Statistics in Medicine WILEY

Dynamic downscaling and daily nowcasting from influenza surveillance data

Rajib Paul[1] | **Dan Han[2]** | **Elise DeDoncker[3]** | **Diana Prieto[4,5]**

[1]Department of Public Health Sciences, University of North Carolina at Charlotte, Charlotte, North Carolina USA

[2]Department of Mathematics, University of Louisville, Louisville, Kentucky, USA

[3]Department of Computer Science, Western Michigan University, Kalamazoo, Michigan, USA

[4]Carey School of Business, Johns Hopkins University, Baltimore, Maryland, USA

[5]School of Industrial Engineering, Pontificia Universdad de Catòlica de Valparaìso, Valparaìso, Chile

Correspondence
Rajib Paul, Department of Public Health Sciences, University of North Carolina at Charlotte, Charlotte, NC 28223, USA.
Email: Rajib.Paul@uncc.edu

Funding information
National Science Foundation, Grant/Award Number: 1537379

Abstract

Real-time trends from surveillance data are important to assess and develop preparedness for influenza outbreaks. The overwhelming testing demand and limited capacity of testing laboratories for viral positivity render daily confirmed case data inaccurate and delay its availability in preparedness. Using Bayesian dynamic downscaling models, we obtained posterior estimates for daily influenza incidences from weekly estimates of the Centers for Disease Control and Prevention and daily reported constitutional and respiratory complaints during emergency department (ED) visits obtained from the state health departments. Our model provides one-day and seven-day lead forecasts along with 95% prediction intervals. Our hybrid Markov Chain Monte Carlo and Kalman filter algorithms facilitate faster computation and enable us to update our estimates as new data become available. Our method is tested and validated using the State of Michigan data over the years 2009-2013. Reported constitutional and respiratory complaints at the EDs showed strong correlations of 0.81 and 0.68 respectively, with influenza rates. In general, our forecast model can be adapted to track an outbreak with only one respiratory virus as a causative agent.

KEYWORDS

Bayesian methods, influenza, Kalman filter, semiparametric modeling, temporal analysis

1 | INTRODUCTION

Every year, it is estimated that around *one billion* individuals fall sick with flu like symptoms worldwide,[1] and between 291 243 and 645 832 individuals die from illnesses associated with seasonal influenza.[2] When viruses mutate, immunity mechanisms fall short in responding to the mutation, which may yield a rapid spread of pandemic proportions, like the H1N1 outbreak in 2009. Due to the continuously changing and fast spreading nature of respiratory viruses, the resulting diseases pose constant threats to human health.

Surveillance systems are key to addressing uncertainties in disease detection, monitoring, and control for influenza.[3] However, due to disparities in reporting mechanisms, daily data are often noisy, subjected to delay in reporting, and unreliable. Traditional surveillance occurs passively when symptomatic cases report their symptoms to the healthcare systems. As an example, the state of Michigan works with three different sources to collect data from symptomatic cases: (1) the Michigan Syndromic Surveillance System (MSSS), which collects daily chief complaint data from individual registrations

in emergency departments (EDs) and urgent cares, (2) the Michigan component of the Centers for Disease Control and Prevention (CDC) ILINet (influenza like illness network), which collects weekly aggregates from primary and urgent care practitioners,[4] and (3) the MSSS, which captures weekly totals and individual case reports from EDs, primary, and urgent care providers, as well as any other type of individual submission (eg, from schools or flu testing labs). Although non-traditional data sources have been proposed for influenza surveillance (from Google search queries, Twitter messages, Facebook, Wikipedia article reviews, restaurant reservations, non-prescription pharmacy sales, etc.), there remain challenges in system validation and implementation,[5] which render traditional surveillance as the most reliable source of uncertainty management.

Current research on influenza surveillance points at real-time influenza incidences forecasting as an ideal tool for uncertainty management, and resource planning during influenza outbreaks.[6] A handful of forecasting studies are presented in the literature, where forecasts vary in their estimation method, timeframe, settings, outcomes, and data types used, and are reviewed in their scope by Chretien et al[7] and Nsoesie et al.[8] Since the 2013/2014 influenza season, the CDC has organized a challenge where modeling groups are allowed to submit real-time weekly forecasts of influenza incidences. A recent cross-validation study compares the performance of 22 of the models producing forecasts for the CDC challenge[9]

Due to the wide variability across seasons, and the unpredictable nature of pandemic outbreaks, several modeling groups have accommodated uncertainty in their prediction models, a feature that is becoming increasingly prevalent in the more recent models with demonstrated real-time usefulness. There exist Bayesian networks with probabilistic nodes,[10,11] empirical Bayes,[12] mechanistic models coupled with Bayesian-based filtering procedures to assimilate data,[13] dynamic Bayesian hierarchical models embedding a mechanistic model,[14] kernel density estimation to model autoregressive dependency on previous observations,[15] and kernel density estimation of incidence distributions for the forward looking weeks of the season, with copulas to model the dependence in incidence across different weeks.[16] There are also superensemble methods that combine the predictions of multiple models into a single estimate. Most of these superensembles also incorporate probabilistic components.[13,15,17]

Using a continuous time SIR (susceptible-infected-recovery) modeling framework and simulated data, Rhodes et al[18] provided a strategy for detecting the start date of an epidemic and predicting the behavior of the epidemic on a daily scale. Roberts et al[19] also used an SIR modeling framework to estimate the prevalence and incidence during epidemics and its peak. Lytras et al[20] used a flexible Bayesian model to predict probabilities for five distinct phases during an epidemic, such as, pre-epidemic, epidemic growth, epidemic plateau, epidemic decline, and post-epidemic. While predicting epidemics and understanding its behavior is paramount, very little work has been done on how to get a hold on the behavior of influenza viruses on a daily basis, utilizing the resources available to health departments.

Most of the existing forecasting methods are designed to predict on a weekly basis.[7,9,11] This is the case since publicly available data sources (eg, influenza incidences Sentinel Network and Google Flu trends) either accumulate case counts by week or need to refine their daily counts as more data become available. In addition, many methods are made for seasonal flu, in which the weekly resolution may suffice for planning purposes. The situation is different with pandemic outbreaks; daily forecasts are an ideal feature to support decisions with high cost-benefit trade-off where daily incidence rates may grow exponentially, and daily forecasts can be helpful to timely detect and stop chains of transmission (eg, by closing schools and workplaces) that would be harder to control if observed on a weekly basis. Out of the four models in the literature that provide forecasts on a daily basis, three assumed data that were collected daily and it included all daily cases of the outbreak analyzed.[18,19,21] Some improvements are presented in Jiang et al,[10] which assumed incomplete daily counts of data, and used an additional data trend (daily ER chief complaints of respiratory symptoms) to complete the daily estimate. However, we were unable to find articles that fuse data with multiple time resolutions.

Our work focuses on a tool for dynamic downscaling from weekly to daily rates using data resources that are already available to the state health departments, and provides one-day and seven-day forecasts of influenza incidences that county and state health departments can use for timely detection of fast growing transmission chains. Dynamic downscaling is a numerical procedure where data collected at coarser time resolution (eg, weekly scale) are used to interpolate at a finer resolution (eg, daily scale). Dynamic downscaling methods are popular in regional climate predictions based on information at a coarser scale from general circulation models.[22] However, it has been hardly applied for disease predictions and forecasts.

Our modeling framework is purely data-driven and requires little to no strong assumptions that are often needed for SIR/SEIR models. In addition to using prior seasonal outbreak data to predict the incidence in upcoming flu seasons, our model fuses data collected in real-time, which broadens the applicability of our approach to single-event pandemic outbreaks.

Our model has several interesting features and it fuses two different data sources: (1) weekly confirmed flu cases from CDC, as reported based on the World Health Organization (WHO) and National Respiratory Enteric Virus Surveillance Systems (NREVSS), (2) MSSS data on daily reported symptoms aggregated in two categories: constitutional symptoms (ie, fever, headache, malaise, fatigue, and diarrhea), and respiratory symptoms (ie, shortness of breath or difficulty breathing, cough, sore throat, and runny nose). Constitutional and respiratory symptoms can characterize other diseases, but are a strong indicator of influenza like illness (ILI) when appearing concurrently. We assume that the true underlying daily influenza incidence trend is a latent variable that cannot be observed, but these datasets provide information as predictors. Hybrid Markov Chain Monte Carlo (MCMC)[23] and Kalman filter algorithms[24] are developed for model fitting. Thus, the true underlying daily influenza incidence trend is some type of shrinkage estimate between the weekly CDC rates and daily respiratory and constitutional symptoms reported.

The remainder of this article is organized as follows: Section 2 gives details on our datasets; Section 3 provides the modeling framework; Section 4 presents the results along with model evaluation and sensitivity analysis; Sections 5 and 6 present our discussion and conclusions.

2 | DATA

Weekly incidence data were retrieved from the CDC, which aggregates the reports from the NREVSS for the state of Michigan, and the WHO (https://www.who.int). The NREVSS receives data on influenza incidence cases tested for influenza and reported laboratory confirmed cases of influenza from a network of clinics and hospitals, and then reports these both unweighted and weighted against the state's population. The WHO's Global Influenza Surveillance and Response System (GISRS), a global network of laboratories that examine influenza focused disease trends and viral surveillance, also contributes to the NREVSS (influenza incidence cases tested and confirmed influenza cases).[25]

On the daily scale, IRB exempted MSSS data were obtained from the Michigan Department of Health and Human Services (MDHHS). The MSSS data included constitutional and respiratory symptoms of patients attending EDs in Michigan from 1 January 2009 until 25 December 2013. The MSSS system automatically captures descriptive features of symptoms from the chief complaint data in the EDs, and classifies them into symptom categories, being constitutional and respiratory a couple of the categories used for classification. Hence, not all the reports in either the constitutional or respiratory categories correspond to an influenza incidence case. In addition, there exists underreporting, as just around 61% of the Michigan hospitals were participating in the MSSS at the time of data retrieval (exact numbers of hospitals participating were not provided by the MDHSS and underreporting rates could not be calculated). However, this information plays an important predictor for estimating the daily influenza incidences.

3 | BAYESIAN DYNAMIC DOWNSCALING MODEL

Combining disparate data sources to improve estimates at a granular level is widely popular in survey statistics and small area estimation. For example, Rose et al[26] combined small area census records and household microdata via an iterative proportional algorithm to find refined estimates for infant mortality and household attributes (such as the percentage of house with electricity, tap water etc.) in Bangladesh provinces. Multilevel regressions and poststratification techniques were used by Zhang et al[27] on Behavioral Risk Factor Surveillance System data, Missouri County-Level Study, and American Community Survey data to obtain county-level estimates of health indicators such as cardio obstructive pulmonary disease and uninsured rates. Spatial downscaling approaches using regression-based models that are widely used in Geosciences also incorporate data fusion techniques based on exogenous variables.[28-30] However, it is imperative to keep in mind that validity and sensitivity analyses based on scientific reasoning are essential to avoid bias and "garbage in and garbage out" scenarios. In what follows, we describe a Bayesian dynamic downscaling and nowcasting procedure based on data fusion and semiparametric approaches.

Our analyses were conducted in log-transformed scale, which means that all rates were modeled as log rates. We start with some notations. Let $Y_{d,w}$ denote the influenza incidence for the dth day of the wth week. This quantity is latent, as we do not know this from the available data, and our goal is to obtain this from weekly CDC influenza incidence Z_w, daily constitutional rates $Y_{d,w}^C$, and daily respiratory rates $Y_{d,w}^R$. We also provide a mechanism for nowcasting the influenza incidence. We adopt Bayesian methods for their natural benefits of hierarchical modeling, forecasting, and prediction intervals. The quintessential structure of our hierarchical models is comprised of four levels and prior distributions for parameters as:

[Weekly CDC influenza incidences] $\equiv \mathbf{A} \times$ Daily influenza incidences (Downscaled)

+ White noise

[Daily respiratory] $\equiv$ Autoregression + Seasonal component + White noise

[Daily constitutional] $\equiv$ Autoregression + Seasonal component + White noise

[Daily influenza incidences (Downscaled)] $\equiv$ h (Denoised daily respiratory and daily constitutional)

+ White noise

$[\boldsymbol{\theta}] \equiv$ Prior distributions,

where $\mathbf{A}$ is a 1×7 vector of coefficients that maps weekly observations to latent daily influenza incidences, *a priori* we assume that it is uniformly distributed across the week and set the values to 1/7, and $[\boldsymbol{\theta}]$ denotes prior distributions for all unknown parameters estimated in the model.

At the top level of our modeling framework, we link the weekly CDC rates with downscaled daily influenza incidences as:

$$Z_w = \mathbf{A}\,\mathbf{Y}_w + e_w^z, \tag{1}$$

where $\mathbf{Y}_w$ is a 7×1 matrix with elements of daily influenza incidences during a week and the white noise, e_w^z, follows a zero mean Gaussian distribution with variance σ_z^2.

We assume Bayesian Structural Time Series (BSTS) models[24,31-33] for $Y_{d,w}^C$ and $Y_{d,w}^R$. The BSTS model decomposes a time series into *trend, autoregression, seasonal component, cyclical component*, and *white noise*. Specifically, we consider a structural time series model in the state space form as:

$$Y_t^j = \mu_t^j + \tau_t^j + \varepsilon_t^j, \tag{2}$$

$$\mu_{t+1}^j = \mu_t^j + \nu_t^j + \xi_t^j, \tag{3}$$

$$\nu_{t+1}^j = \nu_t^j + \eta_t^j, \tag{4}$$

$$\tau_{t+1}^j = -\sum_{i=1}^{s-1} \tau_{t+1-i}^j + \omega_t^j, \quad j \in \{C, R\}, \quad t = 7(w-1) + d, \tag{5}$$

where $\{d, w\}$ represents the dth day in the wth week; μ_t^j is the unobserved underlying trend of the time series that evolves through an autoregressive process; ν_t^j is a slope term characterized by a random walk distribution, which controls a steady upward or downward movement of the time series and imposes a drift into the structural model for the trend; τ_t^j denotes the seasonal component, where s is the number of seasons for Y_t^j, and $\varepsilon_t^j, \xi_t^j, \eta_t^j$, and ω_t^j are independent zero-mean Gaussian white noise processes with variance parameters $\sigma_{\varepsilon,j}^2$, $\sigma_{\xi,j}^2$, $\sigma_{\eta,j}^2$, and $\sigma_{\nu,j}^2$, respectively. Note that we excluded the cyclical component from the BSTS model in Equation (2). From our datasets, we could observe that influenza outbreaks exhibit strong seasonal variations but no cyclical/periodic variations out of regular season.

At the next level of our hierarchical modeling, we characterize the relation between $Y_{d,w}$ and the mean respiratory rates $x_1 \equiv \mu_{d,w}^R + \tau_{d,w}^R$ and constitutional rates $x_2 \equiv \mu_{d,w}^C + \tau_{d,w}^C$ as:

$$Y_{d,w} = h(x_1, x_2) + e_{d,w}^y, \tag{6}$$

where $h(\cdot)$ is an unknown function of the trends and seasonal components of constitutional and respiratory symptoms, which will be specified semiparametrically using basis functions as:

$$h(x_1, x_2) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \sum_{k=1}^{K} u_k \, \|\mathbf{x} - \boldsymbol{\eta}_k\| \, \log \|\mathbf{x} - \boldsymbol{\eta}_k\|, \quad \mathbf{x} = (x_1, x_2), \tag{7}$$

where K is the number of knots, and $\boldsymbol{\eta}_k$ are selected knots on the bivariate domain that control the smoothness of the function $h(\cdot)$. The knots were selected using exploratory analyses based on a Clara algorithm,[34] where we consider a

uniform distribution of CDC daily confirmed rates by dividing the weekly rates with an initial value of seven. The Clara algorithm selected the knots based on a two-dimensional surface of the observed daily time series of constitutional and respiratory rates via restricted maximum likelihood estimation using connections between penalized splines and linear mixed models. Clara algorithms are available in R packages `SemiPar`[35] and `Cluster`.[36] White noise, $e_{d,w}^{y}$, follows a zero mean Gaussian distribution with variance σ_y^2.

We use Markov Chain Monte Carlo methods and Kalman filter techniques to fit our proposed model to the data. Under our model assumptions, Z_w follows a univariate Gaussian distribution with mean $\mathbf{A}\,\mathbf{Y}_w$ and variance σ_z^2, and $\mathbf{Y}_w$ follows a multivariate Gaussian distribution with mean vector $\mathbf{h}_w$ and covariance matrix $\sigma_w^2 \mathbf{I}_7$. Here $\mathbf{h}_w$ is a 7×1 vector of h function values computed over the wth week. The full conditional distribution for generating $\mathbf{Y}_w$ is multivariate Gaussian with mean vector $\mu_{Y|Z}$ of dimension 7×1 and covariance matrix $\mathbf{\Sigma}_{Y|Z}$ of dimension 7×7 are obtained as:

$$\mathbf{\Sigma}_{Y|Z} = \left(\frac{1}{\sigma_w^2}\mathbf{I} + \frac{\mathbf{A}'\mathbf{A}}{\sigma_z^2} \right)^{-1},$$

$$\mu_{Y|Z} = \mathbf{\Sigma}_{Y|Z} \left(\frac{\mathbf{h}_w'\mathbf{I}}{\sigma_w^2} + \frac{Z_w\mathbf{A}}{\sigma_z^2} \right).$$

Thus, the downscaled daily ILI trends are shrinkage estimates of the $\mathbf{h}_w'$ function and $Z_w\mathbf{A}$. Note that the $h(\cdot)$ function is a function of constitutional and respiratory rates after distilling their measurement errors.

We impose improper priors on the real line for all variance parameters (σ^2), that resulted in full conditional distributions for σ_w^2 as proper inverse-gamma with shape parameter $n/2 + 1$ and scale parameter $0.5(\mathbf{Y} - \hat{\mathbf{Y}})'(\mathbf{Y} - \hat{\mathbf{Y}})$, where $\mathbf{Y}$ denotes the $n \times 1$ vector of all daily downscaled influenza incidences over our study period. Similarly, in our MCMC algorithm, σ_z^2 was generated from an inverse-gamma distribution with rate parameter $n_w/2 + 1$ and scale parameter $0.5(Z_w - \mathbf{A}\,Y_w)'(Z_w - \mathbf{A}\,Y_w)$, where n_w is the number of weeks.

Let $\boldsymbol{\theta} = [\beta_0, \beta_1, \beta_2, \mu_k, k = 1, 2, \ldots, K]'$ denote the coefficients in the semi-parametric regression; we impose a multivariate Gaussian prior with mean vector $\boldsymbol{\mu}_\theta$ and covariance matrix $\mathbf{\Sigma}_\theta$ on $\boldsymbol{\theta}$. The full conditional distribution for $\boldsymbol{\theta}$ results in a Gaussian distribution with mean vector $\mu_{\theta|\text{Rest}}$ and covariance matrix $\mathbf{\Sigma}_{\theta|\text{Rest}}$ as:

$$\mathbf{\Sigma}_{\theta|\text{Rest}} = \left(\frac{\mathbf{G}\mathbf{G}'}{\sigma_w^2} + \Sigma_\theta^{-1} \right)^{-1}$$

$$\mu_{\theta|\text{Rest}} = \mathbf{\Sigma}_{\theta|\text{Rest}} \left(\frac{\mathbf{G}'\mathbf{Y}}{\sigma_w^2} + \Sigma_\theta^{-1}\boldsymbol{\mu}_\theta \right).$$

The full conditional distributions for mean respiratory rates $x_1 \equiv \mu_{d,w}^R + \tau_{d,w}^R$ and constitutional rates $x_2 \equiv \mu_{d,w}^C + \tau_{d,w}^C$ are unavailable in closed form due to the nonlinearity in the semiparametric regression, Equation (7). We use a hybrid method that comprises a Kalman filter and a Metropolis-Hastings step to sample these parameters. Now we focus on the Bayesian Kalman filter algorithm that is widely used for BSTS models.

We rewrite Equations (2)-(4) for constitutional and respiratory rates as:

$$Y_t^j = \mathbf{X}_t^j \boldsymbol{\alpha}_t^j + \epsilon_t^j, \tag{8}$$

$$\boldsymbol{\alpha}_{t+1}^j = \mathbf{T}_t^j \boldsymbol{\alpha}_t^j + \mathbf{M}_t^j \boldsymbol{e}_t^j, \tag{9}$$

where the state-space parameters are $\boldsymbol{\alpha}_t^j = (\mu_t^j, \gamma_t^j, \tau_t^j, \tau_{t-1}^j, \ldots, \tau_{t-s+1}^j)'$, s is the number of seasons, and $\boldsymbol{e}_t^j$ is a zero-mean multivariate Gaussian random vector of dimension 3×1 with diagonal covariance matrix $\boldsymbol{Q}_t^j \equiv \text{Diag}\left(\sigma_{\xi,j}^2, \sigma_{\eta,j}^2, \sigma_{\nu,j}^2 \right)$.

In Equation (8), $\mathbf{X}_t^j$ is $\text{Diag}\left(\boldsymbol{X}_{[1]}^j, \boldsymbol{X}_{[2]}^j \right)$, where $\boldsymbol{X}_{[1]}^j = (1, 0)$ and $\boldsymbol{X}_{[2]}^j$ is a $1 \times s$ vector with the first element 1, and the remaining elements are 0. The matrix $\mathbf{T}_t^j$ is of dimension $(s + 2) \times (s + 2)$ and can be written as $\text{Diag}\left(\boldsymbol{T}_1^j, \boldsymbol{T}_2^j \right)$, where the 2×2 matrix $\boldsymbol{T}_1^j$ and the $s \times s$ matrix $\boldsymbol{T}_2^j$ are:

$$T_1^j \equiv \begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix} \quad T_2^j \equiv \begin{bmatrix} -1 & -1 & \cdots & -1 & -1 \\ 1 & 0 & \cdots & 0 & 0 \\ 0 & 1 & \cdots & 0 & 0 \\ 0 & 0 & \cdots & 1 & 0 \end{bmatrix}$$

In Equation (9), the $(s+2) \times (s+2)$ matrix M_t^j can be written as $\mathrm{Diag}\left(M_{[1]}^j, M_{[2]}^j\right)$, with $M_{[1]}^j$ being an identity matrix of rank 2 and $M_{[2]}^j$ is an s-dimensional vector with the first element one and the remaining entries zero.

A Kalman filter algorithm,[32] is used assuming the initial state α_1^j follows a Gaussian distribution with mean a_1^j and covariance P_1^j. Let y_t^j denote the observations up to time t as $y_t^j = ((Y_1^j)', (Y_2^j)', \dots, (Y_t^j)')'$. Denoting $a_{t|t}^j = E[\alpha_t^j | y_t^j]$, $a_{t+1}^j = E[\alpha_{t+1}^j | y_t^j]$, $P_{t|t}^j = Var(\alpha_t^j | y_t^j)$, and $P_{t+1}^j = Var(\alpha_{t+1}^j | y_t^j)$, the derived Kalman-Filter equations are:

$$a_{t|t}^j = a_t^j + P_t^j X_t^{j\top} F_t^{j-1} v_{t,j},$$

$$P_{t|t}^j = P_t^j - P_t^j X_t^{j\top} F_t^{j-1} X_t^j P_t^j,$$

$$a_{t+1}^j = T_t^j a_t^j + K_t^j v_{t,j},$$

$$P_{t+1}^j = T_t^j P_t^j (T_t^j - K_t^j X_t^j)' + M_t^j Q_t^j (M_t^j)' \text{ for } t = 1, 2, \dots, T, \quad j \in \{C, R\},$$

where $v_{t,j} = Y_t^j - X_t^j a_t^j$, $F_t^j = X_t^j P_t^j X_t^{j\top} + H_t^j$, and $K_t^j = T_t^j P_t^j X_t^{j\top} F_t^{j-1}$. The state vector α_{t+1}^j is sampled from a Gaussian distribution with mean $a_{t+1}^j = T_t^j a_{t|t}^j$ and covariance matrix $P_{t+1}^j = T_t^j P_{t|t}^j (T_t^j)' + M_t^j Q_t^j (M_t^j)'$. This distribution is considered as a proposal distribution for the Metropolis-Hastings step to sample the *trend* and *seasonal components* $\mu_t^j, \tau_t^j, j \in C, R$, $t = d, w$, respectively. Pseudocode of the algorithms is provided. For details, see Algorithms 1 and 2.

4 | RESULTS

4.1 | Real data analysis

Recall that our primary goal is obtaining influenza incidence trends on a daily scale, $Y_{d,w}$, based on weekly CDC reported influenza incidence trends and MDHHS reported daily constitutional and respiratory trends. We also obtain 95% prediction intervals for the downscaled influenza incidences.

Since we fitted a semiparametric model, the nonparametric function $h(\cdot)$ estimation involves a large number of parameters. In Table 1, we only showed the posterior estimates of parameters from the linear portion of $h(\cdot)$ function and all variance parameters. These parameters were summarized in the table using posterior medians, means, and 95% credible intervals. According to our model, the $h(\cdot)$ function is the *a priori* mean of the downscaled ILI rates on log scale. In Figure 1, we plotted the posterior median of $h(\cdot)$ function with respect to log-transformed respiratory and constitutional rates. From this plot, we see an overall increasing trend of the mean surface of daily downscaled ILI rates with constitutional and respiratory rates.

Figure 2 shows four time series, three of which are on daily scale and one on weekly scale. The green and greenish blue lines indicate daily reported respiratory and constitutional log-transformed rates per 100 000 population. The red dots indicate weekly log-transformed CDC influenza incidences per 100 000 population. From these three time series, one can clearly see the seasonal patterns. Also, the 2009 H1N1 pandemic is clearly observed. The Pearson's correlations between weekly influenza incidences and weekly constitutional and respiratory rates are 0.81 and 0.68, respectively. These high correlation values confirm that constitutional and respiratory rates are strong predictors of influenza incidences. Whenever there is a peak in weekly influenza incidences, we also see peaks for constitutional and respiratory rates.

The black lines in Figure 2 indicate the posterior medians of downscaled *daily* influenza incidences, with gray-shaded 95% prediction intervals. The downscaled time series follow similar patterns of weekly CDC rates. During the 2009 pandemic, there are many confirmed cases in a typical week. For example, based on our downscaled influenza incidence trends, between 12 October 2009 and 28 November 2009, 43 741 population got affected with an average of 1121 people

---

Algorithm 1. Algorithm for downscaling influenza incidences with available data

---

1. Step 1: From BSTS models get $niter_0 = 30000$ realizations of constitutional daily rates and respiratory daily rates. Use the parts after burn-in (*burnin*). Extract trend and seasonal components $\mu_{d,w}^C$, $\tau_{d,w}^C$, $\mu_{d,w}^R$, and $\tau_{d,w}^R$ for ($niter_0 - burnin$) realizations.
2. Step 2: Set loop number n_1 as ($niter_0 - burnin$), set initial value of the weight $w[1] = -1.15 \times 10^{13}$.

  for $k = 1$ to n_1 **do** xc[k]=$\mu_{d,w}^C[k] + \tau_{d,w}^C[k]$, xr[k]=$\mu_{d,w}^R[k] + \tau_{d,w}^R[k]$
      if $k = 1$ **then** $p = 1$
      else
          rate=$\min(1, \frac{w[k]}{w[k-1]})$, $p \sim binomial(1, rate)$
      end if
      if $p = 1$ **then**
          Constitutional rates[k]=xc[k], Respiratory rates[k]=xr[k]
      else
          Constitutional rates[k]=xc[k-1], Respiratory rates[k]=xr[k-1]
      end if
      Run semi-parametric regression to get h_w, and fitted value for CDC daily rates: *fitted_value* and set up the initial prior of θ: μ_θ, which is the mean of coefficients in semi-parametric regression, and Σ_θ, which is the covariance matrix in the semi-parametric regression.
      for $i = 1$ to n_w **do**
  $\Sigma_{Y|Z} = (\frac{1}{\sigma_w^2}\mathbf{I} + \frac{\mathbf{A}'\mathbf{A}}{\sigma_z^2})^{-1}$, $\mu_{Y|Z} = \Sigma_{Y|Z}(\frac{\mathbf{h}_w'\mathbf{I}}{\sigma_w^2} + \frac{Z_w\mathbf{A}}{\sigma_z^2})$, influenza incidences Daily[, i]= $\mathbf{Y}_w|Z_w \sim N_7(\mu_{Y|Z}, \Sigma_{Y|Z})$
      end for
      Transform influenza incidences Daily ($7 \times n_w$) into a vector of daily rates:
      $\mathbf{Y}[k] = vec(\text{influenza incidences Daily}[k])$
      Get the scale for the truncated inverse gamma distribution of σ_z^2, *scale_sum* $= 0$
      for $i = 1$ to n_w **do**
          $scale\_sum = scale\_sum + (Z_w[i] - A \times \text{influenza incidences Daily}[, i])^2$
      end for
      Generate σ_w^2 and σ_z^2: $\sigma_z^2 \sim$ inverse gamma$(1, n_w/2 + 1, \frac{2}{scale\_sum})$
      $\sigma_w^2 \sim$ inverse gamma$(1, n/2 + 1, \frac{2}{(\mathbf{Y}[k]-fitted\_value)'(\mathbf{Y}[k]-fitted\_value)})$
      Generate θ: $\Sigma_{\theta|R} = $ positive definite$((\frac{\mathbf{G}\mathbf{G}'}{\sigma_w^2} + \Sigma_\theta^{-1})^{-1})$, $\mu_{\theta|R} = \Sigma_{\theta|R}(\frac{\mathbf{G}'\mathbf{Y}[k]}{\sigma_w^2} + \Sigma_\theta^{-1}\mu_\theta)$
      $\theta \sim N(\mu_{\theta|R}, \Sigma_{\theta|R})$
      Get weights w by calculating natural logarithm of the multivariate normal density of $\mathbf{Y}[k]$:
      $w[k + 1] = dmvnorm(\mathbf{Y}[k], G \times \theta, \sigma_w^2 * diag(n), \log=\text{TRUE})$
  end for
  Take 2.5% and 97.5% percentile as the upper bound and lower bound of the downscaled daily influenza incidences confidence interval; and take the median as the downscaled daily influenza incidences. (If the value is below 0, then take 0.)

---

per day. During a regular flu season, between 1 January 2012 and 30 April 2012, 25 084 population got affected with an average of 207 people per day. Note that our downscaled series have several desirable properties: (1) they provide some degree of shrinkage, and (2) they are more stable and less noisy. If we look closely at the time period between 1 January 2013 and 1 March 2013, we see that the respiratory rates are substantially high; however, our estimates are balanced by the other two time series. Our estimates are impacted less by the noise present in the respiratory data during that time period.

We set aside the last 15 days of data and predicted for those days using the remaining data from 1805 days. Our predicted daily influenza incidence trends (blue lines) and the associated prediction intervals (gray shades) are shown in Figure 3. The weekly CDC data is also shown (red dots). Comparing the CDC data with the predictions, we can clearly see that our predictions follow similar patterns. Additionally, the widths of prediction intervals ensure that the model is able to capture the uncertainties associated with the data.

Algorithm 2. Algorithm for predicting downscaled influenza incidences

---

1. Step 1: Use BSTS models to predict n_1 days of constitutional daily rates and respiratory daily rates with N realizations. Use the parts after burn-in (*burnin*). Extract trend and seasonal components of predicted n_1 days: $\mu_{d,w}^C$, $\tau_{d,w}^C$, $\mu_{d,w}^R$ and $\tau_{d,w}^R$ for $n = N - burnin$ realizations.

2. Step 2: Set the initial value of the weight $w[1] = -1.15 \times 10^{13}$

   for $k = 1$ to n **do**
   xc[k]=$\mu_{d,w}^C[k] + \tau_{d,w}^C[k]$, xr[k]=$\mu_{d,w}^R[k] + \tau_{d,w}^R[k]$
   　if $k = 1$ **then** $p = 1$
   　else rate=min(1, $\frac{w[k]}{w[k-1]}$), $p \sim binomial(1, rate)$
   　end if
   　if $p = 1$ **then**
   　　Constitutional rates[k]=xc[k], Respiratory rates[k]=xr[k]
   　else
   　　Constitutional rates[k]=xc[k-1], Respiratory rates[k]=xr[k-1]
   　end if
   Predict initial CDC daily rates *y_hat_predict* by the semi-parametric model in Algorithm 1. Run semi-parametric regression for *y_hat_predict* with constitutional rates [k] and respiratory rates [k] to get h_w, and fitted value for CDC daily rates in the prediction periods: *fitted_value*. Set up the initial prior of θ: μ_θ is the mean of coefficients and Σ_θ is the covariance matrix in the semi-parametric regression of predicted periods. *n_w_pred* is the number of predicted weeks.
   　for $i = 1$ to *n_w_pred* **do**
   $\Sigma_{Y|Z} = (\frac{1}{\sigma_w^2}\mathbf{I} + \frac{\mathbf{A'A}}{\sigma_z^2})^{-1}$, $\mu_{Y|Z} = \Sigma_{Y|Z}(\frac{\mathbf{h}_w'\mathbf{I}}{\sigma_w^2} + \frac{Z_w\mathbf{A}}{\sigma_z^2})$
   influenza incidences Daily[, i]= $\mathbf{Y}_w|Z_w \sim N_7(\mu_{Y|Z}, \Sigma_{Y|Z})$
   　end for
   Transform influenza incidences Daily ($7 \times$ *n_w_pred*) matrix into a vector of daily rates:
   $\mathbf{Y}[k] = vec($influenza incidences Daily[k]$)$
   Get the scale for the truncated inverse gamma distribution of σ_z^2, *scale_sum* $= 0$
   　for $i = 1$ to *n_w_pred* **do** *scale_sum* $=$ *scale_sum* $+ (Z_w[i] - A \times$ influenza incidences Daily[, i]$)^2$

   　end for
   Generate σ_w^2 and σ_z^2: $\sigma_z^2 \sim$ inverse gamma$(1, n_w/2 + 1, \frac{2}{scale\_sum})$
   $\sigma_w^2 \sim$ inverse gamma$(1, n/2 + 1, \frac{2}{(\mathbf{Y}[k]-fitted\_value)'(\mathbf{Y}[k]-fitted\_value)})$
   Generate θ:
   $\Sigma_{\theta|R} =$ positive definite$((\frac{\mathbf{GG'}}{\sigma_w^2} + \Sigma_\theta^{-1})^{-1})$, $\mu_{\theta|R} = \Sigma_{\theta|R}(\frac{\mathbf{G'Y}[k]}{\sigma_w^2} + \Sigma_\theta^{-1}\mu_\theta)$ $\theta \sim N(\mu_{\theta|R}, \Sigma_{\theta|R})$
   Get weights w by calculating natural logarithm of the multivariate normal density of $\mathbf{Y}[k]$:
 $w[k + 1] = dmvnorm(\mathbf{Y}[k], G \times \theta, \sigma_w^2 * diag(n), \log=TRUE)$

   end for
   Take 2.5% and 97.5% percentile as the upper and lower bound of predicted downscaled daily influenza incidences confidence interval; and take the median as predicted downscaled daily influenza incidences. (If the value is below 0, then take 0.)

TABLE 1 Posterior medians, means, and 95% credible intervals (LB: lower bound and UB: upper bound) for model parameters

Parameters	Median	Mean	95% CrI LB	95% CrI UB
β_0	−9.12	−9.11	−10.09	−8.09
β_1	2.26	2.26	2.12	2.41
β_2	0.44	0.44	0.26	0.6
σ_z	0.24	0.12	1.9×10^{-4}	0.25
σ_y	0.27	0.27	0.26	0.28
$\sigma_{\epsilon,C}$	0.05	0.05	0.04	0.07
$\sigma_{\eta,C}$	0.06	0.06	4.9×10^{-3}	0.06
$\sigma_{v,C}$	1.9×10^{-4}	3.9×10^{-4}	9.2×10^{-5}	4.3×10^{-3}
$\sigma_{\xi,C}$	0.015	0.02	2.1×10^{-4}	0.07
$\sigma_{\epsilon,R}$	0.04	0.04	0.03	0.08
$\sigma_{\eta,R}$	0.07	0.07	4.3×10^{-3}	0.08
$\sigma_{v,R}$	2.2×10^{-4}	4.1×10^{-4}	1.01×10^{-4}	3.5×10^{-3}
$\sigma_{\xi,R}$	2.6×10^{-3}	7.6×10^{-3}	1.6×10^{-4}	0.04
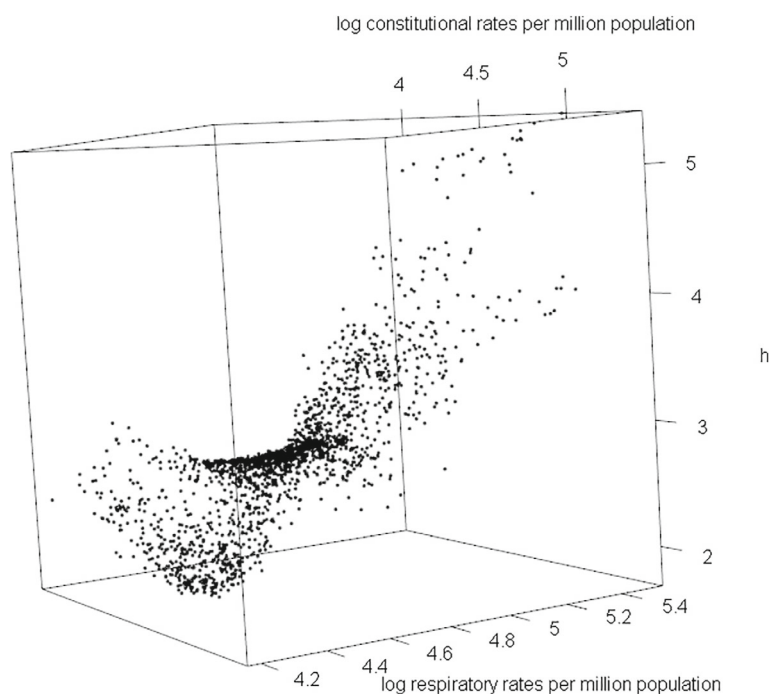
FIGURE 1 Posterior median of $h(\cdot)$ function with respect to constitutional and respiratory rates

Since there is no gold standard to validate our model, we conducted sensitivity analysis by setting aside portions of data, and cross-validation by comparing downscaled influenza incidences with full data and with partial data.

In Figure 4A, we set aside data for 90 days from 1 January 2011 through 1 April 2011. The orange lines show the downscaled influenza incidences with partial data and the black lines show the downscaled influenza incidences with complete data. The red dots are weekly CDC confirmed cases. We can easily see that both downscaled time series are close enough to assure us that the model developed based on the three time series provides meaningful dynamic downscaling and predictions. To further ensure, we also set aside data from peaks and dips of the time series, obtained downscaled influenza incidence trends from those missing portions and compared them with downscaled influenza incidences with
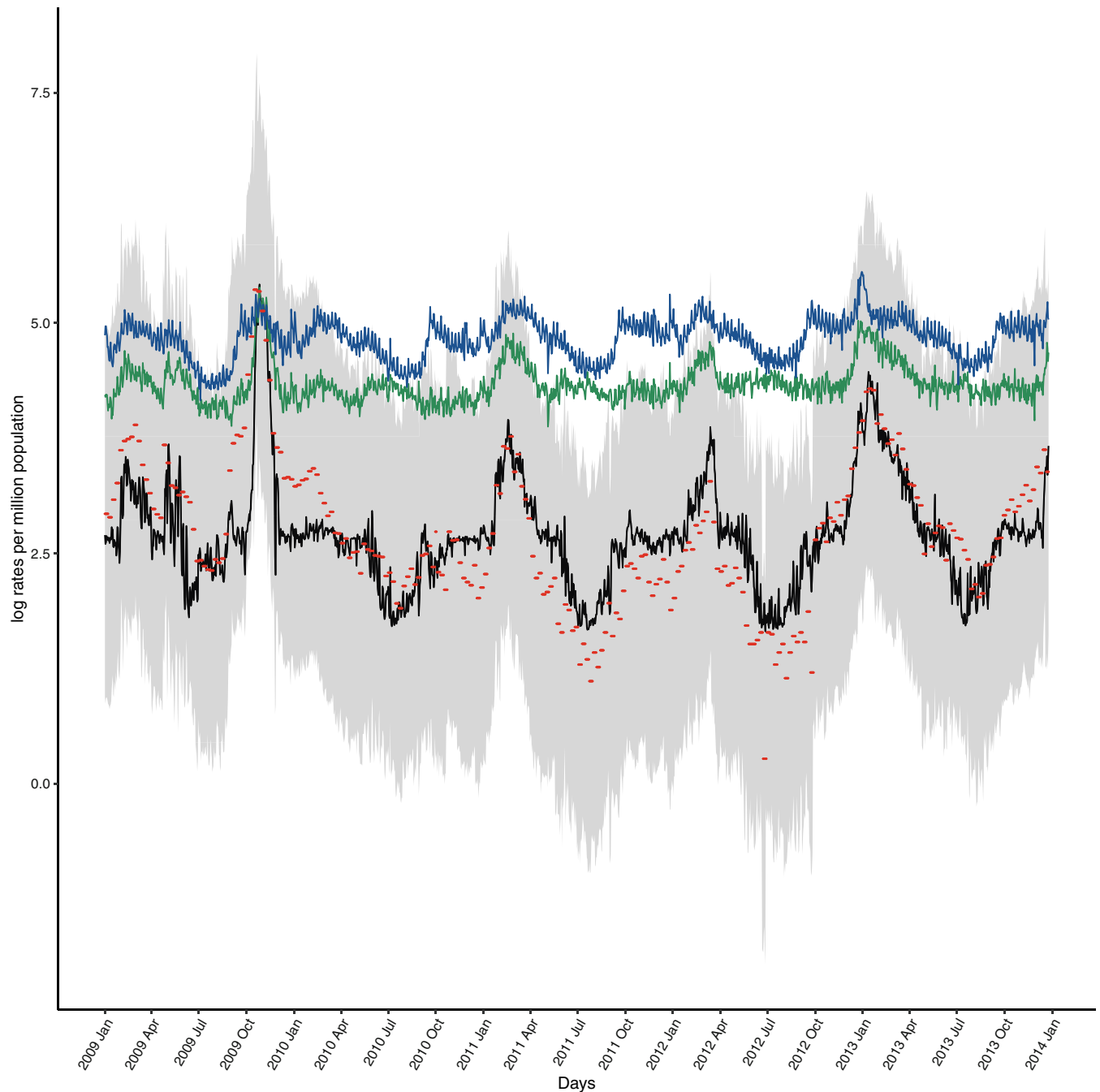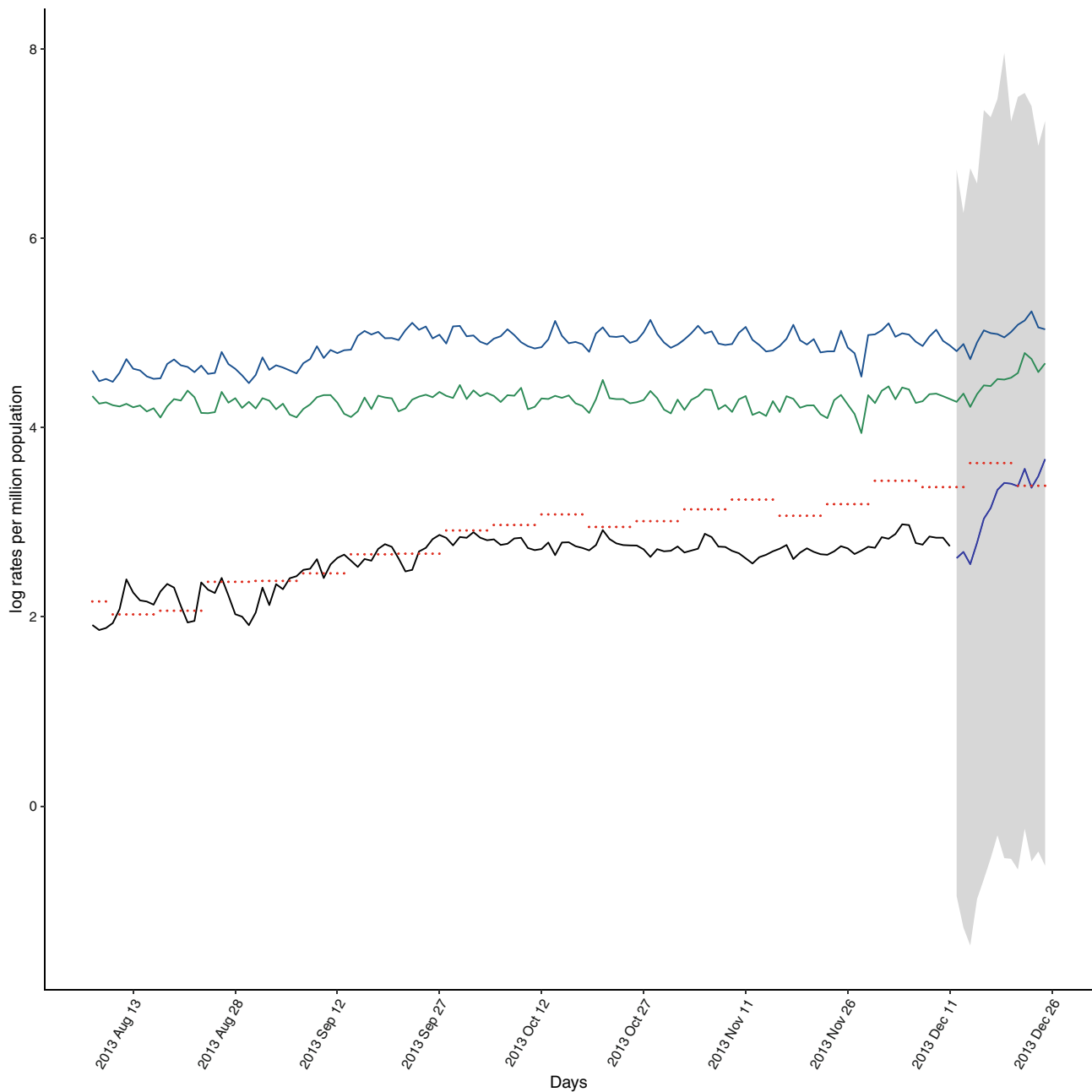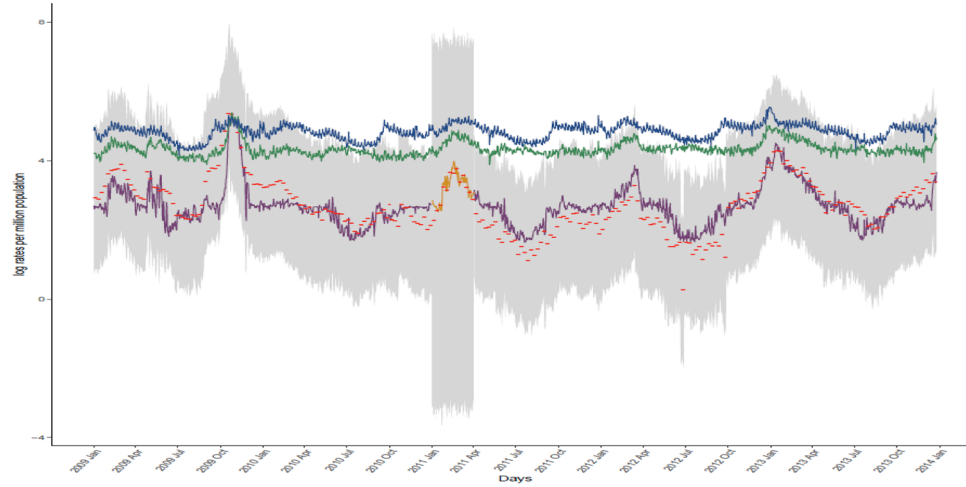
FIGURE 2 Analyses were conducted in log-scale. Downscaled daily log influenza incidences are denoted by black lines, red dots denote the log CDC reported weekly influenza incidences, the gray shades indicate 95% prediction intervals, green lines indicate log daily respiratory rates, and greenish blue lines indicate log daily constitutional rates

full data. Figure 4B singles out a dip in the time frame of June through the end of August 2011, and Figure 4C does the same for a peak in the range from January through the end of June 2012. In each case, we find that the predictions show meaningful trends as we compare them with CDC data.

4.2 | Simulation results

Since we do not have daily influenza incidence rates in the real dataset, we simulated these rates and use them for model fitting. We generated daily constitutional and respiratory rates using the BSTS model described in Equations (2)-(5). The
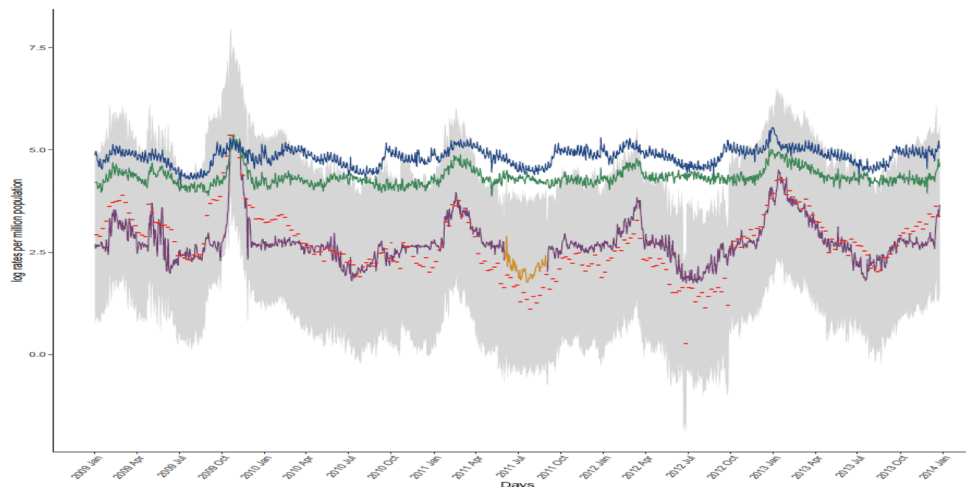
FIGURE 3 Analyses were conducted in log-scale. Downscaled daily log influenza incidences are denoted by black lines, red dots denote the CDC reported weekly log influenza incidences, the gray shades indicate 95% prediction intervals, green lines indicate daily log respiratory rates, and greenish blue lines indicate daily log constitutional rates. The blue lines indicate 14 day predictions for log influenza incidences

generated respiratory and constitutional cases were used in the Equation (6) to get daily cases. Then, weekly rates were generated by adding the daily cases over a week and dividing by population sizes. We simulated 30 unique datasets where we set the variance parameters as follows: $\sigma_{\varepsilon,C} = 0.0428$, $\sigma_{\xi,C} = 0.006$, $\sigma_{\eta,C} = 0.00013$, $\sigma_{v,C} = 0.003$, $\sigma_{\varepsilon,R} = 0.039$, $\sigma_{\xi,R} = 0.071$, $\sigma_{\eta,R} = 0.000156$, $\sigma_{v,R} = 0.003$, $\sigma_{z} = 0.0001$, $\sigma_{y} = 0.0001$. The $\{\mu\}$ terms and $\{v\}$ terms were set at $\mu_0^C = 4.3$, $\mu_0^R = 4.9$, $v_0^C = v_0^R = \tau_0^C = \tau_0^R = 0$. These initial values of μ_0^C and μ_0^R were set as the observed values of constitutional rates and respiratory rates at time 0. The level and slope parameters $\sigma_{\varepsilon,C}$, $\sigma_{\eta,C}$ $\sigma_{\varepsilon,R}$, and $\sigma_{\eta,R}$ were chosen from one realization of MCMC based on real data after burn-in. The noise parameters $\{\sigma_{v,C}, \sigma_{v,R}, \sigma_{z}, \sigma_{y}\}$ were chosen to avoid the overlap in the seasonal perturbations of constitutional and respiratory rates.
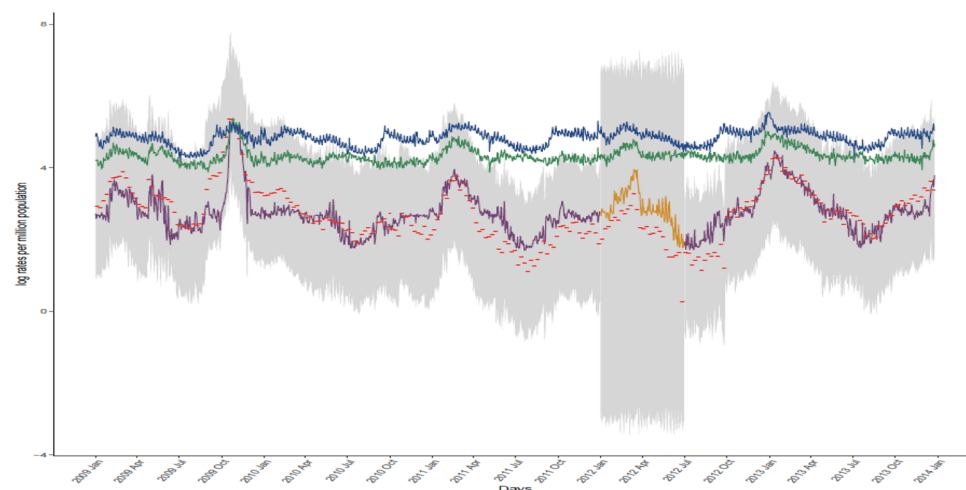
We fitted our models to these 30 unique simulated datasets, each over $T = 1820$ days. The boxplot in Figure 5 shows the summary of mean squared errors (MSE) from these 30 simulated datasets in log rates per million population. The

(A)

(B)

(C)

F I G U R E 4 Analyses were conducted in log scale. In all figures, log downscaled daily influenza incidences are denoted by dark purple lines, red dots denote the CDC reported weekly log influenza incidences, the gray shades indicate 95% prediction intervals, green lines indicate daily log respiratory rates, and greenish blue lines indicate daily log constitutional rates. The orange lines indicate the predicted daily log influenza incidences for the withheld time period
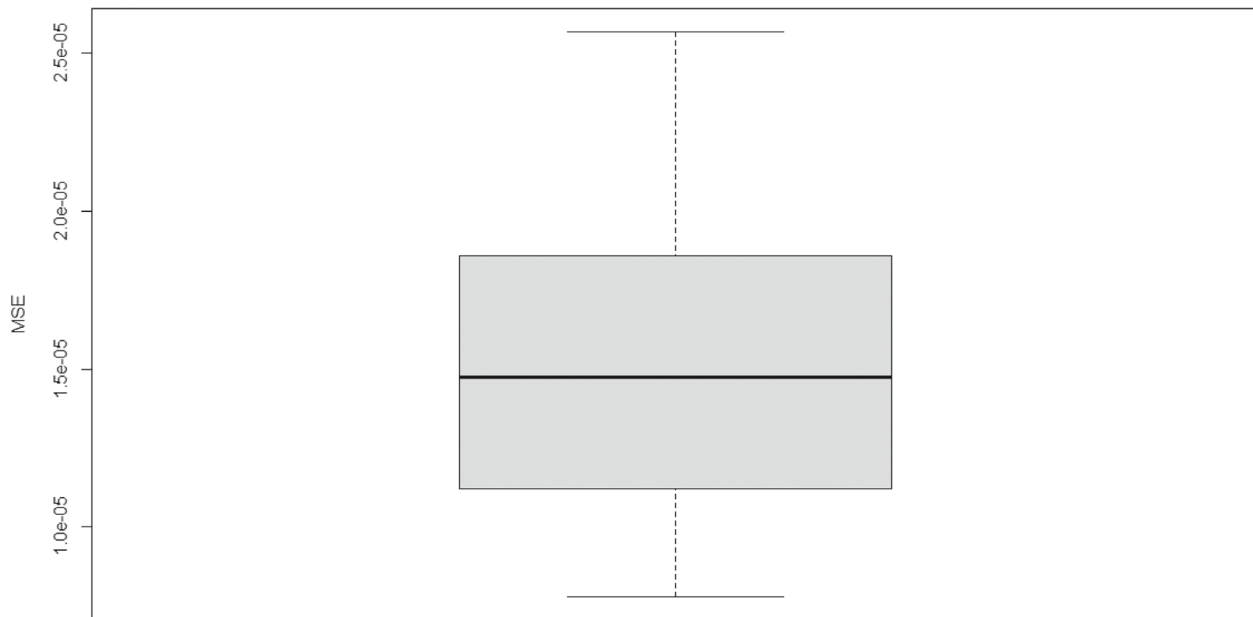
FIGURE 5   Boxplot of mean squared error from 30 simulated datasets

MSE was calculated using the differences between posterior medians and simulated daily ILI rates. Figure 6 shows the results of six out of 30 simulated datasets, where estimated (posterior medians) daily log influenza incidence rates (in black) closely resemble the simulated log daily rates (in red). These figures clearly indicate reasonable model fitting with acceptable uncertainty estimates measured by prediction intervals. We also use posterior predictive checks to assess model fitting.

5 | DISCUSSION

In this article, we present a dynamic downscaling method for estimating daily influenza incidences from weekly data. These estimates can support preparedness by the state and county health systems during flu emergencies. In the United States, the healthcare burden for flu is high and the county health departments are at the forefront to get their communities ready by providing enough vaccines, monitoring severity by tracking hospitalization and death rates, and creating awareness. The granularity of the method in this article allows identifying the peaks and will help in channeling appropriate resources whenever necessary. Bayesian hierarchical modeling[37] provides a natural platform for combining information from disparate data sources in a coherent mathematical framework. Semiparametric modeling[38] boosts the flexibility that was needed to characterize the *unknown* relations between the daily observed respiratory and constitutional symptoms, and the actual influenza incidence. Markov Chain Monte Carlo[23] and Kalman filter[32] based hybrid algorithms provided the efficiency for better convergence and continuous updating of the estimates as new data become available on a daily basis. Posterior distribution based inferences rely on the shrinkage between the data and the prior beliefs of the unknown parameters. We selected noninformative priors for robust Bayesian inference. Additionally, the 95% prediction intervals provide the uncertainty associated with the estimates.

Our algorithm contributes to timely and accurate information that the data collection systems are currently unable to provide due to delays in submission of testing results to the electronic reporting systems at the state level. In the Michigan case, the electronic reporting system is the MDSS, which expects information from health-care providers in real time. It is well known, however, that due to the variety of testing methods and processing criteria, queuing priorities for testing, and exacerbated demand for testing during influenza emergencies, submission of confirmed cases can take up to 2 weeks.[39] Therefore, daily trends of influenza incidence are currently
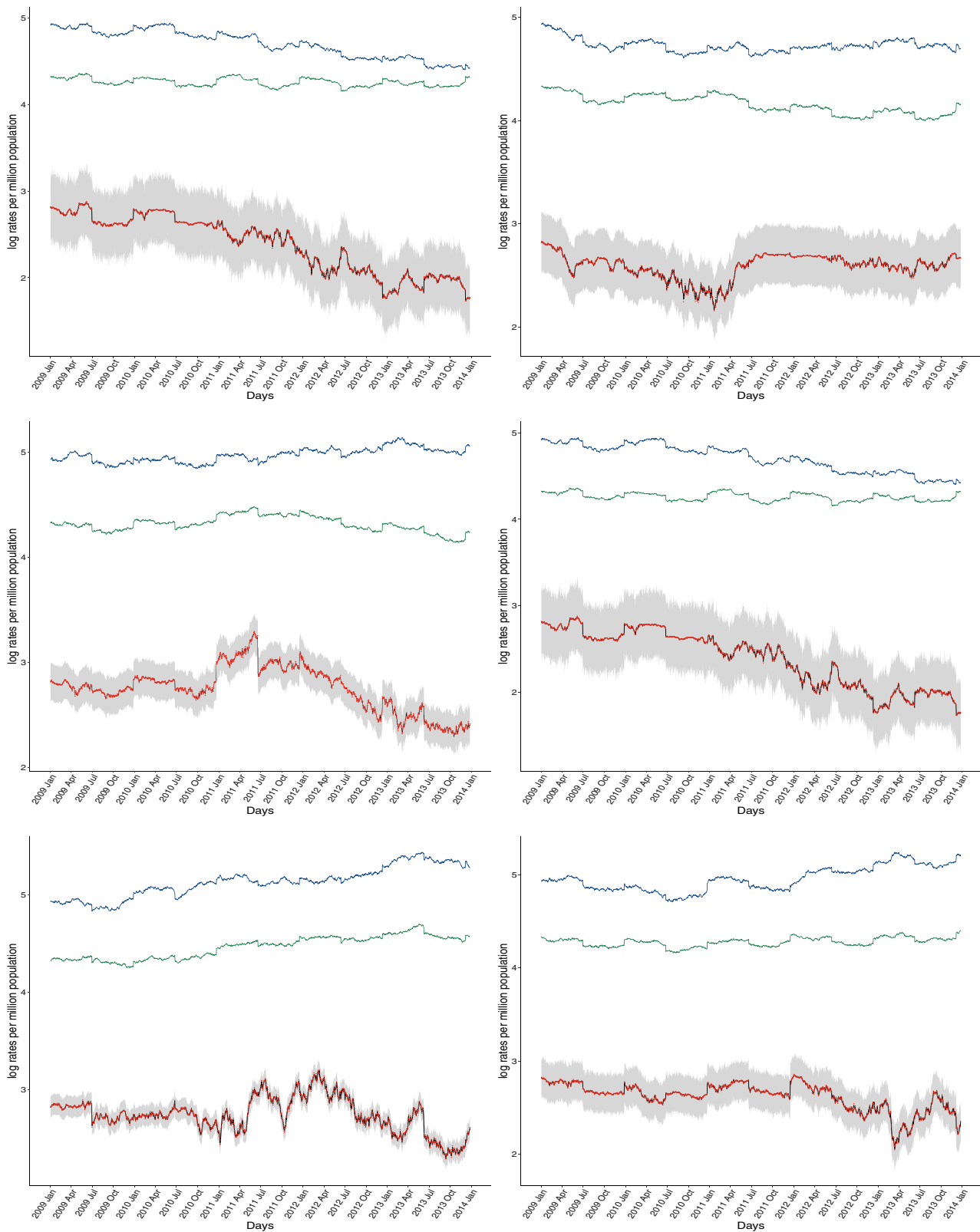
FIGURE 6 Results from 6 simulations out of 30 simulations done. Black line: Estimated downscaled daily log influenza incidences per 100 000 million. Red dotted line: Simulated log daily influenza incidences per 100 000 million. Gray shades: 95% prediction intervals. Green line: Simulated log daily respiratory rates. Greenish blue line: Simulated log daily constitutional rates

observed weekly and corrected as more confirmed specimens are submitted to the system. Our algorithm is suitable to accommodate the time resolution from weekly to daily, which corrects for the operational delays in the submission.

From our work, we found that MDHHS constitutional and respiratory trends are good proxies for estimating benchmark daily trends of influenza incidence. There are several advantages in the use of constitutional and respiratory trends: They are retrieved with a one-day delay only, and they are based on reported symptoms and not on confirmatory testing, which makes them independent from the specificity, sensitivity, timeliness, and variability of the testing infrastructure built in each reporting healthcare provider. Although the constitutional and respiratory trends present underreporting, our work shows that they are meaningful predictors contributing their fluctuation patterns in the estimation of the daily influenza incidence trend.

In this article, a semi-parametric function was estimated to fuse the daily constitutional and respiratory trends. We found that the shrinkage contribution associated with the constitutional trends in the basis function was higher than the contribution of the respiratory trends. This finding is justifiable as the respiratory trends might be subject to more noise during the influenza seasons than the constitutional. The EDs receive many patients with respiratory symptoms from non-flu related diseases (eg, heart failure, chronic obstructive pulmonary disease, asthma, acute coronary syndrome, and arrhythmia). In contrast, the constitutional burden might be mostly attributed to flu and flu related diseases during the influenza seasons.

Our modeling framework considers that the constitutional and respiratory trends are proxies of influenza activity, which is a reasonable assumption during the flu seasons or during an outbreak from a recognized causative flu strain. During flu seasons, however, the trends are only helpful to track ILI, and determining which virus is driving the increasing trend is a recognized challenge.

Constitutional and respiratory trends might also work to track an outbreak caused by a respiratory virus, as long as the outbreak does not overlap with a flu related case surge. Previous research demonstrates how increasing ILI trends were heavily correlated with the respiratory syncytial virus outbreak in Washington State during December 2019, and with the SARS-CoV-2 outbreak in the United States during March 2020.[40]

6 | CONCLUSION

We developed a dynamic downscaling method for estimating daily influenza incidences from weekly data. Despite the assumptions and limitations described in the discussion section, our work possesses several strengths: (1) our method is based on two powerful data sources, which are also widely and easily available to the MDHHS. Most of the state health departments in the U.S. have an established infrastructure to retrieve influenza incidence trends like the constitutional and respiratory. In addition, the CDC's weekly trends of cases confirmed with a flu virus are publicly available; (2) since we used structured surveillance data/systems to build our framework, public health officials are better suited to question and validate the results of the down-scaled trends, if our algorithm is implemented; (3) our fast algorithms facilitates updated predictions as the new data become available; and (4) our modeling framework has the ability to capture adaptive seasonal effects.

When infrastructure limits the capacity of testing, EDs report constitutional and respiratory cases that are useful proxies for estimating surges and the true incidences during epidemics. At the time of writing this article, it is unknown how SARS-CoV-2 and flu will interact during flu season. However, our model accounts for symptoms that are common in both diseases, and hence our model contributes to the groundwork that is needed to understand such interactions in the future.

DATA AVAILABILITY STATEMENT

The article used publicly available data from the Centers for Disease Control and Prevention (CDC) and data obtained from the Michigan Department of Health and Human Services (MDHHS). The data from the MDHHS cannot be shared. However, we created a simulated dataset and shared it on Github along with our algorithm written in R. Github Link: https://github.com/dan-han-mathematics/dowscaling-ili.

ORCID

Rajib Paul ⓘ https://orcid.org/0000-0002-5120-7525

REFERENCES

1. World Health Organization WHO. *8 things to know about pandemic influenza; 2019*. World Health Organization. https://www.who.int/news-room/feature-stories/detail/8-things-to-know-about-pandemic-influenza. Accessed January 19,2019.
2. Danielle Iuliano A. Estimates of global seasonal influenza-associated respiratory mortality: a modelling study. *Lancet*. 2018;391(10127):1285-1300.
3. Pearce N, Vandenbroucke JP, VanderWeele TJ, Greenland S. Accurate statistics on COVID-19 are essential for policy guidance and decisions; 2020.
4. Centers for Disease Control and Prevention. U. S. influenza surveillance system: purpose and methods; 2021. https://www.cdc.gov/flu/weekly/overview.htm#anchor_1539281266932. Accessed January 13, 2021.
5. Althouse BM. Enhancing disease surveillance with novel data streams: challenges and opportunities. *EPJ Data Sci*. 2015;4(1):17.
6. Viboud C, Vespignani A. The future of influenza forecasts. *Proc Nat Acad Sci*. 2019;116(8):2802-2804.
7. Chretien J-P, George D, Shaman J, Chitale RA, McKenzie FE. Influenza forecasting in human populations: a scoping review. *PLoS One*. 2014;9(4):e94130.
8. Nsoesie EO, Brownstein JS, Ramakrishnan N, Marathe MV. A systematic review of studies on forecasting the dynamics of influenza outbreaks. *Influenza Other Respirat Virus*. 2014;8(3):309-316.
9. Reich NG, Brooks LC, Fox SJ, et al. A collaborative multiyear, multimodel assessment of seasonal influenza forecasting in the United States. *Proc Nat Acad Sci*. 2019;116(8):3146-3154.
10. Jiang X, Wallstrom G, Cooper GF, Wagner MM. Bayesian prediction of an epidemic curve. *J Biomed Inform*. 2009;42(1):90-99.
11. Sebastiani P, Mandl KD, Szolovits P, Kohane IS, Ramoni MF. A Bayesian dynamic model for influenza surveillance. *Stat Med*. 2006;25(11):1803-1816.
12. Logan C. Brooks, David C. Farrow, Sangwon Hyun, Ryan J. Tibshirani, and Roni Rosenfeld. Flexible modeling of epidemics with an empirical Bayes framework. *PLoS Comput Biol*, 11(8):e1004382, August 2015.
13. Teresa K. Yamana, Sasikiran Kandula, and Jeffrey Shaman. Individual versus superensemble forecasts of seasonal influenza outbreaks in the United States. *PLOS Comput Biol*, 13(11):e1005801, 2017.
14. Osthus D, Gattiker J, Priedhorsky R, Del Valle SY. Dynamic Bayesian influenza forecasting in the United States with hierarchical discrepancy (with Discussion). *Bayesian Anal*. 2019;14(1):261-312.
15. Logan C. Brooks, David C. Farrow, Sangwon Hyun, Ryan J. Tibshirani, and Roni Rosenfeld. Nonmechanistic forecasts of seasonal influenza with iterative one-week-ahead distributions. *PLoS Comput Biol*, 14(6):e1006134, June 2018.
16. Ray EL, Sakrejda K, Lauer SA, Johansson MA, Reich NG. Infectious disease prediction with kernel conditional density estimation: infectious disease prediction with kernel conditional density estimation. *Stat Med*. 2017;36(30):4908-4929.
17. Ray EL, Reich NG. Prediction of infectious disease epidemics via weighted density ensembles. *PLoS Comput Biol*. 2018;14(2):e1005910.
18. Rhodes CJ, Hollingsworth TD. Variational data assimilation with epidemic models. *J Theor Biol*. 2009;258(4):591-602.
19. M. G. Roberts. Epidemic models with uncertainty in the reproduction number. *J Math Biol*, 66(7):1463–1474, June 2013.
20. Lytras T, Gkolfinopoulou K, Bonovas S, Nunes B. Fluhmm: a simple and flexible Bayesian algorithm for sentinel influenza surveillance and outbreak detection. *Stat Methods Med Res*. 2019;28(6):1826-1840.
21. Aguirre A, Gonzalez E. The feasibility of forecasting influenza epidemics in cuba. *Memórias do Instituto Oswaldo Cruz*. 1992;87(3):429-432.
22. Xue Y, Janjic Z, Dudhia J, Vasic R, De Sales F. A review on regional dynamical downscaling in intraseasonal to seasonal simulation/prediction and major factors that affect downscaling ability. *Atmosph Res*. 2014;147:68-85.
23. Robert C, Casella G. *Monte Carlo Statistical Methods*. New York, NY: Springer Science & Business Media; 2013.
24. Scott SL, Varian HR. Predicting the present with Bayesian structural time series. *Int J Math Modell Numer Optim*. 2014;5(1-2):4-23.
25. Centers for Disease Control and Prevention. Cdc's world health organization (who) collaborating center for surveillance, epidemiology and control of influenza; 2017. Accessed July 24, 2018. https://www.cdc.gov/flu/weekly/who-collaboration.htm
26. Rose AN, Nagle NN. Validation of spatiodemographic estimates produced through data fusion of small area census records and household microdata. *Comput Environ Urban Syst*. 2017;63:38-49.
27. Zhang X, Holt JB, Yun S, Hua L, Greenlund KJ, Croft JB. Validation of multilevel regression and poststratification methodology for small area estimation of health indicators from the behavioral risk factor surveillance system. *Am J Epidemiol*. 2015;182(2):127-137.
28. Gevaert CM, García-Haro FJ. A comparison of starfm and an unmixing-based algorithm for landsat and modis data fusion. *Remote Sens Environ*. 2015;156:34-44.

29. Rundel CW, Schliep EM, Gelfand AE, Holland DM. A data fusion approach for spatial analysis of speciated pm2. 5 across time. *Environmetrics*. 2015;26(8):515-525.

30. Bai L, Long D, Yan L. Estimation of surface soil moisture with downscaled land surface temperatures using a data fusion approach for heterogeneous agricultural land. *Water Resour Res*. 2019;55(2):1105-1128.

31. Andrew C Harvey. *Forecasting, structural time series models and the Kalman filter*. Cambridge, UK: Cambridge University Press; 1990.

32. Durbin J, Koopman SJ. *Time Series Analysis by State Space Methods*. Oxford: Oxford University Press; 2012.

33. Qiu J, Jammalamadaka SR, Ning N. Multivariate Bayesian structural time series model. *J Mach Learn Res*. 2018;19(1):2744-2776.

34. Kaufman L, Rousseeuw PJ. *Finding Groups in Data: An Introduction to Cluster Analysis*. Vol 344. New York: John Wiley & Sons; 2009.

35. Wand MP, Coull BA, French JL, Ganguli B, Kammann EE, Staudenmayer J, Zanobetti A. Semipar 1.0. r package; 2005. http://cran.r-project.org

36. Maechler M. cluster: finding groups in data: cluster analysis extended Rousseeuw et al., 2021. R package version 3.4.0.

37. Gelman A, Carlin JB, Stern HS, Dunson DB, Vehtari A, Rubin DB. *Bayesian Data Analysis*. Boca Raton, FL: CRC Press; 2013.

38. Ruppert D, Wand MP, Carroll RJ. *Semiparametric Regression*. Vol 12. New York, NY: Cambridge University Press; 2003.

39. Prieto DM, Das TK, Savachkin AA, Uribe A, Izurieta R, Malavade S. A systematic review to identify areas of enhancements of pandemic simulation models for operational use at provincial and local levels. *BMC Public Health*. 2012;12(1):251.

40. Silverman JD, Hupert N, Washburne AD. Using influenza surveillance networks to estimate state-specific prevalence of SARS-CoV-2 in the United States. *Sci Transl Med*. 2020;12(554):eabc1126.