SANDIA REPORT

SAND2016-abcd
Unlimited Release
Printed September 2016

Online Mapping and Forecasting of Epidemics Using Open-Source Indicators

J. Ray, S. Lefantzi, J. Bauer, M. Khalil, A. Rothfuss, K. R. Cauthen, P. D. Finley and H. Smith

Prepared by
Sandia National Laboratories
Albuquerque, New Mexico 87185 and Livermore, California 94550

Sandia National Laboratories is a multi-program laboratory managed and operated by Sandia Corporation, a wholly owned subsidiary of Lockheed Martin Corporation, for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-AC04-94AL85000.

Sandia National Laboratories

Issued by Sandia National Laboratories, operated for the United States Department of Energy by Sandia Corporation.

Available to DOE and DOE contractors from
U.S. Department of Energy
Office of Scientific and Technical Information
P.O. Box 62
Oak Ridge, TN 37831

Telephone:	(865) 576-8401
Facsimile:	(865) 576-5728
E-Mail:	reports@adonis.osti.gov
Online ordering:	http://www.osti.gov/bridge

Available to the public from
U.S. Department of Commerce
National Technical Information Service
5285 Port Royal Rd
Springfield, VA 22161

Telephone:	(800) 553-6847
Facsimile:	(703) 605-6900
E-Mail:	orders@ntis.fedworld.gov
Online ordering:	http://www.ntis.gov/help/ordermethods.asp?loc=7-4-0#online

SAND2016-abcd
Unlimited Release
Printed September 2016

Online Mapping and Forecasting of Epidemics Using Open-Source Indicators

J. Ray, S. Lefantzi, J. Bauer, M. Khalil, A. Rothfuss
Sandia National Laboratories, P. O. Box 969, Livermore CA 94551

K. R. Cauthen, P. D. Finley, H. Smith
Sandia National Laboratories, P. O. Box 5800, Albuquerque NM 87185-0751

G. Lambert
Apple Inc, 1 Results Way, Cupertino, CA 95014

{jairay,slefant,josbaue,mkhalil,arothfu,kcauthe,pdfinley,hsmith}@sandia.gov,
gregory_lambert@apple.com

Abstract

Open-source indicators have been proposed as a way of tracking and forecasting disease outbreaks. Some, such are meteorological data, are readily available as reanalysis products. Others, such as those derived from our online behavior (web searches, media article etc.) are gathered easily and are more timely than public health reporting. In this study we investigate how these datastreams may be combined to provide useful epidemiological information. The investigation is performed by building data assimilation systems to track influenza in California and dengue in India. The first does not suffer from incomplete data and was chosen to explore disease modeling needs. The second explores the case when observational data is sparse and disease modeling complexities are beside the point. The two test cases are for opposite ends of the disease tracking spectrum.

We find that data assimilation systems that produce disease activity maps can be constructed. Further, being able to combine multiple open-source datastreams is a necessity as any one individually is not very informative. The data assimilation systems have very little in common except that they contain disease models, calibration algorithms and some ability to impute missing data. Thus while the data assimilation systems share the goal for accurate forecasting, they are practically designed to compensate for the shortcomings of the datastreams. Thus we expect them to be disease and location-specific.

Acknowledgment

This work was funded under LDRD (Laboratory Directed Research and Development) Project Number 173112 and Title "Online Mapping and Forecasting of Epidemics Using Open-Source Indicators". Sandia National Laboratories is a multi-program laboratory managed and operated by Sandia Corporation, a wholly owned subsidiary of Lockheed Martin Corporation, for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-AC04-94AL85000.

Contents

1 Motivation, hypothesis and tests 9

2 Data assimilation for influenza 11

2.1 Introduction . . . 11

2.2 Materials and Methods . . . 13

2.2.1 Data . . . 13

2.2.2 Temporal Prediction . . . 13

2.2.3 Spatial Prediction . . . 15

2.3 Results . . . 17

2.4 Conclusions . . . 19

3 Data assimilation for dengue 25

3.1 Introduction . . . 25

3.2 Materials and methods . . . 26

3.2.1 Data . . . 26

3.2.2 Conditionally auto-regressive models . . . 27

3.2.3 Boosting . . . 28

3.3 Results . . . 28

3.4 Conclusions . . . 29

4 Follow-on applications 33

4.1 Data assimilation for wildfires . . . 33

4.2 Data assimilation and disease modeling . . . 35

5 Conclusions 37

References 38

List of Figures

2.1 Data assimilation for San Francisco for 2014-2015 influenza season, starting September 21, 2014. At the very top, we plot $V^{(o)}$ (symbols) and V; the data is available every week. In the second plot, we illustrate the inferred evolution of $I(t)$. The third plot shows the estimate of $\beta(t)$ and the plot at the very bottom shows the convergence of the value of τ over time measured in days. 16

2.2 17

2.3 Left: Error between $Y^{(o)}$ and $\bar{Y}$, the modeled forecast value of ILI+, normalized by $Y^{(o)}$ i.e., $\eta = (\bar{Y} - Y^{(o)})/Y^{(o)}$. Here $\bar{Y}$ is the week-ahead mean forecast. The horizontal solid green lines are the $\pm 10\%$ error bounds. Each symbol denotes one of the 11 Californian cities tracked by GFT. Right: The same test of forecasting accuracy, but $\bar{Y}$ is the two-week-ahead predictions. Results are for the 2014-2015 influenza season. The start and end of the influenza season in California is denoted by the dashed blue line, and spans January to March. 18

2.4 Left: Error between $Y^{(o)}$ and $\bar{Y}$, the modeled forecast value of ILI+, normalized by 3ς. Here $\bar{Y}$ is the week ahead forecast. The horizontal solid green lines are show whether the observed data fall within the 99% credibility interval. Each symbol denotes one of the 11 Californian cities tracked by GFT. Right: The same test of forecasting accuracy, but performed for two-week-ahead predictions. Results are for the 2014-2015 influenza season. The start and end of the influenza season in California is denoted by the dashed blue line, and spans January to March. 19

2.5 Nowcast ILI+ intensity map, computed using the spatial prediction method described in Sec. 2.2.3. 20

2.6 Test of accuracy of the spatial prediction described in Sec. 2.2.3. On the left, we nowcast ILI+ activity in San Mateo County in the San Francisco Bay Area (solid black line) using ILI+ data from nearby municipalities (dashed lines). On the right, we perform a spatiotemporal forecast, using one-week-ahead forecasts of ILI+ activity of nearby municipalities. Vertical green lines show the approximate start and end of the intense influenza activity in 2013-2014. 21

2.7 ILI+ intensity maps developed for the last week of March 2015. Left: A nowcast. Right: A one-week-ahead forecast. 21

2.8 A snapshot of the output of the data assimilation system, displayed as a web page on an internal Sandia server. 22

3.1 Plots of the binned counts of HM articles on dengue for India for October 2011 (left), 2012 (middle) and 2013 (right). The blank states recorded no data. States with data are shaded with a color corresponding to the lower bound of their bin. We see missing data occurs at random. 27

3.2 A comparison of the predictive error for the boosted (red) and non-boosted STCAR for filling in missing data. Results are plotted for the 15 states where there is some HM data. Overall boosting improves performance 29

3.3 Top: The raw HealthMap data with gaps in it. Bottom: Filled in version of the HealthMap dataset. 30

3.4 Left: Forecast of the HM data, for the state of Kerala. The blue dot is the mean prediction, with the shaded error being the 90% and 99% credibility intervals. The red dot is the true (not filled in) value. Right: The same, but for the state of Maharashtra. 31

4.1 The wildfire after two hours. No observational data have been assimilated. The spread in forecast fire fronts (red contours) are due to our ignorance of wind and initial conditions. The blue contour is the true fire front. 34

4.2 The wildfire after five hours. Data assimilation has occurred and reduces the uncertainty spread. The snapshot shows results displayed in SUMMIT's web client. The red contours are the 5 hour forecasts without data assimilation. A very narrow ensemble of purple fire fronts, very near the blue (true) fire front (and difficult to see) is the forecast ensemble, after assimilating data available after two hours. 35

4.3 The wildfire after five hours, in detail. The uncertainty spreads show risk to human habitation and reinforce the need for data assimilation and probabilistic forecasting using filtering methods. The purple ensemble is easy to see and shows the enormous decrease in forecasting uncertainty effected by the assimilation of data after two hours. 36

Chapter 1

Motivation, hypothesis and tests

Fast, dependable forecasting of disease activity can revolutionize medical planning and response. Collection of public health (PH) data, traditionally used for this purpose, is slow and thus not useful for effective response. Due to its voluntary nature, epidemiological reporting typically has irregular and incomplete spatial coverage. Thus, real- time mapping and forecasting of epidemiological activity is still not feasible.

Online, open-source indicators (OSI) of disease activity e.g., disease-related searches, media reports etc. and meteorology can serve as strong covariates and leading indicators of outbreaks. They are readily available, timely, and have far superior spatiotemporal resolution than PH data, especially in developing countries. Currently there are few data assimilation (DA) methods that can fuse disparate datastreams to compensate for delayed/unavailable PH data, nor meteorology-driven disease models for accurate spatiotemporal forecasting. We propose to develop the methods and models and integrate them into a DA framework. Such a framework would be invaluable for disease tracking in the US and globally.

The key hypothesis behind this study is that OSI are sufficiently rich to calibrate a high-resolution spatial representation of disease activity, modeled on weather patterns. Within the DA framework, the spatial model will interpolate sparse disease data. OSI are noisy datastreams, and the spatial model will allow noise suppression by pooling of information across monitored sites (generally large cities). The spatial model, along with the meteorology-driven disease model, will allow OSI-calibrated forecasts in regions outside OSI coverage. Scalable ensemble Kalman filters will provide the mathematical underpinnings of data fusion so that the framework can be applied to country-sized problems. The game-changing potential of data assimilation has not been applied to disease forecasting, because it has relied on sparse PH data. OSI, and our data assimilation framework, would be a novel development with impact in data- poor regions.

We will demonstrate this via a three step process. First, we will develop a data assimilation system to perform spatiotemporal forecasting of influenza in California, where data is plentiful. Next we will develop a system for tracking the evolution of the annual dengue outbreak in India using OSI data from HealthMap (`https://www.healthmap.org`). This second effort is expected raise data issues (e.g., missing data) which are not expected in case of Californian influenza. We conclude with a test of generality. We show how the algorithms and capabilities developed during the construction of the data assimilation systems can be leveraged in other problems, especially where the data worth of open-source data streams (or the models that ingest them) have to be assessed.

Chapter 2

Data assimilation for influenza

2.1 Introduction

Seasonal influenza results in between a quarter to half a million deaths worldwide every year, with 3-5 million cases of severe illness [1]. Many countries have influenza surveillance networks of sentinel physicians who track cases of influenza-like illness (ILI, defined as a fever over 37.8°C plus cough and/or sore throat; patients for whom the etiology is known not to be influenza are not classified as ILI) ; for example, United States' Center for Disease Control (CDC) Outpatient Influenza-like Illness Surveillance Network (ILINet) consists of a group of 2900 outpatient healthcare professionals who voluntarily provide information on total and ILI-related visits that they receive. The data is compiled, processed and publicly reported on a weekly basis [2], with a 1-2 week delay [3]. Note that since ILI symptoms can be caused by a number of diseases other than influenza e.g., rhinovirus and respiratory syncytial virus, CDC also reports data on laboratory samples testing positive for influenza at the national and regional levels [4]. Note that many countries and indeed, many counties in the United States itself, do not have such a comprehensive influenza surveillance infrastructure.

Epidemiological outbreaks leave an imprint on our online lives as we search for information on the disease. Consequently, there have been attempts to track disease activity using web search query logs [3, 5], Twitter posts [6, 7], Wikipedia article views [8, 9] and clinician and medical databases [10]. In most cases, the underlying hypothesis in these digital disease detection (DDD) techniques is that the intensity of disease activity is correlated with the intensity of the activity in the diverse digital datastreams. Statistical models have been devised to relate ILINet data (taken as the ground truth of ILI activity) to the easily observed digital proxies. Since these digital proxies are timely, they are used to "nowcast" the current influenza/ILI activity level, 1-2 weeks ahead of the release of ILINet data.

Google Flu Trends (GFT; [3]) is one of the oldest and better known DDD efforts to track ILI. Originally, it used a set of 45 keywords to determine if a Google search was related to ILI and gathered a time-series of the volume of ILI-related searches. This time-series was found to be correlated to ILINet data. A model was built by regressing ILINet data to a normalized version of the web search time-series. The method was dependent on the choice of keywords whose usage could evolve with time; consequently, GFT has often been wrong in its nowcasts [11, 12]. It has been revised a few times [13, 14] and was discontinued in mid-2015. The revisions were mainly about changing/extending the set of keywords used to pick ILI-related web searches. There have been attempts to improve on this basic method - in [15, 16] the authors hypothesized that the picking power of diverse keywords was variable (and also changed over time) and determined weights for various keywords by regressing ILINet data to time-series of web search queries picked by each of the keywords. They used shrinkage regression to eliminate keywords with negligible predictive power and performed this calibration for different time periods. They also released the set of keywords. Other studies have hypothesized that DDD datastreams (GFT, medical databases, microblog posts etc.) are weak predictors of ILI activity but could be combined, in a weighted manner, to increase their predictive power; studies that do so in a variety of ways can be found in [17, 18, 19].

One need not depend on digital proxies for forecasting influenza activity - Tamerius et al [20] showed that relative humidity and temperature were good predictors of influenza activity, though their effect was muted in the tropics. Soebiyanto et al [21] modeled influenza activity in Hong Kong, using precipitation, temperature and relative humidity as the predictors. In [22, 23], the authors used a SIRS (susceptible-infected-recovered-susceptible) model of influenza, with an absolute humidity-based basic reproduction number, to forecast influenza (*not* ILI) activity in approximately 100 US municipalities using GFT data. They used an ensemble adjustment Kalman filter to calibrate the humidity-dependent SIRS model to GFT data, modified (using ILINet's data on laboratory samples testing positive for influenza) to reflect influenza, not ILI activity (see [24] for a description). They produced 1-4 week-ahead forecasts. Cities on the US East Coast were predicted better than the West Coast municipalities. The Bayesian nature of the assimilation allowed them to also estimate various disease characteristics e.g., incubation period, and produce forecasts as Gaussian distributions, thus capturing the estimation/prediction uncertainties. Particle filters have also been used for this purpose [25, 23].

Epidemics also display spatial patterns i.e., epidemiological activity at nearby locations tend to be similar. This may be due to population mixing or similarity in latent epidemiological factors such as demographics, socioeconomic conditions etc. Disease data is usually areal in nature i.e., it is collected at the municipal/provincial/national scale and correlation in disease activity in neighboring areal units are used to develop disease maps or fill-in missing data. A review of disease mapping techniques is in [26]. Typically, a measure of disease activity e.g., death rate, in an areal unit is modeled using a deterministic and a random term. The deterministic term is generally modeled as a regression to underlying latent factors. The random term can be a multivariate normal distribution [27] or a conditional autoregressive model [28, 29, 30]. The model can also include a temporal autoregressive term to capture the time evolution of an outbreak [31], though there have been recent efforts to model the entire spatio-temporal dataset as a Gaussian process [32]. The same concepts have been used to capture the spatiotemporal patterns on ILI behavior. In [33] a SIRS model for influenza was applied to all 50 states of the US, which included a parameterized model of inter- and intra-state population mixing and a disease spread rate that depended on population density and summer and winter temperatures. These parameters were estimated by fitting their model to GFT data, which was available for each state. In [34] the authors noticed that non-contiguous areas in US (which are well-linked via air travel) showed correlated ILI activity, as captured using GFT. Using historical GFT data, they developed a correlation matrix whose structure was modeled using airline network data in the US. The covariance matrix so obtained was used to constrain/modulate current GFT predictions. The model performed better than GFT data when compared against laboratory-confirmed influenza case.

To summarize, GFT and other digital proxies of disease activity are approximate predictors of ILI and influenza activity. They can be combined into stronger predictors and have been jointly assimilated with meteorological data to provide good temporal predictions of disease evolution. There seems to be little work on using the known dependence of influenza incidence on specific humidity and temperature to perform spatial prediction despite the availability of reanalysis products at fine spatiotemporal resolutions e.g., National Land Data Assimilation System [35, 36]. Coupled with temporal forecasts at locations where GFT data is available, such disease mapping techniques applied to DDD datastreams hold the potential to provide forecasts of disease activity in regions where such data is not collected. In this paper, we present a spatio-temporal prediction technique to do so, and test it in the San Francisco Bay Area.

2.2 Materials and Methods

In this section we formulate the spatial and temporal prediction problems. We also describe the observational data that we used in this work.

2.2.1 Data

The work employed meteorological data and open-source indicators of influenza activity downloaded from the Web. We used 2-meter-above-ground estimates of air temperature and specific humidity, as extracted from the National Land Data Assimilation System (NLDAS) Project-2 [35]. It provides reanalysis products at an hourly resolution on a 0.125 degree grid. Data is available since 1979 for the continental US. The data was time-averaged over each day, before constructing a daily climatology over 1990-2010. Data for each municipality was then extracted and used in Sec. 2.2.3.

We used Google Flu Trends (GFT) as an estimate for ILI activity. The data provides a measure of the number of cases with ILI symptoms in every 100,000 physician visits for the cities of San Francisco, Oakland, Berkeley, San Jose and Sunnyvale (henceforth, the SFBA cities). We will refer to this set of municipalities as $\mathbb{N}$. We omit Santa Clara (available in the dataset), as it seems to have an anomalously low level of ILI activity. The data spans September 2003 to August 2015. The model used to generate GFT was developed in 2008 [3], and updated in 2009, 2013, and 2014. The data used contains 2009-model estimates up until July 2013; 2013-model estimates from Aug. 2013 to July 2014; and 2014-model estimates from Aug. 2014 onward; it was downloaded from [37]. We convert the ILI cases into influenza cases in the manner described in [23], by multiplying the GFT values by the fraction of laboratory samples (from patients with ILI symptoms) that test positive for influenza. Borrowing the terminology from [23], we will refer to this estimate as "ILI+". This data on laboratory samples testing positive for influenza is distributed, in the form of a time-series of weekly resolution, by the US Center for Disease Control (CDC) [38]. It has a two-week reporting lag. We used the values for CDC's Pacific census division which contains California.

2.2.2 Temporal Prediction

Our data assimilation scheme is similar to the one in [23] and many methodological details are shared. We provide a summary of our method below.

The model : The assimilation of ILI+ data is performed separately for each influenza season and each municipality. We assume a perfectly mixed population and uses an SIR (susceptible-infectious-removed) model.

$$\begin{aligned} \dot{S} &= -\tfrac{\beta(t)SI}{N} - \alpha; \quad \dot{I} = \frac{\beta(t)SI}{N} - \frac{I}{\tau} + \alpha \\ \dot{R} &= \tfrac{1}{\tau}; \qquad \dot{V} = \frac{\beta(t)SI}{N}. \end{aligned} \tag{2.1}$$

Here S, I and R are the susceptible, Infectious and Recovered cohorts. α is the number of infections imported into the municipality per week and is set to 1 infection every 10 days (as in [23]). Unlike [23], we ignore loss of immunity and re-introduction into the susceptible cohort, as the timescale of loss of immunity is difficult to estimate over a season and does not affect results much [23]. The set $Z = \{S(t), I(t), \beta(t), \tau\}$ is deemed unknown and estimated from ILI+ observational data. The variable $V(t)$ tracks the number of people turning sick over a week, a fraction of whom seek medical care and are thus captured in ILI+ data. An observed value of V, $V^{(o)}$, is derived from ILI+ data and is used to calibrate the SIR model. Note that unlike [22, 23], we do not impute a dependence of β on humidity. This decision arises from two results documented in [23]. First, the rather mild climate in the SFBA does not experience the large variations observed in the MidWest or US East Coast, where the data assimilation system (DAS) described in [23] did well. Secondly, the same forecasting system was seen to be less accurate on the West Coast. Consequently, we simplified the model and in the process removed two parameters that the DAS in [23] also estimates.

Observational data : ILI+ is measure of the number of influenza cases (per 100,000 physician visits per week) and is not analogous to I, the number of infectious (and symptomatic) cohort at any time t. This led us to define V which captures the new weekly cases of influenza. As in [23], we relate $V^{(o)}$ to ILI+ data as $V^{(o)} = \gamma Y^{(o)}$, where $Y^{(o)}$ denotes the ILI+ data. We tested $\gamma \in [2, 10]$ in data assimilation tests and $2 \leq \gamma \leq 4$ provided the best results. Following [23], we use $\gamma = 2.5$ for the results here. The observational error is modeled as a zero mean Gaussian whose variance for week k is modeled as in [22]:

$$\sigma(k)^2 = \left(1.0 \times 10^5 + \frac{\left(\frac{1}{3}\sum_{j=k-3}^{k-1} Y^{(o)}(j)\right)^2}{5} \right) I^2$$

where I is the number of infected people per 100,000 population.

The Bayesian filter : We use an ensemble transform Kalman filter (ETKF,[39]) to assimilate $V^{(o)}(t)$ and update $Z(t)$. Since $V^{(o)}(t)$ is a time-series, Z is updated sequentially, as data becomes available. The process is called filtering. Sequential filtering computes the probability density over state $Z(t_e)$, at time t_e, using $V^{(o)}(t), 0 \leq t \leq t_e$. By Bayes theorem,

$$\begin{aligned} f(Z(t_e)|V^{(o)}(t_e), V^{(o)}(t_e-1), \cdots) \quad \propto \quad & \underbrace{f(V^{(o)}(t_e)|Z(t_e))}_{\text{Likelihood}} \\ & \underbrace{f(Z(t_e)|V^{(o)}(t_e-1), V^{(o)}(t_e-2), \cdots)}_{\text{Prior at } t_e}. \end{aligned} \tag{2.2}$$

The left-hand term is the posterior density for $Z(t_e)$ that includes the information observed up to $V^{(o)}(t_e)$ while the last term is our prior belief of $V^{(o)}$ based on previously observed data and a model prediction for $t = t_e$. Kalman filters model $f(:|:)$ as Gaussian distributions, reducing the problem to the estimation and evolution of the mean and covariance of the distributions. Ensemble filtering do not construct the covariance matrix explicitly, but rather preserve and evolve an ensemble of samples drawn from the prior distribution. The precise manner in which the samples are evolved/updated with the information in $V^{(o)}$ sets the various ensemble filtering methods apart.

As the ensemble of Z is updated over time, the correlated behavior of the elements of Z becomes apparent. This information is present in the entire ensemble which consists of possible realizations of model states, conditional on the $V^{(o)}$ assimilated. Rigorous methods for assimilating $V^{(o)}(t)$ and modifying all elements of Z depend on knowledge of this co-varying behavior of Z's elements; observations for every element of Z are not required.

Kalman filters update the ensemble in a manner such that only the mean and covariance are correctly evolved. Kalman filters are optimal only for linear problems, but are quite efficient for weakly nonlinear models like Eq. 2.1. Further, ETKF can be implemented independent of the SIR model; further, the SIR model does not have to supply information on the co-variability of the elements of Z e.g., gradients, to the ETKF. We used the public-domain implementation of ETKF in [40].

Running the filter : The filter's ensemble consists of 250 members, each driving an SIR model. We start with a population $N = 100,000$. The influenza season starts every year on week 40, but influenza activity in California does not reach significant levels before mid-December. We start our data assimilation process on week 35, using the most conservative starting point from [23]. Since the initial condition i.e., $Z(t = 0)$ is not known, we populate the ensemble from samples drawn from our prior belief regarding $Z(t = 0)$. We initialize the number of infected cases as $I(t = 0) \sim \mathcal{N}(100, 25^2)$, where $\mathcal{N}(\mu, \sigma^2)$ denotes a Gaussian distribution with mean μ and standard deviation σ. The initialization for other variables in Z are: $\tau \sim \mathcal{N}(0.2, 0.2^2)$, $\beta(t = 0) \sim \mathcal{N}(0.8, 0.16^2)$ and $S(t = 0) = N - I(t = 0)$. This prior distribution, as represented by the 250 members, is integrated for a week, after which $V^{(o)}$ is assimilated to update the vector Z. This provides the posterior distribution for Z at the end of the week. The filter can sometimes collapse i.e., the variability of the members in the ensemble can become spuriously low, at which point the influence of $V^{(o)}$ on the update of Z become weak and the filter can diverge from the true epidemic trajectory. We correct this by inflating the variance of I by a factor of 1.05. The inflated ensemble then becomes the prior ensemble for the following week. An n-week-ahead forecast at the end of week k is made by simply running the posterior ensemble forward for n weeks without any filter updates.

2.2.3 Spatial Prediction

We assume that influenza activity in a given municipality is a function of certain meteorological variables which govern virus survival and transmission [20, 41, 42] and a discrepancy δ that accounts for epidemiological processes not completely governed by meteorology. Thus for any week k, the ILI+ data for a municipality $\mathbf{x}$, $Y^{(o)}(t, \mathbf{x})$ is given by:

$$\begin{aligned} Y^{(o)}(k, \mathbf{x}) = M(T, Q; W) + \delta(k, \mathbf{x}) \quad = \quad & w_0 + w_0^{(T)} T(k, \mathbf{x}) + w_1^{(T)} T(k-1, \mathbf{x}) + \\ & w_0^{(Q)} Q(k, \mathbf{x}) + w_1^{(Q)} Q(k-1, \mathbf{x}) + \\ & \delta(k, \mathbf{x}), \end{aligned} \tag{2.3}$$

where $T(k, \mathbf{x})$ and $Q(k, \mathbf{x})$ are weekly averaged temperatures and specific humidity at 2 meters above ground level. The weekly averages are computed using the daily climatologies described in Sec. 2.2.1. $w_l^{(j)}, j \in \{T, Q\}, l \in \{0, 1\}$ are weights. This linear model is fitted via step-wise regression, and simplified using bidirectional elimination and Akaike Information Criterion. $Y^{(o)}(t, \mathbf{x})$ data spans 2006 to 2011, starting mid-year. The process is repeated for all municipalities. The regression minimizes the norm of $\delta(k, \mathbf{x})$, over the learning period, for each municipality. The step-wise regression removes the lagged variables, yielding a weight vector $W = \{w_0, w_0^{(T)}, w_0^{(Q)}\}$. Further, these weights are different for each municipality i.e., the weights $W(\mathbf{x})$ are a function of location.

In order to predict ILI+ at any arbitrary location $\mathbf{x}^*$, the coefficients $W(\mathbf{x}^*)$ and discrepancy $\delta(t, \mathbf{x}^*)$ have to be spatially predicted using the computed values $W(\mathbf{x}_i)$ and $\delta(t, \mathbf{x}_i), i \in \mathbb{N}$. We perform this prediction using a Nadaraya-Watson smoother [43], using a Gaussian kernel. For any spatial quantity $\phi(\mathbf{x}^*)$, an approximation

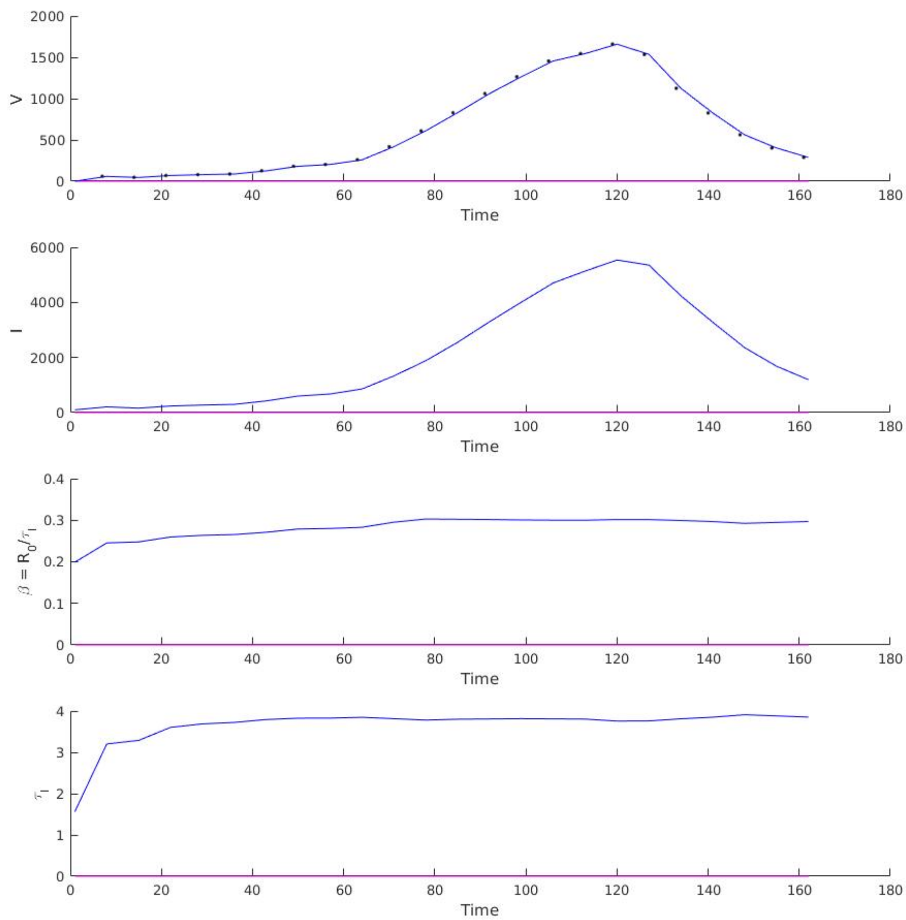

Figure 2.1: Data assimilation for San Francisco for 2014-2015 influenza season, starting September 21, 2014. At the very top, we plot $V^{(o)}$ (symbols) and V; the data is available every week. In the second plot, we illustrate the inferred evolution of $I(t)$. The third plot shows the estimate of $\beta(t)$ and the plot at the very bottom shows the convergence of the value of τ over time measured in days.

$\widehat{\phi(\mathbf{x}^*)}$ is computed as

$$\widehat{\phi(\mathbf{x}^*)} = \sum_{i \in \mathbb{N}} K\left(\frac{\|\mathbf{x}^* - \mathbf{x}_i\|}{\lambda}\right)\phi(\mathbf{x}_i), \tag{2.4}$$

where $\|\mathbf{x}^* - \mathbf{x}_i\|$ is the great-circle distance between two locations $\mathbf{x}^*$ and $\mathbf{x}_i$, and $K(:;\lambda)$ is the smoothing kernel, with λ as its length-scale. We compute an optimal λ from data via leave-one-out cross-validation. Eq. 2.4 is used to obtain a spatially predicted $\widehat{W(\mathbf{x}^*)}$ and $\widehat{\delta(t, \mathbf{x}^*)}$ as approximations to $W(\mathbf{x})$ and $\delta(t, \mathbf{x})$ in Eq. 2.3, and thus obtain a prediction $\widehat{Y^{(o)}(t, \mathbf{x}^*)}$.

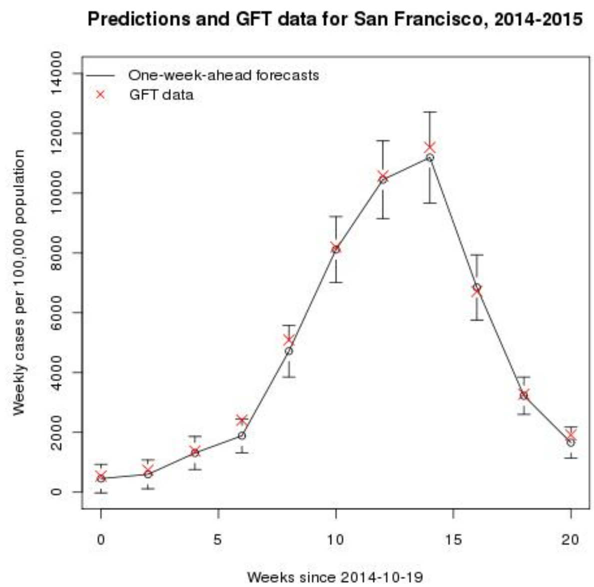

Figure 2.2

2.3 Results

We first illustrate the solution of Eq. 2.1 using an ETKF with $\sigma(k)^2$ acting as the variance of the Gaussian observational error for time indexed by k days. Fig. 2.1 plots the data assimilation for San Francisco for the 2014-2015 influenza season. The assimilation is started on September 21, 2014, which corresponds to $k = 1$ on the time (horizontal) axis. In the top plot we observe $V^{(o)}$, the number of influenza cases per 100,000 physician visits (as symbols) and the modeled value in Eq. 2.1. The infectious (and infected) cohort is plotted in the second figure and follows the same basic profile. In the last subplot in Fig. 2.1, we see that the value of τ, the infectious period for influenza, is inferred to be about 4 days. Note that these values are the means over 250 members of the ensemble. In Fig. 2.2 we plot GFT (*not* ILI+) and its (modeled and) one-week-ahead forecast value using the ETKF. The line denotes the mean prediction, and the error bars the $\pm 3\varsigma$ bounds (ς is the standard deviation of the forecasts produced by the 250 members of the ensemble). We see the mean agrees well with the GFT data from which ILI+ and $V^{(o)}$ is derived. However, there is considerable scatter/uncertainty in ensemble as ς in the prediction is quite significant.

Next we perform a check for forecasting accuracy for all the 11 CA cities tracked by GFT. In Fig. 2.3 we plot $\xi = (\bar{Y} - Y^{(o)})/Y^{(o)}$ as a function of time for the 2014-2015 influenza season. The data assimilation starts on September 21, 2014. Here $\bar{Y}$ is the mean of the 250 forecasts produced by the ensemble. Each city is denoted by a symbol. Tho horizontal lines denote the $\pm 10\%$ error bounds. We see that early in the season (when there is not much of an influenza outbreak signal in the ILI+ data), the SIR model neither calibrates nor forecasts well for all Californian cities. However, after late December and till March, the mean forecasts are quite accurate, with less than 10% forecasting error (for one-week-ahead forecasts, Fig. 2.3 (left)). If one increases the forecasting horizon to two weeks, the accuracy degrades but is within 20% error (green horizontal dashed lines).

In Fig. 2.4 we plot $\eta = (\bar{Y} - Y^{(o)})/3\varsigma$, where ς is the standard deviation of the 250 forecasts produced by the ensemble. We see that the deviation between mean forecast and observations lie between the $\pm 3\varsigma$ bounds

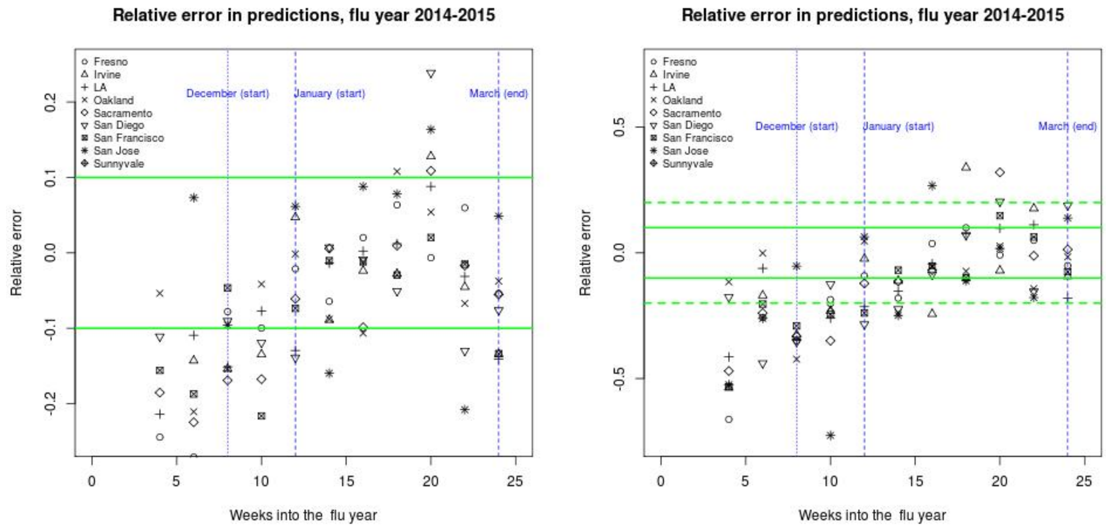

Figure 2.3: Left: Error between $Y^{(o)}$ and $\bar{Y}$, the modeled forecast value of ILI+, normalized by $Y^{(o)}$ i.e., $\eta = (\bar{Y} - Y^{(o)})/Y^{(o)}$. Here $\bar{Y}$ is the week-ahead mean forecast. The horizontal solid green lines are the $\pm 10\%$ error bounds. Each symbol denotes one of the 11 Californian cities tracked by GFT. Right: The same test of forecasting accuracy, but $\bar{Y}$ is the two-week-ahead predictions. Results are for the 2014-2015 influenza season. The start and end of the influenza season in California is denoted by the dashed blue line, and spans January to March.

between January and March when the outbreak signal is strong in the GFT data. Two-week-ahead forecasts are more accurate than one-week-ahead forecasts. Figs. 2.3 and 2.4 show that the mean forecasts are quite accurate and the predictive uncertainty bounds (ς) correctly bounds the prediction error.

Next we address spatial interpolation as described in Sec. 2.2.3. We apply the spatial prediction method to the ILI+ data from the last week of March, 2013 and plot the results in Fig. 2.5. The six SFBA municipalities that constitute $\mathbb{N}$ are plotted with red crosses. An ILI+ intensity is predicted for every grid cell and plotted, producing a map. The figure uses current ILI+ data (not forecasts) and consequently the figure contains a nowcast map. Note that the color map shows the number of cases per 100,000 physician visits; the actual number of case counts will be proportional to physician visits, which in turn should be proportional to population density.

Next, we check the accuracy of the spatial prediction. In Fig. 2.6 we plot estimates of ILI+ for San Mateo County (which is not tracked by GFT). Redwood City, San Mateo city and Daly City are the primary population centers of San Mateo County, with San Mateo city lying approximately in the center; consequently, its location (latitude/longitude) was used for $\mathbf{x}^*$ in Eq. 2.4. In Fig. 2.6, we plot the inferred ILI+ behavior for San Mateo County (thick black line) using ILI+ data from $\mathbb{N}$; these are plotted with dashed lines. The symbols are data (samples testing positive for influenza) reported by the San Mateo County public health department for the influenza year of 2013-2014. The figure on the left is a nowcast (no forecasting errors), whereas the figure on the right is computed using one-week-ahead forecasts (computed using the ETKF-based forecasting described in Sec. 2.2.2). We see that the predictions capture the trend seen in the public health data, though the results are better for nowcasts. The difference between the two figures captures the impact of forecasting error.

GFT tracks 11 Californian municipalities, which form three clusters - the SFBA (municipalities constituting

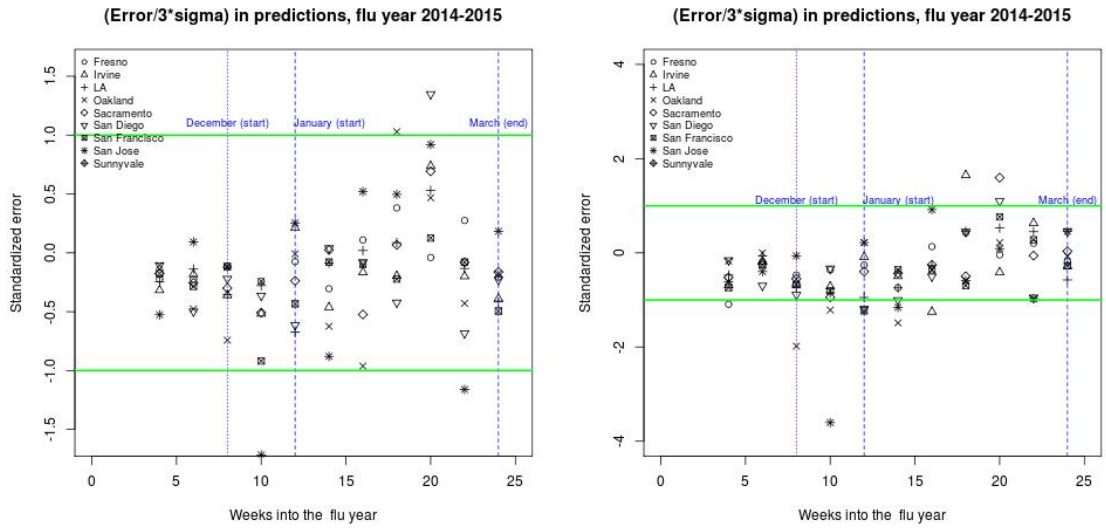

Figure 2.4: Left: Error between $Y^{(o)}$ and $\bar{Y}$, the modeled forecast value of ILI+, normalized by 3ς. Here $\bar{Y}$ is the week ahead forecast. The horizontal solid green lines are show whether the observed data fall within the 99% credibility interval. Each symbol denotes one of the 11 Californian cities tracked by GFT. Right: The same test of forecasting accuracy, but performed for two-week-ahead predictions. Results are for the 2014-2015 influenza season. The start and end of the influenza season in California is denoted by the dashed blue line, and spans January to March.

ℕ), the Los Angeles - San Diego corridor (including Irvine) and the Central Valley (Sacramento and Fresno). The spatiotemporal prediction described above for SFBA can be performed for the other two clusters, though they will certainly be less accurate due to the paucity of data. Fig. 2.7 shows such maps developed for the last week of March 2015. We do not have public health data from any municipality or county in these clusters and are unable to validate the predictions, unlike for SFBA.

Finally, the data assimilation system - temporal forecasting and spatial prediction - was implemented and run weekly (and automatically) every week between August 2014 and August 2015, when GFT ceased to provide data publicly. The prototypical implementation - a combination of Matlab, R and shell scripts - download GFT data, processed and assimilated them to produce nowcasts and one-week-ahead forecasts for SFBA. These predictions were displayed on an internal Sandia web page as zoom-able maps (implemented using JavaScript), and provided a quantitative measure of influenza activity around Sandia National Laboratories, Livermore, CA. A screen-shot of a map is in Fig 2.8.

2.4 Conclusions

In this study, we have investigated whether open-source indicators (OSI) can be used to track and forecast influenza activity in a small, well mixed population such as the San Francisco Bay Area (SFBA). This was motivated by the fact that OSI of disease activity that are based on our online behavior (or, in fact, any digital interaction/media) can be collected very quickly and could provide a far better measure of disease activity than data collected by health surveillance networks. Unfortunately, these OSI are proxies of disease activity, and may be inaccurate. Consequently, this is a drawback of the study presented here. However, the

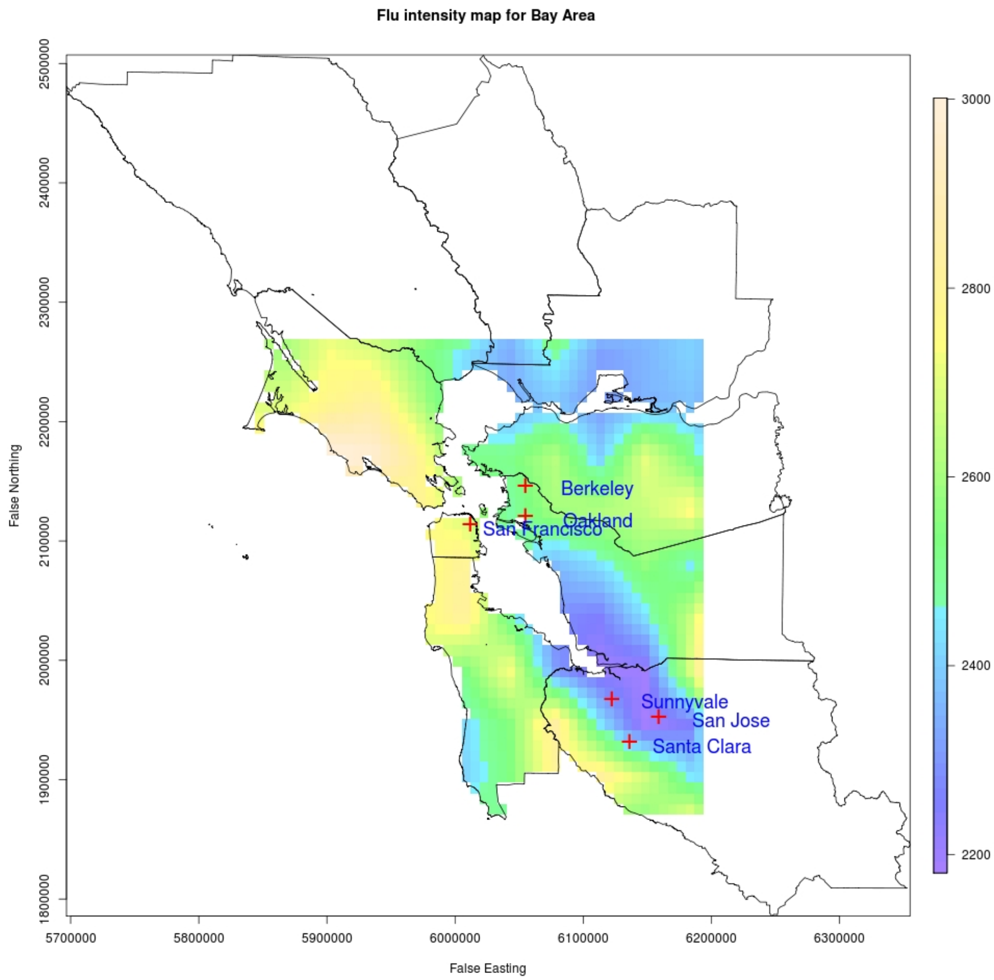

Figure 2.5: Nowcast ILI+ intensity map, computed using the spatial prediction method described in Sec. 2.2.3.

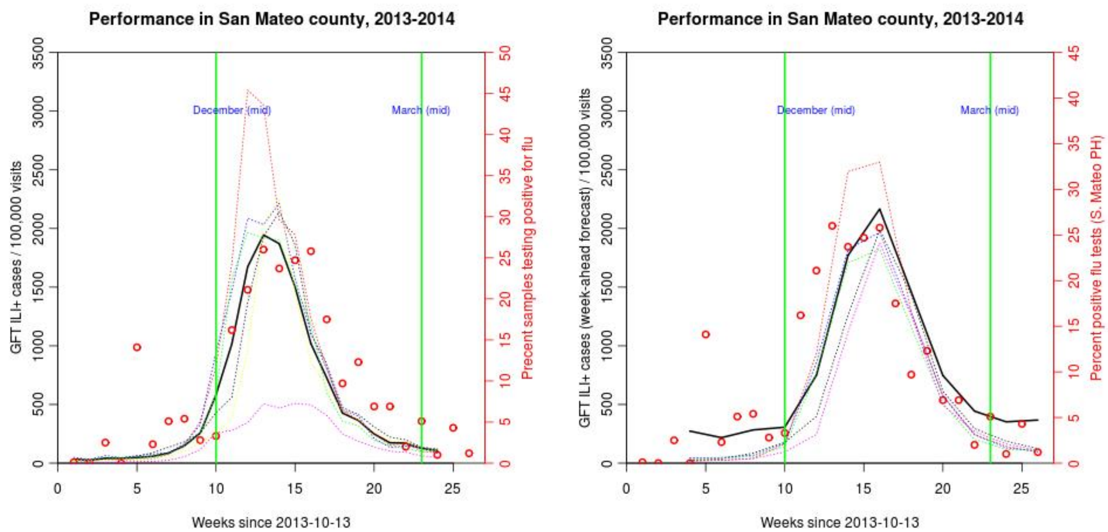

Figure 2.6: Test of accuracy of the spatial prediction described in Sec. 2.2.3. On the left, we nowcast ILI+ activity in San Mateo County in the San Francisco Bay Area (solid black line) using ILI+ data from nearby municipalities (dashed lines). On the right, we perform a spatiotemporal forecast, using one-week-ahead forecasts of ILI+ activity of nearby municipalities. Vertical green lines show the approximate start and end of the intense influenza activity in 2013-2014.

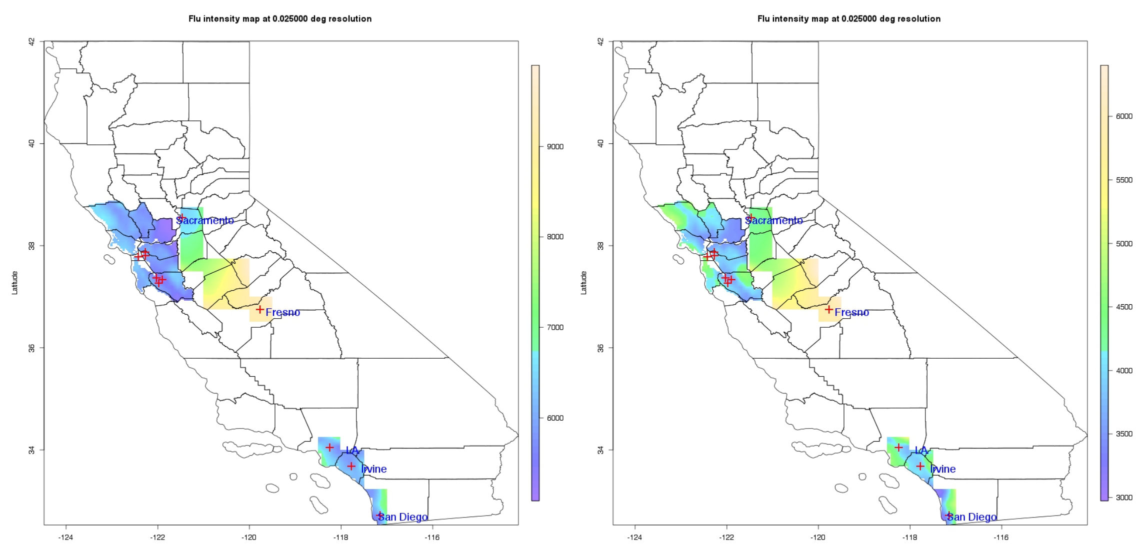

Figure 2.7: ILI+ intensity maps developed for the last week of March 2015. Left: A nowcast. Right: A one-week-ahead forecast.

OSI (Google Flu Trends in this study) can be replaced by public health data (which are usually delayed by a couple of weeks) and two-weeks-ahead forecasts could provide a nowcast of influenza activity. Such a system would not be affected by the approximate nature of GFT.

We also investigated whether the well-known dependence of influenza activity on meteorological variables [20] could be used to spatially predict said activity from municipalities with data and produce influenza

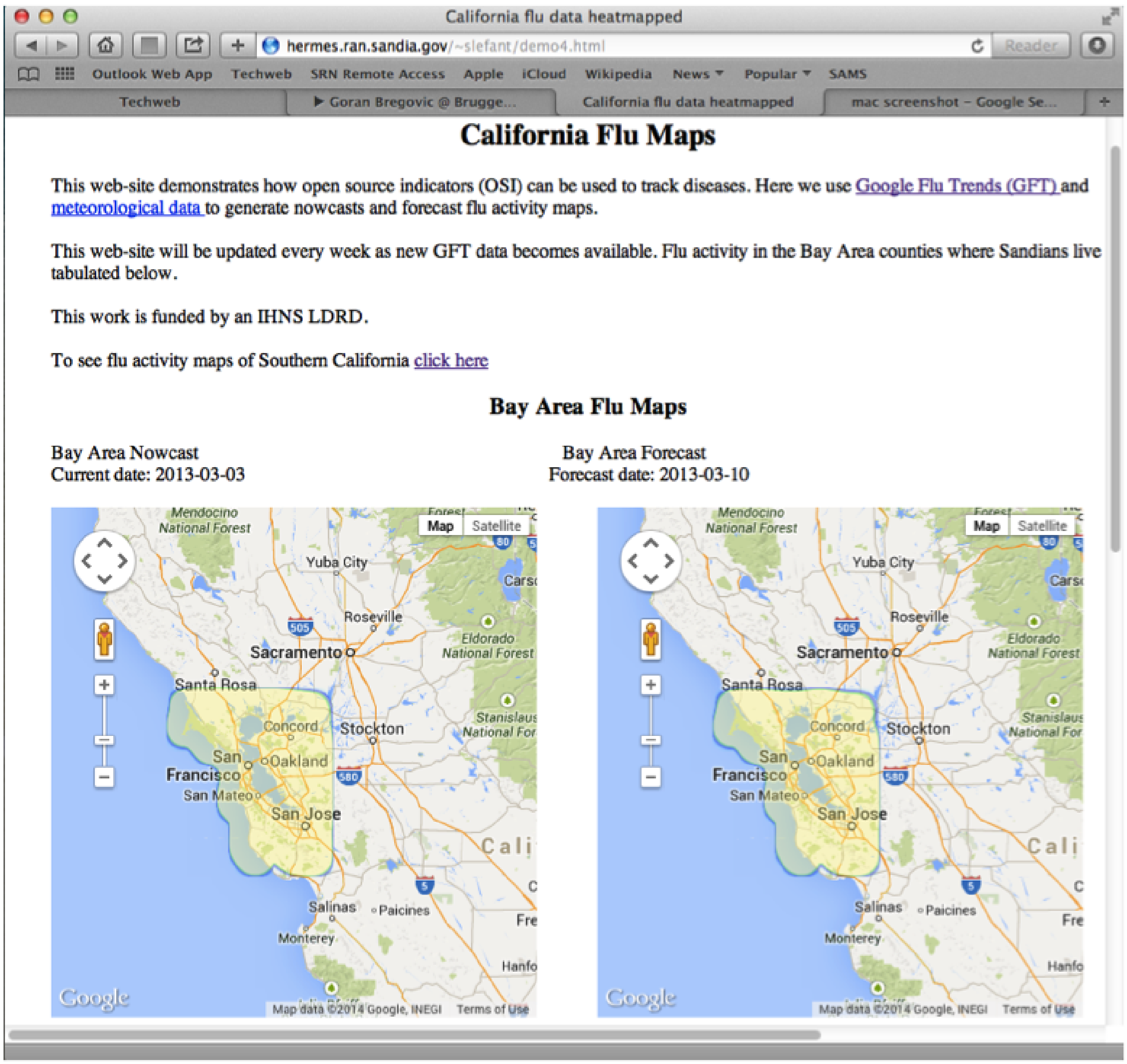

Figure 2.8: A snapshot of the output of the data assimilation system, displayed as a web page on an internal Sandia server.

activity maps. It required a new spatial prediction scheme using climatologically averaged temperature and specific humidity, as well as kernel smoothing. We tested this hypothesis in SFBA and our preliminary results have been encouraging when compared to public health data from San Mateo County. The method was extended to target 3 population centers in California (SFBA being one) and produce influenza activity maps, but we do not have independent public health data to test the accuracy of the spatial prediction method in the other two clusters.

The data assimilation system was implemented as a prototype and run weekly on a Sandia server for a year. It produced influenza activity maps for a year (2014-2015) and was stopped when GFT stopped publishing its data. It required little manual intervention and demonstrates that as long as OSI are available, such localized data assimilation systems can be constructed and deployed in the cloud. This could be very helpful in countries/areas with poor public health reporting. Collecting digital proxies of disease activity (web search logs, media articles pertaining to an outbreak etc.) is a well-established activity and global meteorological reanalysis products are easily available (e.g., Goddard Earth Sciences Data and Information Sciences Center, http://disc.sci.gsfc.nasa.gov/mdisc/data-holdings). Our method provides a way of combining these datastreams, along with disease models, to provide information of epidemiological and public health relevance.

Chapter 3

Data assimilation for dengue

3.1 Introduction

Dengue is a tropical disease spread by mosquitoes, typically *Aedes aegypti* [44]. In India, the dengue season follows the rainy monsoon seasons in the middle of the year. The monsoons sweep in from the Arabian Sea, in a Southwest to Northeast direction, leading to rains after the Indian summer. Mosquitoes and dengue soon follow [45, 46, 47, 48].

There have been attempts to track dengue activity using open source indicators (OSI). Web search logs have been used to track dengue in southeast Asia [49], using a model that linearly related the frequency of search to dengue incidence, as obtained from public health reports. The same study found that data obtained at a weekly resolution was preferable. Google Dengue Trends (GDT) provided a service (which lasted till August 2015) that tracked dengue in a number of countries [50]. Data was provided aggregated at the national scale, though for certain countries e.g., Mexico, provincial data was also available. In [51], the authors used provincial GDT data from Mexico to (1) compare against public health data and (2) elicit the effect of climate on dengue incidence. GDT data was found to be in good agreement with public health data, and a dependence on climatic factors was identified, with a lag of 8 weeks i.e., meteorological variables were a leading indicator. In a follow-on publication [52], they showed that the agreement between GDT and public health data was better in regions with intense dengue activity; including climatic factors in the model played a small part in improving predictive skill. In [53], the authors used alerts from HealthMap (`https;//www.healthmap.org`; HM), caused by an excessive number of media articles, open-source documents etc. on dengue, to track dengue in Latin America. They used kernel smoothing to create proxies of dengue incidence maps, but did not exploit dengue's dependence on climatic factors in the spatial prediction models.

Dengue's dependence on the presence of its vector limits it to regions where the *Aedes aegypti* mosquito is present. The mosquito does not have much of a flight range and consequently dengue cases tend to be cluster. A temporal variation is imposed by weather which controls the breeding of mosquitoes. Dengue incidence data (time and location of individual cases) have been subjected to clustering analysis in space-time [54, 55], using Knox-like tests to detect clusters [56]. The analysis is used to compute dengue risk to the population by convolving it with population density maps [57]. Incidence data collated over areal units have been used to make risk maps via kernel smoothing [58, 59]. Socioeconomic, environmental and land-use patterns are also factors in dengue prediction, and serve as exogenous variables in dengue risk mapping. The continuous variables (socioeconomic and environmental factors) are generally included via linear regression [60], while land-use categories are assimilated using logistic regression [61]. Conditionally auto-regressive (CAR) models with Poisson noise (to model dengue incidence counts) have been used for the spatial mapping of dengue risks [62], as have Gaussian processes in space-time [63]. However, there do not seem to be any studies that use statistical methods on provincial GDT or HM data to compute dengue risk maps (using climatic factors as exogenous variables), to impute missing data or to forecast a risk/dengue incidence map.

In this study we investigate a method to create maps (nowcasts and forecasts) of dengue activity in India. The OSI of relevance is obtained from HM. HM scrapes the web for documents - media articles, Ministry of Health publications, ProMed Mail articles - that concern dengue and makes the data available dis-aggregated by date and state. Only articles in English are collected and thus we obtain a small sampling of the total dengue-related media activity. Many of the media articles are duplicates, reprints from the same article obtained from news agencies such as Reuters (`https://www.reuters.com`). These duplicates are retained. We also use re-analysis products, temperature and precipitation fields obtained from[64].

Our mapping method is based on the hypothesis, similar to the one behind GFT and GDT, that dengue-related media activity is correlated to people's interest in the topic, which in turn could be caused by a dengue outbreak. Further, dengue activity would be correlated with temperature and rainfall, perhaps with a time lag required by mosquitoes to breed and spread dengue. Collating the data on a state-by-state basis could allow one to provide forecasts using time-series methods, with precipitation and temperature acting as exogenous factors/predictors.

3.2 Materials and methods

In this section we describe the data, its shortcomings and the modeling requirements for constructing dengue activity maps.

3.2.1 Data

The MERRA (Modern Era Retrospective-analysis for Research and Applications) meteorological data used in this study - temperature at 2m above ground and precipitation - are obtained from [64]. These reanalysis gridded datasets (0.5×0.67 degree resolution) are available for every hour and are averaged to their monthly values. Data from HM was purchased and provided us with dengue data for 2011 - 2013. The data was disaggregated by state and we computed monthly figures for the counts of media articles etc. The data was sparse and few states, over a month, exceeded 50 media mentions. Further, no data was available before August 2011, providing us with 29 months of data, in all.

The HM data was gappy i.e., there were months in the dengue season where a state ostensibly recorded no media article on the topic. Further there was no pattern in the "missingness" of the data. Fig. 3.1 plots the number of HM articles for India for October 2011, 2012 and 2013. The north Indian states (left blank) recorded no media articles. For the rest, we binned the HM counts and shaded the states accordingly. Only 15 states had any data at all, and we will focus our forecasting efforts on this subset. It is clear that the HM data is missing at random (MAR). However, meteorological data is available everywhere and is correlated with HM data. Filling the missing HM data is a pre-requisite for performing time-series predictions on a state-by-state level for the subset of 15 Indian states. Of the $15 \times 29 = 435$ possible data points, about 60% were missing.

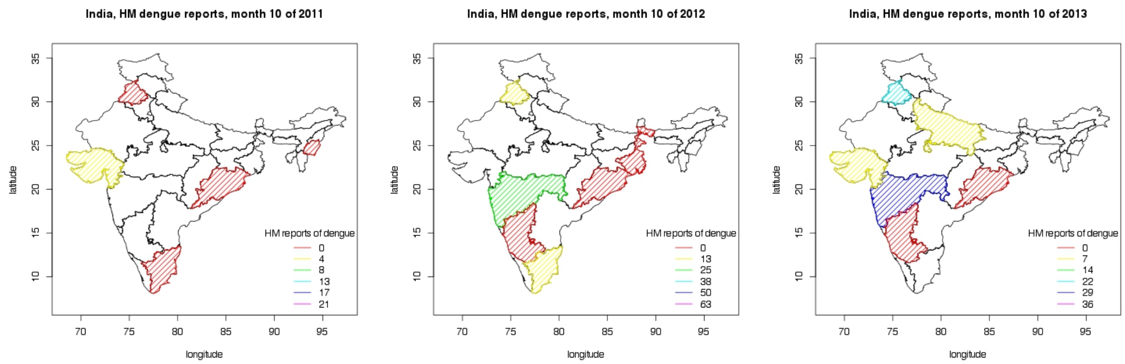

Figure 3.1: Plots of the binned counts of HM articles on dengue for India for October 2011 (left), 2012 (middle) and 2013 (right). The blank states recorded no data. States with data are shaded with a color corresponding to the lower bound of their bin. We see missing data occurs at random.

3.2.2 Conditionally auto-regressive models

The HM dataset contained "holes" - missing data points which were surrounded in time by observed data points. We constructed a neighborhood matrix for each state - any state that shared a border was deemed a neighbor. This 15×15 matrix $\mathbf{W}$ has an "1" entry if two states abut each other, else the matrix element is zero. This revealed that a missing data point often had observed data in a few of its neighbors. In order to impute a value for a missing data point, it was necessary to impose a spatio-temporal model. We performed this using conditionally auto-regressive models [29, 30, 31]. We provide a summary of the space-time conditionally auto-regressive model (STCAR) below.

Let Y_{kt} be the count for state k and time t. We model it as a draw from a normal distribution $\mathcal{N}(:,:)$

$$Y_{kt}|\boldsymbol{\mu}_{kt} \sim \mathcal{N}(\boldsymbol{\mu}_{kt}, \nu^2), \quad \text{where} \quad \boldsymbol{\mu}_{kt} = X_{kt}\boldsymbol{\beta} + \phi_{kt}, \; \boldsymbol{\beta} \sim \mathcal{N}(\mu_{\boldsymbol{\beta}}, \Sigma_{\boldsymbol{\beta}}), \; k = 1 \ldots K, \quad \text{and} \quad t = 1 \ldots T. \tag{3.1}$$

Here X_{kt} is a 1×2 vector containing temperature and precipitation for the state k and time t and $\boldsymbol{\beta} = \{\beta_T, \beta_P\}$, the regression weights for the meteorological variables. ϕ_{kt} models the correlation in space-time and ν^2 is the observation error variance, modeled with an Inverse-Gamma (InvGamma(1.0, 0.01)) prior.

Let $\boldsymbol{\phi}_t = \{\phi_{kt}\}, \; k = 1 \ldots K, \; K = 15$. The model for $\boldsymbol{\phi}$ is

$$\begin{aligned}
\boldsymbol{\phi}_t|\boldsymbol{\phi}_{t-1} &\sim \mathcal{N}(\rho_1\boldsymbol{\phi}_{t-1}, \tau^2 Q(\mathbf{W}, \rho_2)^{-1}) \\
\boldsymbol{\phi}_1 &\sim \mathcal{N}(0, \tau^2 Q(\mathbf{W}, \rho_2)^{-1}) \\
\tau^2 &\sim \text{InvGamma}(1, 0.01) \\
\rho_1, \rho_2 &\sim \mathcal{U}(0, 1)
\end{aligned} \tag{3.2}$$

Thus temporal correlation is modeled by the mean $\rho_1\boldsymbol{\phi}_{t-1}$ and spatial autocorrelation by the variance $\tau^2 Q(\mathbf{W}, \rho_2)$. The precision matrix $Q(\mathbf{W}, \rho_2)$ is given by

$$Q(\mathbf{W}, \rho_2) = \rho_2 \left(\text{diag}(\mathbf{W}1) - \mathbf{W}\right) + (1 - \rho_2)\mathbf{I}$$

where $\mathbf{I}$ is a 15×15 identity matrix.

This model is fitted i.e., $\boldsymbol{\beta}, \rho_1, \rho_2, \tau^2$ are estimated from the available data using an Markov chain Monte Carlo approach. The `R` [65] package CARBayesST [66] was used for the purpose.

3.2.3 Boosting

The STCAR model did not provide very accurate estimates and we resorted to boosting using Friedman's gradient boosting [67]. Consider a response Y with a set of predictors $\mathbf{X}$. Our aim to estimate a mapping $Y = \widehat{f(\mathbf{X})}$ by minimizing the expectation of a loss function $\Psi(Y, f)$

$$\widehat{f(\mathbf{X})} = \text{argmin}_{f(\mathbf{X})} \left(E_{y,x} \Psi(y, f)\right)$$

The procedure is as follows. We set $\widehat{f(\mathbf{x})}$ to a constant. Then for $t = 1 \ldots T$, do the following

1. Compute the negative gradient

$$z_i = -\frac{\partial}{\partial f(\mathbf{X})} \Psi\left(y, f(\mathbf{X})\right)\Big|_{f(\mathbf{X}_i)}$$

2. Fit a model $g(\mathbf{X})$ that predicts z_i from $\mathbf{X}_i$
3. Choose a gradient descent step Δ

$$\Delta = \text{argmin}_{\Delta} \sum_{i=1}^{N} \Psi\left(Y, f(\mathbf{X}_i) + \Delta g(\mathbf{X}_i)\right)$$

4. Update the estimate of $f(\mathbf{X})$

$$\widehat{f(()}\mathbf{X}) = \widehat{f(\mathbf{X})} + \Delta g(\mathbf{X})$$

5. Repeat the steps above till the successive difference between $\widehat{f(()}\mathbf{X})$ become smaller than a tolerance.

The boosted STCAR model provide a means of obtained "filled-in" datasets for each of the states. Thereafter, we fit a seasonal auto-regressive integrated moving average model with exogenous inputs (SARIMAX; see Chapter 8 in [68]) to each state's data and provide a forecast, along with a predictive error bound.

3.3 Results

Imputation accuracy was compared between a traditional, non-boosted STCAR approach and the boosted approach, where the base learner was an STCAR model. The data from the last three months of 2013 were held-out and used for testing the boosted STCAR models. Tests were performed for each state separately and then averaged for an overall performance figure. The resultant RMSEs (root mean square errors) from the cross-validation procedure are given in Fig. 3.2. Overall, the boosted setting performs better than the non-boosted setting by a factor of nearly two. We observe variability in the relative performance of the non-boosted and boosted settings across states, but we were unable to identify the phenomenon responsible for this variability. It does not appear that these differences in relative performance are related to amount of missingness or the extremity of the values we were trying to predict.

A comparison between the observed data set with missing values and the resultant fully-imputed data set using boosting are shown in Fig. 3.3. The seasonal trends are expected given the seasonal nature of dengue

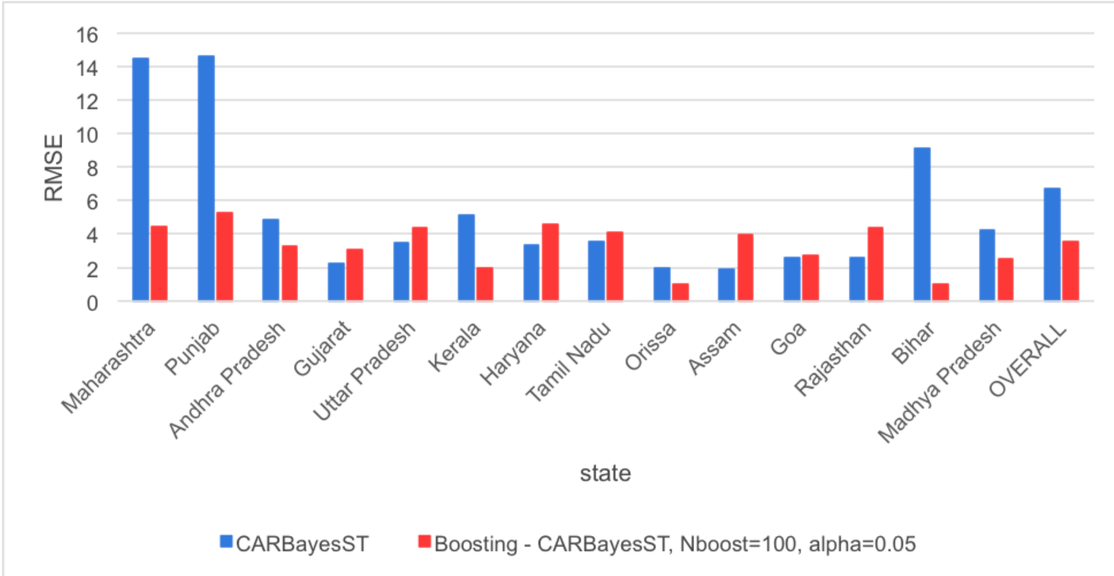

Figure 3.2: A comparison of the predictive error for the boosted (red) and non-boosted STCAR for filling in missing data. Results are plotted for the 15 states where there is some HM data. Overall boosting improves performance

outbreaks over the course of the year. Likewise, the time series for the states are not fully aligned with one another due to the geographically-varying monsoon onset times across India.

The SARIMAX forecasting model was tested using the filled-in datasets. The last three months of 2013 were held back and used to check the predictive skill of the model. Fig. 3.4 shows results for 2 states. The results are mixed - forecasting does not necessarily provide very good estimates. This arises mainly from the short time-series data, which does not allow us to learn the time-series model well. Further, the coarse time resolution does not reveal much smoothness in the evolution.

3.4 Conclusions

In this chapter, we investigated whether HM data could be used in forecasting dengue activity in Indian states. HM data was gappy and available only for 15 states. Dengue evolution is dependent on meteorological values and we investigated whether space-time conditionally auto-regressive model could be used to fill in (impute) the missing. They proved disappointing because of the large (60%) degree of missingness.

Next we sought to determine if a boosting approach to imputing missing data could out-perform the traditional, single imputation method. Boosted models result in improved imputation compared to non-boosted models, both in the simulation study and the application. The performance metrics were about 50% better in the boosted setting. These findings held even when the rate of missingness was very high.

The boosted imputation methodology is a promising one for future applications. It could be applied to a dataset with virtually any structure, and it does not necessarily require parametric assumptions, depending upon the learner chosen. Additionally, this methodology can use information from all cases, not just those that are complete.

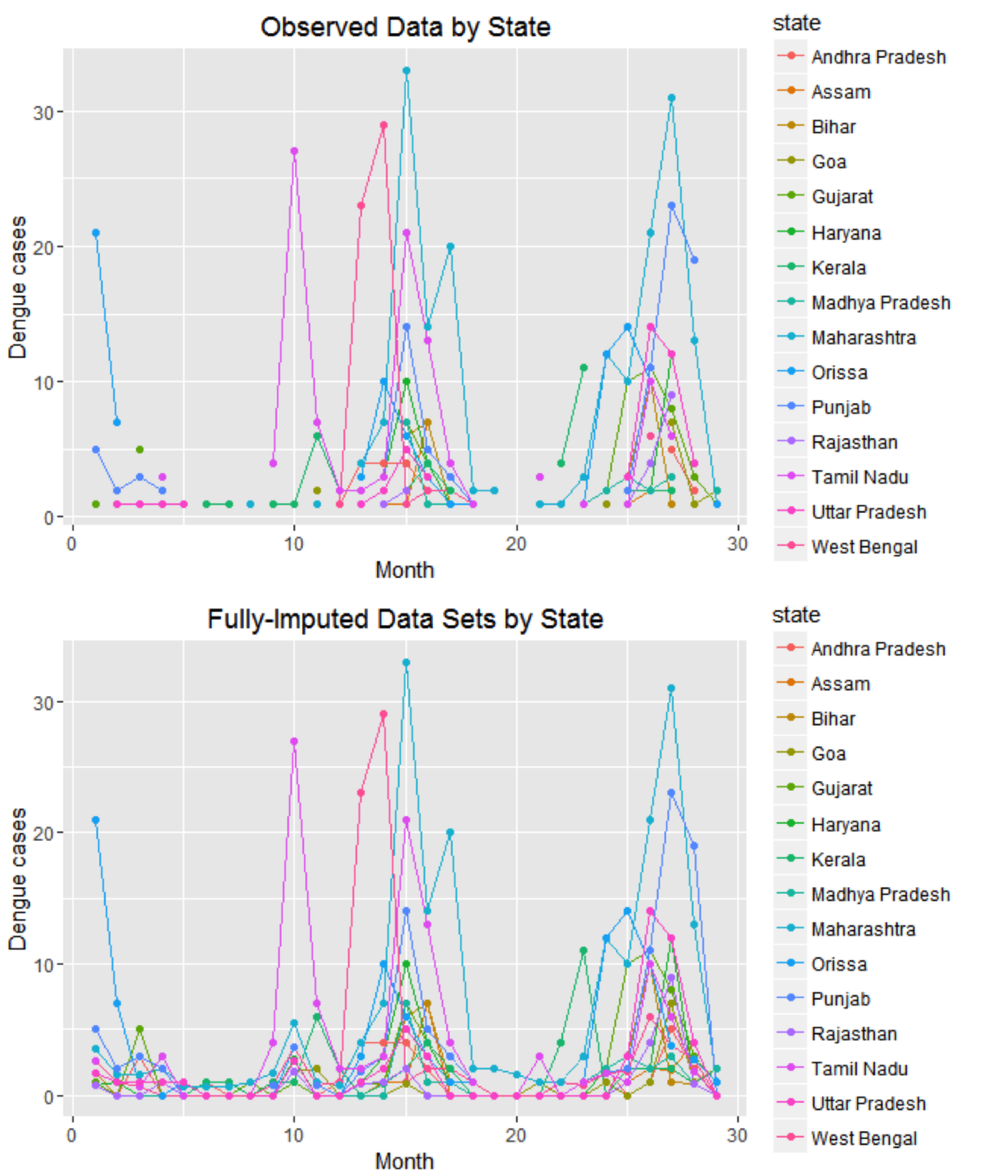

Figure 3.3: Top: The raw HealthMap data with gaps in it. Bottom: Filled in version of the HealthMap dataset.

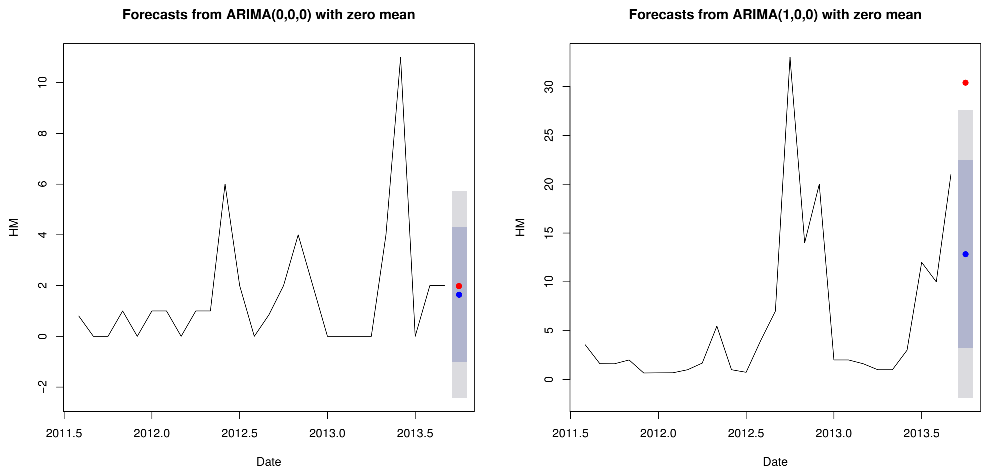

Figure 3.4: Left: Forecast of the HM data, for the state of Kerala. The blue dot is the mean prediction, with the shaded error being the 90% and 99% credibility intervals. The red dot is the true (not filled in) value. Right: The same, but for the state of Maharashtra.

One limitation of the boosting imputation methodology is that the time required to obtain boosted imputations is linearly related to the number of boosting iterations specified. If a base learner model takes a substantial amount of time to fit, then boosting it could potentially be memory and time consuming. More work is required to ascertain the robustness of these methods under various data conditions such as missingness mechanism, data structure, distribution, and missingness in multiple variables. Future efforts should focus on assessment of the robustness of the method and potential improvements that might be made by using adaptive boosting algorithms

The imputed datasets were then used in SARIMAX modeling. Results were not encouraging primarily because the datasets were too short to learn a SARIMAX model. If the HM articles were more copious, allowing us to aggregate on a weekly rather than monthly basis, and if the dataset spanned a larger duration, we believe forecasting results would have displayed smaller predictive errors.

Chapter 4

Follow-on applications

In this chapter, we discuss some work that incorporated or built on the methods or approaches developed in our study.

4.1 Data assimilation for wildfires

The data assimilation architecture described in Chp. 2 is sufficiently general that the models and data streams can be replaced so that the same general philosophy can be used in a very different setting. One such setting is crisis management where sparse data is usually available but not with sufficient accuracy or completeness to allow good situational awareness. Models of many types of crises and their consequences to society do exist (mostly for planning a response). It raises the question whether data, along with a model, could "fill in" the missing information. This could lead to a situational awareness toolkit containing both nowcasting and forecasting capabilities.

SUMMIT (Standard Unified Modeling and Mapping Integration Toolkit, [69]) is a Sandia framework for integrating models used in crisis management. It is also designed to wrap third-party models and allow communication and interactions between them, mainly to allow response planning. It also has the capability to to collect live datastreams and contains numerous visualization tools to display the results of model executions. To date, it has not had the capability to exploit data streams, perhaps with models, to infer information about the crisis that is not readily apparent from the raw data.

To that end, we are incorporating data assimilation capability (algorithmically similar to Chp. 2) to fill in unobserved data (the aim for Chp. 3). As a first step, we enabled SUMMIT to provide access to the results of the temporal data assimilation capability described in Chp. 2 (ETKF and SIR model of influenza). This required extension of SUMMIT to accommodate the peculiar interaction between models and data that are hallmark of ETKF and visualization of probabilistic forecasts. It also allowed us set out design enhancements SUMMIT would require to evolve into a crisis management tool where uncertainty in situational awareness (and the consequent impact on response planning) are fully and rigorously accommodated.

The next step was a more difficult exercise where we performed data assimilation and probabilistic forecasting for wildfires using a model that SUMMIT hosts. FARSITE [70] is a wildfire modeling tool developed by the US Department of Agriculture. It uses meteorological inputs (wind, humidity etc.) and vegetation (fuel) parameters to provide time-resolved evolution of a fire front, modeled as a set of perimeter points.The model adaptively refines the fire front as it expands so that the spatial resolution does not suffer. FARSITE is generally used to compute one-day-ahead forecasts of the fire front, using an infra-red image of the fire (provided by National Oceanographic and Atmospheric Administration's satellites) as the initial condition. The initial condition is assumed deterministic i.e., neither ensemble simulations nor data assimilation are

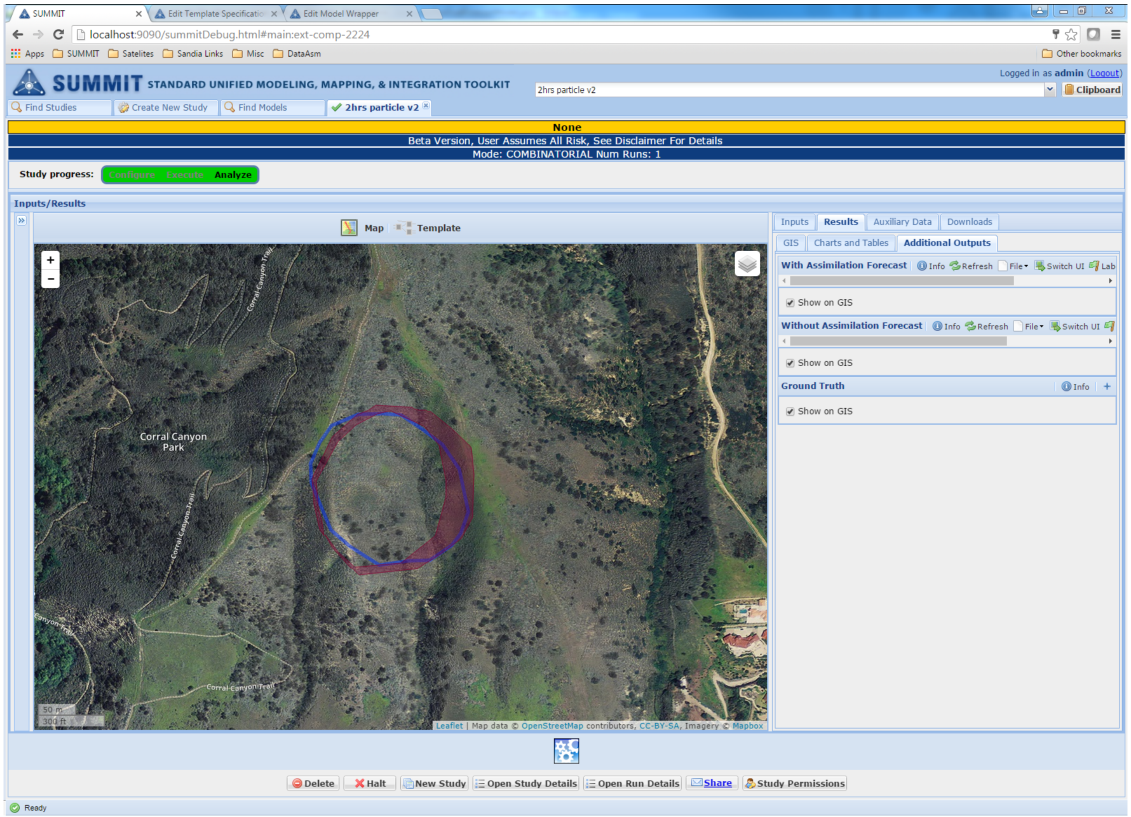

Figure 4.1: The wildfire after two hours. No observational data have been assimilated. The spread in forecast fire fronts (red contours) are due to our ignorance of wind and initial conditions. The blue contour is the true fire front.

currently performed to obtain forecasts. We enabled data assimilation with FARSITE by encoding a particle filter [71] to drive FARSITE. The data assimilation system ingests sparse observational data on the location of the fire front (i.e, data on where the fire has been observed) and meteorological data (wind and humidity available at a few locations) to provide updated (i.e., data informed) estimates of the full fire front and wind information. The ensemble of approximately 100 FARSITE instances (that reflects our ignorance of initial conditions, meteorology and fuel/vegetation parameters) is continuously updated to agree with sparse observational data and then used to provide forecasts.

We tested the data assimilation system using synthetic data from a fire near Santa Monica, California. Fig. 4.1 shows results the fire front after two hours; both the true perimeter and the spread of 100 ensemble simulations are shown overlaid on a satellite image. Fig. 4.2 shows the true and forecasted fire front after five hours. The importance of probabilistic forecasting is seen in Fig. 4.3 where the uncertainty spread in forecasts includes residences.

SUMMIT's development (which included the ability to wrap and orchestrate the interaction between models, access and incorporate data streams as well as visualization) was funded over a period of time (approximately five years) by US Department of Homeland Security, S&T Division. The data assimilation capability is novel, but still prototypical. It has not been tested with real data, which will inevitably require new models for observational errors. The sensitivity of the data assimilation system to sparseness of observational data as well as the impact of outlier observations is not known. Further, it is unknown how large an ensemble

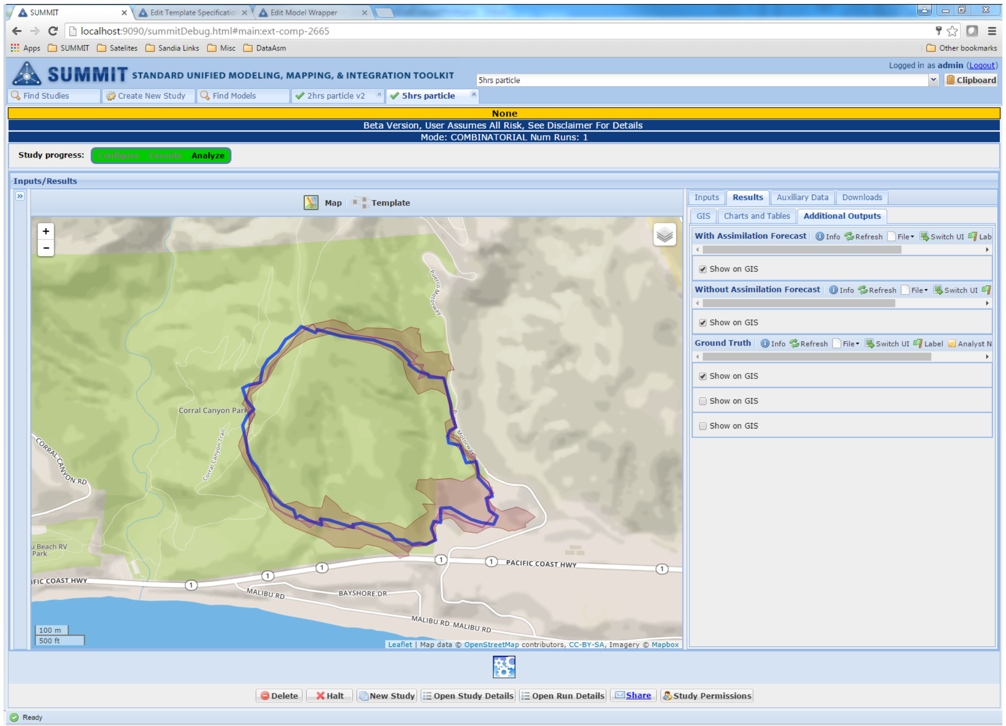

Figure 4.2: The wildfire after five hours. Data assimilation has occurred and reduces the uncertainty spread. The snapshot shows results displayed in SUMMIT's web client. The red contours are the 5 hour forecasts without data assimilation. A very narrow ensemble of purple fire fronts, very near the blue (true) fire front (and difficult to see) is the forecast ensemble, after assimilating data available after two hours.

should be to deliver a given forecasting accuracy, as well as to be numerically stable. In order to develop these capabilities, we will explore funding opportunities with California Department of Forestry and Fire Protection (`https://www.fire.ca.gov`) in 2017.

4.2 Data assimilation and disease modeling

The description in Chapters 2 and 3 provide a glimpse of the various approaches being pursued to exploit open-source indicators to track and forecast outbreaks. However, it has been noted that such models often produce conflicting forecasts. Further, an ensemble of such models can provide even more confusion since the disagreement between model predictions are not uniform i.e., subsets of models might agree. However, it is unclear whether the agreement is due to a similarity/correlation between the data streams being assimilated and the structure of the models (in which case, the agreement between models produces no new information) or if the agreement reflects some underlying truth. Sandia has been funded to address this problem. Our ability to calibrate disease models to open-source data-stream (as described in Chp. 2 and 3) and thus construct an ensemble is a prerequisite for performing this work, and the ease with which we could construct the ensemble played a role in the success of our proposal to the Defense Threat Reduction Agency. The work starts in FY17.

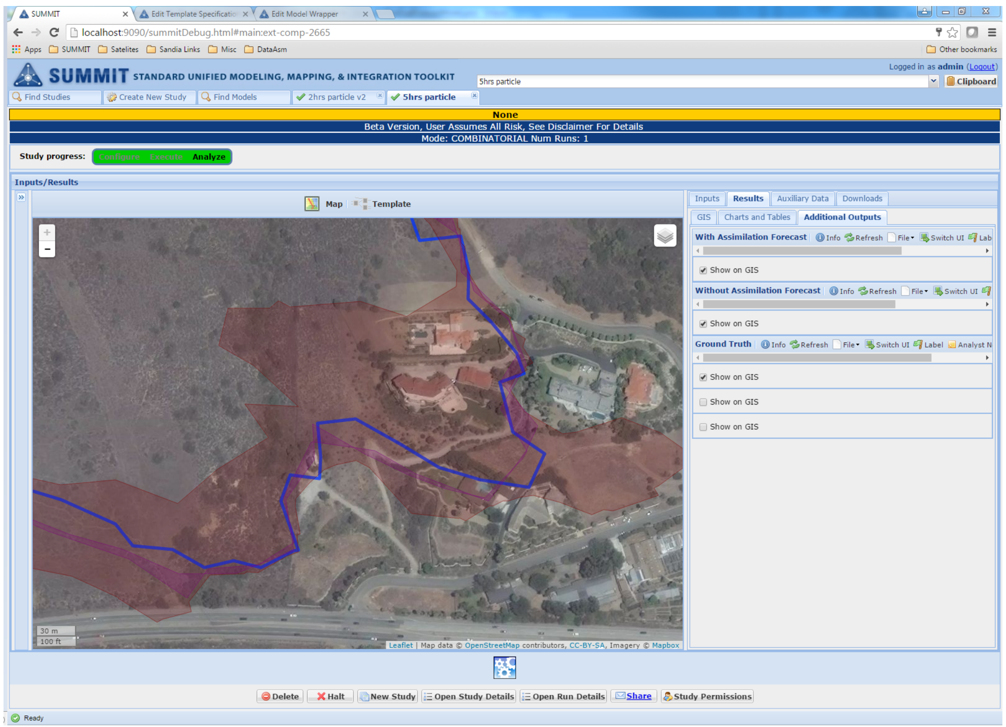

Figure 4.3: The wildfire after five hours, in detail. The uncertainty spreads show risk to human habitation and reinforce the need for data assimilation and probabilistic forecasting using filtering methods. The purple ensemble is easy to see and shows the enormous decrease in forecasting uncertainty effected by the assimilation of data after two hours.

Chapter 5

Conclusions

In this study we have investigated the use of open-source indicators to track and forecast epidemiological activity. In particular, we have explored the types of data assimilation systems that might be required to infer useful information about epidemiological dynamics. Our first foray, described in Chp. 2, may be considered one end of the spectrum. Here completeness of the observational data is not an issue. Rather, one questions whether the data stream i.e., web search logs serving as proxies of disease activity is justified, and how disease models may be pressed into inferring disease dynamics and forecasting. The emphasis lies on sophisticated data science methods, disease models and forecasting accuracy.

Data assimilation for dengue, as described in Chp. 3, forms the other end of the spectrum where the observational data is spotty and the emphasis lies in filling in the missing data. We have developed a technique that uses two open-source indicators - data from HealthMap and meteorology, to complete observational datasets. The actual forecasting is performed using simple time-series methods. Both the data assimilation methods (Chp. 2 and 3) produce disease activity maps, but the two systems have nothing in common in their need for methodological sophistication. This is because the data assimilation systems, practically, are designed to compensate for the shortcomings of the data streams when inferring disease activity maps. They will therefore be disease and location-specific. However, structurally, they will all contain disease models, calibration methodologies and missing-data imputation technologies, but their particular implementations (and algorithmic choices) will vary depending on the datastreams at hand.

The techniques developed in this study find many uses. The ETKF and thereafter particle filters have been implemented in a software framework for crisis management. The ability to construct calibrated disease models that ingest open-source information has lead to a funded project to assess the worth of disease models that "work off" different data streams (Chp. 4). Thus while the question of data assimilation for disease forecasting is indeed an interesting one, the importance (and worth) of the project lies in the development of methodological sophistication in Sandia's data science's capabilities. It is a fundamental strength and, as shown in Chp. 4, is being leveraged in myriad unexpected ways.

References

[1] World Health Organization Inflenza (Seasonal) factsheet. http://www.who.int/mediacentre/factsheets/fs211/en/.

[2] CDC Flu View. https://www.cdc.gov/flu/weekly/.

[3] J. Ginsberg, M. H. Mohebbi, R. S. Patel, L. Brammer, and M. S. Smolinski. Detecting influenza epidemics using search engine query data. *Nature*, 457:1012–1014, 2009.

[4] Influenza National and Regional Level Graphs and Data. http://gis.cdc.gov/grasp/fluview/fluportaldashboard.html.

[5] P. M. Pohlgreen, Y. Chen, D. M. Pennock, F. D. Nelson, and R. A. Weinstein. Using Internet searches for influenza surveillance. *Clinical Infectious Diseases*, 47(11):1443–1448, 2008.

[6] M. J. Paul, M. Dredze, and D. Broniatowski. Twitter improves influenza forecasting. *PLOS Currents Outbreaks*, (1), 2014.

[7] V. Lampos and N. Cristianini. Nowcasting events from the Social Web with statistical learning. *ACM Transactions on Intelligent Systems and Technology*, 3(4), 2012. Article 72.

[8] D. J.McIver and J. S. Brownstein. Wikipedia usage estimates prevalence of influenza-like illness in the united states in near real-time. *PLoS Comput Biology*, 10(4):1–8, 04 2014.

[9] N. Generous, G. Fairchild, A. Deshpande, S. Y. Del Valle, and R. Priedhorsky. Global disease monitoring and forecasting with wikipedia. *PLoS Comput Biology*, 10(11):1–16, 11 2014.

[10] M. Santillana, E. O. Nsoesie, S. R. Mekaru, D. Scales, and J. S. Brownstein. Using clinicians search query data to monitor influenza epidemics. *Clinical Infectious Diseases*, 59(10):1446–1450, 2014.

[11] D. R. Olson, K. J. Konty, M. paladini, C. Viboud, and L. Simonsen. Reassessing Google Flu Trends data for detection of seasonal and pandemic influenza: A comparative epidemiological study at three geographic scales. *Public Library of Science, Computational Biology*, 9(10), 2013. e1003256.

[12] D. Lazer, R. Kennedy, G. King, and A. Vespignani. The parable of google flu: Traps in big data analysis. *Science*, 343(6176):1203–1205, 2014.

[13] Patrick Copeland, Raquel Romano, Tom Zhang, Greg Hecht, Dan Zigmond, and Christian Stefansen. Google disease trends: An update. In *International Society of Neglected Tropical Diseases 2013*, page 3, 2013.

[14] S. Cook, C. Conrad, A. Fowlkes, and Matthew H. M. H. Mohebbi. Assessing google flu trends performance in the united states during the 2009 influenza virus a (h1n1) pandemic. *PLoS ONE*, 6(8):1–8, 08 2011.

[15] S. Yang, M. Santillana, and S. C. Kou. Accurate estimation of influenza epidemics using google search data via argo. *Proceedings of the National Academy of Sciences*, 112(47):14473–14478, 2015.

[16] Mauricio Santillana, D. Wendong Zhang, Benjamin M. Althouse, and John W. Ayers. What can digital disease detection learn from (an external revision to) google flu trends? *American Journal of Preventive Medicine*, 47(3):341 – 347, 2014.

[17] 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 Biology*, 11(10):1–15, 10 2015.

[18] P. Chakraborty, P. Khadivi, B. Lewis, A. Mahendiran, J. Chen, P. Butler, E. O. Nsoesie, S. R. Mekaru, J. S. Brownstein, M. V. Marathe, and N. Ramakrishnan. *Forecasting a Moving Target: Ensemble Models for ILI Case Count Predictions*, chapter 30, pages 262–270. 2014.

[19] Z. Wang, P. Chakraborty, S. R. Mekaru, , J. S. Brownsteinand J. Ye, and N. Ramakrishnan. Dynamic poisson autoregression for influenza-like-illness case count prediction. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '15, pages 1285–1294, New York, NY, USA, 2015. ACM.

[20] J. D. Tamerius, J. Shaman, W. J. Alonso, K. Bloom-Feshbach, C. K. Uejio, A. Comrie, and C. Viboud. Environmental predictors of seasonal influenza epidemics across temperate and tropical climates. *PLoS Pathog*, 9(3):1–12, 03 2013.

[21] R. P. Soebiyanto, F. Adimi, and R. K. Kiang. Modeling and predicting seasonal influenza transmission in warm regions using climatological parameters. *Public Library of Science, One*, 5:e9450, 2010.

[22] Jeffrey Shaman and Alicia Karspeck. Forecasting seasonal outbreaks of influenza. *Proceedings of the National Academy of Science of the United States*, 109(50), 2012. doi:10.1073/pnas.1208772109.

[23] J. Shaman, A. Karspeck, W. Yang, J. Tamerius, and M. Lipsitch. Real-time influenza forecasts during the 2012-2013 season. *Nature Communications*, 4, 2013. 2837.

[24] E. Goldstein, C. Viboud, V. Charu, and M. Lipsitch. Improving the estimation of influenza-related mortality over a seasonal baseline. *Epidemiology*, 23(6):829838, 2012.

[25] W. Yang, A. Karspeck, and J. Shaman. Comparison of filtering methods for the modeling and retrospective forecasting of influenza epidemics. *PLoS Comput Biology*, 10(4):1–15, 04 2014.

[26] L. Waller and B. Carlin. Disease mapping. In A. E. Gelfand, P. J. Diggle, M. Fuentes, and P. guttorp, editors, *Handbook of Spatial Statistics*. Chapman & Hall / CRC Press, 2010.

[27] J. Besag, J. C. York, and A. Mollie. Bayesian image restoration, with two applications in spatial statistics (with discussion). *Annals of the Institute of Statistical Mathematics*, 43, 1991.

[28] L. Held, M. Hohle, and M. Hofmann. A statistical framework for the analysis of multivariate infectious disease surveillance counts. *Statistical Modelling*, 5(3):187–199, 2005.

[29] Duncan Lee. A comparison of conditional autoregressive models used in bayesian disease mapping. *Spatial and Spatio-temporal Epidemiology*, 2(2):79 – 89, 2011.

[30] D. Lee, A. Rushworth, and S. K. Sahu. A bayesian localized conditional autoregressive model for estimating the health effects of air pollution. *Biometrics*, 70(2):419–429, 2014.

[31] A. Rushworth, D. Lee, and R. Mitchell. A spatio-temporal model for estimating the long-term effects of air pollution on respiratory hospital admissions in greater london. *Spatial and Spatio-temporal Epidemiology*, 10:29 – 38, 2014.

[32] Ransalu Senanayake, Simon OCallaghan, and Fabio Ramos. Predicting spatio–temporal propagation of seasonal influenza using variational gaussian process regression. In *Thirtieth AAAI Conference on Artificial Intelligence*, 2016.

[33] M. B. Hooten, J. Anderson, and L. A. Waller. Assessing north american influenza dynamics with a statistical SIRS model. *Spatial and Spatio-temporal Epidemiology*, 1(23):177 – 185, 2010. GEOMED Conference.

[34] M. W. Davidson, D. A. Haim, and J. M. Radin. Using networks to combine big data and traditional surveillance to improve influenza predictions. *Scientific Reports*, 2015. Published 2015/01/29/online.

[35] Y. Xia and et al. Ncep/emc, nldas primary forcing data l4 hourly 0.125 x 0.125 degree v002, version 002, 2009. http://disc.sci.gsfc.nasa.gov/uui/datasets/NLDAS_FORA0125_H_V002/summary?keywords=NLDAS#prod-summary.

[36] B. A. Cosgrove, D. Lohmann, K. E. Mitchell, P. R. Houser, E. F. Wood, J. C. Schaake, A. Robock, C. Marshall, J. Sheffield, Q. Duan, L. Luo, R. W. Higgins, R. T. Pinker, J. D. Tarpley, and J. Meng. Real-time and retrospective forcing in the north american land data assimilation system (nldas) project. *Journal of Geophysical Research: Atmospheres*, 108(D22), 2003. 8842.

[37] Google Flu Trends. https://www.google.org/flutrends/about.

[38] National and regional level outpatients illness and viral surveillance. http://gis.cdc.gov/grasp/fluview/fluportaldashboard.html.

[39] C. H. Bishop, B. J. Etherton, and S. J. Majumdar. Adaptive sampling with the ensemble transform kalman filter. part i: Theoretical aspects. *Monthly Weather Review*, 129(3):420–436, 2001.

[40] Pavel Sakov. Enkf-matlab: Matlab code for ensemble kalman filter, with a number of simple models, version 0.31. https://enkf.nersc.no.

[41] J. Shaman and M. Kohn. Absolute humidity modulates influenza survival, transmission and seasonality. *Proceedings of the National Academy of Science, USA*, 106:3243–3248, 2009.

[42] J. Shaman, V. E. Pitzer, C. Viboud, B. T. Grenfell, and M. Lipsitch. Absolute humidity and the seasonal onset of influenza in the continental United States. *Public Library of Science Biology*, 8:e1000316, 2010.

[43] T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning*. Springer Series in Statistics. Springer New York Inc., New York, NY, USA, 2008.

[44] Dengue fever, wikipedia, the free encyclopedia. https://en.wikipedia.org/wiki/Dengue_fever.

[45] P. Aggarwal, P. Gupta, S. D. Kandpal, B. Kakati, and D. Gupta. Post-monsoon season surveillance a must for curtaling annual dengue epidemic in rural india. *Journal of Community Medicine*, 5(1):153–155, 2014.

[46] S. Chandy, K. Ramanathan, A. Manoharan, D. Mathai, and K. Baruah. Assessing effect of climate on the incidence of dengue in Tamil Nadu. *Indian Journal of Medical Microbiology*, 13(3):283–286, 2013.

[47] K. Durani, J. Dund, H. Shingala, and M. Sinha. Epidemiological trend analysis of dengue virus infection in western part of Gujarat. *Indian Journal of Research*, 3(6):146–148, 2014.

[48] S. Jeelani and S. Sabesan. Aedes vector population dynamics and occurence of dengue fever in relation to climate variables in Puducherry, south india. *International Journal of Current Microbiology and Applied Sciences*, 2(12):313–322, 2013.

[49] Emily H Chan, Vikram Sahai, Corrie Conrad, and John S Brownstein. Using web search query data to monitor dengue epidemics: a new model for neglected tropical disease surveillance. *PLoS Negl Trop Dis*, 5(5):e1206, 2011.

[50] Google Flu Trends. https://www.google.org/flutrends/about.

[51] Rebecca T Gluskin, Mauricio Santillana, and John S Brownstein. Using google dengue trends to estimate climate effects in mexico. *Online journal of public health informatics*, 5(1), 2012.

[52] Rebecca Tave Gluskin, Michael A Johansson, Mauricio Santillana, and John S Brownstein. Evaluation of internet-based dengue query data: Google dengue trends. *PLoS Negl Trop Dis*, 8(2):e2713, 2014.

[53] Anne G Hoen, Mikaela Keller, Aman D Verma, David L Buckeridge, and John S Brownstein. Electronic event-based surveillance for monitoring dengue, latin america. *Emerg Infect Dis*, 18(7):1147–50, 2012.

[54] S. THAMMAPALO, V. CHONGSUVIVATWONG, A. GEATER, and M. DUERAVEE. Environmental factors and incidence of dengue fever and dengue haemorrhagic fever in an urban area, southern thailand. *Epidemiology and Infection*, 136(1):135–143, 03 2007.

[55] Camilo Rotela, Florence Fouque, Mario Lamfri, Phillipe Sabatier, Virginia Introini, Mario Zaidenberg, and Carlos Scavuzzo. Spacetime analysis of the dengue spreading dynamics in the 2004 tartagal outbreak, northern argentina. *Acta Tropica*, 103(1):1 – 13, 2007.

[56] Martin Kulldorff and Ulf Hjalmars. The knox method and other tests for space-time interaction. *Biometrics*, 55(2):544–552, 1999.

[57] Mohammad Ali, Yukiko Wagatsuma, Michael Emch, and Robert F Breiman. Use of a geographic information system for defining spatial risk for dengue transmission in bangladesh: role for aedes albopictus in an urban outbreak. *The American journal of tropical medicine and hygiene*, 69(6):634–640, 2003.

[58] Florisneide R Barreto, Maria Gloria Teixeira, N Costa Maria da Conceição, Marilia S Carvalho, and Mauricio L Barreto. Spread pattern of the first dengue epidemic in the city of salvador, brazil. *BMC Public Health*, 8(1):1, 2008.

[59] João B Siqueira, Celina MT Martelli, Ivan J Maciel, Renato M Oliveira, Maria G Ribeiro, Fluvia P Amorim, Bruno C Moreira, Divina DP Cardoso, Wayner V Souza, and Ana Lucia SS Andrade. Household survey of dengue infection in central brazil: spatial point pattern analysis and risk factors assessment. *The American journal of tropical medicine and hygiene*, 71(5):646–651, 2004.

[60] Adriano Mondini, Francisco Chiaravalloti Neto, Gallo Y Sanches, José Carlos Cacau Lopes, et al. Spatial analysis of dengue transmission in a medium-sized city in brazil. *Revista de Saúde Pública*, 39(3):444–451, 2005.

[61] Birgit HB Van Benthem, Sophie O Vanwambeke, Nardlada Khantikul, Chantal Burghoorn-Maas, Kamolwan Panart, Linda Oskam, Eric F Lambin, and Pradya Somboon. Spatial patterns of and risk factors for seropositivity for dengue infection. *The American journal of tropical medicine and hygiene*, 72(2):201–208, 2005.

[62] Gustavo S Ferreira and Alexandra M Schmidt. Spatial modelling of the relative risk of dengue fever in rio de janeiro for the epidemic period between 2001 and 2002. *Brazilian journal of Probability and Statistics*, pages 29–47, 2006.

[63] Hwa-Lung Yu, Shang-Jen Yang, Hsin-Ju Yen, and George Christakos. A spatio-temporal climate-based model of early dengue fever warning in southern taiwan. *Stochastic Environmental Research and Risk Assessment*, 25(4):485–494, 2011.

[64] Goddard Earth Sciences Data and Information Sciences Center. Modern era retrospective-analysis for research and applications (merra). http://disc.sci.gsfc.nasa.gov/mdisc/data-holdings.

[65] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2016.

[66] Duncan Lee, Alastair Rushworth, and Gary Napier. *CARBayesST: Spatio-Temporal Generalised Linear Mixed Models for Areal Unit Data*, 2016. https://CRAN.R-project.org/package=CARBayesST.

[67] J. H. Friedman. Stochastic gradient boosting. *Computational Statistics and Data Analysis*, 38(4):367–378, 2002.

[68] R. Hyndman and G. Athanasopoulos. *Forecasting: principles and practice*. OTexts, 2013. https://www.otexts.org/book/fpp.

[69] Lynn I. Yang. Summit: Standard unified modeling and mapping integration toolkit. SAND2016-4881C.

[70] FARSITE: Fire, Fuel and Smoke Science Program. https://www.firelab.org/project/farsite.

[71] K. Law, A. Stuart, and K. Zygalakis. *Data Assimilation: A mathematical introduction*. Springer, 2015. Texts in Applied Mathematics.

DISTRIBUTION:

1	Jaideep Ray, 08959	MS 9159
1	Katherine Cauthen, 06132	MS 1137
1	P. Finley, 06131	MS 1188
1	H. Smith, 06826	MS 1363
1	S. Lefantzi, 08959	MS 9152
1	Technical Library, 08944 (electronic)	MS 0899

Sandia National Laboratories