my future ambitions and whose fastidious approach, expertise, and understanding added immensely to my graduate experience. I appreciate his support throughout my Masters degree program without which this thesis would never have been possible. The door to Prof. Osgood's office was always open whenever I ran into a trouble spot or had a question about my research and he steered me in the right direction whenever he thought I needed it.

I would also like to thank my spouse Aydin Teyhouee and my family, for their spiritual support throughout my life.

Contents

List of Tables

List of Figures

List of Abbreviations

WHO	The World Health Organization
H1N1	Influenza Al
SMC	Sequential Monte Carlo
MCMC	Markov Chain Monte Carlo
API	Application Program Interface
MLE	Maximum Likelihood Estimation
S	Susceptible
E	Exposed
I	Infective
R	Removed
V	Vaccinated
GFT	Google Flu Trend

Chapter 1

Introduction

Infectious diseases are one of the leading cause of death in the world. While many of the old infections are with us still, new infections also continue to emerge today and are a dominant feature of public health considerations. A lot of public health resources and researches are committed to modeling the spread of infectious disease based on mathematical theories and also presenting some illustrative applications of these models to improve early epidemiological assessment of epidemics.

A large variety of studies have been conducted to understand and describe the dynamics of influenza-like diseases such as H1N1 influenza. Models provide public health professionals and policy-makers with tools to examine tradeoffs between alternative strategies for clinical resource management and possible interventions. For example, dynamic models help with prioritizing vaccination initiatives and addressing effectiveness of interventions such as social distancing measures, including school closure and suspension of public activities.

Inevitably, dynamic models are simplifications of real systems and are make use of parameter values that are either uncertain or themselves evolving stochastically over time. Stochastic transitions associated with societal and economic behaviours along with the lack of information about the natural history of diseases make it hard to anticipate the progression of outbreaks, particularly fast-breaking ones or those involving emerging infectious diseases. For example, in infection transmission models, it can be challenging to obtain estimates for parameters such as contact patterns, fraction of total incident cases that are reported, and initial values of model states. While calibration of dynamic models can help with short-term projection of pandemics, they often fail to accurately predict incidence rates across longer time-frames. Calibration accuracy depends heavily on the size of available historical data and calibration methods thus offer limited benefit for early prediction of outbreaks. They also lack the capability to adapt based on the latest and new-arrived data points, reflecting the fact that re-calibration typically requires substantial manual effort. This adaptation to new data is of critical importance when learning about new diseases. The calibration approaches support estimation of parameters, but not model states. While parameter values may be updated to the best estimate, the model's estimate of state often increasingly diverge over time from real-world state.

Within this thesis, we use a statistical filtering method in the form of particle filtering to overcome shortcomings of the traditional calibration methodology. Particle filtering is used to run particles simultaneously offering different hypotheses concerning the entirety of underlying model state. A likelihood test is performed to identify the particles that best match with the observed measurements based on survival of the fittest.

While empirical data may be limited to matching small pieces of the model, particle filtering allows for estimating (via sampling) the full extent of the state of the model, whether latent or observed. Because accurate understanding of the latent state is a fundamental enabler for accurate assessment of intervention tradeoffs, not only does particle filtering enhance predictive accuracy, but it also supports elevated understanding the current situation (via latent state estimation) and provides the capacity to accurately assess intervention tradeoffs. It bears emphasis that each particle estimates the full state of the model, including all latent variables.

To evaluate the effectiveness of particle filtering, we examined a case study based on the second peak of pandemic H1N1 in the province of Manitoba during 2009-2010. For one investigation, we also considered the second peak of H1N1 in Quebec during 2009-2010.

Since reporting of clinically confirmed cases is subject to significant inaccuracy, we further sought to investigate trade-offs between employing less frequent but more stable data sampling and more frequent but noisy measurements. We also explored the validity of the particle filter to assumptions underlying the method and also about the behavioural change in population.

Clinically-observed data offer rich information concerning individuals who seek medical treatments, but fail to catch information about those who do not present for medical care. A large portion of population – either infected or not infected, but anxious about a pandemic – might use search engines to obtain information about vaccination, treatments and news about the pandemic. Google Trends and specifically Google Flu Trends (GFT) provide data about search data trends and search volumes [2, 3]. Although there are some criticisms on GFT algorithms [4], several studies have been conducted regarding classifying and analyzing search data, finding correlation between clinical datasets and search query volumes and also examining the capability of search data for pandemic predictions via statistical methods. To our knowledge, none of the studies have investigated the ability of aggregate compartmental or System Dynamics models to be recurrently re-grounded using both search query data and clinical data. Reflecting this opportunity, in the final investigation of this thesis, we investigated the possibility of using search volumes along with clinical data during a pandemic for projection of pandemic progress.

1.1 Motivation

Despite much progress on the public health front, the burden of communicable diseases remains globally high. Of particular importance in a world where human development increasingly encroaches upon natural ecosystems are emerging disease, which are diseases that are new, or has been detected in a new region or with manifestations that differ from what was previously recognized. Emerging infectious diseases or pathogens exert high burdens in the public health area. The large avian influenza (A) H7N7 outbreak in the Netherlands in 2003, the global severe acute respiratory syndrome (SARS) outbreak in 2003, Marburg virus importation in 2008, pandemic influenza A (H1N1) during the 2009-2010 influenza season and other avian influenza virus

outbreaks when the influenza (A) H5N1 viruses started to spread from China, Ebola virus disease, Middle East respiratory syndrome (MERS) and Zika are some examples of recent emerging pandemics. Figure 1.1 demonstrates the cumulative confirmed cases of H1N1 for different countries during 2009 pandemic. It shows that the total cases of H1N1 reaches to more than 65,000 at the end of pandemic.

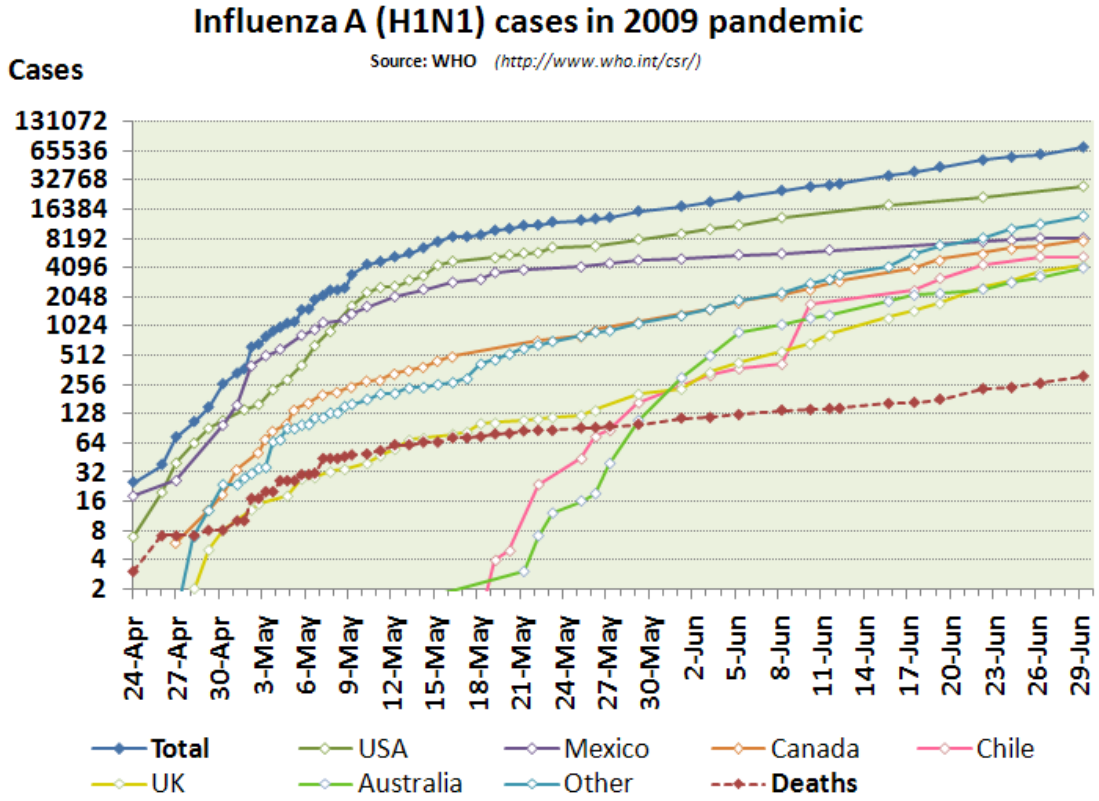


Figure 1.1: Cumulative H1N1 cases in the 2009 pandamic (The image is taken from [1]).

In addition to the costs associated with vaccination and hospitalization, according to a national survey in Canada, high levels of absenteeism amongst critical infrastructure workers during a pandemic influenza outbreak would create a substantial and immediate negative influence on the economy. Guy Holburn, Associate Professor at the University of Western Ontario's Richard Ivey School of Business, estimates that absenteeism amongst workers during an influenza pandemic could cost (via absenteeism) the Canadian economy $9 billion [5].

In addition to financial burdens, key challenges to control of an outbreak of an emerging disease include how fast such events can develop, the difficulty of working with an unknown disease, the challenge of ensuring a coordinated response between general population and public health experts, public health resource constraints and manpower requirements.

Mathematical modelling plays an increasingly important role in helping to guide policy-makers to overcome these challenges. Modelling studies are increasingly performed to address questions about the effec-

tiveness of interventions. Modelling may also be useful in the context of prioritizing and planning clinical trials. Finally, mathematical modelling can be used in economic evaluations of clinical and public health interventions and in assessing long-term outcomes.

Observational data can be used to evaluate a public health policy after it is underway, but have little value in helping to project the future impact of a proposed program. Furthermore, when an emerging outbreak occurs, it is often required to respond to new threats, for which there is limited or no previous data on which to assess the threat.

Computational and mathematical models can aid assessment of potential impacts early in the process. Models can also help in interpreting data from complex systems; however, there are a number of challenges in achieving a successful model.

Model projections depend on underlying assumptions and model parameters. Problematic assumptions can lead to flawed public health policies [6]. Models of the time-course of infection, in particular, make use of parameters which are central to predicting infection trajectories for individuals, thereby movement between population categories [7]. If the underlying assumptions and model parameters are poorly defined, subsequent modelling can be conjectural. This issue is particularly challenging in emerging infections, where there is a shortage of observations regarding the time-course of disease [8].

The unknown natural history and pathogenesis of diseases affected by previous interventions (e.g., via selective pressures) or other changes to the system can also significantly challenge the model predictions.

Identifying approaches to improve the value of model predictions despite model inaccuracies and highlighting predictions that are close to real-world observations can help overcome the challenge of unknown history and parameters involved in models. The development of such methods can be possible by updating and checking model assumptions as more data becomes available. Such methods have the potential to create models that can be used in ongoing planning.

Combining dynamic models of pandemics with new generations of computational statistics algorithms makes it feasible to reformulate models as new observations become available. These algorithms can also help in estimating unknown parameters and latent state of the model, and thus help with more accurate prediction of future outcomes based on current observations.

1.2 Problem

To investigate whether joining mathematical transmission models for influenza and empirical data can improve the accuracy of model predictions of incident case counts.

To resolve this problem, there are a number of technical problems that need to be addressed in order to create reliable simulations models:

1. Pandemic simulation models are commonly involved with unknown or little-known parameter values, such as contacts per day and fraction of incident cases reported, and vaguely known aspects of the

natural history of the diseases; e.g., parameters associated with initial values of recovered and susceptible states. At the same time, there is often medium- to large-amounts of data that relates not to specific parameter values isolation, but which instead describes the emergent behaviour of the system or subsystems. Traditionally, many modelers employ calibration methodologies to estimate the unknown parameters and tune models in the presence of such emergent empirical data. However, such approaches are not flexible in evolving the parameter values over time or estimating the latent state of the model state variables. Therefore, other techniques should be examined to leverage the presence of empirical data and help models to reground model state estimates as new data becomes available.

2. Although combining computational statistics algorithms such as particle filtering with incoming observations provides the possibility of regrounding dynamic models with empirical data, the effectiveness of such algorithms is prone to be affected by configurations such as frequency of sampling from data and representation of parameter change.

3. Traditional empirical data such as clinically-confirmed case reports can provide valuable information about infected population who seek medical care; however, such traditional data is burdened by some shortcomings, including delays in reporting infected cases and a failure to capture infected individuals who do not present for clinical services.

1.3 Solution

In this section, we briefly describe our solutions to the problems mentioned above. Each problem is addressed in this section briefly, and then with complete details and results in the following chapters.

1. Several studies used specified formulation by public health officials to identify the unknown parameters of pandemic models [9, 10]. Traditionally, transmission modeling is used to unravel epidemic development in the area of infectious diseases. Some studies account for the effect of behavioural changes on disease transmission [11, 12, 13]. Such previous models, to deal with parameters that change over the course of a pandemic, considered the parameters to be static, either not taking advantage of empirical data for estimating unknown parameter values at all or by using traditional calibration methodologies, which are not reliable means of keeping current with the latest in empirical data. It would be more desirable to allow those parameters to be adapted dynamically as new data points are available during a pandemic. To address this problem, we used particle filtering methodology to enable the model to learn from ongoing real-world data in a dynamic fashion in order to estimate model state evolution as well as stochastics associated with selected model parameters. We further investigated whether particle filtering is more capable than calibration in estimating the uncertain parameters and in predicting model elements associated with data from real-world outbreaks.

2. To overcome the second problem, we formulated a set of scenarios to explore how changes to particle

filter configurations and also empirical data would affect the error associated with particle filtered model predictions. Specifically, we investigated the choice of the values of the dispersion parameter associated with the negative binomial likelihood formulation, the contact rate volatility parameter, the total period for which empirical observations were available so that the model could learn from them, and the frequency of aggregation associated with empirical data observations provided to the model. Choices of such values are especially important for health decision makers to obtain robust guidance when anticipating outbreak evolution for emerging infectious diseases by combining preliminary models with particle filtering techniques.

3. Taking advantage of the increasing tendency of many individuals to post and tweet about their illnesses and to use search engines such as Google to obtain information about diseases and their treatments, we evaluated the gains secured from the use of an online source of data to complement clinical datasets. Time series of volumes of Google searches over time can be used to explore the presence of influenza-like illnesses in the population that are not necessarily included in the empirical records provided by clinics. We used search query volumes previously provided by Google Trends and Google Flu Trends. To address this task, we adapted the compartmental model developed by Epstein et al., which explored the effect of behavior changes such as social distancing based on fear in epidemic dynamics [14]. We combined particle filtering with this adapted model to help model learn from both clinical and search volume datasets. This approach exploited the fact that the large volume of data available from communicational activities of the population can offer early information to the model regarding disease activity. This can help policy makers respond quickly to reduce the impact of pandemic and seasonal influenza like diseases.

1.4 Contributions

The main contributions of this thesis are as follows:

1. **Investigating the performance of particle filtering in predicting pandemic influenza using empirical data.** We used empirical data obtained from Manitoba Health, Healthy Living and Seniors, which indicated weekly confirmed cases and vaccine delivery rates of pandemic H1N1 in 2009-2010. We further compared the accuracy of particle filtering to that obtained via calibration methodology in terms of their ability to project pandemics in a compartmental model with stochastic parameters.

2. **Implementation of different scenarios to obtain an optimum range for configuration parameters, the sampling period and observation frequency**. To explore the pattern of change in contact rate over the period of an outbreak, and how it affects the spread of infection, we performed particle filtering examining different values for the contact rate volatility parameter over a broad range. The performance of particle filtering to projected infected case counts is also sensitive to the type

of likelihood function, and specifically in our work to the dispersion parameter associated with the negative binomial distribution. Retaining the mean value to be constant, lower values for dispersion parameter elevate the dispersion associated with the likelihood function. We examined how different values of dispersion parameter affect particle filtering performance. The noise in the clinically observed data is often pronounced. Aggregating data over a longer period – more than one day – between observations reduces the proportional size of the noise associated with such data; however, aggregation yields fewer data points, and hence particle filtering learns from fewer observations. To investigate the trade-off between employing more aggregated but less noisy data when compared to less aggregated but more noisy data, we examined different inter-observation aggregation intervals. With an original data source supplying daily data, we examined the effects of using data daily, aggregated over three days and aggregated over seven days for the purpose of sampling in particle filtering. A common scenario anticipated for application of particle filtering would be one in which the procedure is used throughout an outbreak. At any one time, particle filtering can only take into account data observed from the start of the outbreak until that timepoint (a timepoint which we denoted as T^*). We examined the performance of particle filtering for different data points in which particle filtering would be able to re-sample and learn from empirical data. We considered T^* equivalent to predictions made at 5, 6, 7 and 8 weeks into the outbreak.

3. **Examining the performance of a particle filtered compartmental model in prediction of pandemics progression using search volumes during an outbreak.** We implemented particle filtering using two different datasets and examined whether combining dataset that moves beyond purely a clinically observed dataset to also includes a time series of search volumes can enhance prediction of pandemic progression, which can help with earlier warning, and hence earlier prevention and control measures during an outbreak.

1.5 Thesis Outline

In this section, we will present an overview of the remainder of the thesis.

- Chapter 2 presents background information on topics that are necessary for understanding the methodologies employed in this thesis, including particle filtering and transmission models for influenza-like illnesses.

- Chapter 3 examines a Susceptible, Exposed, Infected, Recovered and Vaccinated (SEIRV) compartmental model for influenza. Particle filtering is tested and compared with calibration in terms of the discrepancy between empirical data and model output.

- Chapter 4 describes the unknown and stochastic parameters in a particle filtered SEIRV model, the reason for considering parameters such as contacts per unit time and fraction reported incidence to be

states of the particle filtered model. The chapter introduces scenarios for studying different configurations of particle filtering. We provided a detailed exploration that explores whether and how different configurations affect the performance of particle filtering, and measured by the same discrepancy introduced in Chapter 3. Some limitations of particle filtering are also discussed in this chapter.

- Chapter 5 describes the use of particle filtering to adapt a previously developed model – which considered fear among people during a pandemic – to be informed by a time series of search volumes. This chapter provides details regarding the use of data from online communicational behaviour that can improve pandemic predictions. Some other limitations of particle filtering are also discussed in this chapter.

- Chapter 6 provides a summary of the thesis. This chapter presents the main contributions of the thesis, limitations of the work, and directions for future work that can improve the results of this thesis.

1.6 Publications

- Chapter 3 includes a manuscript entitled "Particle Filtering in a SEIRV Simulation Model of H1N1 Influenza" by Anahita Safarishahrbijari (AS), Trisha Lawrence (TL), Richard Lomotey (RL), Juxin Liu (JL), Cheryl Waldner (CW) and Nathaniel D Osgood (NDO), published in Proceedings of the 2015 Winter Simulation Conference [15]. Authors' contributions are as follows:

  AS drafted the manuscript; TL helped with drafting the introduction and "Motivation for Calibration and Particle Filtering" sections; RL helped with drafting the "Introduction" section; NDO designed and supervised the study, provided the basic skeletal SMC structure, provided help in adapting it to the H1N1 context, and modified the manuscript; AS and TL and RL contributed in modeling and adapting the SMC framework; AS obtained results; AS contributed in obtaining empirical data from the source website; CW and NDO gave advice about the model parameters and validity of the results; JL provided advice regarding SMC.

- Chapter 4 includes a manuscript entitled "Predictive Accuracy of Particle Filtering in Dynamic Models Supporting Outbreak Projections" by AS, Aydin Teyhouee (AT), JL, CW and NDO, published in BioMed Central Infectious Diseases Journal [16]. Authors' contributions are as follows:

  AS, CW and NDO drafted the manuscript; NDO designed and supervised the study and helped advise on adaptation of SMC machinery reused from the model characterized in the previous chapter; AS and CW performed the statistical analysis; AS and AT contributed in modeling and obtaining results; AS contributed in obtaining empirical data from the source website; CW and NDO gave advice about the model parameters and validity of the results; JL gave advice about SMC.

- Chapter 5 includes a manuscript entitled "Social Media Surveillance Improves Outbreak Projection via Transmission Models" by AS and NDO, submitted to the Journal of Medical Internet Research. Authors' contributions are as follows:

AS and NDO drafted the manuscript; NDO designed and supervised the study and helped advice on adaptation of the SMC algorithm used from the previous chapter; AS contributed in modeling, adapting SMC to the model and obtaining results; NDO provided advice about the model and the relationship between the empirical data and the model; AS contributed in obtaining empirical data from the source websites.

Chapter 2

Background

This chapter focuses on techniques that are applied to develop predictive models. In this chapter, we provide background on mathematical description of particle filtering. Section 2.1.1 describes the basics of Monte Carlo methods. Section 2.1.2 will include an introduction to particle filtering and provides those characteristics of particle filtering that are shared between all chapters in this thesis. Section 2.2 briefly explains influenza, particularly the H1N1 strain and section 2.3 describes a simple transmission model for communicable diseases. The chapter further discusses a background of the role of data on predictive models in the area of public health (section 2.4).

2.1 Particle Filtering

Particle filtering is a broad and popular class of Monte Carlo algorithms to provide approximate numerical solutions to problems for non-Gaussian non-linear state-space models – problems which typically cannot be solved analytically. The particle filtering algorithm was first introduced in 1993 by Gordon et al.[17]. Different methods of filtering algorithms have been used in different fields from computer vision and navigation to economics and mathematical finance [18, 19, 20, 21].

2.1.1 Sequential Monte Carlo Methods

Advanced particle methods for filtering and smoothing are amongst the most common techniques for approximation derived from the general sequential Monte Carlo (SMC) algorithm. This technique is useful for online inference in dynamic systems and overcomes the limitations associated with analytically tractable solutions, which are available for linear Gaussian models, but not for complex models. SMC is a subclass of Monte Carlo algorithms that sequentially samples from a sequence of target probability densities $\pi_n(x_{1:n})$ so as to compute the posterior distributions. Each target probability density ($\pi_n(x_{1:n})$) is defined on the product space $\mathcal{X}^n$ – n refers to the time and is a natural number – and can be written in the form of:

$$\pi_n(x_{1:n}) = \frac{\gamma_n(x_{1:n})}{Z_n} \tag{2.1}$$

where $\gamma_n(x_{1:n})$ is a distribution defined on the product space $\mathcal{X}^n$ and Z_n is the normalizing constant:

$$Z_n = \int \gamma_n(x_{1:n})\, dx_{1:n} \tag{2.2}$$

An approximation of $\pi_1(x_1)$ and an estimate of Z_1 at time 1 are provided by SMC. Then an approximation of $\pi_2(x_{1:2})$ and an estimate of Z_2 are provided at time 2. This approximation continues to time n. For filtering techniques, if we choose $\gamma_n(x_{1:n})$ to be $p(x_{1:n}, y_{1:n})$ and Z_n to be $p(y_{1:n})$, then $\pi_n(x_{1:n})$ would be $p(x_{1:n}|y_{1:n})$ [22]. For the case considered here, $x_{1:n}$ represents the latent state of dynamic model.

Monte Carlo Methods – The Basics

This section closely follows [23]. To approximate a generic probability density $\pi_n(x_{1:n})$, we can sample N random variables $X^i_{1:n}$ (where $1 \leq i \leq N$) is distributed according to that distribution and approximate the distribution as follows:

$$\widehat{\pi}_n(x_{1:n}) = \frac{1}{N} \sum_{i=1}^{N} \delta_{X^i_{1:n}}(x_{1:n}) \tag{2.3}$$

where $\delta_{x_0} x$ denotes the Dirac delta mass (informally, impulse function) at x_0. The expectation of a function φ_n of a random variable X that has a density $\pi_n(x_{1:n})$ is given by:

$$E(\varphi_n) = \int \varphi_n(x_{1:n})\, \pi_n(x_{1:n})\, dx_{1:n} \tag{2.4}$$

yielding a Monte Carlo estimation of the expectation as:

$$E(\varphi_n) := \int \varphi_n(x_{1:n})\, \widehat{\pi}_n(x_{1:n})\, dx_{1:n} = \frac{1}{N} \sum_{i=1}^{N} \varphi_n(X^i_{1:n}) \tag{2.5}$$

1. While sampling is readily achieved for simple (e.g., uniform or normal) distributions or for unidimensional distribution (via computation of the cumulative distribution), it is challenging to sample from high dimensional probability distributions which are of complex character.

2. Even if we could easily sample from an arbitrary high-dimensional probability distribution $\pi_n(x_{1:n})$, the computational complexity of such a sampling increases linearly as the dimensions n increases.

Importance sampling and sequential importance sampling are two functional Monte Carlo methods that address both of the problems above, respectively[23].

Importance Sampling

Importance sampling is an approach that addresses the first problem above using a two-phased approach to sampling from a target distribution. In the first phase, the approach draws samples generated from a different distribution from which it is easy to sample, such as a multivariate Gaussian distribution, or an exponential distribution, but weights those samples in a manner that takes into account the features of the

target distribution. The second phase then samples from these samples with a probability given by the weight. Here we cover the first phase of importance sampling; resampling is covered in section 2.1.1 below.

Importance sampling requires a density called the importance density, proposal or instrumental density, $q_n(x_{1:n})$, from which it is easy to sample, and which is guaranteed to be of non-zero density for all points $x_{1:n}$ for which the target density has non-zero density.

We then make use of weights $\omega_n(x_{1:n})$ and the relations 2.1 and 2.2 [24] to give

$$\pi_n(x_{1:n}) = \frac{\omega_n(x_{1:n})\, q_n(x_{1:n})}{Z_n} \tag{2.6}$$

Which implies that the un-normalized weight function is given by $\omega_n(x_{1:n}) = \frac{\gamma_n(x_{1:n})}{q_n(x_{1:n})}$.

It follows follows that:

$$Z_n = \int \omega_n(x_{1:n})\, q_n(x_{1:n})\, dx_{1:n} \tag{2.7}$$

If we draw N samples $X^i_{1:n}$ from the importance density $q_n(x_{1:n})$, we can then consider 2.5 and 2.7 to obtain the approximation:

$$\widehat{Z}_n = \frac{1}{N} \sum_{i=1}^{N} \omega_n(X^i_{1:n}) \tag{2.8}$$

In turn, by inserting the Monte Carlo approximation of $q_n(x_{1:n})$ into 2.6 and 2.8, we have:

$$\widehat{\pi}_n(x_{1:n}) = \sum_{i=1}^{N} W^i_n\, \delta X^i_{1:n} \tag{2.9}$$

where the normalized weights are given as follows:

$$W^i_n = \frac{\omega_n(X^i_{1:n})}{\sum_{j=1}^{N} \omega_n(X^j_{1:n})} \tag{2.10}$$

If we were interested in computing the expectation of a function φ_n, then we can use the estimate:

$$E^{IS}(\varphi_n) := \int \varphi_n(x_{1:n})\, \widehat{\pi}_n(x_{1:n})\, dx_{1:n} = \sum_{i=1}^{N} W^i_n\, \varphi_n(X^i_{1:n}) \tag{2.11}$$

Sequential Importance Sampling

Sequential importance sampling is an algorithm that can address problem 2 above by lowering the computational complexity at each time step [25] through recursive characterization of weights. In this algorithm, we elect to adopt an importance distribution that can be characterized as follows:

$$q_n(x_{1:n}) = q_{n-1}(x_{1:n-1})\, q_n(x_n|x_{1:n-1}) = q_1(x_1) \prod_{k=2}^{n} q_k(x_k|x_{1:k-1}) \tag{2.12}$$

To obtain particles $X^i_{1:n} \sim q_n(x_{1:n})$ at time n, we first sample $X^i_1 \sim q_1(x_1)$ at time 1. Then we sample $X^i_k \sim q_k(x_k|X^i_{1:k-1})$ at time k and for $k = 2, ..., n$. By virtue of selecting an importance distribution using

the structure assumed above, we can compute the unnormalized weight recursively at each timepoint l in a way that considers just the new data for time l – rather than having to consider all of the data for time $1 \leq k \leq l$:

$$\omega_n(x_{1:n}) = \frac{\gamma_n(x_{1:n})}{q_n(x_{1:n})} = \frac{\gamma_{n-1}(x_{1:n-1})}{q_{n-1}(x_{1:n-1})} \frac{\gamma_n(x_{1:n})}{\gamma_{n-1}(x_{1:n-1})\, q_n(x_n|x_{1:n-1})} \tag{2.13}$$

It is can be recognized that the first quotient in the equation 2.13 is simply $\omega_{n-1}(x_{1:n-1})$ or it can be written recursively as:

$$\omega_n(x_{1:n}) = \omega_{n-1}(x_{1:n-1})\, \alpha_n(x_{1:n}) \tag{2.14}$$

where

$$\alpha_n(x_{1:n}) = \frac{\gamma_n(x_{1:n})}{\gamma_{n-1}(x_{1:n-1})\, q_n(x_n|x_{1:n-1})} \tag{2.15}$$

Alternatively, the above can be unpacked in an iterative fashion as

$$\omega_n(x_{1:n}) = \omega_1(x_1) \prod_{k=2}^{n} \alpha_k(x_{1:k}) \tag{2.16}$$

Following [26], sequential importance sampling can thus be summarized as follows:

For the initial time (time $n = 1$)

- Draw X_1^i from $q_1(x_1)$
- Calculate weights $\omega_1(X_1^i)$, and normalized weights $W_1^i = \frac{\omega_n(X_1^i)}{\sum_{j=1}^{N} \omega_n(X_1^i)}$. By default, we can elect to impose uniform weights at time 1, and thus $\omega_1(X_1^i) = 1$, and $W_1^i = \frac{1}{N}$.

For time $n \geq 2$

- Sample X_n^i from $q_n(x_n|X_{1:n-1}^i)$
- Compute the weights recursively according to 2.15, as $\omega_n(X_{1:n}^i) = \omega_{n-1}(X_{1:n-1}^i)\, \alpha_n(X_{1:n}^i)$

Re-sampling

Section above 2.1.1 noted that importance sampling involves two successive phase; in the first phase, the importance sampling approximation $\widehat{\pi}_n(x_{1:n})$ of a target distribution $\pi_n(x_{1:n})$ is generated by weighted sampling from $q_n(x_{1:n})$. In the second phase, to draw approximate samples from the target distribution $\pi_n(x_{1:n})$, we can sample from its importance sampling approximation $\widehat{\pi}_n(x_{1:n})$ by selecting $X_{1:n}^i$ with the probability of W_n^i. Since we sample from an approximation $\widehat{\pi}_n(x_{1:n})$, which was itself generated from sampling, this process is called resampling.

Generic Sequential Monte Carlo Algorithm

Sequential Monte Carlo algorithms are developed by joining sequential importance sampling method and resampling. At time 1, we collect some weighted particles (W_1^i, X_1^i) and generate an importance sampling approximation $\widehat{\pi}_1(x_1)$ of $\pi_1(x_1)$. In the next step, we resample particles, drawing, as usual, each with a probability proportional to its weight. As a result, the particles with low weights tend to perish, and those with high weights tend to reproduce. We then associate a weight of $\frac{1}{N}$ with each particle. We denote the equally-weighted re-sampled particles by $(\frac{1}{N}, \bar{X}_1^i)$. In the next step, according to sequential importance sampling, we sample $X_2^i \sim q_2(x_2|\bar{X}_1^i)$. Hence $(\bar{X}_1^i, X_2^i)$ is distributed according to $\pi_1(x_1)\, q_2(x_2|x_1)$. As a result, we can then compute the corresponding importance weights simply as $\alpha_2(x_{1:2})$. We then resample particles based on these weights [25, 27]. A summary of the sequential Monte Carlo algorithm is as follows. This formulation closely follows that of [26].

For time $n = 1$

- Sample X_1^i from distribution $q_1(x_1)$.
- Compute the weights $\omega_1(X_1^i)$ and correspondingly normalized weights $W_1^i \propto \omega_1(X_1^i)$.
- Re-sample $\{W_1^i, X_1^i\}$ to obtain N equally-weighted particles $\{\frac{1}{N}, \bar{X}_1^i\}$.

For times $n \geq 2$

- Sample X_n^i from distribution $q_n(x_n|\bar{X}_{1:n-1}^i)$ and set $X_{1:n}^i \leftarrow (\bar{X}_{1:n-1}^i, X_n^i)$
- Recursively compute the weights $\alpha_n(X_{1:n}^i)$ and their normalized analogues $W_n^i \propto \alpha_n(X_{1:n}^i)$
- Re-sample $\{W_n^i, X_{1:n}^i\}$ to obtain N equally-weighted particles $\{\frac{1}{N}, \bar{X}_{1:n}^i\}$.

2.1.2 Sequential Monte Carlo Methods and Particle Filtering

In filtering approaches for a state-space model with state transition function $f(x_n|x_{n-1})$, we aim to compute a numerical approximation to the distribution $p(x_{1:n}|y_{1:n})$ sequentially in time. Particle filtering is an application of the sequential Monte Carlo algorithm described in the previous section.

Sequential Monte Carlo for Filtering

Consider the simple case of $\gamma_n(x_{1:n}) = p(x_{1:n}, y_{1:n})$ and hence yielding $\pi_n(x_{1:n}|y_{1:n}) = p(x_{1:n}|y_{1:n})$ and $Z_n = p(y_{1:n})$. For this case, we only need to select the importance distribution, $q_n(x_n|x_{1:n-1})$. It can be demonstrated that the optimal form of importance distribution in the sense of minimizing the variance in the important weights at time n and thus maximize the effective sample size would be $q_n^{opt}(x_n|x_{1:n-1}) = \pi_n(x_n|x_{1:n-1})$ [28, 29], where

$$\pi_n(x_n|x_{1:n-1}) = p(x_n|y_n, x_{n-1}) = \frac{g(y_n|x_n)\, f(x_n|x_{n-1})}{p(y_n|x_{n-1})}. \tag{2.17}$$

and the incremental importance weight is $\alpha_n(x_{1:n}) = p(y_n|x_{n-1})$. Whether it is possible to sample from this distribution or we need to approximate it, rather than making q_n dependent on previous values of y (i.e., $y_{1:n-1}$) or earlier values of x (i.e., $x_{1:n-2}$), it is sufficient to use an importance distribution adhering to the following structure:

$$q_n(x_n|x_{1:n-1}) = q(x_n|y_n, x_{n-1}) \tag{2.18}$$

Considering 2.18, 2.15 and 2.14, we obtain an incremental weight update as the following.

$$\alpha_n(x_{1:n}) = \alpha_n(x_{n-1:n}) = \frac{g(y_n|x_n)\, f(x_n|x_{n-1})}{q(x_n|y_n, x_{n-1})} \tag{2.19}$$

It is notable that in computing the weight at time n, this formulation only considers the state of the model at times n and $n-1$, and the observed data at time n.

Particle Filter and its Characteristics in Proposed Models

Within this thesis, we used the particle filtering method for performing inference in state-space models. For these models, the state of a system evolves across time, and the state x_t of the system at time t depends only on the state at time $t-1$, that is, $p(x_t|x_{t-1}, x_{t-2}, ..., x_1) = p(x_t|x_{t-1})$. The state vector x_t is assumed to be latent or unobservable. Information about x_t is obtained through noisy observations y_t, which are governed by the observation component for the probabilistic model conditional on the state variable x_t, denoted by $g(y_t|x_t)$. The general particle filter algorithm leverages the approach of importance sampling which utilizes the fact that if one wishes to sample from a target distribution $p(x)$ but is unable to do so directly, one can sample instead from an importance proposal distribution $q(x)$ which holds the key features of $p(x)$. By maintenance of a series of weights together with corresponding samples from $q(x)$, the net effect of sampling from $p(x)$ can be obtained. The algorithm can be summarized as follows and is following [30, 26]. Let N be the number of particles.

1. At time $t = 1$, for $i = 1, 2, ..., N$

   i) Sample $X_1^{(i)}$ from $q_1(x_1|y_1)$

   ii) In light of sample y_1, compute a weight for each particle $w_1^{(1)} = \frac{g(y_t|x_t)f(x_1)}{q(x_1^{(i)}|y_1)}$.

2. At time $t \geq 2$, perform a recursive update as follows:

   i) Advance the sampled state by sampling $X_t^{(i)} \sim q(x_t^{(i)}|y_t, x_{1:t-1})$. Further, record the trajectory by setting $X_{1:t}^{(i)} to (\bar{X}_{1:t-1}^{(i)}, X_t^{(i)})$.

   ii) Update the weights to reflect the probabilistic and state update models $w_t^{(i)} = w_{t-1}^{(i)} \frac{g(y_t|x_t^{(i)})\, f(x_t^{(i)}|x_{t-1}^{(i)})}{q(x_t^{(i)}|y_t, x_{t-1}^{(i)})}$, where x_t possesses the Markov property and x_t and y_t are conditionally independent.

ii) Normalize the weights: $w_t^{(i)} = \frac{w_t^{(i)}}{\sum_{i=1}^{N} w_t^{(i)}}$

3. Re-sampling step: For any time t, if the effective sample size is too small (i.e., if the variance of the weights is too high, $\frac{1}{\sum_{i=1}^{N}(w_t^{(i)})^2} < k$), re-sample $X_t^{(i)}$ and set $w_t^{(i)} = \frac{1}{N}$. Here k is a threshold value for the variation of the weights [31]. In our models, we use the simplest and most widely used proposal distribution, $q(x_t|y_t, x_{t-1}) = f(x_t|x_{t-1})$, and the weight update simplifies to $w_t^{(i)} = w_{t-1}^{(i)}\, g(y_t|x_t^{(i)})$. Here the weights are not restricted to being updated by considering later measurements but are obtained for a given observation point t by multiplying the weight associated with each particular particle at t by the likelihood of observing the measured data conditional on the state of that particle. This approach, used in our models and termed the condensation algorithm, does possess some vulnerabilities but is a well-established and highly popular sub-type of particle filtering [30, 32].

In our models, each particle at a point in time t is associated with all state variables (thus completely characterizing x_t); such a particle can be viewed as embodying a hypothesis concerning the underlying state of the model at time t. It is notable that because the suggested dynamic models include parameters such as contact rate and fraction of reported incidents which are associated with (evolving) state variables, the particle includes the state of such evolving parameters as well.

A key element of the particle filtering algorithm used here consists of the definition of the likelihood function $g(y_t|x_t^{(i)})$, which is the likelihood of observing observation y_t given the state of a given particle. For this thesis, the likelihood function was based on the negative binomial distribution, which was preferred as being a more robust distribution than the binomial distribution for the particle filtering methodology. This reflects the fact that for situations where all particles are simply a number of binomial trials (e.g., count of incident cases) smaller than the corresponding empirical datum observed, weights identically equal to zero would be triggered across all such particles, causing a singularity during weight renormalization [33]. The likelihood functions used in each model are explained in details in the corresponding chapters.

Limitations of Particle Filtering

The particle filtering algorithm is associated with several limitations. Despite sampling from the optimal importance distribution $p(x_n|y_n, x_{n-1})$, the variance of the resulting approximation depends on the variance of $p(y_n|x_{n-1})$. At a practical level, this implies a need to resample frequently and the approximation $\widehat{p}(x_{1:n}|y_{1:n})$ of the distribution $p(x_{1:n}|y_{1:n})$ may not be reliable. Particularly, for $k << n$, the distribution $\widehat{p}(x_{1:k}|y_{1:n})$ will sometimes be approximated by only a few particles (because the algorithm has resampled very frequently between times k and n). The problem associated with this approach is that it is just the variables $\{X_n^i\}$ that are sampled at time point n, while previous values along the path $\{X_{1:n-1}^i\}$ are unchanged. One can improve this algorithm by modifying the values of the path in addition to sampling the last value $\{X_n^i\}$ at time n [23]. In addition, despite being parallelizable, particle filter requires a lot of particles and is comparatively computationally expensive, although not so much so to prevent real-time updates for data arriving at rates

characteristic of epidemiological data streams. Also, Particle filtering cannot be used to sample from the value of static parameters (in contrast to PMCMC, which can be used in this way). Since particle filtering relies on the accuracy of the underlying state space model, and because the state space models examined here posit random mixing within a population, there can be limitations associated with spatial scalability. Although particle filtering can correct model states and parameters, the state space model dynamics and projections can be inaccurate at very local or large scale levels. This could contribute to significant model deviations from the underlying situation between observations, and in the course of model-based projections.

2.2 Influenza

Influenza, also colloquially known as "the flu", is a respiratory illness of varying types and pathophysiology. It is a contagious viral infectious disease spread by the coughs and sneezes of and even via touching an infected person. Since influenza is viral, it can not be treated by antibiotics, and the best way to prevent influenza is vaccination [34]. Adults are contagious 1-2 days before observing symptoms and up to 7 days after becoming ill.

There are three types of human influenza viruses: A, B and C. While virus A can cause both seasonal epidemics and emerging, new and very different influenza A infections, virus B causes only seasonal epidemics and virus C typically causes mild respiratory illness not leading to epidemics. The sub-types of Influenza A virus are defined based on two proteins on the surface of the virus: the hemaglutinin(H) and the neuraminidase(N). There are 18 different hemagglutinin sub-types and 11 different neuraminidase subtypes. There are also different strains of influenza A. Currently there are influenza A H1N1 and H3N2 viruses affecting humans.

In 2009, a new influenza A (H1N1) virus emerged, which was different from the circulating H1N1 at that time [35]. In Canada, about 3.5 million – about 10% of the population – were infected, resulting in 428 confirmed deaths [36].

2.3 Introduction to Communicable Disease Transmission Models

Many types of models can be used to forecast the progress of infectious diseases. The first mathematical model of epidemics was introduced by Bernoulli in 1766 to analyze the progress of mortality caused by smallpox in England [37]. After Bernoulli, many publications addressed epidemics modeling, but the first modern mathematical model in epidemiology was developed by Ross in 1911 [38]. He used a set of equations to describe the discrete-time dynamics of malaria. Following Ross's work, Kermack and McKendrick developed a deterministic compartmental model for epidemics by suggesting that the probability of infection of a susceptible increases with the number of its contacts with infected people. He introduced a SIR – where S, I and R represent the size of the population of Susceptible, Infected, and Recovered individuals, respectively – model by giving the rate at which susceptible people are infected as kSI. Kermack and McKendrick also

considered the rate at which infected individuals become recovered as λI and the rate at which recovered people become susceptible again to be μR, where λ and μ are constants [39, 40, 41]. Different mathematical modeling and simulation approaches can be used in epidemiology according to different perspectives in looking at the situation, and particularly when seeking to investigate different questions. Statistical methods for epidemic surveillance can be used for early identification of spatial patterns that can aid in controlling the spread of outbreaks. [42] State-space models – mathematical models within the context of dynamical systems – can be used to project the evolution of an ongoing outbreaks or pandemic or to help with forecasting potential epidemics. Based on the complexity of the problem and the precision of approximation of real-world systems, state-space models can be divided into compartmental (including System Dynamics), discrete event, and agent-based models. Compartmental models, characterized in the form of differential equations, describe the coarse-grained dynamics of outbreaks. For example, considering the evolution of an epidemic as a function of time or age can be described in a compartmental model. The population is usually divided in stocks such as susceptible (S), Exposed (E), Infected (I), Recovered (R) or even Vaccinated (V) based on their health state as an extension of Kermack and McKendricks's SIR model. Through stratification, SIR-type models can also be extended to describe demographics such as mortality, migration, age distributions, aging and gender. [43, 44, 45].

Discrete event models describe the operation of a system as a discrete sequence of events in time. These model are usually at individual level and emphasize queuing, waiting times and waiting length size in structured workflows typically limited by capacity. There can be transitions within this associated with health status change. For example for a stochastic SIR model, in case of a physical communication, an infected individual (I) infects a susceptible (S) with a probability. Several SIR stochastic models within the context of discrete event have been developed, considering age structure, environmental transmission of virus and even a combination of epidemic and economic models [46, 47].

Agent-based modeling helps with simulating interactions of agents, including individual, organizations and groups, considering the effects of such interactions on the system as a whole and vice-versa. These models are particularly powerful for capturing certain effects (e.g., heterogeneity, network patterns, history-dependence, and in representing individual-level decision making). Agent-based models have been used to assess spatiotemporal pattern of pandemics, considering population mobility, details about households, location of schools, workplaces, and hospital units [48, 49, 50, 51].

In this work, we have applied a System Dynamics approach, which constitutes a subtradition of compartmental modeling focused around feedbacks and accumulations. In chapters 3 and 4, we used a SEIR model and added a vaccinated (V) state to the model. In 5, we explicated and applied a previously published coupled contagious dynamic model, which incorporated states representing the level of fear in population (e.g., scare state). Description of models and their parameters is presented in the relevant chapter.

2.4 Social Media Data

In recent years, data extracted from search engines and online communication platforms have been employed to investigate social trends.

Many studies have examined whether data obtained from Google can be used to develop statistical forecast models. This subset of research evaluated the degree to which GFT data in combination with statistical (rather than dynamic) models can support accurate predictions. For example, Dugas et al. designed statistical forecast models to predict one week in advance from weekly counts of confirmed influenza cases over seven seasons from 2004 to 2011. They employed the Box-Jenkins method, generalized linear models, and generalized linear auto-regressive moving average (GARMA) methods to assess the contribution of external variables such as Google Flu Trends, meteorological data, and temporal information. According to their results, GARMA with a Pascal distribution integrating GFT data provided the most accurate predictions of weekly incident infuenza case counts [52]. Moreover, Pollett et al. abstracted weekly proportions of positive influenza-tests for eight countries in Latin America from FluNet for the period of January 2011 to December 2014. They also obtained concurrent weekly Google-predicted influenza activity in the same countries from GFT [53]. They determined the Pearson correlation coefficients between observed and Google-predicted influenza activity trends for each country. They further used permutation tests to examine background seasonal correlation between FluNet and GFT for each country. The investigators reported substantial discrepancies between FluNet and GFT-predicted influenza activity throughout Latin America. Also, Araz et al. performed correlation analyses to understand temporal correlations between several predictors of ILI-related emergency department (ED) visits. They used the clinical data available for Douglas County, for Omaha within that County, and for a major hospital in Omaha. They further used GFT for both Nebraska and Omaha, total ED visits in Douglas County attributable to ILI, and a ILI surveillance network data for Douglas County and Nebraska as the predictors and data for the hospital's ILI-related ED visits as the dependent variable. They used Seasonal Autoregressive Integrated Moving Average and Holt Winters methods with linear regression models to forecast ILI-related ED visits at the hospital and evaluated model performance by comparing the root mean square errors (RMSEs). Their research suggested that GFT data statistically improved the performance of predicting ILI-related ED visits in Douglas County, and that this result could be generalized to other communities [54].

Some lines of previous research have investigated the correlation between real time empirical data and data obtained from Google. For example, Thompson et al. evaluated the relationship between GFT estimates and syndromic indicators of influenza disease activity developed using ED data – total ED visits attributed to ILI and percentage of visits attributed to ILI. They found the correlation among these indicators and between these indicators and weekly counts of clinically-confirmed influenza in Manitoba. They used linear regression models and concluded that both ED and GFT data performed well as syndromic indicators of influenza activity, and were highly correlated with each other in real time [55].

In an important subset of public health research, investigators jointly leveraged influenza data drawn from both traditional and novel data sources. Santillana et al. used five different sources: near real-time hospital visit records from a medical practices management company, Google Trends time series, influenza-related Twitter microblogging posts, and FluNearYou, a participatory surveillance system to self-report ILI and GFT, to monitor and autonomously update their statistical models. They applied machine learning approaches such as Stacked linear regression, Support Vector Machine regression, and AdaBoost with Decision Trees Regression as their modeling approaches to leverage data sources and provide real-time and forecast estimates of influenza activity in the US. According to their results, the information from multiple data sources complement one another and lead to the most robust flu predictions [56]. Moreover, Sharpe et al. collected data from the CDC, GFT, HealthTweets, and Wikipedia for the 2012-2015 influenza seasons. Google, Twitter and Wikipedia were compared using Bayesian change point analysis to detect seasonal changes, or change points, in each of the data sources [57].

Unlike the studies mentioned in the previous paragraph, which used online data sources in statistical prediction models, we used both clinical data and search volume data in a System Dynamics model simulating the contagion dynamics of both disease and fear. Our work suggests that frequent reporting of clinical data and availability of social media surveillance can be used to reconstruct the state of dynamic models as new data about the real world arrives to project evolution of outbreaks at their early stages. The early projection of outbreaks would particularly be useful in the context of emerging infectious diseases with unknown or little-known parameters. So informed, the development of well-established dynamic models can offer strong guidance for health policy makers by providing them with key information about the risk and magnitude of outbreaks.

In this work, specifically in chapter 5, we used normalized daily Google search counts from Google trends and un-normalized weekly search counts from GFT for the provinces of Manitoba and Quebec for the period of the second wave of the 2009-2010 H1N1 pandemic. Specifically, we used search terms related to flu for the period of October 6^{th}, 2009 through January 18^{th}, 2010 for Manitoba and October 6^{th}, 2009 and December 19^{th}, 2010 for Quebec [58, 59].

We investigated whether the predictions of a System Dynamics model assisted by particle filtering can improve through the use of both Google search data and clinical data compared to results from using only clinical data.

Chapter 3

Particle Filtering in a SEIRV Simulation Model of H1N1 Influenza

This chapter includes text drawn from a manuscript entitled "Particle Filtering in a SEIRV Simulation Model of H1N1 Influenza" by Anahita Safarishahrbijari, Trisha Lowrence, Richard Lomotey, Juxin Liu, Cheryl Waldner and Nathaniel D Osgood, published in Proceedings of the 2015 Winter Simulation Conference [15]. The author's contributions are described in chapter 1.

Numerous studies have been conducted using simulation models to predict the epidemiological spread of H1N1 and understand intervention trade-offs. However, existing models are generally not very accurate in H1N1 model predictions, in the sense that their predictions mis-estimate what actually happens in the real world. In this chapter, we examine the impact of using particle filtering in a compartmental SEIRV (susceptible, exposed, infected, recovered and vaccinated) model which considers the impact of vaccination on the outbreak in the province of Manitoba. For the purpose of evaluating the performance of the particle filtering method, this work further compares the ability of particle filtering and traditional calibration to anticipate the evolution of the outbreak. Preliminary simulated results indicate that the particle filtering approach outperforms the calibration method in terms of the discrepancy between empirical data and model data.

3.1 Introduction

The emergence and subsequent spread of pandemic H1N1 present several challenges to public health professionals and policy makers, including as planning vaccination schedules and clinical resource constraints. Epidemiological time series by themselves fail to offer much assistance for these tasks. This reflects the fact that they are not only extremely noisy, but – more importantly – fail to provide insight into counterfactuals, such as how an outbreak will play out in the absence of further intervention. Dynamic modeling for outbreak analysis plays a significant role in the planning of the public health reaction to infectious disease outbreaks. Statistical and mathematical models aid in understanding the role of social distancing measures such as school closure and in evaluating the value of the vaccination programs and establishing priorities to target populations for vaccination, prioritizing data collection, addressing application of antiviral therapy and in easing collaboration between policy-makers and analysts. One of the most essential planning tools is

to anticipate outbreak progression in light of empirical time series data. While models offer strong benefits, there is the inevitable need to omit or approximate some processes and factors. Inevitably – and particularly for fast-breaking outbreaks of emerging pathogens – this leads to simplification and misestimation of the dynamic models. These shortcomings – together with stochastic transitions associated with human and economic behavior – inevitably lead the model forecasts to diverge from empirical data [60, 61, 62].

This quandary has attracted many and diverse studies from the research community. Seasonal influenza viruses, including H1N1, cause 3 to 5 million cases resulting severe illness each year with between $250,000$ and $500,000$ deaths (according to the WHO reports). Each year the vaccine is modified to include currently circulating strains thought to present the greatest risk to public health. Antiviral drugs can also be used to limit the severity of complications and risk of death. However, the virus is constantly changing and is an ongoing source of uncertainty in public health. Simulation modeling is an important tool in predicting the behavior of the virus and planning intervention strategies. Hence, Manchanda et al. proposed an immune system mathematical modeling methodology that focuses on the explanation of variations in influenza kinetics caused by virus strains in mice. Using ordinary differential equations, the authors model considers several variables and parameters to conduct sensitivity and identifiability analysis. The model is able to predict the outcome of infection, and simulate and interpret the cause of outcomes. However, the work offers little contribution at the epidemiological level, such as with regards to the impact of vaccination, and the spread of infection with exposure to the virus and so on [63]. Furthermore, the need to understand influenza H1N1s transmission motivated by Chao et. al to model a colony of agents representing virtual humans termed the "artificial community". The authors defined connections between the agents at three ordinal levels, such that the agents can be described as having strong ties, ordinary ties, and weak ties with each other. By adopting the SEIR model, the authors seek to pay attention to critical flow constraints, such as the natural history characterized by a latent period and treatment-receiving period. The authors, however, did not compare the model results with any empirical data and the sensitivity analysis is not sufficiently detailed to guarantee reproducibility of the model outcome [64]. Moreover, the global spread of the H1N1 virus caught the attention of Shubin et. al who studied the impact of the outbreak of H1N1 in Finland. The work considered prior and posterior distributions factors such as severity. The model predictions show that the severity of the outbreak in the second season is almost half of the first epidemic. Although the authors in this case looked into the effect of vaccination, their primary model is the susceptible, infective, and recovered (SIR) model, rather than focusing on secondary infection risks [65]. Also, Pongsumpun and Tang have seen the need to study the impact of H1N1 virus transmission using the SEIQR (susceptible, exposed, infected, quarantine, and recovered) model. The proposed model also took into account the incidents of death in the population and the impact of repetitive contacts. The work showed that when the repetitive contacts increase, the number of susceptible people decreases. The authors, however, did not consider vaccination and its impact on the population [66]. Particularly for diseases with nonspecific symptoms, several factors obstruct the tracing and prediction of emerging epidemics: the disconnect between transparent epidemic

dynamics and what is discernible from noisy and incomplete surveillance data and the imperfectly observed system. Also, behavior changes compound this through altering both true dynamics and reporting patterns. Birrell et. al seek to unravel these effects to resolve the hidden dynamics of the 2009 influenza A/H1N1 pandemic in London. To disclose significant changes in contact patterns and health-seeking behavior, they embed an age-structured model into a Bayesian synthesis of multiple evidence sources. As the result, this approach is capable of real-time learning about model parameters during the epidemic progress, and provides a sequence of nested projections to reflect the epidemic [67]. Conway et. al in their model, represented the Greater Vancouver Regional District and surrounding residential areas with a population of 2 million and investigated the effect of timing of different vaccination strategies in estimating the transmission of the pandemic H1N1 [68]. With the development of a compartmental susceptible-infected-recovered (SIR)-type epidemic model, different distribution strategies were initiated. For each vaccination strategy, the effect of varying the vaccination strategy under various baseline transmission parameter values were tested. It was found that the model output was consistent with provincial surveillance data and that vaccine efficacy had an important impact on depleting the size of the susceptible population and consequently reducing the outbreak size. Their work could further be improved by considering the addition of a vaccination stock in their compartmental model. Tuite et. al developed a compartmental model of influenza transmission in the Canadian population and sought to obtain the optimal strategy for prioritization of vaccine distribution in order to minimize morbidity and mortality rates [69].

To yield a more accurate consensus estimate [33] used sequential Monte Carlo methods in the form of particle filtering to combine intuitions from dynamic models containing systematic errors and noisy empirical data, and to aid in parameter estimation. To demonstrate the advantages from particle filtering, parameters and variables in an aggregate systematically biased SEIR model, they compared particle filtering against synthetic ground truth produced by an agent-based model. In this chapter, in addition to introducing a model of H1N1 in which vaccinated percentage has been considered, we use clinical data from the Midwestern Canadian province of Manitoba for H1N1 pandemic 2009 to evaluate the application of particle filtering approach, using a temporally-based cross-validation approach. Specifically, we compare the performance of the particle filter with a traditional calibration method in anticipating the future evolution of counts of reported cases.

3.2 Motivation for Calibration and Particle Filtering

For emergent conditions such as H1N1, there is an acute need to plan and mathematical modeling through outbreak analysis plays a significant role in the planning. The corresponding parameter values, the current situation, and even the natural history of the infection, are frequently unknown or poorly known in the early stages of an emergent condition. In this context, a model that supports a wide range of interpretations is particularly valuable. In our model we sought to obtain empirical estimates for various parameters,

for example, contacts per week multiplied by probability of infection transmission given exposure ($c\beta$), mean latent time (τ), fraction reported incidence (f), fraction initially susceptible, fraction initially exposed, fraction initially infective and fraction initially recovered by calibrating the model to the empirical data obtained by Manitoba Health, Healthy Living and Seniors. To predict shorter-term projection of the existing conditions or intervention scenarios, well-calibrated dynamic models are frequently accurate, but for longer term projections they tend to diverge from empirical patterns and also, generally, there exist a shortage of reliable and automated means of keeping current with the latest in empirical data. Particle Filtering was introduced as a method that builds on well-studied statistical techniques to join together dynamic models and empirical data, while decreasing the inherent weakness of both. While calibration processes often require much time and typically entail manual oversight and intervention, the particle filtering process was executed in considerable less time and proved to be more accurate in model predictions. Particle filtering however, has been applied to comparatively few previous applications in the public health area, specifically in predicting infectious diseases.

3.3 Scheme of the Model

We present the formulation of a compartmental model, which includes Susceptible, Exposed, Infectious, Recovered and Vaccinated stocks (SEIRV). We present here a comparison between the applications of a particle filter and a calibration method for a System Dynamics transmission model for H1N1 influenza, and then evaluate the performance of that particle filter compared to that of traditional calibration when operating using empirical data from Manitoba Health, Healthy Living and Seniors.

3.3.1 Empirical Data

The empirical data obtained from Manitoba Health, Healthy Living and Seniors indicated weekly confirmed cases of pH1N1 and vaccine delivery rates for the period of October 6^{th}, 2009- January 18^{th}, 2010 [58].

3.3.2 Dynamic Model

We describe here our dynamic model to be used with the particle filter and calibration. Figure 1 demonstrates all stocks, flows and parameters.

The aggregate compartmental state equations for the model are given as follows:

$$\dot{S} = -c\beta \frac{I}{S+E+I+R+V} S - abS \tag{3.1}$$

$$\dot{E} = c\beta \frac{I}{S+E+I+R+V} S + c\beta \frac{I}{S+E+I+R+V} V - \frac{E}{\tau} \tag{3.2}$$

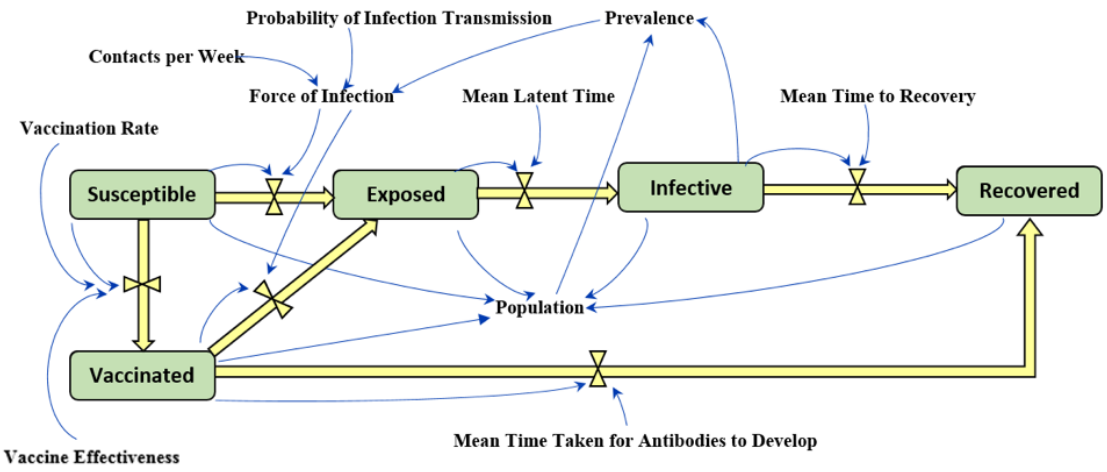


Figure 3.1: System dynamics model

$$\dot{I} = \frac{E}{\tau} - \frac{I}{\mu} \tag{3.3}$$

$$\dot{R} = \frac{I}{\mu} + \frac{V}{v_a} \tag{3.4}$$

$$\dot{V} = abS - \frac{V}{v_a} - c\beta \frac{I}{S+E+I+R+V} V \tag{3.5}$$

In comparison with the previous work, we have added a vaccinated stock to the model. We added this stock to capture the impact of vaccination, which is known to strongly influence the dynamics of many infectious diseases; such a stock is routinely incorporated into many contemporary compartmental models of influenza transmission [70, 71, 72]. We have defined the input of this stock as the multiplication of Susceptible, vaccine effectiveness parameter (b) and the per-capita vaccination rate (a), where the vaccine effectiveness parameter refers to the ability of the vaccine to bring about the intended beneficial effects on vaccinated individuals and the vaccination rate (a) is defined to be the fraction of newly vaccinated people taken over the entire population per unit time (i.e., the vaccination rate parameter is a variable of time). For this parameter (a), we made use of the empirical data obtained from Manitoba province. The outputs of the "Vaccinated" stock are the number of people vaccinated divided by mean time taken for antibodies to develop (ν_a) which enter "Recovered" stock and number of people vaccinated multiplied by force of infection which enter "Exposed" stock. The model runs for 15 weeks and the primary model output examined here are reported infectives which is a multiplication of the size of "Infective" stock and the fraction of reported incidence. The compartmental parameters are specified in Table 1. In this work, we did not conduct sensitivity analysis to examine the sensitivity of our model to static parameters. However, some of the previous works from which we extracted the value of static parameters have considered sensitivity analysis [69].

It is notable that the model includes a stochastic process associated with Contacts per Week and Fraction of Reported Incidents. In the particle filtering model, these parameters are initially uniformly distributed be-

Variable name	Notation	Value	Source	Units
Probability of infection transmission given exposure	β	0.06	Expert opinion	Unit
Mean time to recovery	μ	1	[69]	Week
Vaccine effectiveness	b	0.9	[68]	Unit
Mean time taken for antibodies to develop	ν_a	2	Expert opinion	Week
Total population size	N	1214403	[73]	Person
Mean latent time	τ	Uniformly distributed (0.4, 0.8)	[69]	Week
Vaccination rate	a	Extracted from empirical vaccinated percentage		1/Week

Table 3.1: Table showing parameters

tween maximum and minimum parameter values, however, these parameters are calibrated in the calibration model.

3.3.3 Particle Filter Characteristics in Proposed Model

In our model, each particle at a point in time t is associated with all state variables (S, E, I, R, V). Moreover, the suggested dynamic model includes parameters such as contact rate and fraction of reported incidents which are associated with state variables evolving over time. To use particle filtering to adapt to values of such parameters, we further associate each particle with a value for the parameters c and f. Each particle is thus associated with a vector $[S, E, I, R, V, c, f]$. The results presented in this chapter are based on model runs employing 10000 particles, which was judged to be enough because it appears to yield a well-behaved distribution in most cases, and is clearly enough according to the judgment of statistician colleagues.

In estimating the likelihood formulation for observing y_t individuals per week given an estimated weekly count of i_t becoming cases, we employ the negative binomial distribution $p(y_t|i_t)$, where $p = \frac{i_t}{i_t+r}$, r is a dispersion parameter and $i_t = \frac{E}{\tau}$.

3.3.4 Comparison between Particle Filter and Calibration

In this contribution, we investigate the degree to which the model is efficient in robust estimation and prediction of model states with and without particle filtering. Since the knowledge of the situation is imperfect there is frequently a need to estimate model parameters based on available empirical data regarding phenomena that are emergent within the model.

In this section, we investigate the capabilities of particle filter and calibration methodologies in mitigating the effects of aggregation and prognosticating model states in the context of data from a real-world outbreak.

We defined a variable, "check time", which indicates the time t up to which the particles weights are updated based on observation, where $0 \leq t \leq T^*$. After $t = T^*$, the particle filtering ceases, in that particle weights is no longer updated using the empirical data, and no further re-sampling occurs. In this experiment, we utilized the parameter "fraction reported incidence" to account for the fact that reported counts only included a subset of the persons infected. For uncertain parameters such as "probability of transmission given exposure" and "mean latent time", we define a function that takes a range of uniformly distributed values from minimum to a maximum. For the calibration method, we ran the model for $20,000$ iterations (For more iterations, the objective function did not appear to be substantially decreasing). In order to ensure robustness in the context of the stochastic evolution of model parameters c and f, we further ran 10 realizations (replications) per iteration. For this optimization experiment, the objective function involved minimizing the average of square of difference between linearly interpolated datasets which are model data and empirical data. The integration range is the intersection of argument ranges of datasets. In calibration, we considered the empirical data up to time $t = T^*$ and after calibrating the parameters, we were able to obtain simulation results for the entire time range (including time points $t > T^*$) for the model based on those parameters. Specifically, we assess particle filter and calibration by comparing their estimates of reported new infections against corresponding quantities from empirical data.

3.4 Results

According to Figures 3.2,3.3, 3.4 and 3.5 we have demonstrated the performance of particle filter and calibration for $T^* = 14$ and $T^* = 6$. In Figures 3.3 and 3.5, we have plotted all sampled particles. We have defined the discrepancy as a function which focuses on the average per-time-unit error during the time $t > T^*$. In this case, we are only considering how accurate it is in predicting data about which it has not been told (a form of cross-validation). Besides, we have divided the discrepancy over the time period $t > T^*$ by the length of that time period to have comparability of results. The function below calculates the value for discrepancy found for the particle filtering process. This function was defined as $\frac{\sum_{i=T^*+1}^{T_f}(x_i^M - x_i^E)^2}{T_f - T^*}$ for the calibration process, where is the end time, is the data extracted from the model at time i and is the respective empirical data. For the particle filtering methodology, by sampling n particles with larger weights, the discrepancy value is obtained via below formula while is data pertaining sampled particle j at time i. In figures 3.4 and 3.5, the red data items (prior to and including T^*) were incorporated for both particle filtering and calibration, and the black data items (after T^*) were incorporated into neither the calibration nor the particle filtering.

Figure 3.6 presents a histogram showing the discrepancies from Particle Filter and Calibration for $T^* = 6, 7, 8, 9$ and 10. For all values of $t = T^*$, the discrepancy from particle filter is less than the discrepancy from calibration. However, for both the particle filter and calibration methods, the discrepancy increases as the value of decreases. Put another way, as the window of empirical data considered by both the particle filter and calibration methods grow in size, the accuracy of those approaches in predicting the entire time series

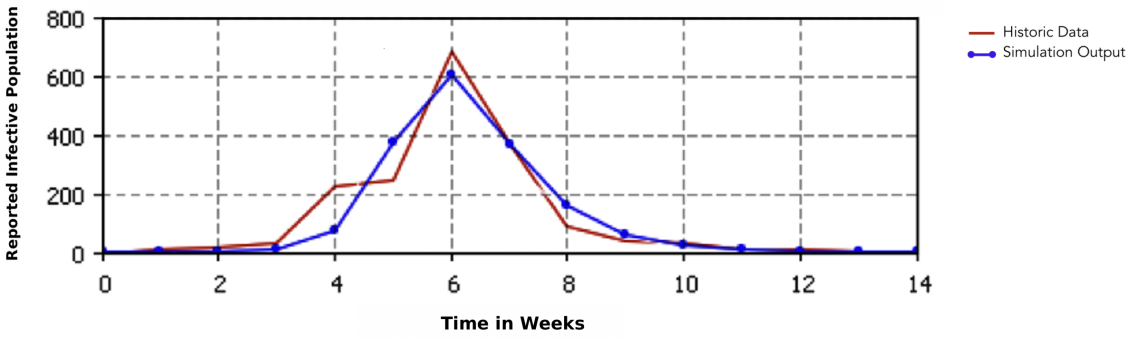


Figure 3.2: Calibration results for $20,000$ iterations and for $T^* = 14$.

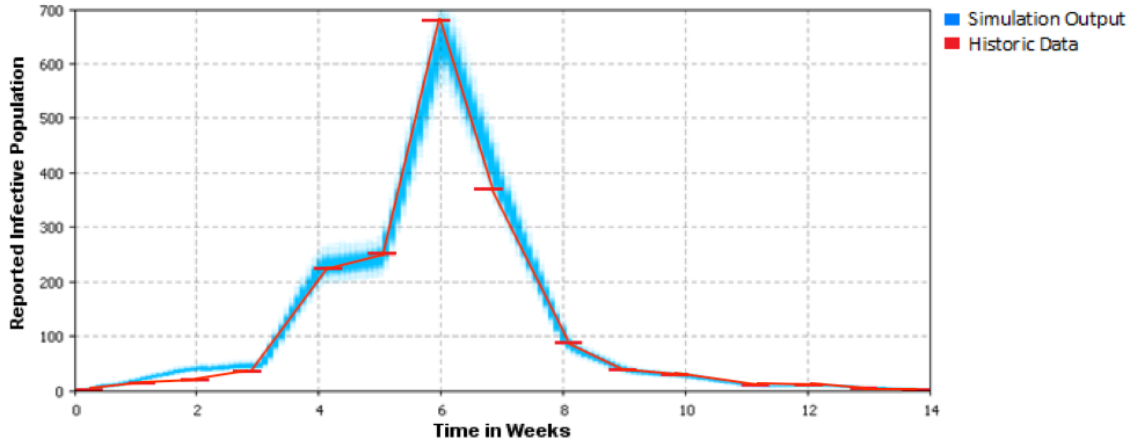


Figure 3.3: Particle filtering results for $T^* = 14$.

rises.

3.5 Conclusion

In this work, we explored the performance of particle filtering and calibration in a System Dynamics model against empirical data from an H1N1 outbreak. The particle filtering was put forward to readily read data and further correct the model output using historic data. In addition to particle filtering contributing to the estimation of model states, particle filtering also aided in estimating the model parameters. It was well adapted to evolution in the effective value of dynamic parameters that would otherwise be treated as static. For example, by applying a distribution to the Contacts per week parameter, a more accurate estimate was achieved during the model simulation.

The work examines the SEIRV (susceptible, exposed, infected, recovered and vaccinated) model and provides an extension to many existing SEIR models to capture the pronounced impact of vaccination on the dynamic of infectives. Moreover, the proposed model is similar in structure to the models that do consider the effects of vaccination [70, 72, 71]. The discrepancy for the particle filtering was found to be less than the

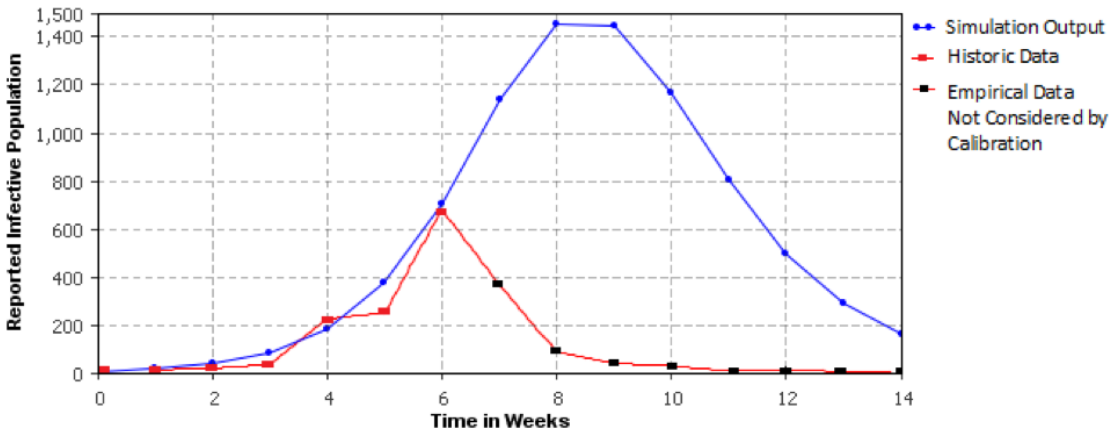


Figure 3.4: Calibration results for $20,000$ iterations and for $T^* = 6$.

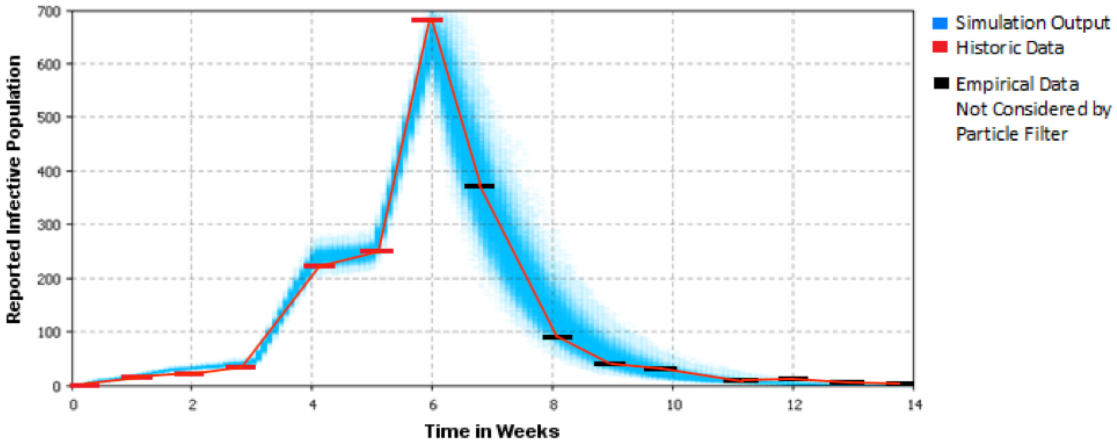


Figure 3.5: Particle filtering results for $T^* = 6$.

discrepancy associated with the calibration method when compared to existing empirical data. In addition to this phenomenon being true for different time scenarios, the particle filtering methodology was observed to better predict the model outcome when using observable data. The calibrated parameters and their values for check time 14 are specified in Appendix A.

The main contributions of our work include the proposal of the SEIRV model, the comparison of particle filtering and calibration methodologies and the prediction of future outcome based on current empirical data. Many priorities remain for future work. It will be important to incorporate heterogeneity within our model by observing various age groups and also anti-viral treatments. We further hope to investigate the impact of relaxing the constraints of the condensation algorithm on model accuracy.

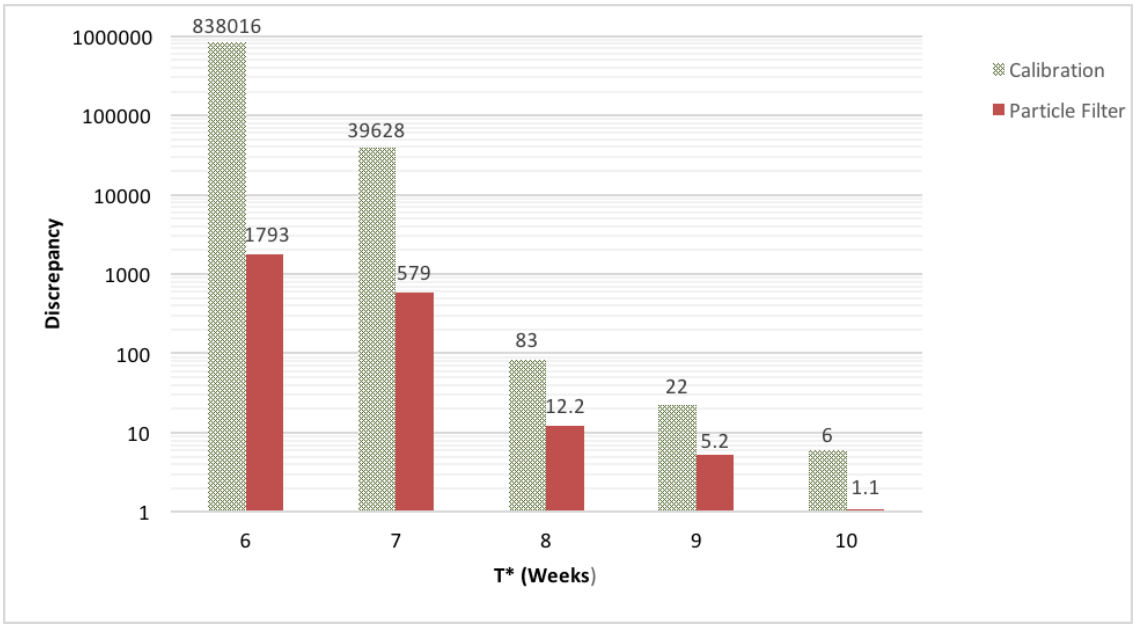


Figure 3.6: Logarithmic Graph Showing Discrepancy for Calibration and Particle Filtering vs T^*.

Chapter 4

Predictive Accuracy of Particle Filtering in Dynamic Models Supporting Outbreak Projections

The text of this chapter is largely drawn from a manuscript entitled "Predictive Accuracy of Particle Filtering in Dynamic Models Supporting Outbreak Projections" by Anahita Safarishahrbijari, Aydin Teyhouee, Juxin Liu, Cheryl Waldner and Nathaniel D Osgood, published in BioMed Central Infectious Diseases Journal [16]. Author's contributions are described in chapter 1.

While a new generation of computational statistics algorithms and availability of data streams raises the potential for recurrently regrounding dynamic models with incoming observations, the effectiveness of such arrangements can be highly subject to specifics of the configuration (e.g., frequency of sampling and representation of behaviour change), and there has been little attempt to identify effective configurations.

Combining dynamic models with particle filtering, we explored a solution focusing on creating quickly formulated models regrounded automatically and recurrently as new data becomes available. Given a latent underlying case count, we assumed that observed incident case counts followed a negative binomial distribution. In accordance with the condensation algorithm, each such observation led to updating of particle weights. We evaluated the effectiveness of various particle filtering configurations against each other and against an approach without particle filtering according to the accuracy of the model in predicting future prevalence, given data to a certain point and a norm-based discrepancy metric. We examined the effectiveness of particle filtering under varying times between observations, negative binomial dispersion parameters, and rates with which the contact rate could evolve.

We observed that more frequent observations of empirical data yielded super-linearly improved accuracy in model predictions. We further found that for the data studied here, the most favourable assumptions to make regarding the parameters associated with the negative binomial distribution and changes in contact rate were robust across observation frequency and the observation point in the outbreak.

Combining dynamic models with particle filtering can perform well in projecting future evolution of an outbreak. Most importantly, the remarkable improvements in predictive accuracy resulting from more frequent sampling suggest that investments to achieve efficient reporting mechanisms may be more than paid back by improved planning capacity. The robustness of the results on particle filter configuration in this case study suggests that it may be possible to formulate effective standard guidelines and regularized

approaches for such techniques in particular epidemiological contexts. Most importantly, the work tentatively suggests potential for health decision makers to secure strong guidance when anticipating outbreak evolution for emerging infectious diseases by combining even very rough models with particle filtering method.

4.1 Introduction

According to World Health Organization (WHO), seasonal influenza viruses cause 3 to 5 million cases of severe illness, with about 250,000 to 500,000 deaths each year, with emerging-strains sometimes significantly increasing this burden. An important example of this was high-burden emergence of pandemic influenza A (H1N1) during the 2009-2010 influenza season. Vaccination and intervention strategies such as school closures for early mitigation of pandemic influenza spread may reduce severe complications and deaths [74]. Key concerns during an outbreak include staffing requirements for implementation of a pandemic response, clinical resource constraints [75], managing individuals expectations and behaviors, which often relate their risk perception [76], and mobilization of health resources [77]. Rapid or ideally real-time reporting of surveillance data provide a clear picture of what has happened, but fail to provide clarity on how the epidemic will evolve. Simulation modeling can be an important tool to anticipate what is most likely to happen in the near future, to ask questions concerning interventions and identify desirable policies.

Mathematical models describing the dynamic of epidemiological infections can be useful for projection purposes [63, 65, 64, 66, 78], but often the fundamental challenge in leveraging models for emerging communicable diseases and strains is that there is limited epidemiological knowledge regarding the natural history of infection and the values needed for model parameters [79]. While a well-formulated model can be useful for planning, often the knowledge needed to build that model is lacking at the time when it is the most urgently needed. In this situation, a precisely calibrated and highly tuned model can play an important role, but is often infeasible to build in a time compatible with planning needs. Even for models of endemic infections such as seasonal influenza in which refined estimates of parameter values and understanding of natural history are available, model predictions secured early in an outbreak inevitably diverge from observations [80, 61, 62]. This reflects the fact that all models are simplifications (and thus inevitably omit factors). In addition, stochastics are involved in real-world systems, which depend on unpredictable or hard-to-predict factors such as shifting vaccine attitudes and risk perception that can impact contact patterns [81, 82, 83], as well as the vagaries of transmission and the health system response. This divergence is made more likely by the fact that many such factors including changes in human contact patterns are believed to play a substantial role in disease transmissions [82, 83, 14] and are often not captured in models. Statistical filtering and estimation methods for dynamic models, such as Sequential Monte Carlo (SMC) and Markov Chain Monte Carlo (MCMC) methods, provide an attractive tool to not only create model predictions based on where we are right now, but to use empirical observations from continuing surveillance to reground that model on an ongoing basis [61, 84, 67, 85, 86, 31].

Among estimation algorithms, Kalman filtering has long been used for creating estimates based on consensus of empirical data and model predictions using Maximum Likelihood Estimation (MLE) [87, 88, 89, 90, 91]. However, it is hampered by stiff distributional assumptions regarding process and measurement error. The Kalman filters reliance on gaussian assumption and MLE further limits its accuracy, particularly in the context of non-linear systems. The reliance of Kalman filtering on linearization of nonlinear distributions both raises strong challenges for accurate state estimation in the context of infrequent observations and limits the applicability of such models to an important but circumscribed subset of transmission models for which linearization is possible [33].

As a SMC, particle filtering offers similar overall types of benefits as Kalman filtering while relaxing such constraints. Particle filtering deals with less restrictive assumptions concerning the noise and process model, and samples from a joint distribution of state trajectories rather than conforming to a MLE approach. This method [92] samples from the posterior distribution of model state trajectories, combining empirical data and model dynamics. Key mechanics of particle filtering are drawn from the importance sampling method. With importance sampling, we sample from a particular distribution from which sampling is difficult (target distribution) in a two-phased approach in which we first draw weighted samples from an alternative distribution (importance proposal distribution) that retains the major properties of the target distribution, and then sample from those weighted samples with a probability proportional to their weight. Similar to importance sampling, in a particle filter, sampling is performed from the particles based on their weights. When new empirical data arrive, the filter further updates the weights to reflect the fitness of particles to these observations (as quantified by the ratio of the target distribution to the proposal distribution). The method that we use here to update the weight of particles is based on the condensation algorithm [30, 32], in which the weight of each particle is updated at each observation time by multiplying it by the likelihood of observing the observed data given the state of that particle at that point in time. Following [93], and our previous success in applying this approach for previous transmission models [33, 15, 94], we assume that the likelihood distribution is characterized by a negative binomial distribution:

$$P(y_t|i_t) = \binom{y_t + r - 1}{y_t} p^{y_t}(1-p)^r \tag{4.1}$$

where $p = \frac{i_t}{i_t+r}$, r is a dispersion parameter, y_t is the model observation (number of incident cases reported for time t), and i_t is the incident case count recorded over a scenario-specific interval.

The objective of this study was to apply particle filtering to predictive models of emerging communicable diseases, which are often built in the presence of limited information about underlying parameters. In light of the growing availability of epidemiological data streams, we seek here to investigate the impact on model accuracy of varying the inter-observation interval, studying the tradeoff between pursuing more frequent but more noisy sampling and less frequent but more stable estimates. We further examine the robustness of the particle filter to different assumptions concerning behaviour change and assumptions regarding observational error.

4.2 Methods

We formulated a transmission model for an influenza-like disease in a classic compartmental fashion and used it with the SMC method of particle filtering. The dynamic model includes the same states as the model presented in 3. Given that the R state includes not just those who are recovered, but also those who are now fully protected via vaccination, we called them "Removed" rather than "Recovered" in the model presented in this chapter. Thus the model includes Susceptible (S), Exposed (E), Infective (I), Removed (R), and Vaccinated (V) stocks (Fig. 4.1). It bears noting that the Vaccinated state represents a transient set of individuals who have received the vaccine but have not yet attained immunity; upon achieving immunity, such individuals transition to the Removed state. The aggregate compartmental state equations describing the model stocks are the same as compartmental state equations described in 3.

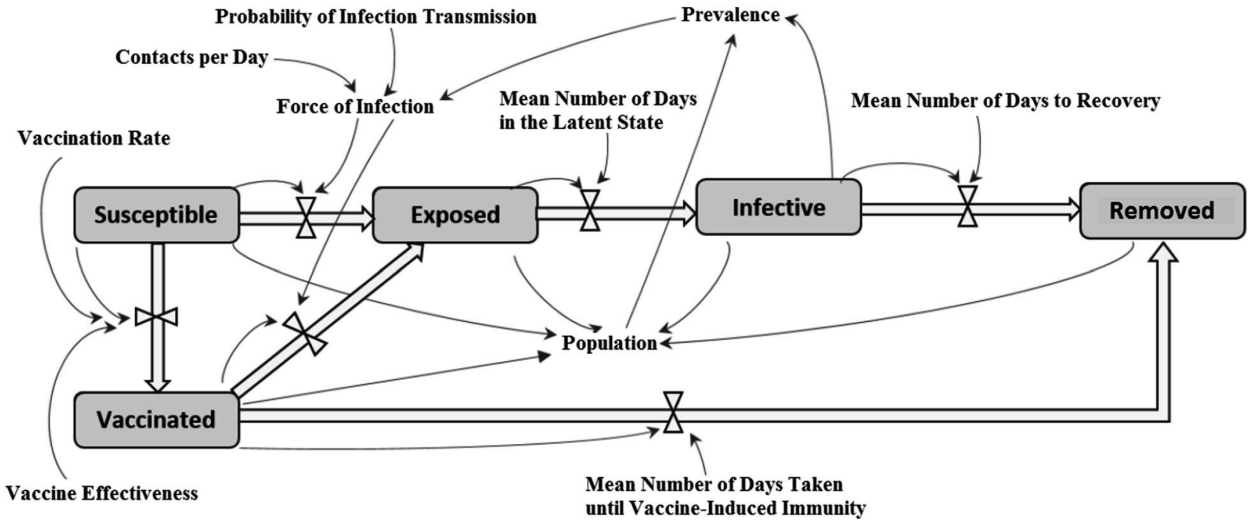


Figure 4.1: Transmission model

In our model, each particle is associated with a complete copy of model state, including the state of two evolving parameters of the model: contact rate (c) and fraction of reported incidents (f) fI accounts for fractional actual reporting, which are associated with evolving state variables whose values can be sampled by particle filtering. Thus, each particle is associated with a vector of model states $[S, E, I, R, V, c, f]$. Following [33, 93], a negative binomial distribution is assumed to link the observed incident case count for a specified time period to the underlying count of individuals emerging from latency in the model. We preferred a negative binomial distribution over the binomial distribution due to the robustness of negative binomial distribution for the particle filtering methodology [33]. It particularly avoids the risk of a situation in which all particles are associated with zero weights, causing a singularity during weight re-normalization. As the model runs and learns from the empirical data over time, the particles associated with the stocks that exhibit the greatest fitness – in terms of explaining the observed data – survive, are replicated and henceforth evolve independently.

This work builds on previous work by Osgood and Liu evaluating particle filtering against ground truth from an agent-based model [33] and our previous work evaluating particle filtering in terms of its ability to predict future reported real world prevalence in the absence of a ground truth model [15]. In this work, we seek to examine the impact on model predictive accuracy of the inter-observation interval of empirical data, and the robustness of ranges of plausible values for the dispersion parameter and the parameters associated with the random walk associated with c and f. Such variations are examined for a number of different observation points during the outbreak.

The prediction of particle filtering was evaluated against empirical data publicly available from Manitoba Health, Healthy Living and Seniors, which included daily confirmed cases of pandemic H1N1 for the period of October 6$^{\text{th}}$, 2009 through January 4$^{\text{th}}$, 2010. To judge the deviation of particle filtering prediction from observations, we defined the discrepancy metric as the expected value of the L^2 norm of the difference between sampled particles. We sampled n particles (n=1000). Given that several dozen samples is often viewed as the minimum number to reliably estimate a sample mean, 1000 was judged to be well sufficient to capture a narrow distribution in the mean discrepancy. The discrepancy value was obtained from the collection of such sampled particles using the following equation:

$$discrepancy = \frac{\sum_{i=T^*+1}^{T_f} (\frac{\sum_{j=1}^{n} (x_{ij}^P - x_i^E)^2}{n})}{T_f - T^*} \tag{4.2}$$

where x_{ij}^P is the expected sample associated with sampled particle j at observation i, x_i^E is the respective empirical data at observation i. T_f is the end time being set equal to 91 and T^* indicated the time t up to which the particles weights were updated based on observation, where $0 \leq t \leq T^*$. In other words, the data before and equal to this time was taken into account for particle filtering based on the observed data; after time T^*, particle weights were no longer updated using the empirical data, no further re-sampling occurred, and we evaluated how well particle filtering predicted the remaining empirical data.

4.2.1 Parameter values

Initial values:

We set the initial value of Susceptible and Removed stocks based on sampling from a truncated normal distribution instead of considering the initial values as a static number. Figure 4.2 gives curves for Susceptible and Removed stocks. Detailed information about initial values is provided in Appendix B.

Contacts per unit time (c)**:**

In this work, particle filtering contributes to the estimation of this dynamic parameter over time through particle selection. This parameter – which carries a non-negative value – is log transformed, with the logarithm evolving stochastically according to an (unbounded) zero-mean Gaussian random walk with standard deviation (γ). This is characterized according to the notations of Stratonovich stochastic differential equa-

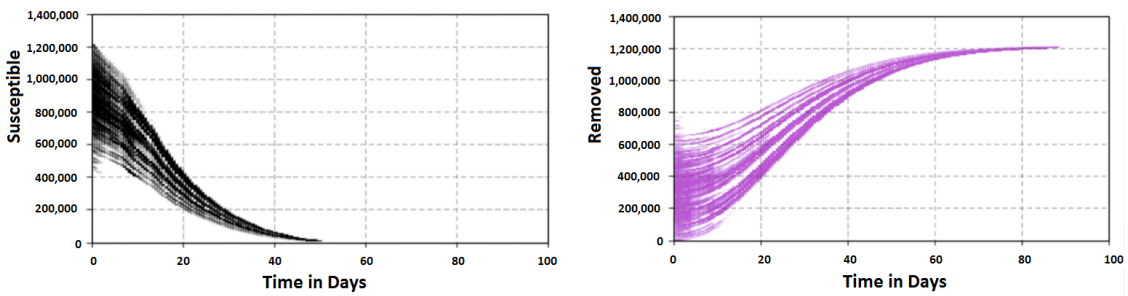


Figure 4.2: Progress of susceptible and removed stocks over time, initializing with a range of values.

tions in 4.3. The term dW_t at the right side of equation 4.3 is a standard Wiener process following a normal distribution with mean of 0 and variance of 1 [95], which leads $\frac{d(ln(c))}{dt}$ in any infinitesmal interval to follow an independent draw from a normal distribution with mean of 0 and variance of γ^2. High values of γ allow the contact rate to evolve more quickly, while low values of γ would be associated with the assumptions of comparatively slow changes in contact rate. In this work, we examined model behavior over a wide range of γ to identify appropriate ranges for this important parameter. The initial value of the stock associated with the logarithm of c is set to the logarithm of the uniform distribution on the interval between minimum contacts per day and maximum contacts per day which have been considered as 1 and 300, respectively (4.4).

$$d(\log\, c) = \gamma\, dW_t \tag{4.3}$$

$$(\log\, c)|_{t=t_0} = \ln\, (U(c_{min}, c_{max})) \tag{4.4}$$

Fraction reported incidence:

The other stochastic parameter included here represents the fraction of reported incidents (f). The fraction of people who present for care (and are reported to public health authorities) when emerging from the latent state is an uncertain value. It is also likely to evolve according to risk perception on the part of the population and provider perception of the importance of reporting. As for c, we considered (a transformed value of) this parameter as a state of the model and thus associated each particle with a value for this parameter. We considered the transformed version of this parameter as evolving according to a zero-mean gaussian random walk with a standard deviation given by a parameter (η). Since f is a fraction varying between 0 and 1, the (unbounded) random walk was conducted on the logit of this parameter (4.5, again shown according to Stratonovich calculus notation) which was itself the aspect of model state and the initial value of this state is set to the logit of fraction reported incidence sampled from a continuous uniform distribution on the interval between 0 and 1 (4.6). As for c, we have examined stochastics for this parameter to induce variability among particle trajectories, both to let these quantities evolve during outbreaks, and to provide requisite variability in particles to allow for the existence of considerably "fitter" particles.

$$d(logit(f)) = d(ln(\frac{f}{1-f})) = \eta\, dW_t \tag{4.5}$$

$$(\text{logit } f)|_{t=t_0} = \text{logit } (U(0,1)) \tag{4.6}$$

The other parameters of the model are considered as static and are shown in Table 1.

Variable name	Notation	Value	Source	Units
Probability of infection transmission given exposure	β	0.06	Expert opinion	Unit
Mean time to recovery	μ	7	[69]	Day
Vaccine effectiveness	b	0.9	[68]	Unit
Mean time taken for antibodies to develop	v_a	14	Expert opinion	Day
Total population size	N	1214403	[73]	Person
Mean latent time	τ	Uniformly distributed (2, 4)	[69]	Day
Vaccination rate	a	Extracted from empirical vaccinated percentage		1/Day

Table 4.1: Table showing parameters

4.3 Scenarios

We formulated a set of scenarios to explore how the error associated with the particle filtered model predictions would respond to changes in the total period for which empirical data was available to the model (T^*), the frequency of and degree of aggregation associated with empirical data observations supplied to the model, contact rate volatility parameter (γ) and dispersion parameter (r).

Adequacy of empirical data (T^*)

We examined the impact of the particle filter on model predictive accuracy at various time points during the progression of an outbreak. This simulated a situation in which a health authority is partway through an outbreak and can only take into account data observed until this point when making predictions for coming weeks. Specifically, in each scenario, particle filtering used data from the start of the outbreak up to and equal to a time T^*; the accuracy of particle filter was then evaluated in predicting the data for all times after T^*. We considered T^* equal to 35, 42, 49, and 56, equivalent to predictions made at 5, 6, 7 and 8 weeks into the outbreak.

Inter-observation aggregation interval / Frequency of data observations

Clinically observed data commonly contains noise, for example, due to errors introduced by measurement tools and random errors introduced by processing or by clinical experts when the data is gathered. As a result, there is a trade-off between employing more frequently observed (but less aggregate) data and reducing the noise associated with each data point via observations that are aggregated over longer periods of time. Employing more frequent sampling – by using shorter time intervals between observations – yields more numerous data points, but each such datum will typically exhibit greater proportional variability. By contrast, employing less frequent sampling during training (thereby aggregating data over a longer period between observations) leads to fewer but proportionately less noisy individual data points. To examine the impact of the frequency of data observations on filtered model accuracy, we investigated the impact of aggregating empirical data used in particle filtering observations at three levels. First, we considered daily data i.e., the number of people clinically confirmed as infected per day to update the particles weights during particle filtering. Because the original data source specifies data on a daily basis, no further aggregation was required for this case. Second, data was aggregated over three days for the purposes of particle filtering. In the third and final alternative setting, the particle filtering used data aggregated on a weekly basis. It should be emphasized that such aggregation affected only the model observations, and not the calculation of discrepancies between model results and empirical data.

Random walk standard deviation parameter (γ)

To explore the changes in contact per unit time patterns during an outbreak, and its effect on the spread of infection, we performed particle filtering using alternative values for the contact rate variability parameter (γ). In order to explore a broad dynamic range, we examined parameter values at successive powers of two of the smallest value: 0.125, 0.25, 0.5, 1, 2, 4 and 8.

Dispersion parameter (r)

The ability of particle filtering to project incident case counts is sensitive to the dispersion parameter value associated with the negative binomial distribution. Increasing the dispersion parameter makes the negative binomial distribution tighter, while retaining the same mean value [96]. We compared the discrepancy resulting from running the model with alternative values of the dispersion parameter to developing an understanding as to how this parameter affects predictive accuracy. To ensure the comparability of scenarios when running the models using three-day and weekly observations, we considered the r parameter respectively three times and seven times as great as the r that we used when observing daily data. This linear scaling of the dispersion parameter r with sampling period reflects the fact that as the inter-observation interval rises, the likelihood function is operating with observed values for incident case counts that are correspondingly larger, and the resulting dispersion would also be expected to scale in the same way. To identify the way in

which model discrepancy changes with the dispersion parameter, and to identify the dispersion parameter that offers the greatest accuracy, we ran scenarios considering different values of this parameter. Values 1, 2, 4, 8, 16 and 32 were examined for experiments regarding the daily scenario, while values 3, 6, 12, 24, 48 and 96 were used for three-day experiments and values 7, 14, 28, 56, 112 and 224 were used for weekly experiments.

Statistical analysis discrepancy results

To provide an objective assessment of the differences in discrepancy associated with each of the variables considered in the above scenarios, we employed Box-Cox multivariable regression analysis [97]. Box-Cox analysis was selected rather than traditional multiple linear regression as the discrepancy results were not normally distributed and routinely used transformations did not adequately address the assumptions of normality or homogeneous variance. The adequacy of empirical data (T^*), inter-observation interval or frequency of data observations, contact rate random walk standard deviation parameter (γ), and dispersion parameter (r) were evaluated as categorical variables as none of the parameters appeared to have a linear association with discrepancy based on data visualization exercises and there was also interest in understanding the specific differences among the chosen parameter values. Differences with p values < 0.05 were considered statistically significant.

Results

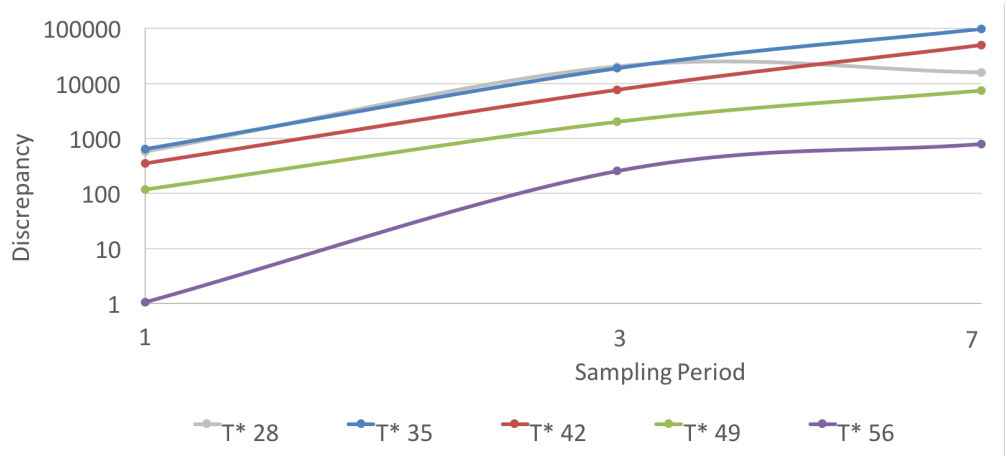


Figure 4.3: Log of discrepancy vs. log of sampling period for different observation times (r=32, γ= 0.125).

On the basis of running the model using daily, accumulated three days and accumulated weekly empirical data, particle filtering observing daily data performed consistently and markedly better than while observing three-day and weekly data. Particle filtering using successively larger sampling periods yielded super-linearly higher levels of discrepancy (Fig. 4.3, Table 4.2, 4.3 and 4.4). The exact difference in discrepancy between

sampling periods varies by the amount of data available (as given by T^*), but consistently the discrepancy extending from particle filtering using daily data was orders of magnitude smaller than for the larger sampling periods. Tables showing the discrepancy of particle filtering predictions in frequency scenarios for different observation times and $\gamma = 0.125$ and $\gamma = 2$ are included in Appendix C. The observed super-linear scaling of error with inter-observation interval was similar when comparing three day vs. weekly sampling.

Frequency scenarios (γ = 0.25)	$T^* = 35$	$T^* = 42$	$T^* = 49$	$T^* = 56$
PF using daily data, r=2	380	225	69	0
PF using three-day data, r=6	11453	5667	1646	205
PF using weekly data, r=14	80850	39578	6291	482
PF using daily data, r=8	384	213	29	0
PF using three-day data, r=24	14044	6452	1249	79
PF using weekly data, r=56	104043	42248	5484	447
PF using daily data, r=32	230	196	45	0
PF using three-day data, r=96	13617	4701	1096	86
PF using weekly data, r=224	149164	39232	4945	250

Table 4.2: Discrepancy of particle filtering predictions in frequency scenarios for different observation times and $\gamma = 0.25$

Frequency scenarios ($\gamma = 0.5$)	$T^* = 35$	$T^* = 42$	$T^* = 49$	$T^* = 56$
PF using daily data, r=2	474	270	80	0
PF using three-day data, r=6	13038	6577	1637	128
PF using weekly data, r=14	97325	38652	6661	592
PF using daily data, r=8	337	230	66	0
PF using three-day data, r=24	14900	6482	1264	67
PF using weekly data, r=56	126163	43288	5761	418
PF using daily data, r=32	635	188	13	0
PF using three-day data, r=96	13868	4590	766	44
PF using weekly data, r=224	156099	45808	4231	277

Table 4.3: Discrepancy of particle filtering predictions in frequency scenarios for different observation times and $\gamma = 0.5$

After accounting for differences across all of the examined scenarios for the adequacy of empirical data (T^*), random walk standard deviation parameter (γ), and dispersion parameter (r), the average discrepancy was significantly greater for data collected over three-day ($p < 0.001$) and seven-day ($p < 0.001$) intervals than for daily data.

Frequency scenarios ($\gamma = 1$)	$T^* = 35$	$T^* = 42$	$T^* = 49$	$T^* = 56$
PF using daily data, r=2	3327	695	87	0
PF using three-day data, r=6	43931	12590	1630	39
PF using weekly data, r=14	645037	154916	16362	976
PF using daily data, r=8	1568	241	18	0
PF using three-day data, r=24	35024	6251	682	4
PF using weekly data, r=56	1216215	129467	6072	376
PF using daily data, r=32	904	104	5	0
PF using three-day data, r=96	25452	4199	393	0
PF using weekly data, r=224	1243398	129629	4580	254

Table 4.4: Discrepancy of particle filtering predictions in frequency scenarios for different observation times and $\gamma = 1$

Frequency scenarios	Discrepancy
Without PF using daily data	101942842
Without PF using three-day data	386532229
Without PF using weekly data	575977188

Table 4.5: Discrepancy without particle filtering in frequency scenarios

The effect of the standard deviation for the random walk in the log of the contact rate (γ) also exhibited pronounced scaling patterns. Plotting three dimensional surfaces to represent the change of discrepancy in terms of this parameter γ and dispersion parameter r, we observed that for all daily, every-three-day and weekly scenarios, a γ parameter in the range of 0 to 2 yields markedly reduced discrepancy compared with γ values above 2 (Fig. 4.4, 4.6, 4.7 and 4.8). After accounting for differences across all of the examined scenarios for the frequency of data collection, adequacy of empirical data (T^*), and dispersion parameter (r), the average discrepancy was significantly greater for random walk standard deviation values of 4 ($p < 0.001$) and 8 ($p < 0.001$) compared to the baseline value of 0.125. However, there was no significant difference between random walk standard deviation values of 0.25 (p=0.97), 0.5 (p=0.99), 1 (p=0.97), or 2 (p=0.42) and the baseline random walk standard deviation of 0.125.

Figure 4.5 presents the discrepancies from particle filtering for different values of standard deviation associated with fraction reported incidence parameter (η). It appears that a System Dynamics model combined with particle filtering to learn from empirical data behaves robustly to changes in η for daily, every-three-day and weekly scenarios. The value for η was set to 1 for all of the scenarios reported in this work.

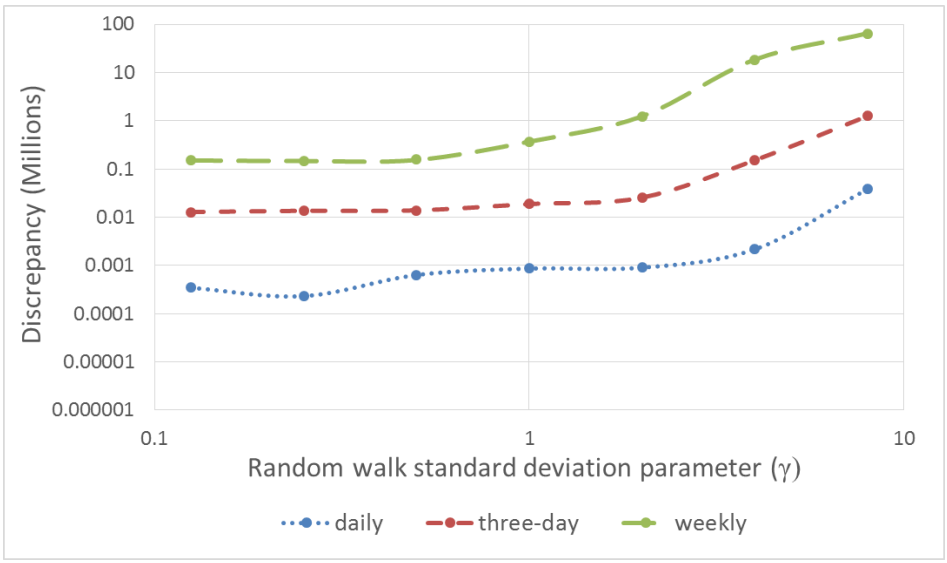


Figure 4.4: Discrepancy versus random walk standard deviation using daily, three-day and weekly observations ($T^* = 35$ and $r = 32$ for daily, 96 for three-day, and 224 for weekly observations)

As shown in figures 4.9 and 4.10, results suggest that increasing the dispersion parameter does not appear to strongly affect the performance of particle filtering at smaller values of contact rate random walk standard deviation parameter (γ). However, at larger values of γ, the impact of the dispersion parameter become more apparent (Fig. 4.6, 4.7 and 4.8). After accounting for differences across all of the examined scenarios for the frequency of data collection, adequacy of empirical data (T^*), and the contact rate random walk standard deviation parameter (γ), the average discrepancy was significantly smaller for each increasing dispersion parameter (r) from 1 to 32 ($p < 0.001$) as compared to the baseline value of 1.

Table 4.5 shows the discrepancy for the model without particle filtering. The discrepancy for particle filtering scenarios was found to be less than the discrepancy associated with the model without particle filtering.

4.4 Discussion and Future Work

The particle filtering method explored here offers considerable potential. The value offered by this approach seems likely to be particularly pronounced when used in the context of emerging communicable diseases in which limited parameter information is available to inform available models, but where frequent (e.g., daily) reporting of case counts are available. Particle filtering supports an adaptive response updating the current state and stochastic parameter values involved in dynamic models. In this way, the models are kept current with the latest evidence, which can be used to predict forward and to be used to then anticipate possible trade-offs between interventions. The key finding in this work is that particle filtering can perform orders of magnitude more accurately in case the daily clinical reports are available. For public health authorities seeking to employ accurate projection systems for communicable disease outbreaks, this finding suggests a

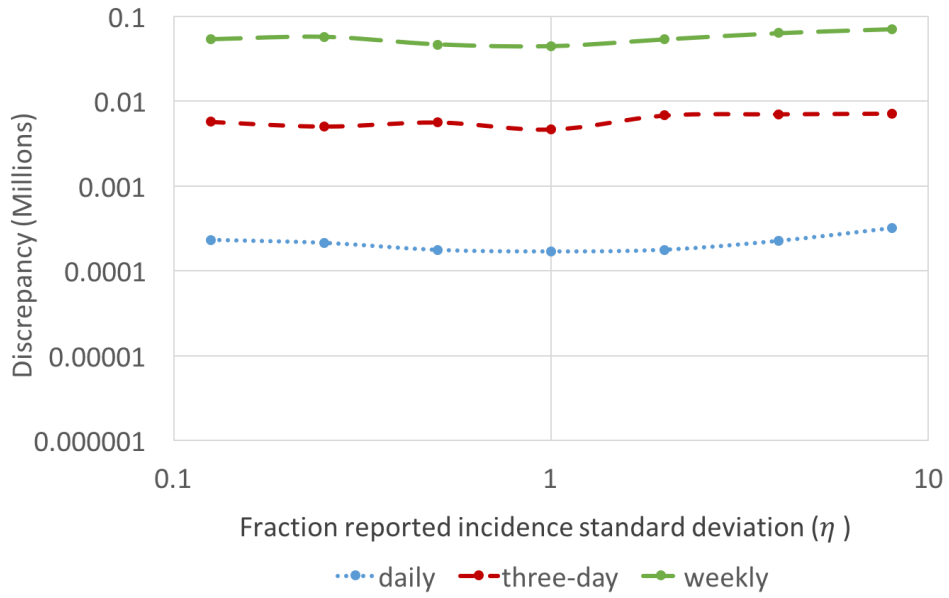
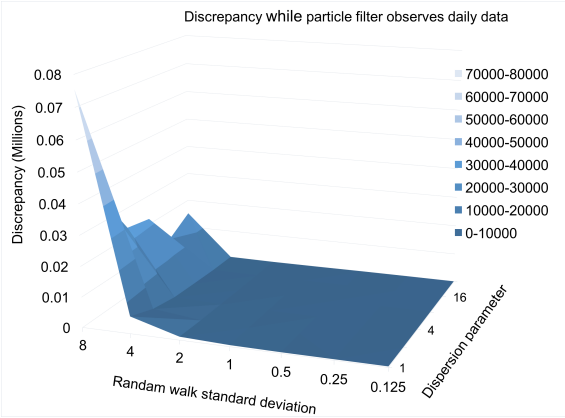


Figure 4.5: Discrepancy versus fraction reported incidence standard deviation using daily, three-day and weekly observations ($T^* = 35$, $\gamma = 0.125$ and $r = 32$ for daily, 96 for three-day, and 224 for weekly observations)

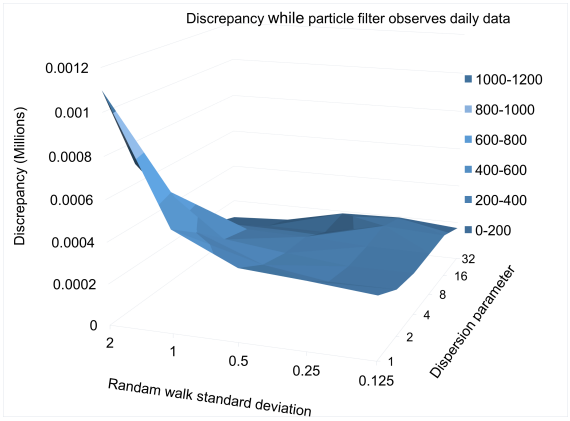
premium on putting in place efficient reporting schemes.

A second set of findings relates to the high robustness of preferred particle filtering parameter assumptions as we change the observation time in the outbreak and the inter-observation interval. While the assumption made for dispersion parameter associated with the negative binomial likelihood formulation does exert some impact on the accuracy of particle filtering, the results are far less sensitive to variations in this parameter beyond an inter-observation interval specific threshold. By contrast, while the results are highly sensitive to the assumptions regarding the rate of potential evolution of contacts per unit time (γ), the findings across different inter-observation intervals and time of observation are consistent in suggesting a specific range of low values for this parameter. While the particulars of these values are likely to differ somewhat for distinct epidemiological contexts (e.g., pathogens), populations and types of data, the consistency of these results suggests the potential for simpler guidelines to govern the application of particle filtering in specific epidemiological contexts. Importantly, given this robustness and daily reporting, these results suggest favorable starting assumptions for application of this approach to similar pathogens in developed countries. For different epidemiological contexts, the robustness of the results also suggest that a much simpler variant of the methodology used here might be applied in the opening days and weeks of an outbreak to estimate favorable parameter values for the dispersion parameter and rate of contact rate evolution for that particular context.

Research progress is needed to adequately realize particle filtering on other types of models, including agent-based and discrete-event models [98]. Since these modeling techniques are widely used in public health, and since implementing particle filtering in the presence of these types of models is not as straightforward



(a) Random walk standard deviation in the range of [0.125-8]



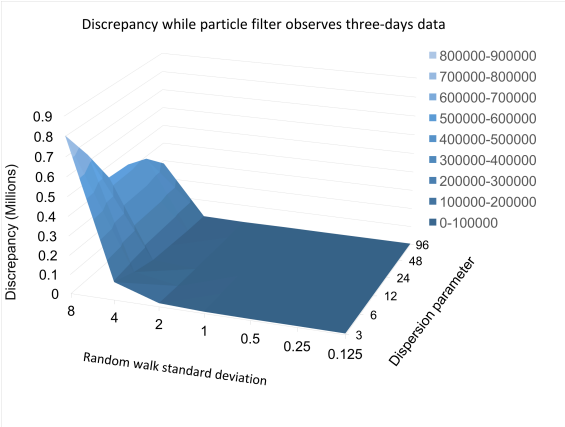
(b) Random walk standard deviation in the range of [0.125-2]

Figure 4.6: Discrepancy in terms of dispersion parameter and random walk standard deviation – daily empirical data and $T^* = 42$

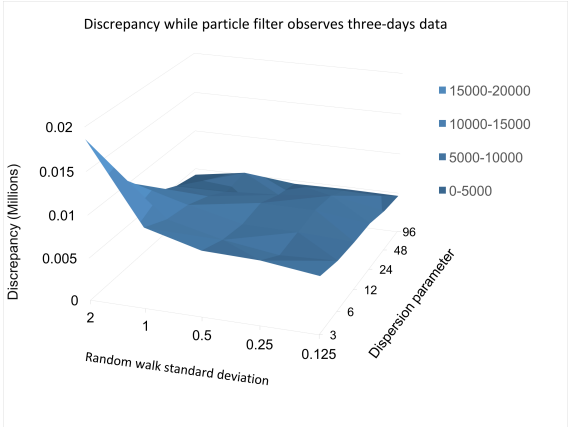
due to software limitations, advances are urgently required to improve software support for particle filtering for such models.

4.5 Conclusion

The findings presented here demonstrate that in the presence of simple models, particle filtering in combination with dynamic models can develop accurate predictive systems in the context of emerging communicable diseases, particularly when models lack information about parameters, but frequent reporting of empirical data is available. The results suggest that more frequent sampling improves predictive accuracy remarkably. The robustness of particle filtering in this case study also suggests that it may be possible to apply a variant of the method presented here to estimate unknown parameters of an emerging outbreak – specifically a new pathogen that is not well-known – in its opening days and weeks. According to the findings in this work, even very rough models can be combined with particle filtering to project the evolution of emerging infectious diseases and secure strong guidance for health policy makers.

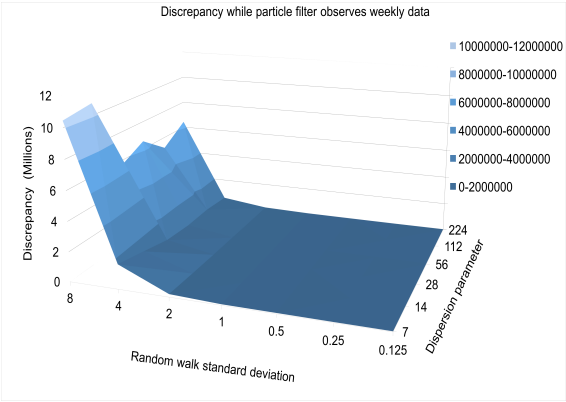


(a) Random walk standard deviation in the range of [0.125-8]

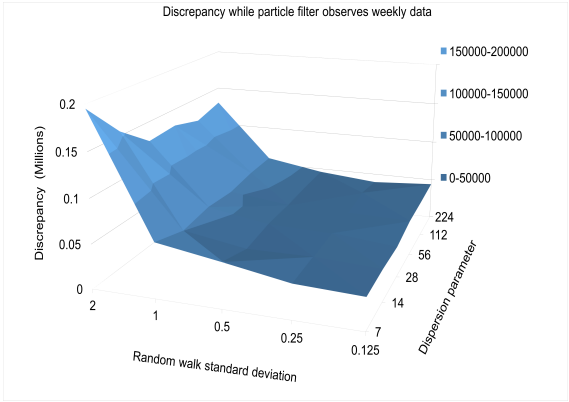


(b) Random walk standard deviation in the range of [0.125-2]

Figure 4.7: Discrepancy in terms of dispersion parameter and random walk standard deviation empirical data available every three-days and $T^* = 42$



(a) Random walk standard deviation in the range of [0.125-8]



(b) Random walk standard deviation in the range of [0.125-2]

Figure 4.8: Discrepancy in terms of dispersion parameter and random walk standard deviation weekly empirical data and $T^* = 42$

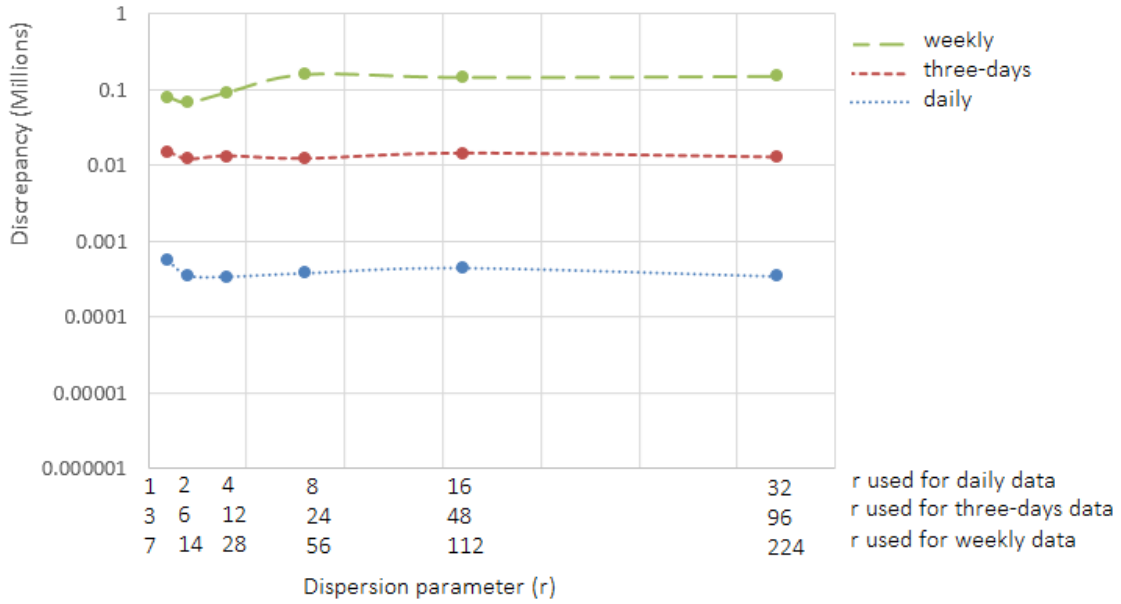


Figure 4.9: Discrepancy versus dispersion parameter using daily, three-day and weekly observations ($T^* = 42$ and $\gamma = 0.125$

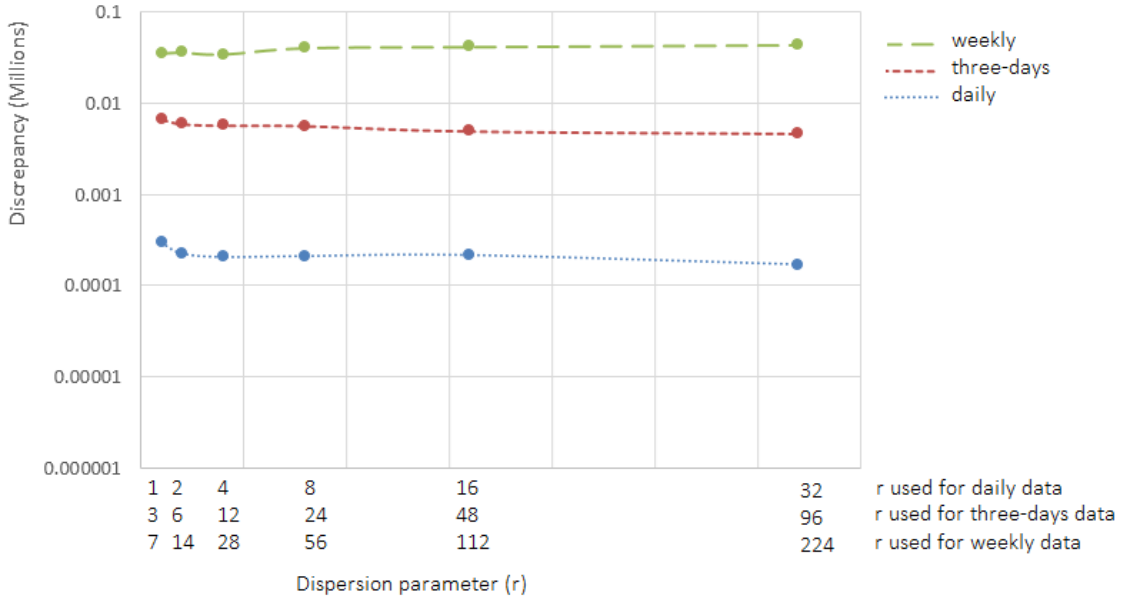


Figure 4.10: Discrepancy versus dispersion parameter using daily, three-day and weekly observations ($T^* = 35$ and $\gamma = 0.125$)

Chapter 5

Social Media Surveillance Improves Outbreak Projection via Transmission Models

The text of this chapter is largely drawn from a manuscript entitled "Social Media Surveillance Improves Outbreak Projection via Transmission Models" by Anahita Safarishahrbijari and Nathaniel D Osgood, submitted to the Journal of Medical Internet Research. Authors' contributions are described in 1.

While dynamic models are increasingly used by decision makers as a source of insight to guide interventions to control communicable disease outbreaks, such models have long suffered from a risk of rapid obsolescence due to a failure to keep updated with emerging epidemiological evidence. The application of statistical filtering algorithms to high-velocity data streams has recently demonstrated effectiveness in allowing such models to be automatically re-grounded by each new set of incoming observations. The attractiveness of such techniques has been enhanced by the emergence of a new generation of geospatially specific, high-velocity data sources, including daily counts of relevant searches and social media posts. The information available in such electronic data sources complements that of traditional epidemiological data sources.

This chapter seeks to evaluate the degree to which the predictive accuracy of pandemic projection models re-grounded via machine learning in daily clinical data can be enhanced by extending such methods to leverage daily search counts.

We combined a previously published influenza A (H1N1) pandemic projection model with the sequential Monte Carlo technique of particle filtering so as to reground the model on a daily basis using confirmed incident case counts and search volumes. The effectiveness of particle filtering was evaluated using a norm discrepancy metric via predictive- and dataset specific- cross-validation.

Results suggested that despite the data quality limitations of daily search volume data, the predictive accuracy of dynamic models can be strongly elevated by the inclusion of such data in filtering methods.

The predictive accuracy of dynamic models can be notably enhanced by tapping a readily accessible, publically available high-velocity data source. This work highlights a low-cost, low-burden avenue for strengthening model-based outbreak intervention response planning using low-cost public electronic datasets.

5.1 Introduction

The capacity to accurately project communicable disease outbreak evolution is of great value in public health planning prevention and control activities. Use of such information can inform resource allocation, including surge-capacity planning and planning of the timing of outbreak response immunization campaigns, and – when applied across distinct scenarios – provides a basis for evaluating tradeoffs between intervention strategies. While dynamic models are increasingly widely used to conduct such scenario projection, the construction of such models for new and rapidly evolving pathogens commonly faces significant barriers due to uncertainties regarding important factors governing the natural history of the disease, such as durations of latent, incubation and infectious phases, the probability of asymptomatic carriage, rates of waning immunity, contact rates and per-discordant-contact transmission probabilities. Moreover, even the most intricate models face strict limitations in their ability to project evolution of factors treated as stochastic, such as weather-related variables and the timing of arrival of exogenous infections due to global travel. Using computational statistical estimation methods such as sequential Monte Carlo techniques, researchers have in recent years contributed approaches to elevate the predictive accuracy of dynamic transmission models by updating their state estimates at the time of appearance of each new observation. The predictive accuracy of methods have thus far been evaluated purely in the context of models which make use of traditional surveillance datasets, such as laboratory and clinically confirmed case reports [61, 31, 33, 15, 16, 94].

While such traditional surveillance datasets offer high-quality, rich information concerning individuals who present for medical care, they suffer from notable shortcomings, including delayed reporting and a failure to include counts of infective individuals who elect not to present. In a separate stream of work from the dynamic modeling work noted above, researchers have in recent years sought to compensate for the limitations of traditional epidemiological data sources more generally by exploiting information related to online communicational behavior, and particularly the growing tendency of many users to search, post, and tweet about their illnesses. Specifically, such researchers have assessed the health insights that can be gained from public health surveillance applications employing a variety of online sources of information.

A prominent line of this work has focused on time sequences of search query volumes, such as those previously captured in Google Flu Trends (GFT) [3] and (on a more generic and continuing basis) Google Trends [2]. Within this sphere, a wide variety of investigations have utilized statistical and machine learning methods to perform classification and analysis on such Google search volume data and volumes of social media postings, including for communicable illnesses. Many researchers have investigated biomedical and health related knowledge obtained from twitter platform, suggesting the opportunities and limitations associated with different machine learning classifiers and training models for tweet mining [99, 80, 100]. Other case studies have reported significant correlation between Tweets and clinical reports and concluded that social media text mining can improve public health communication efforts by providing insight into major themes of public concerns in the health sphere [101, 102].

An important subset of research in this area has leveraged data obtained from Google to develop statistical forecast models and evaluated the degree to which GFT data in combination with statistical models can support accurate predictions [52, 53, 54] and correlation with real time empirical data [55]. Some investigators jointly used multiple data sources, including GFT and Twitter, and compared the performance of statistical prediction models using each data source and also in scenarios where different data sources complement one another [56, 57].

The prediction of epidemic outbreaks by dynamic models often involves significant error and generally needs to consider both underlying dynamics and noise related to both measurement and process evolution. While older techniques based on Kalman Filtering and variants [87] have long provided a computationally frugal means of filtering stochastic dynamic models, such MLE-based approaches are impaired by strong distributional assumptions concerning measurement and process noise, and limited accommodation for non-linearity in the system being characterized. This challenge in handling non-linearity is felt most keenly in terms of an inability to capture the effects of probability distributions across multiple basins of attraction, and a requirement for model linearization that is problematic for important modeling formalisms, such as agent-based models. For these and other reasons, recent researchers have increasingly turned to stronger filtering methods. Several authors have applied the sequential Monte Carlo (SMC) technique of particle filtering (PF) as an effective tool in support of both model estimation and predictions from real world data. Ong et al. established a real-time surveillance system in Singapore to feed data into a stochastic model of influenza-like disease dynamics, which was refitted daily using PF [61]. Osgood and Liu used a synthetic ground truth model to evaluate the effectiveness of PF for an H1N1-like infection in the presence of noisy data and systematic model simplifications [33]. Safarishahrbijari et al. evaluated the effectiveness of PF subject to specifics of the configuration, such as frequency of data sampling and representation of behaviour change in the form of an evolving contact rate for H1N1 [15, 16]. Oraji et al. developed a system dynamics model for studying the tuberculosis transmission, and applied PF to estimate the latent state of the system, including many epidemiological quantities that are not directly measured. Their results suggested an improvement of model accuracy using PF and high additional value extending from consideration of additional epidemiological quantities in the probabilistic model [94]. Li et al. applied particle filtering to a measles compartmental model using reported measles incidence for Saskatchewan. They also performed particle filtering on an age structured adaptation of their model by dividing the population into children and adults age-groups. According to their results, particle filtering can offer high predictive capacity for measles outbreak dynamics in a low vaccination context [103]. The literature characterized in this paragraph indicates that, when used with a suitable dynamic model, particle filtering can offer high predictive capacity for contagious diseases and outbreaks; however, none of these works have used data extracted from online communicational behaviour time series such as those available via GFT, and the underlying models do not consider the transmission of fear between individuals.

Epstein et al. explored the effect of adaptive behaviors such as social distancing based on fear and contact

behaviour in models of epidemic dynamics [14]. They used nonlinear dynamical systems and agent-based computation and integrated disease and fear of the disease contagion processes. Based on their models, individuals anxious ("scared") about or infected by a pathogen can transfer fear through contact with other individuals who are not scared, and scared individuals may isolate themselves, which affects the contact rate dynamic, which is a key parameter in governing outbreak evolution. The authors studied flight as a behavioral response and concluded that even small levels of fear-inspired flight can have a dramatic impact on spatio-temporal epidemic dynamics [14].

Despite the fact that both high velocity search volume and social media data and transmission models share a temporal perspective, data drawn from such internet series has not to our knowledge been previously used as a source of information for filtering (via recurrent re-grounding) compartmental transmission models with the arrival of new data.

In this chapter, we sought to address that gap by combining the transmission model from [14] with the sequential Monte Carlo method of particle filtering, considering the interaction between disease and fear of disease contagion processes for the 2009-2010 H1N1 influenza pandemic. The particle filtered model used time series of both clinically-observed data and daily Google search query volumes to automatically and recurrently re-ground the model as successive data points became available. Based on lessons learned from [15, 16] as to the importance of incorporating higher-velocity rather than time-averaged data, we made use of daily data. In contrast to past PF work grounding transmission models that have used empirical data purely as a comparison with model results reflecting the natural history of infection, the model presented here engaged in such comparisons for the clinical data, but further compared the search query volume data with ideation-related model state (individuals with fear).

5.2 Methods

5.2.1 Particle Filtering Model

As the first stage of characterization of the particle filtered model, we first present the formulation of the existing Epstein compartmental model from [14], which characterizes the population into states according to both their natural history of infection and presence of anxiety regarding influenza. The state variables of the model are as follows: Susceptible to pathogen and fear (S), Infected with fear (in fear) (I_F), Infected with pathogen (I_P), Infected with pathogen and in fear (I_{FP}), Removed due to fear (R_F), Removed due to fear and pathogen (R_{FP}) and Recovered (R). We used an adaptation of the model that included an Exposed (E) state variable (Figure 5.1). In this model, λ_F is the (hazard) rate of removal due to self-isolation of those in fear only, λ_P refers to rate of recovery from infection with pathogen and λ_{FP} represents rate of removal due to self-isolation of infected who are also afraid, while H is the rate of recovery from fear (alone) and return to circulation [14].

The parameters α and β denote transmissibility of fear and pathogen, respectively. Specifically, α rep-

resents the probability that a contact between an individual A who is currently without fear but who is susceptible or infected purely with pathogen and an individual B with either fear or pathogen will cause individual A to grow afraid. By contrast, β denotes the probability that a contact between an individual A who has never been infected with pathogen and an individual B specifically infected with pathogen will infect individual A with pathogen. Given that α and β are probabilities (and are thus of unit dimension), it bears emphasis that simple dimensional analysis demonstrates that the original authors assume an effective per-person-per-unit time mixing rate holding a value of unity. While not considered within the scope of the original article, this mixing rate can itself be characterized in accordance with longtime mathematical epidemiology practice as the product of a per-unit-time contact rate c and disease transmissibility divided by the (constant) total population N. Because we consider changes to the value of c within this work, this quantity is shown explicitly in the equations below. To explain this term required for dimensional consistency, we note that each transmission term (such as $\beta\alpha\frac{c}{N}\,c\,I_{FP}$) can be considered as characterizing the rate of transmission (in terms of persons per unit time) from possible transmitters in category Y (here, I_{FP}) to persons in at-risk category X (here, S). Each such at-risk person X is assumed to engage in an average of c contacts per unit time. Those overall contacts are then assumed to be spread proportionally among the compartments in the population, with the fraction taking place with those in a category Y of possible transmitters being the count of people in Y divided by the total population N. The probability in the prefix of the term (here, $\beta\,\alpha$) indicates the probability that each such potentially-transmitting contacts does in fact lead to the type of transmission being considered in that term (either fear, pathogen, or, as in this example, both).

In adapting the model, we took advantage of the previously demonstrated [15, 16] capacity of particle filtering to support stochastic evolution of designated parameters (captured as state variables). One of the stochastic parameters included in this model represents the fraction of reported incidents (f_P). This represents the fraction of people who are reported to public health authorities when emerging from the latent state, and is a value that is both uncertain and evolving over time. Likewise, the fraction of people becoming afraid who search Google upon infection – named the fraction of Google search incidents (f_F) – is further treated as a dynamic uncertain parameter.

Other parameters also treated as stochastic are the contact rate (c), removal rate from those with fear to self-isolation (γ_F) and removal rate from those with fear who are also infected (γ_{FP}). To support this, such dynamic parameters are associated with state variables evolving over time according to stochastic differential equations. Because variable c is a non-negative quantity, we performed a log-transform on this variable according to the Brownian Motion, so that it varied over the full real numbers. The stochastic differential equation of contact rate c is described using Stratonovich notation as:

$$d(ln(c)) = \gamma\, dW_t \tag{5.1}$$

where dW_t is a standard Wiener process following a normal distribution with mean of 0 and variance of

1 [95]. Thus, $d(ln(c))$ is is subject to Gaussian perturbations. We also performed a log-transform on λ_F; the stochastic differential equation of λ_F is formulated in Stratonovich notation as:

$$d(ln(\lambda_F)) = s_{\lambda_F}\, dW_t \tag{5.2}$$

The initial value of c and λ_F is drawn uniformly from the interval between 0 and 100 per day and between 0.4 and 1 per day, respectively. The standard deviation of γ and s_{λ_F} were both selected to be 1. By contrast, reflecting the fact that f_P and f_F represent fractions, such parameters were logit-transform, with the initial value for each varying between 0 and 0.2. We described the stochastic differential equations of fractions f_P and f_F according to Brownian Motion as the following, again following Stratonovich notation for each:

$$d(logit(f_P)) = d(ln(\frac{f_P}{1-f_P})) = \eta\, dW_t \tag{5.3}$$

$$d(logit(f_F)) = d(ln(\frac{f_F}{1-f_F})) = s_{fF}\, dW_t \tag{5.4}$$

Within the model, the parameter f_P is multiplied by inflows to state variables Infective (I) and Scared Infective (I_{FP}) to account for fractional actual reporting. Similarly, the parameter f_P is multiplied by inflows to state variables Scared (I_F), Scared Infective (I_{FP}), Removed due to Fear and Infection (R_{FP}) and Removed due to Fear (R_F) accounts for the fractional actual scared population.

We treated γ_{FP} as $\frac{1}{meanlatenttimetorecovery \times \lambda^{'}_{FP}}$ and then considered $\lambda^{'}_{FP}$ as a fraction and performed a logit-transform on it. This parameter varies over the range from 0 to 1 and the dynamic process for $\lambda^{'}_{FP}$ is similar to f_P and f_F, specifically:

$$\frac{d(logit(\lambda^{'}_{FP}))}{dt} = \frac{d(ln(\frac{\lambda^{'}_{FP}}{1-\lambda^{'}_{FP}}))}{dt} = s_{\lambda^{'}_{FP}}\, dW_t \tag{5.5}$$

The standard deviations η, s_{fF} and $\lambda^{'}_{FP}$ are selected to be 5, 5 and 1, respectively. The initial values of f_P, f_F and $\lambda^{'}_{FP}$ are set on the intervals [0, 0.2), [0, 0.2) and [0, 0.5), respectively.

By applying random walks to these parameters, a more accurate estimate was achieved during model simulation. As such, in our model, each particle at each point in time is associated with all state variables and state variables associated with stochastic parameters $(S, E, I_F, I_P, I_{FP}, R_F, R_{FP}, R, c, f_P, f_F, \lambda_F, \lambda^{'}_{FP})$.

$$\frac{dS}{dt} = -\beta(1-\alpha)\frac{c}{N}SI_P - (1-\beta)\alpha\frac{c}{N}SI_P - \beta\alpha\frac{c}{N}SI_F - \beta(1-\alpha)\frac{c}{N}SI_{FP} - (1-\beta)\alpha\frac{c}{N}SI_{FP} - \beta\alpha\frac{c}{N}SI_{FP} \tag{5.6}$$

$$\frac{dS}{dt} = \beta(1-\alpha)\frac{c}{N}SI_P + \beta(1-\alpha)\frac{c}{N}SI_{FP} - \frac{E}{\tau} \tag{5.7}$$

$$\frac{dI_P}{dt} = \frac{E}{\tau} - \alpha I_P I_P - \alpha I_P I_F - \alpha I_P I_{PF} - \lambda_P I_P + H R_{PF} \tag{5.8}$$

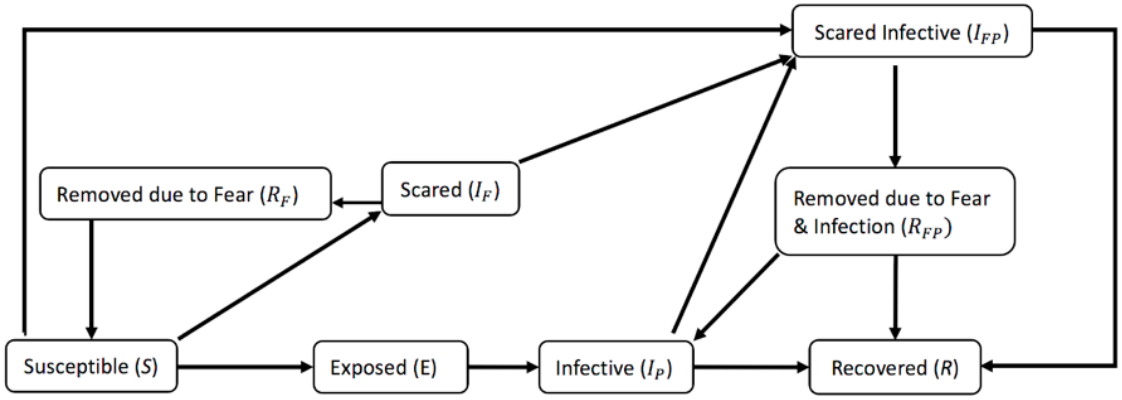


Figure 5.1: Coupled contagion dynamics of fear and disease

$$\frac{dI_F}{dt} = (1-\beta)\alpha\frac{c}{N}SI_P + \alpha\frac{c}{N}SI_{FP} + (1-\beta)\alpha\frac{c}{N}SI_{FP} - \beta\frac{c}{N}I_F I_{FP} - \lambda_F I_F \tag{5.9}$$

$$\frac{dI_{FP}}{dt} = \beta\alpha\frac{c}{N}SI_P + \beta\alpha\frac{c}{N}SI_{FP} + \beta\frac{c}{N}I_F I_P + \beta\frac{c}{N}I_F I_{FP} + \alpha\frac{c}{N}I_P I_P + \alpha\frac{c}{N}I_P I_F + \frac{c}{N}\alpha I_P I_{FP} - \lambda_P I_{FP} - \lambda_{FP} I_{FP} \tag{5.10}$$

$$\frac{dR_F}{dt} = \lambda_F I_F - HR_F \tag{5.11}$$

$$\frac{dR_{FP}}{dt} = \lambda_{FP} I_{FP} - \lambda_P^{'} R_{FP} - HR_{FP} \tag{5.12}$$

$$\frac{dR}{dt} = \lambda_P I_P + \lambda_P I_{FP} + \lambda_P^{'} R_{FP} \tag{5.13}$$

5.2.2 Description of Data Sources

We evaluated the prediction of the above-described dynamic model assisted by particle filtering against two publicly available empirical datasets. The first was from Manitoba Health, Healthy Living and Seniors and included daily laboratory-confirmed case counts of pandemic H1N1 influenza for the period of October 6^{th}, 2009 through January 18^{th}, 2010 for the province of Manitoba [58]. The second dataset was from the Institut National de Santé Publique du Québec (INSPQ) – a public health expertise and reference centre in Quebec – and included daily confirmed case counts of pandemic H1N1 influenza between October 6^{th}, 2009 and December 19^{th}, 2010 [59].

In addition to the daily clinical case count data noted above, we obtained normalized daily Google search counts from Google trends and weekly normalized data from Google flu trends for Manitoba and Quebec during the second pandemic wave. Reflecting the linguistic differences between the two provinces, the search terms used for each were distinct. In Manitoba, we used search terms "flu" and "H1N1", while for Quebec,

we used "flu", "Influenza A virus sub-type H1N1", "h1n1 vaccination", "ah1n1", "ah1n1 vaccin", "grippe" and "grippe ah1n1" categories – the most frequent search queries related to this topic suggested by Google during that period.

5.2.3 Particle Values and Parameter Values

In defining the likelihood function for observing the empirical data given the state of a given particle, the exact variant of the likelihood used varied across three different scenarios examined. The first scenario evaluated the impact of assuming a likelihood formulation that considered purely clinical data, termed $\mathcal{L}_{Infection\,with\,Pathogen}$. The likelihood being used in the second scenario considered only the likelihood of observing the empirical data regarding Google search counts for the appropriate province in light of the count of individuals posited to be currently in fear within the model, a likelihood denoted as $\mathcal{L}_{Infection\,with\,Fear}$.

Following several past contributions [33, 15, 94, 93], we assume that each epidemiological quantity follows a Pascal distribution function [95]. Thus, given y_t and i_t as representing observed individuals per day and particle-posited daily rate (count per day) of new cases, respectively,

$$\mathcal{L}(y_t|i_t) = \binom{y_t + r - 1}{r - 1} p^r (1-p)^{y_t} \tag{5.14}$$

In the formulation for the likelihood function, r is a dispersion parameter and $p = \frac{i_t}{i_t + r}$.

$$\mathcal{L}_{Infection\,with\,Pathogen} = \binom{y_{Pt} + r_P - 1}{r_P - 1} p_P^{r_P} (1-p_P)^{y_{Pt}} \tag{5.15}$$

$$\mathcal{L}_{Infection\,with\,Fear} = \binom{y_{Ft} + r_F - 1}{r_F - 1} p_F^{r_F} (1-p_F)^{y_{Ft}} \tag{5.16}$$

where y_{Pt} and y_{Ft} represent number of lab confirmed incident cases reported for day t and number of Google search incidents for that day, respectively. The probabilities p_P and p_F follow $\frac{i_{P_t}}{i_{P_t} + r_P}$ and $\frac{i_{F_t}}{i_{F_t} + r_F}$, respectively; where i_{P_t} is a fraction of the flow of new cases of infection and i_{F_t} is a fraction of the flow of new cases of scared. The dispersion parameter $\mathcal{L}_{Infection\,with\,Pathogen}(r_P)$ was considered as 40, while $\mathcal{L}_{Infection\,with\,Fear}(r_F)$ was considered as 25. This reflects the larger noise that we believed to be associated with Google search data, in light of the fact that a larger dispersion parameter leads to a more narrowly dispersed distribution.

The third scenario considered a total likelihood function $\mathcal{L}_t$ consisting of a combination of $\mathcal{L}_{Infection\,with\,Pathogen}$ and $\mathcal{L}_{Infection\,with\,Fear}$. For defining the total likelihood function, the simplifying assumption was made that deviations with respect to one measure was independent of the other, and thus the total multivariate likelihood function could be treated as a multiplication of two univariate likelihood functions, given as:

$$\mathcal{L}_T = \mathcal{L}_{Infection\,with\,Pathogen} \times \mathcal{L}_{Infection\,with\,Fear} \tag{5.17}$$

The purpose of running this third scenario was to compare the effectiveness of a univariate likelihood function with the multivariate likelihood function, when evaluated in terms of a calculated discrepancy of model predictions against the epidemiologically confirmed case count.

Parameter	Notation	Value for Quebec	Value for Manitoba	Unit
Probability of infection transmission given exposure	β	0.04	0.04	Unit
Probability of fear transmission given exposure	α	0.02	0.02	Unit
Mean latent time	τ	Uniformly distributed (2, 4)	Uniformly distributed (2, 4)	Day
Mean time to recovery	μ	7	7	Day
Total population of province	N	7843475	1214403	Person
Rate of recovery from fear	H	0.2	0.2	1/Day
Rate of removal to self-isolation from fear	λ_F	Dynamic	Dynamic	1/Day
Fraction of mean time to recovery of going from "Scared Infected" to "Recovered" via "Removed due to Fear & Infection"	$\lambda_{FP}^{'}$	Dynamic	Dynamic	1/Day
Rate of removal to self-isolation from fear and pathogen	λ_{FP}	$\frac{1}{\mu\,\lambda_{FP}^{'}}$	$\frac{1}{\mu\,\lambda_{FP}^{'}}$	1/Day
Rate of recovery from infection with pathogen	λ_P	$\frac{1}{\mu}$	$\frac{1}{\mu}$	1/Day
Rate of recovery from removed due to fear and infection	$\lambda_P^{'}$	$\frac{1}{\mu\,(1-\lambda_{FP}^{'})}$	$\frac{1}{\mu\,(1-\lambda_{FP}^{'})}$	1/Day

Table 5.1: Parameters used in the model

The three scenarios noted above were conducted using particle filtering employing 1000 particles. For each such scenario, reflective of the need make decisions in light of uncertainty about the evolution of an unfolding outbreak in which only information about time points up to the present is available, we sought to examine the impact of right-censoring the empirical data at certain time-point T^*, representing the current time (i.e., the time from which the model is forecasting outbreak evolution). Thus, as the model ran, particle weights were updated based on observations from day one until and including day T^*; after day T^*, particle filtering ceased, particle weights were no longer updated using historic data, and no further particles were re-sampled. Each scenario included a sequence of sub-scenarios, which employed the following distinct values of $T^* : \{25, 30, 35, 40, 45, 50\}$.

To judge the accuracy of particle filter-informed projections for future times against the standard of the reported case counts for those times, we defined a discrepancy metric as the expected value of the L^2 norm of the difference between sampled particles (reporting-rate coefficient times the sum of infected and scared infected states) and reported case count observations calculated after time T^*. We sampled n particles (n=700) according to their weights, and obtained the discrepancy value using the following equation:

$$discrepancy = \frac{\sum_{i=T^*+1}^{T_f} (\frac{\sum_{j=1}^{n} (x_{ij}^P - x_i^E)^2}{n})^2}{T_f - T^*} \tag{5.18}$$

where x_{ij}^P is the value associated with sampled particle j at observation i, x_i^E is the respective reported clinical cases at observation i. T_f is the final observation time and T^* indicates the time from which the projection is being made (i.e., the time up to which the particles' weights were updated based on observation, where $0 \leq t \leq T^*$). Using this formulation, we evaluated how well projections forward predicted the empirical data after T^*, the time at which particle filtering completed.

5.3 Results

In this work, for each scenario (each associated with a particular likelihood function), we plotted the graphs associated with $T^* = 30$ for Manitoba and Quebec. We characterize the results below, organized by scenario.

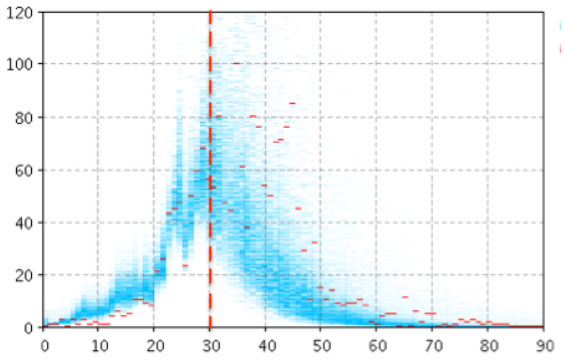
1. Particle filtering using two likelihood functions

   Figure 5.2 and Figure 5.3 depict the empirical data (red and magenta points) superimposed on samples (blue and green) from the model-generated distribution of particles for the model output for the number of reported cases (left panel) and number of searches (right panel) for Manitoba (Figure 5.2) and Quebec (Figure 5.3). For $T^* = 30$, the high posterior density (HPD) are for the projection period is quite localized for cases of pathogen and the number of searches.

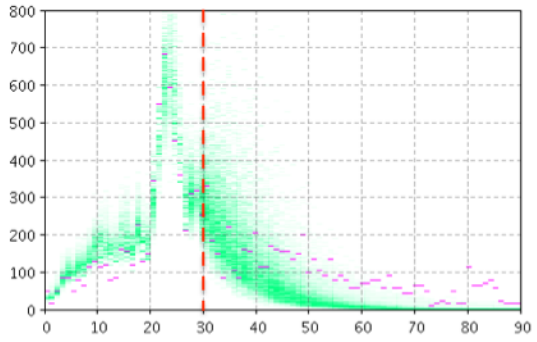
2. Particle filtering using the likelihood function associated with clinical data alone.

   In this configuration, particle filtering was performed using as the sole likelihood function. Figure 5.4 and Figure 5.5 depict empirical data (red and magenta points) superimposed on samples (blue and green) from the model-generated distribution of particles for the model output for the number of reported cases (left panel) and number of searches (right panel) for Manitoba (Figure 5.4) and Quebec (Figure 5.5). Despite the fact that the particle filtering employs reasonably high resolution clinical data, the system exhibits great difficulty both in accurately projecting number of clinical case reports forward from the point where particle filtering ceases (T^*), and in doing so in a fashion where the HPD region is localized. Unsurprisingly, the model informed by the reported clinical case counts alone is unable to accurately characterize the search volume within the population.

3. Particle filtering using the likelihood function associated with search-volume data alone.



(a) Clinical data and particle distribution. Red markers depict empirical data points (clinical empirical data) and blue depicts the 2D histogram of samples from the particle filtered model (reported incidence).

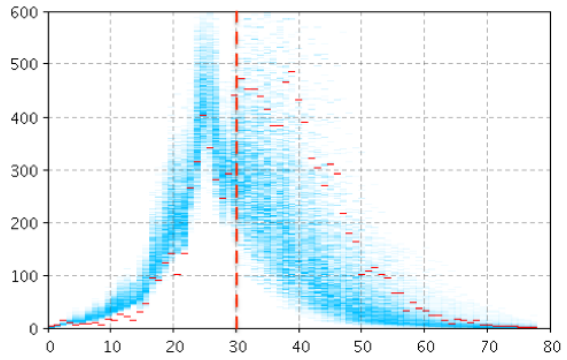


(b) Google search-volume data and particle distribution. Magenta markers depict empirical data points (Google empirical data) and green depict the 2D histogram of samples from the particle filtered model (search volume).

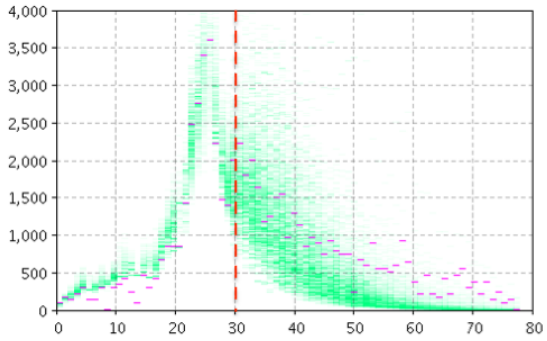
Figure 5.2: Empirical data (red and magenta points) superimposed on samples (blue and green) from the model-generated distribution of particles for the model output for the count of reported cases (left panel) and number of searches (right panel) using two likelihood functions, $T^* = 30$ for Manitoba.

In this configuration, particle filtering was performed using as the sole likelihood function. Figure 5.6 and Figure 5.7 depict empirical data (red and magenta points) superimposed on samples (blue and green) from the model-generated distribution of particles for the model output for the number of reported cases (left panel) and number of searches (right panel) for Manitoba (Figure 5.6) and Quebec (Figure 5.7) While results for both jurisdictions some localization in the projections of the prevalent case count of those living in fear, the failure to consider clinical case count in the particle filtering (and to accordingly update the model estimates for the current number of infectives, susceptibles and the contact rate) leads to poor projection accuracy for the reported clinical case count.

Figure 5.8 and Figure 5.9 depict the (log-scaled) discrepancies between model clinical case predictions and empirical data for different check times (T^*) for Manitoba and Quebec, respectively. Unsurprisingly given the results above, the discrepancy associated with particle filtering informed by both clinical and search volume datasets (Scenario 1) is smaller than the discrepancy associated with either dataset in isolation. In addition, the discrepancy when using particle filtering informed by the (higher-quality) clinical case count data alone is lower than that informed purely by search volume. However, there is a marked difference between Manitoba and Quebec in the levels of discrepancy seen when using clinical case data alone vs. with search volume data. For Manitoba, there is consistently less than an order of magnitude of difference in discrepancies between these two results. By contrast, for Quebec, using the clinical data alone within particle filtering yields a level of discrepancy several orders of magnitude below that resulting from search volume data. Intriguingly, for Manitoba, combining both yields a reduction of discrepancy many orders of magnitude below either, despite the fact that discrepancy is calculated with respect to clinical case reports. This advantage of adding information from the search volume data to that from clinical case counts presumably reflects the fact that



(a) Clinical data and particle distribution. Red markers depict empirical data points (clinical empirical data) and blue depict the 2D histogram of samples from the particle filtered model (reported incidence).



(b) Google search-volume data and particle distribution. Magenta markers depict empirical data points (Google empirical data) and green depict the 2D histogram of samples from the particle filtered model (search volume).

Figure 5.3: Empirical data (red and magenta points) superimposed on samples (blue and green) from the model-generated distribution of particles for the model output for the count of reported cases (left panel) and count of searches (right panel) using two likelihood functions, $T^* = 30$ for Quebec.

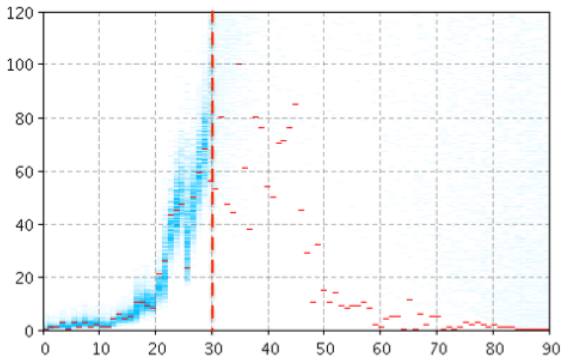
the added search volume information supports particle filtering in more accurately localizing the model state estimates than was the case using purely the reported clinical case counts – a factor manifested in the projections for both clinical case counts. By contrast, for Quebec, using both sources of information reduces the discrepancy significantly, typically by at least one order of magnitude, with the exception of time points $T* = 45$ and 50.

5.4 Discussion

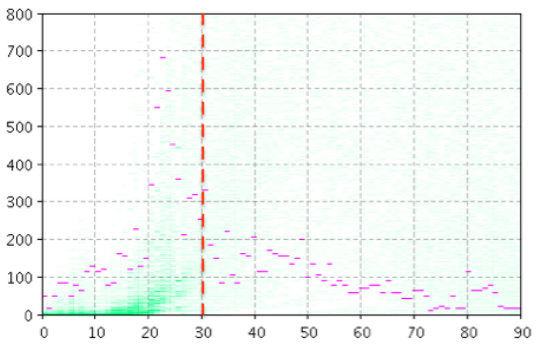
5.4.1 Principal Results

In this contribution, we investigated the predictive accuracy gains from applying particle filtering using both traditional and search volume data to estimate latent states of a compartmental transmission model (including time evolution of stochastic parameters involved in that model). The capacity to perform this estimation then provides support for projection and scenario evolution using the model.

To be able to use search data effectively when particle filtering a transmission model, we found it helpful to move beyond the traditional scope of compartmental transmission models and to adopt a more articulated model of the outbreak, reflecting the fact that causal drivers promoting web searches are not restricted to stages in the natural history of infection, but are additionally driven by factors with distinct but coupled dynamics, such as fluctuations in perceived risk on the part of the population. Responsive to this consideration, we have adapted a previously published model with an explicit consideration of the coupled dynamics of fear and pathogen. While there are challenges associated with assessing perceived risk and anxiety on the part of the population during an outbreak, we found here that projection of outbreak dynamics can be



(a) Clinical data and particle distribution. Red markers depict empirical data points (clinical empirical data) and blue depict the 2D histogram of samples from the particle filtered model (reported incidence).



(b) Google search-volume data and particle distribution. Magenta markers depict empirical data points (Google empirical data) and green depict the 2D histogram of samples from the particle filtered model (search volume).

Figure 5.4: Empirical data (red and magenta points) superimposed on samples (blue and green) from the model-generated distribution of particles for the model output for the count of reported cases (left panel) and count of searches (right panel) using the likelihood function associated with clinical data alone, $T^* = 30$ for Manitoba.

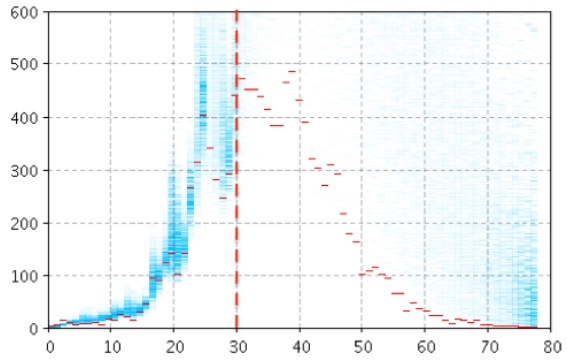
materially enhanced through the inclusion of a surprisingly accessible source of data: Daily relative search query volumes for defined geographic regions on the widely used Google search engine.

The reliable and timely public availability of such data across many areas of the world raises the prospects for significantly enhancing effective outbreak projection using combinations of dynamic modeling and machine learning techniques such as the particle filter.

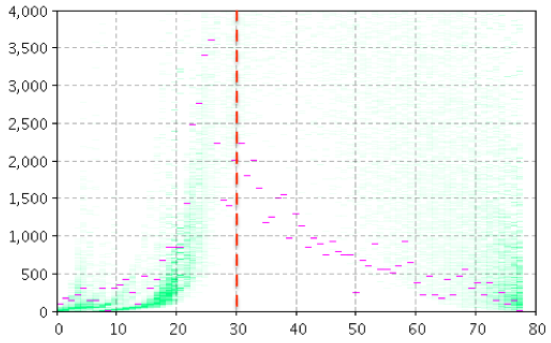
5.4.2 Limitations

The work presented here suffers from significant limitations. Although search trend data provides some indication of topic-specific interest over time in a defined spatial region, from the standpoint of big data, it is often available only with modest (daily) temporal resolution and frequently coarse geographic resolution. It is also affected by many unobserved confounders. Such search trend data is further limited by providing little sense of count of distinct users and no sense of longitudinal progression of a single user. In such regards, the Google search query volume time series compare unfavourably to the richness of information present in other publicly available types of online data, such as region-specific twitter feeds.

In addition to shortcomings in the data sources employed, there are notable methodological limitations of our study. The likelihood function employing two distinct data sources was simplistic in its design, merely serving to multiply each of the dataset-specific likelihood functions. The use of a random walk during particle filtering for no fewer than five distinct parameters likely contributes to a rapid divergence in the model's estimates, compared to the behaviour observed in the previous particle filtered models of influenza [15] and [16]. Further experimentation is required with the parameters governing such random walks. More significant yet, given the limited volatility likely for some of such parameters, a large gain in accuracy may



(a) Clinical data and particle distribution. Red markers depict empirical data points (clinical empirical data) and blue depict the 2D histogram of samples from the particle filtered model (reported incidence).



(b) Google search-volume data and particle distribution. Magenta markers depict empirical data points (Google empirical data) and green depict the 2D histogram of samples from the particle filtered model (search volume).

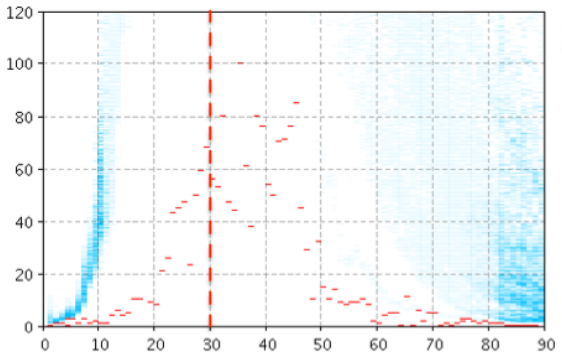
Figure 5.5: Empirical data (red and magenta points) superimposed on samples (blue and green) from the model-generated distribution of particles for the model output for the count of reported cases (left panel) and count of searches (right panel) using the likelihood function associated with clinical data alone, $T^* = 30$ for Quebec.

come from treating such parameters as unknown constants to be sampled for a given simulation from a posterior distribution within Particle Markov Chain Monte Carlo (PMCMC) techniques [92].

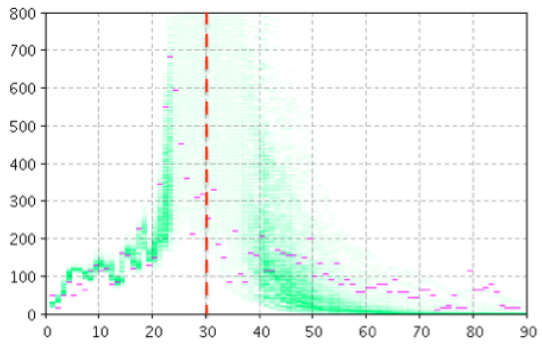
Such limitations point to natural avenues for future work. We expect that the prospects for the sorts of projections explored here will be significantly elevated by combining such data with other public data sources containing distinct sources of information, such as daily or finer resolution time series from Twitter and Tumblr. We further expect the accuracy of the projections to be improved by more powerful machine learning techniques, such as through the use of PMCMC techniques, ensemble techniques supporting inclusion of multiple models, and potentially PMCMC techniques employing multiple models using reverse-jump MCMC strategies.

5.4.3 Conclusion

Pandemic forecasting is important for public health policy making by virtue of its support for judicious planning involving resource allocation. Official statistics typically capture only subsets of the epidemiological burden (e.g., the subset of individuals who engage in care-seeking). Prospects for rapid use of such data to understand outbreak evolution are often further handicapped by reporting delays and a lack of capacity to project epidemiological case count time series forward. Traditional outbreak data have been complemented in recent years by high-resolution datasets from public social media such as Twitter, Tumblr, and time series provided by the Google search API via Google trends and Google flu trends that can be retrieved programmatically and analyzed over time. The results presented in this work suggest that, when combined with traditional epidemiological data sources, social media-drive datasets, machine learning and dynamic



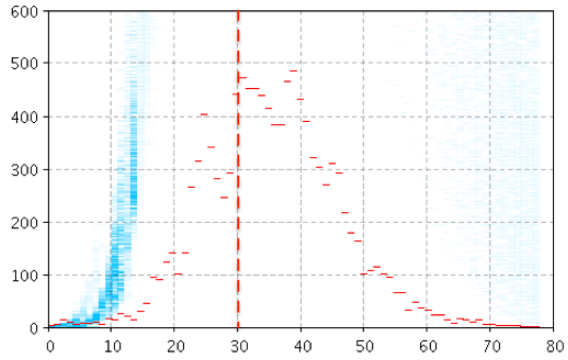
(a) Clinical data and particle distribution. Red markers depict empirical data points (clinical empirical data) and blue depict the 2D histogram of samples from the particle filtered model (reported incidence).



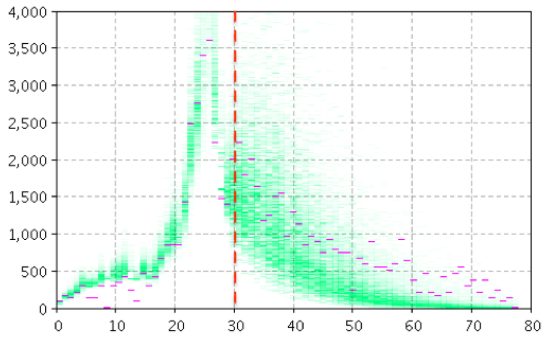
(b) Google search-volume data and particle distribution. Magenta markers depict empirical data points (Google empirical data) and green depict the 2D histogram of samples from the particle filtered model (search volume).

Figure 5.6: 6 Empirical data (red and magenta points) superimposed on samples (blue and green) from the model-generated distribution of particles for the model output for the count of reported cases (left panel) and count of searches (right panel) when using the likelihood function associated with search-volume data alone, $T^* = 30$ for Manitoba.

modeling can offer powerful tools for anticipating future evolution of and assessing intervention tradeoffs with respect to infectious disease outbreaks, particularly for emerging pathogens.



(a) Clinical data and particle distribution. Red markers depict empirical data points (clinical empirical data) and blue depict the 2D histogram of samples from the particle filtered model (reported incidence).



(b) Google search-volume data and particle distribution. Magenta markers depict empirical data points (Google empirical data) and green depict the 2D histogram of samples from the particle filtered model (search volume).

Figure 5.7: 6 Empirical data (red and magenta points) superimposed on samples (blue and green) from the model-generated distribution of particles for the model output for the count of reported cases (left panel) and count of searches (right panel) when using the likelihood function associated with search-volume data alone, $T^* = 30$ for Quebec.

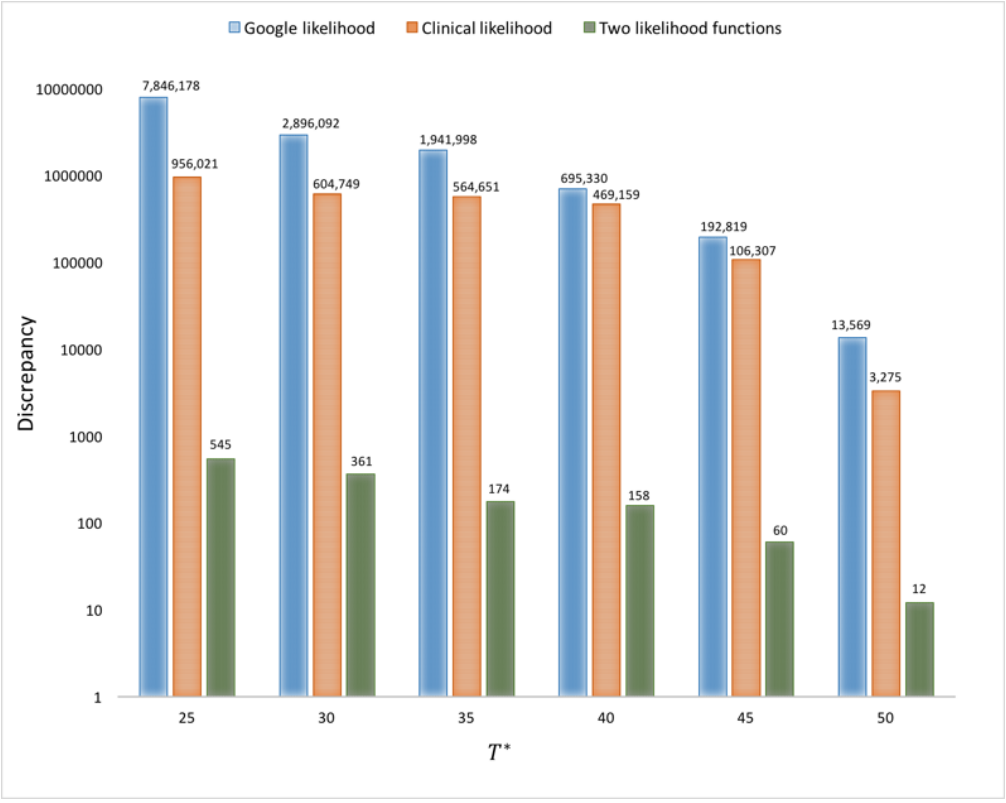


Figure 5.8: Discrepancies associated with different scenarios and different T^* for Manitoba.

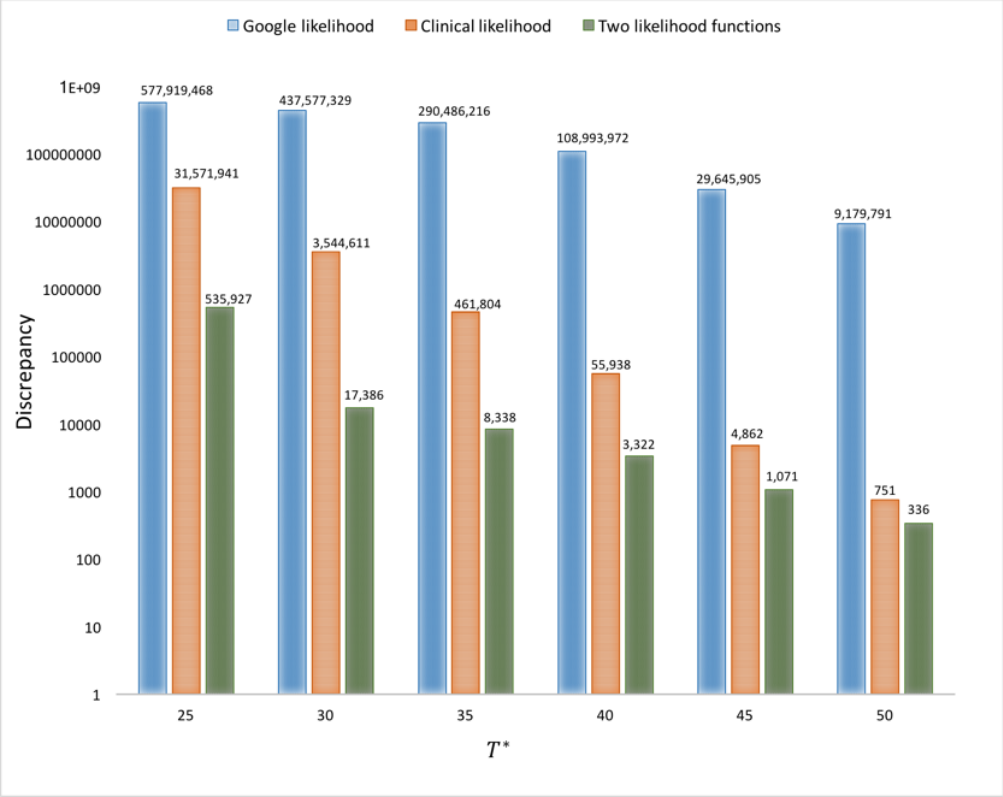


Figure 5.9: Discrepancies associated with different scenarios and different T^* for Quebec.

Chapter 6

Conclusion & Future Work

This thesis described a novel approach for influenza outbreak projection that combines dynamic models with empirical data to better capture outbreak dynamics. Moreover, the thesis demonstrates how traditional surveillance data sources can be complemented with time series characterizing communicational behaviour to inform pandemic influenza models considering adaptive behaviours based on fear. This chapter will provide an overview of thesis contributions and highlight potential directions for future work.

6.1 Summary of Findings

6.1.1 Particle Filtering in a SEIRV Simulation Model

We combined a traditional system dynamics model of epidemics with the sequential Monte Carlo method of particle filtering to enable the model to consider a daily timeseries of reported clinical case counts and correct the model latent state as new observations become available. Particle filtering contributed in estimating the model states as well as evolving model parameters. We evaluated the predictive accuracy of the particle filtered model and compared it with that associated with the calibrated version of the model. The particle filtered model helps overcome that the widespread difficulty of dynamic models that cannot keep current with the latest available empirical data, and supports adaption to evolution in stochastic parameters. In the calibrated model, these parameters are treated as static, and the calibrated model consequently failed to accurately predict long-term projections.

6.1.2 Predictive Accuracy of Particle Filtering in Dynamic Models Supporting Outbreak Projections

We investigated how ranges of parameters values for particle filtered influenza dynamic models influence their predictive accuracy. Factors examined include the frequency of sampling from observations, the dispersion parameter associated with the negative binomial distribution used in particle filtering, and the volatility of stochastically evolving parameters, such as the contact rate associated with the dynamic model. We found that more frequent data observations lead to markedly improved predictions, with a doubling of the sampling rate reducing the discrepancy by more than a factor of two. Our results further suggest that particle filtering

behaves in a robust manner with changes in dispersion parameter and the volatility with which the contact rate evolves across the model time horizon.

6.1.3 Social Media Surveillance Improves Outbreak Projection via Transmission Models

We investigated the degree to which considering daily search counts can enhance the prediction accuracy of particle filtered models. We adapted a previously developed compartmental model for use in particle filtering in such a way that the model can leverage daily search counts associated with the level of anxiety (for "fear") during a pandemic. We learned that combining particle filter and compartmental transmission models while using both clinical observations and search volumes can strongly improve predictive accuracy.

6.2 Contributions

The contributions of this thesis can be summarized into three main areas: 1) by serving as one of the first contributions using particle filtering together with influenza infection transmission models, 2) by serving as the first study to systematically investigate optimum ranges for configuration parameters for transmission models, 3) contributing the first investigation enhancing the results of particle filtering when used with a dynamic model and both traditional data and a member of new generation of electronic epidemiological data source – a search volume time series.

The first contribution was relatively novel as it was the first time that the applications of particle filtering was explored in a SEIRV model. It was also the first model that stochastic parameters were characterized to evolve based on observations of the real world data.

The second contribution was novel in terms of investigating the trade-off between employing more frequent but more noisy data samples and less frequent but less noisy sampling in particle filtering models. We also investigated the impact of varying parameters associated with behavioural change on model accuracy. In addition we sought the impact of different assumptions regarding observational error on particle filtering robustness.

The third contribution was a cutting edge work in terms of applying particle filtering to a coupled contagious (disease and fear) system dynamics model using Google search counts as an evidence of fear in population.

A key benefit of this work compared with the cited literature lies in its investigation of the capacity of particle filtering to estimate the hidden states of the system that cannot be directly measured. While there is a lack of direct empirical data on the values of those hidden states, the dynamics of such states are implied by the combination of the structure mathematical model of the system and the empirical data that is available – the time series of reported case counts and also the time-series of high-velocity online data sources such as GFT. In addition, stochastic parameters have been considered as states so that particle

filtering can aid in estimating these parameters. It bears emphasis that this approach is different than simple forecasting-based contributions in that it can be used in problems of simultaneous (joint) estimation of state and evolving parameters. A fundamental difference is that the particle filter relies upon a mechanistic model of the underlying system – including latent factors, such as the count of susceptibles – whereas simple time-series forecasting methods do not, as they often deal with simple extrapolation of observable quantities. Beyond allowing estimation to support an understanding of the broader state of the system, this difference commonly makes the particle filtering technique more effective in the prediction of uncertain time series in situations where a theory of the underlying dynamical system is available. The other striking difference between particle filtering algorithm and simple forecasting techniques is that particle filtering is involved with updating along with predicting, which enables particle filtering to perform online estimation, recursively – that is, allows particle filtering to update its estimate of the current state of the system as new data arrives. More importantly, the reliance of particle filtering on an underlying model and its capability to perform online estimation makes it suitable for investigating different intervention strategies. Particle filtering can estimate and then be used further to investigate the impact of different interventions by projecting their effects forward, as simulated, for example, by changing parameter values. This approach could serve as a valuable technique to assist public health authorities in estimating size and length of influenza outbreaks. The application of this technique can be extended to the transmission models of other pathogens with understudied dynamics. In the field of public health, the occurrence, development, and prevalence of diseases – communicable or not – can be regarded as faces of the dynamics of an underlying system. Because this underlying system is generally believed to be affected by uncertain factors and stochastics, its investigation in a filtering context can be an appropriate application for particle filtering method. In addition to models of diseases, particle filtering can further be applicable to other spheres of health and health care, including health service delivery (such as the operation of emergency departments), health workforce planning, etc. According to the findings in this thesis, it has been examined that even rough models can take advantage of machine learning methods, clinical data and search data to project evolution of outbreaks at relatively early stages. The robustness of the methods examined here to changes in the parameters associated with that technique suggests that a variant of the methods presented in this work can be used to estimate unknown parameters associated with dynamic models of future emerging outbreaks. We hope that estimating the latent states of the models can further help public health policy makers and their analysts in conducting "what if" scenarios characterizing the effects of interventions and assessing the effects of those interventions.

6.3 Future Work

There are multiple possible research directions that could improve our contributions in this thesis. In this section, we will discuss a few of the limitations in our model and approach. Further studies in these areas contribute marked enhancement to the work presented here.

It will be important to explore system dynamics models which consider other features of a pandemic, such as social behaviour of different age groups. Because of past exposure to different strains of flu, individuals may additionally be subject to differential susceptibility to new flu outbreaks. While capturing this dynamic will require substantial structural changes to the models – essentially, creation of subscripts that capture different combinations of exposure to past infections or vaccines – it may lead to developing more precise models of population response to influenza. It will further make this work more applicable to decision making in public health if the effects of interventions could be assessed by running "what if" scenarios and adding different options of treatment, such as prophylactic antivirals to decrease susceptibility, antiviral administration to those infected to lower the risk of complications and (slightly) the duration of symptoms, and treatment with adjuvanted vaccines. All such interventions could be represented at a basic level with variations of the models examined in this thesis. To develop a decision support system and apply this approach to resolve real-world problems and evaluate different intervention strategies, it will be important to consider the tools that are required to facilitate data gathering on an ongoing – rather than an episodic or one-time – basis. To accomplish this, it is desirable to investigate solutions for streaming data into an "online" model that recursively updates weights of particles when data from a new observation or set of observations arrives. For such a purpose, various models can be used for each task; e.g. one model can be used purely to estimate the current state, another to project forward that state for some time period, and yet other models can be used to project forward the effects of different particular interventions. If necessary, the models could further be restarted periodically, incorporating all data to that point (probably from a database). Further, the models would further have to be equipped with an interface that would allow a decision maker, policy analyst, medical health officer, or other health professional to easily specify hypothesized distributions regarding the initial state, to update or supplement records of observations or change assumed values of parameter values, to run the models in an "online" mode, or to run projection and counterfactual scenarios forward from the current time.

Investigating the performance of other types of probability distributions in particle filtering would also represent a valuable research contribution. Combining particle filtering with other types of models used in public health area, such as agent-based modeling can be valuable. Seeking software and hardware acceleration techniques will be especially important when used with agent-based particle filtering, because of the heavy load imposed by such approaches

To improve the predictive accuracy of models, it will be important to move beyond search query volumes to leverage the vast amount of data from social media platforms such as Facebook, Twitter, and Tumblr for pandemic prediction purposes –self-publishing platforms such as Twitter and Tumblr are specifically attractive due to lower ethics concerns– and to improve the data harvesting methods for a more reliable and higher resolution geographic-specific datasets.

The best design for likelihood functions employing two distinct data sources can be investigated. Further experimentation is required with the stochastic parameters associated with random walks, specifically with

those parameters with high volatility. It will further be important to examine a combination of Markov Chain Monte Carlo methods with sequential Monte Carlo methods in the form of PMCMC. PMCMC can support rapid learning incorporating different lines of evidence related to a system to estimate the system states and principle parameters, and predict system trends.

6.4 Conclusion

Our original hypothesis stated that joining mathematical models and empirical data can improve the predictive accuracy of projection models. In order to investigate this hypothesis, we sought to adapt compartmental models by enabling them to learn from real-time observations. We combined particle filtering with different adapted compartmental models and investigated the effectiveness of different datasets as tools to inform the model regarding the real world. Our results demonstrated that particle filtering in combination with simple dynamic models and particularly in the presence of reporting and high temporal resolution online communicational behaviour data can support robust and accurate projections, and estimation of the latent state of the compartmental model, thereby opening the opportunity for investigation of alternative intervention schemes.

References

[1] "Centers for Disease Control and Prevention." `https://en.wikipedia.org/wiki/2009_flu_pandemic_by_country`. [online; Accessed July 2018.].

[2] "Google Trends." `https://trends.google.com/trends/explore?q=flu&geo=CA`. [online; Accessed July 2018.].

[3] "Google Flu Trends." `http://www.google.org/flutrends/about/data/flu/ca/data.txt`. [online; Accessed July 2018.].

[4] D. Lazer, R. Kennedy, G. King, and A. Vespignani, "Google Flu Trends still appears sick: An evaluation of the 2013-2014 flu season," *SSRN Electronic Journal*, vol. 1, no. 1, pp. 1–12, 2014.

[5] "News Wire." `https://www.newswire.ca/news-releases`. [online; Accessed July 2018.].

[6] B. Cooper, "Poxy models and rash decisions," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 103, no. 33, pp. 12221–12222, 2006.

[7] J. Wallinga and M. Lipsitch, "How generation intervals shape the relationship between growth rates and reproductive numbers," *Proceedings - Royal Society. Biological Sciences*, vol. 274, no. 1609, p. 599604, 2007.

[8] J. Lloyd-Smith, S. Funk, A. Mclean, S. Riley, and J. Wood, "Nine challenges in modelling the emergence of novel pathogens," *Epidemics*, vol. 10, no. 2015, pp. 35–39, 2014.

[9] H. W. Hethcote, "The mathematics of infectious diseases," *Society for Industrial and Applied Mathematics*, vol. 42, no. 4, pp. 599–653, 2000.

[10] P. Magall and W. G, "The parameter identification problem for sir epidemic models: identifying unreported cases," *Journal of Mathematical Biology*, vol. 1, no. 6, pp. 1–20, 2018.

[11] M. C. J. Bootsma and N. M. Ferguson, "The effect of public health measures on the 1918 influenza pandemic in u.s. cities," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 104, no. 18, p. 75887593, 2007.

[12] S. Cauchemez, A. J. Valleron, P. Y. Pierre-Yves Boelle, A. Flahault, and N. M. Ferguson1, "Estimating the impact of school closure on influenza transmission from sentinel data," *Nature*, vol. 452, no. 1, pp. 750–755, 2008.

[13] J. T. Wu, B. J. Cowling, E. H. Y. Lau, D. k. M. Ip, L. M. Ho, T. Tsang, S. K. Chuang, P. Y. Leung, S. V. Lo, S. H. Liu, and S. Riley, "School closure and mitigation of pandemic (H1N1) 2009, hong kong," *Emerging Infectious Diseases*, vol. 16, no. 3, pp. 538–541, 2009.

[14] J. Epstein, J. Parker, D. Cummings, and R. Hammond, "Coupled contagion dynamics of fear and disease: Mathematical and computational explorations," *PLoS One*, vol. 3, no. 12, p. e3955, 2008.

[15] A. Safarishahrbijari, T. Lawrence, R. Lomotey, J. Liu, C. Waldner, and N. D. Osgood, "Particle filtering in a seirv simulation model of H1N1 influenza," in *Proceedings of the 2015 Winter Simulation Conference: 6-9 December; Huntington Beach* (L. Yilmaz, ed.), pp. 1240–1251, IEEE Xplore, 2015.

[16] A. Safarishahrbijari, A. Teyhouee, C. Waldner, J. Liu, and N. D. Osgood, "Predictive accuracy of particle filtering in dynamic models supporting outbreak projections," *BMC Infectious Diseases*, vol. 17, p. 648, 2017.

[17] N. J. Gordon, D. J. Salmond, and A. F. M. Smith, "Novel approach to nonlinear/non-gaussian bayesian state estimation," *IEEE Proceedings F Radar and Signal Processing*, vol. 140, no. 2, p. 107113, 1993.

[18] J. K. Lee and C. Jekeli, "Rao-blackwellized unscented particle filter for a handheld unexploded ordnance geolocation system using imu/gps," *Proceedings of the ASME 2017 International Design Engineering Technical Conferences and Computers and Information in Engineering Conference*, vol. 64, no. 2, pp. 327–340, 2011.

[19] F. Evennou and F. Marx, "Advanced integration of wifi and inertial navigation systems for indoor mobile positioning," *EURASIP Journal on Advances in Signal Processing*, vol. 64, no. 2, pp. 327–340, 2011.

[20] E. Keshavarzi, M. McIntire, and C. Hoyle, "A dynamic design approach using the kalman filter for uncertainty management," *Artificial Intelligence for Engineering Design, Analysis and Manufacturing*, vol. 31, no. 2, pp. 161–172, 2017.

[21] E. Keshavarzi, M. McIntire, T. Y. Tumer, G. K, and C. Hoyle, "Resilient system design using cost-risk analysis with functional models," *Proceedings of the ASME 2017 International Design Engineering Technical Conferences and Computers and Information in Engineering Conference*, vol. 31, no. 2, pp. 1362–1369, 2017.

[22] J. S. Liu and R. Chen, "Sequential monte carlo methods for dynamic systems," *Journal of the American Statistical Association*, vol. 93, no. 443, pp. 1032–1044, 1998.

[23] A. Doucet, S. Godsill, and C. Andrieu, "On sequential monte carlo sampling methods for bayesian filtering," *Statistics and Computing*, vol. 10, no. 0, p. 197208, 2000.

[24] P. W. Glynn and D. L. Iglehart, "Importance sampling for stochastic simulations," *Management Science*, vol. 35, no. 11, pp. 1367–1392, 1989.

[25] P. W. Glynn and D. L. Iglehart, "An overview of existing methods and recent advances in sequential monte carlo," *Proceedings of the IEEE*, vol. 95, no. 5, pp. 899–924, 2007.

[26] A. Doucet and A. M. Johnsen, "A tutorial on particle filtering and smoothing: Fifteen years later." `https://www.stats.ox.ac.uk/~doucet/doucet_johansen_tutorialPF2011.pdf`. [online; Accessed March 2018.].

[27] A. Doucet, N. de Freitas, and N. Gordon, "An introduction to sequential monte carlo methods," *Journal of the Royal Statistical Society*, vol. 52, no. 4, pp. 694–695, 2003.

[28] H. Akashi and K. H, "A random sampling approach to state estimation in switching environments," *Automatica*, vol. 13, no. 0, p. 429434, 1977.

[29] V. S. Zaritskii, V. B. Svetnik, and L. I. Shimelevich, "Monte carlo technique in problems of optimal data processing," *Automation and Remote Control*, vol. 12, no. 0, pp. 95–103, 1975.

[30] K. P. Murphy, *Machine Learning: A Probabilistic Perspective*. Cambridge: MIT Press, 2012.

[31] N. D. Osgood and J. Liu, "Bayesian parameter estimation of system dynamics models using markov chain monte carlo methods: An informal introduction," in *Proceedings of the 30th International Conference of the System Dynamics Society: 22-26 June 2013; New York*, pp. 1391–14008, Curran Associates, 2013.

[32] M. Israd and A. Blake, "Condensationconditional density propagation for visual tracking," *Int J Comput Vis*, vol. 29, pp. 5–28, 1998.

[33] D. N. Osgood and J. Liu, "Towards closed loop modeling: evaluating the prospects for creating recurrently regrounded aggregate simulation models," in *Proceedings of the 2014 Winter Simulation Conference: 7-10 December; Savannah* (A. Tolk, ed.), pp. 829–941, IEEE Xplore, 2014.

[34] J. A. Wilde, J. A. McMillan, J. Serwint, J. Butta, M. A. O'Riordan, and M. Steinhoff, "Effectiveness of influenza vaccine in health care professionalsa randomized trial," *Journal of the American Medical Association*, vol. 28, no. 10, pp. 908–913, 1999.

[35] "Centers for Disease Control, 2009 H1N1 flu." `https://www.cdc.gov/h1n1flu/`. [online; Accessed July 2018.].

[36] "Centers for Disease Control, services and Information." `https://www.canada.ca/en/public-health/services/diseases/flu-influenza.html`. [online; Accessed August 2018.].

[37] S. Blower and D. Bernoulli, "An attempt at a new analysis of the mortality caused by smallpox and of the advantages of inoculation to prevent it," *Reviews in Medical Virology*, vol. 14, no. 5, pp. 275–288, 2004.

[38] Ross, *The prevention of malaria*. London: John Murray, 1911.

[39] W. O. Kermack and A. G. McKendrick, "Contribution to the mathematical theory of epidemics," *Proceedings of the Royal Society of London. Series A, Containing Papers of a Mathematical and Physical Character*, vol. 115, no. 1, pp. 700–721, 1927.

[40] W. O. Kermack and A. G. McKendrick, "Contribution to the mathematical theory of epidemics, part ii," *Proceedings of the Royal Society of London. Series A, Containing Papers of a Mathematical and Physical Character*, vol. 138, no. 1, pp. 55–83, 1932.

[41] W. O. Kermack and A. G. McKendrick, "Contribution to the mathematical theory of epidemics, part iii," *Proceedings of the Royal Society of London. Series A, Containing Papers of a Mathematical and Physical Character*, vol. 141, no. 1, pp. 94–112, 1933.

[42] L. Siettos C I, Russo, "Mathematical modeling of infectious disease dynamics," *Virulence*, vol. 4, no. 4, p. 295306, 2013.

[43] J. Gaudart, O. Toure, N. Dessay, A. L. Dicko, S. Ranque, and L. Forest, "Modelling malaria incidence with environmental dependency in a locality of sudanese savannah area, mali," *Malaria Journal*, vol. 8, no. 1, pp. 61–72, 2009.

[44] M. Ajelli, L. Fumanelli, P. Manfredi, and S. Merler, "Spatiotemporal dynamics of viral hepatitis a in italy," *Malaria Journal*, vol. 8, no. 1, pp. 1–11, 2011.

[45] J. Demongeot, J. Gaudart, J. Mintsa, and M. Rachdi, "Demography in epidemics modelling," *Communications on Pure and Applied Analysis*, vol. 11, no. 1, pp. 61–82, 2012.

[46] D. Bishai, B. Johns, D. Nair, J. Nabyonga-Orem, B. Fiona-Makmot, and E. Simons, "The cost-effectiveness of supplementary immunization activities for measles: a stochastic model for uganda," *Journal of Infectious Diseases*, vol. 204, no. 1, pp. 107–115, 2011.

[47] R. H. Wang, Z. Jin, Q. X. Liu, J. van de Koppel, and D. Alonso, "A simple stochastic model with environmental transmission explains multi-year periodicity in outbreaks of avian flu," *PLoS One*, vol. 7, no. 2, p. e28873, 2012.

[48] S. Eubank, H. Guclu, V. Kumar, M. Marathe, A. Srinivasan, and Z. D. Toroczkai, "Modelling disease outbreaks in realistic urban social network," *Nature*, vol. 429, no. 6988, pp. 180–184, 2004.

[49] N. Ferguson, D. Cummings, S. Cauchemez, C. Fraser, S. Riley, and A. Meeyai, "Strategies for containing an emerging influenza pandemic in Southeast Asia," *Nature*, vol. 437, no. 7056, pp. 209–214, 2005.

[50] D. S. Burke, J. M. Epstein, D. A. Cummings, J. I. Parker, K. C. Cline, and R. M. Singa, "Individual-based computational modeling of smallpox epidemic control strategies," *Academic Emergency Medicine*, vol. 13, no. 11, pp. 1142–1149, 2006.

[51] D. Balcan, V. Colizza, B. Gonalves, H. Hu, J. J. Ramasco, and A. M. Vespignani, "Multiscale mobility networks and the spatial spreading of infectious diseases," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 106, no. 51, pp. 121484–121489, 2009.

[52] A. F. Dugas, M. Jalalpour, Y. Gel, S. Levin, F. Torcaso, T. Igusa, and R. Rothman, "Influenza forecasting with google flu trends," *J Public Health Inform*, vol. 5, no. 1, p. pp.4470, 2013.

[53] S. Pollett, W. J. Boscardin, E. Azziz-Baumgartner, Y. O. Tinoco, G. Soto, C. Romero, J. Kok, M. Biggerstaff, C. Viboud, and G. W. Rutherford, "Evaluating google flu trends in latin america: Important lessons for the next phase of digital disease detection," *Clin Infect Dis*, vol. 64, no. 1, p. pp.3441, 2017.

[54] O. M. Araz, D. Bentley, and R. L. Muelleman, "Using google flu trends data in forecasting influenza-likeillness related ed visits in omaha, nebraska," *Am J Emerg Med*, vol. 32, no. 9, p. pp.10161023, 2014.

[55] L. H. Thompson, M. T. Malik, A. Gumel, T. Strome, and S. M. Mahmud, "Emergency department and google flu trends data as syndromic surveillance indicators for seasonal influenza," *Epidemiol Infect*, vol. 142, no. 11, p. pp.23972405, 2014.

[56] M. Santillana, A. T. Nguyen, M. Dredze, M. J. Paul, E. O. Nsoesie, and J. S. Brownstein, "Combining search, social media, and traditional data sources to improve influenza surveillance," *PLOS Comput Biol*, vol. 11, no. 10, p. pp.e1004513, 2014.

[57] J. D. Sharpe, R. S. Hopkins, R. L. Cook, and C. W. Striley, "Evaluating google, twitter, and wikipedia as tools for influenza surveillance using bayesian change point analysis: A comparative analysis," *JMIR Public Heal Surveill*, vol. 2, no. 2, p. pp.e161, 2016.

[58] "H1N1 flu in manitoba H1N1 report fall 2010. 2009." `http://www.gov.mb.ca/health/documents/h1n1.pdf`. [online; Accessed March 2018.].

[59] "Brousseau n. bilan pidmiologique de la pandmie dinfluenza a(H1N1)." `https://www.inspq.qc.ca/sites/default/files/publications/1212_bilanah1n12009.pdf`. [online; Accessed March 2018.].

[60] T. Lee and H. Shin, "Combining syndromic surveillance and ili data using particle filter for epidemic state estimation," *Flexible Services and Manufacturing Journal*, vol. 28, no. 1, pp. 233–253, 2014.

[61] J. B. S. Ong, I. Mark, C. Chen, A. R. Cook, H. C. Lee, V. J. Lee, R. T. Lin, P. A. Tambyah, and L. G. Goh, "Real-time epidemic monitoring and forecasting of H1N1-2009 using influenza-like illness from general practice and family doctor clinics in singapore," *PloS ONE*, vol. 5, no. 4, pp. 1–21, 2010.

[62] L. H. Chyi, *Evaluation of Real-Time Methods for Epidemic Forecasting*. PhD thesis, National University of Singapore, Department of Statistics and Applied Probability, 2011.

[63] H. Manchanda, N. Seidel, A. Krumbholz, S. A, M. Schmidtke, and R. Guthke, "Within-host influenza dynamics: a small-scale mathematical modeling approach," *Biosystems*, vol. 118, no. 1, p. 5159, 2014.

[64] S. Z. Chao, G. Y. Zheng, D. Hong, M. R. Qing, and Q. Xiao Gang, "The research of influenza H1N1's transmission based on artificial society," *International Journal of Modeling and Optimization*, vol. 4, no. 2, pp. 95–99, 2014.

[65] M. Shubin, M. Virtanen, S. Toikkanen, O. O. Lyytikainen, and K. Auranen, "Estimating the burden of a(H1N1)pdm09 influenza in finland during two seasons," *Epidemiology and Infection*, vol. 142, no. 5, pp. 964–974, 2013.

[66] P. Pongsumpun and I. M. Tang, "Dynamics of new strain of the H1N1 influenza a virus incorporating the effects of repetitive contacts," *Computational and Mathematical Methods in Medicine*, vol. 2014, no. 1, pp. 964–974, 2014.

[67] P. J. Birrell, K. Georgios, N. J. Gay, B. S. Cooper, A. M. Presanis, R. J. Harris, A. Charlett, X. Zhang, P. J. White, R. G. Pebody, and D. De Angelis, "Bayesian modeling to unmask and predict influenza a/H1N1pdm dynamics in london," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 108, no. 45, p. 1823818243, 2011.

[68] J. M. Conway, A. R. Tuite, D. N. Fisman, N. Hupert, R. Meza, B. Davoudi, K. English, P. Van Den Driessche, F. Brauer, J. Ma, L. A. Meyers, M. Smieja, A. Greer, D. M. Skowronski, D. L. Buckeridge, J. C. Kwong, J. Wu, S. M. Moghadas, D. Coombs, R. C. Brunham, and B. Pourbohloul, "Vaccination against 2009 pandemic H1N1 in a population dynamical model of vancouver, canada: timing is everything," *BMC Public Health*, vol. 11, no. 1, pp. 932–932, 2011.

[69] A. Tuite, D. N. Fisman, J. C. Kwong, and A. Greer, "Optimal pandemic influenza vaccine allocation strategies for the canadian population," *PLoS Currents Influenza*, vol. 2, no. 1, pp. RRN1144 –RRN1144, 2010.

[70] B. J. Coburn, B. G. Wagner, and S. Blower, "Modeling influenza epidemics and pandemics: insights into the future of swine flu (H1N1)," *BMC Medicine*, vol. 7, no. 3, pp. 1741–1748, 2009.

[71] J. M. Tchuenche, N. Dube, C. P. Bhunu, J. Smith, and C. T. Bauch, "The impact of media coverage on the transmission dynamics of human influenza," *BMC Public Health*, vol. 11, no. 1, pp. 5–20, 2011.

[72] S. Kim, J. Lee, and E. Jung, "Mathematical model of transmission dynamics and optimal control strategies for 2009 A/H1N1 influenza in the Republic of Korea," *Journal of Theoretical Biology*, vol. 412, no. 1, p. 7485, 2017.

[73] "Estimates of Population, for July 1, Provinces and territories." `http://www5.statcan.gc.ca/cansim/a47`. [online; Accessed March 2017.].

[74] "WHO, Influenza (Seasonal)." `http://www.who.int/mediacentre/factsheets/fs211/en`. [online; Accessed July 2018.].

[75] P. Huston, "Thinking locally about pandemic influenza," *Can J Public Health*, vol. 95, pp. 184–185, 2004.

[76] Y. Ibuka, G. B. Chapman, L. A. Meyers, M. Li, and A. P. Galvani, "The dynamics of risk perceptions and precautionary behavior in response to 2009 H1N1 pandemic influenza," *BMC Infect Dis*, vol. 10, pp. 296–306, 2010.

[77] J. W. Rudge, P. Hanvoravongchai, R. Krumkamp, I. Chavez, W. Adisasmito, P. NgocChao, B. Phommasak, W. Putthasri, C. S. Shih, M. Stein, A. Timen, S. Touch, R. Reintjes, and R. Coker, "Health system resource gaps and associated mortality from pandemic influenza across six asian territories (health systems and pandemic influenza in asia)," *PLoS ONE*, vol. 7, p. e31800, 2012.

[78] D. L. Chao, M. E. Halloran, V. J. Obenchain, and I. M. Longini, "Flute, a publicly available stochastic influenza epidemic simulation model," *PLoS Comput Biol*, vol. 6, p. e1000656, 2010.

[79] D. L. Chao, L. Matrajt, N. E. Basta, J. D. Sugimoto, B. Dean, D. A. Bagwell, B. Oiulfstad, M. E. Halloran, and I. M. Longini, "Planning for the control of pandemic influenza a H1N1 in los angeles county and the united states," *Am J Epidemiol*, vol. 173, pp. 1121–1130, 2011.

[80] K. Lee, A. Agrawal, and A. Choudhary, "Mining social media streams to improve public health allergy surveillance," *Proceedings of the 2015 IEEE/ACM International Conference on advances in social networks analysis and mining*, vol. 2015, no. 1, pp. pp.815–822, 2015.

[81] M. Keeling, "The implications of network structure for epidemic dynamics," *Theor Popul Biol*, vol. 67, pp. 1–8, 2005.

[82] M. Hashemian, W. Qian, K. G. Stanley, and N. D. Osgood, "Temporal aggregation impacts on epidemiological simulations employing microcontact data," *BMC MED INFORM DECIS*, vol. 12, pp. 132–146, 2012.

[83] A. Machens, G. F, R. C, T. A. E, B. A, and C. C, "An infectious disease model on empirical networks of human contact: bridging the gap between dynamic network data and contact matrices," *BMC Infect Dis*, vol. 13, pp. 185–199, 2013.

[84] I. S. Mbalawata, S. Sarkka, and H. Haario, "Parameter estimation in stochastic differential equations with markov chain monte carlo and nonlinear kalman filtering," *Comput Stat*, vol. 28, pp. 1195–1223, 2012.

[85] W. K. Hastings, "Monte carlo sampling methods using markov chains and their applications," *Biometrika*, vol. 57, pp. 97–109, 1970.

[86] F. C. Coelho, C. T. Codeco, and M. G. M. Gomes, "A bayesian framework for parameter estimation in dynamical models," *PloS One*, vol. 6, p. e19616, 2011.

[87] A. Gelb, *Applied Optimal Estimation*. Cambridge: MIT Press, 1974.

[88] W. Qian, N. D. Osgood, and K. G. Stanley, "Integrating epidemiological modeling and surveillance data feeds: a kalman filter based approach," in *Proceedings of the seventh International Conference on Social Computing, Behavioral-Cultural Modeling and Prediction: 1-4 April; Washington DC* (W. G. Kennedy, ed.), pp. 145–152, Springer, 2014.

[89] M. Chiogna and C. Gaetan, "Hierarchical space-time modelling of epidemic dynamics: an application to measles outbreaks," *Stat Method App*, vol. 13, pp. 55–71, 2004.

[90] B. Cazelles and N. Chau, "Using the kalman filter and dynamic models to assess the changing hiv/aids epidemic," *Math Biosci*, vol. 140, pp. 131–154, 1997.

[91] M. Chiogna and C. Gaetan, "Dynamic generalized linear models with application to environmental epidemiology," *J R Stat Soc*, vol. 51, pp. 453–468, 2002.

[92] C. Andrieu, A. Doucet, and R. Holenstein, "Particle markov chain monte carlo methods (with discussion)," *J R Stat Soc*, vol. 72, pp. 269–342, 2010.

[93] I. Dorigatti, S. Cauchemez, A. Pugliese, and N. M. Ferguson, "A new approach to characterising infectious disease transmission dynamics from sentinel surveillance: application to the italian 20092010 a/H1N1 influenza pandemic," *Epidemics*, vol. 4, pp. 9–21, 2012.

[94] R. Oraji, V. Hoeppner, A. Safarishahrbijari, and N. D. Osgood, "Combining particle filtering and transmission modeling for tb control," in *Proceedings of the 2016 International Conference on Health Informatics: 4-7 October; Chicago* (A. Tolk, ed.), pp. 829–941, IEEE Xplore, 2014.

[95] H. Stark and J. Woods, *Probability and Random Processes with Applications to Signal Processing*. New Jersey: Prentice Hall, 2002.

[96] J. M. Hilbe, *Negative Binomial Regression*. Cambridge: Cambridge University Press, 2011.

[97] Stata Corp, College Station, *Stata: Release 14. Statistical Software*, 2015.

[98] K. Kruger and N. D. Osgood, "Particle filtering using agent-based transmission models," in *Proceedings of the 2015 Winter Simulation Conference: 6-9 December; Huntington Beach* (L. Yilmaz, ed.), pp. 737–747, IEEE Xplore, 2015.

[99] S. Tuarob, C. Tucker, M. Salathe, and N. Ram, "An ensemble heterogeneous classification methodology for discovering health-related knowledge in social media messages," *J Biomed Inform*, vol. 49, no. 10, p. 255268, 2014.

[100] X. Zhou, E. Coiera, G. Tsafnat, D. Arachi, M. S. Ong, and A. G. Dunn, "Using social connection information to improve opinion mining: Identifying negative sentiment about hpv vaccines on twitter," *Stud Health Technol Inform*, vol. 216, no. 1, pp. pp.761–765, 2015.

[101] A. J. Lazard, E. Scheinfeld, J. M. Bernhardt, G. B. Wilcox, and M. Suran, "Detecting themes of public concern: A text mining analysis of the centers for disease control and preventions ebola live twitter chat," *Am J Infect Control*, vol. 43, no. 10, p. pp.11091111, 2015.

[102] C. Allen, M. H. Tsou, A. Aslam, A. Nagel, and J. M. Gawron, "Applying gis and machine learning methods to twitter data for multiscale surveillance of influenza," *PLoS One*, vol. 11, no. 7, p. pp.e0157734, 2016.

[103] X. Li, A. Doroshenko, and N. D. Osgood, "Applying particle filtering in both aggregated and age-structured population compartmental models of pre-vaccination measles," *BMC Infectious Diseases*, vol. 49, no. 10, p. 255268, 2018.

Appendix A

The calibrated values of parameters are shown as bellow:

Variable name	Value	Units
Probability of infection transmission given exposure multiplied by contacts per week	4.8	1/Week
Mean latent time	0.278	Week
Fraction of reported incidents	0.001	Unit
Fraction initially susceptible	0.993	Unit
Fraction initially exposed	$9.2\,E-8$	Unit
Fraction initially infective	$1.59\,E-5$	Unit
Fraction initially recovered	0.005	Unit

Appendix B

Detailed information about initial values of compartmental states

S_0: Truncated normal distribution, Mean = 900000, Standard deviation = 150000, Lower bound = 0, Upper bound = N - I_0, Sample size = number of particles = 10000

E_0: 0 for all particles

I_0: 7 for all particles

R_0: N - $S_0 - E_0 - I_0 - V_0$

V_0: 0 for all particles

In this model, V class refers to those receiving vaccination during the pandemic (ongoing vaccination). Those being vaccinated prior to the second wave might be part of R class or S depending on vaccine efficacy. Since the initial values of R and S were unclear, we considered the initial values of these states as distributions.

Appendix C

The discrepancy of particle filtering predictions in frequency scenarios for different observation times and $\gamma = 0.125$ and $\gamma = 2$

Frequency scenarios ($\gamma = 0.125$)	$T^* = 35$	$T^* = 42$	$T^* = 49$	$T^* = 56$
PF using daily data, r=2	354	225	71	0
PF using three-day data, r=6	12109	5945	1593	181
PF using weekly data, r=14	68381	36313	6322	608
PF using daily data, r=8	381	210	44	0
PF using three-day data, r=24	12273	5655	1309	93
PF using weekly data, r=56	162378	40820	5670	476
PF using daily data, r=32	455	169	13	0
PF using three-day data, r=96	12808	4647	1125	90
PF using weekly data, r=224	153010	44106	5224	295

Table C.1: Discrepancy of particle filtering predictions in frequency scenarios for different observation times and γ= 0.125

Frequency scenarios ($\gamma = 2.0$)	$T^* = 35$	$T^* = 42$	$T^* = 49$	$T^* = 56$
PF using daily data, r=2	3327	695	87	0
PF using three-day data, r=6	43931	12590	1630	39
PF using weekly data, r=14	645037	154916	16362	976
PF using daily data, r=8	1568	241	18	0
PF using three-day data, r=24	35024	6251	682	4
PF using weekly data, r=56	1216215	129467	6072	376
PF using daily data, r=32	904	104	5	0
PF using three-day data, r=96	25452	4199	393	0
PF using weekly data, r=224	1243398	129629	4580	254

Table C.2: Discrepancy of particle filtering predictions in frequency scenarios for different observation times and γ= 2.0