Bayesian Analysis (2019) **14**, Number 1, pp. 261–312

Dynamic Bayesian Influenza Forecasting in the United States with Hierarchical Discrepancy (with Discussion)

Dave Osthus[*], James Gattiker[†], Reid Priedhorsky[‡], and Sara Y. Del Valle[§]

Abstract. Timely and accurate forecasts of seasonal influenza would assist public health decision-makers in planning intervention strategies, efficiently allocating resources, and possibly saving lives. For these reasons, influenza forecasts are consequential. Producing timely and accurate influenza forecasts, however, have proven challenging due to noisy and limited data, an incomplete understanding of the disease transmission process, and the mismatch between the disease transmission process and the data-generating process. In this paper, we introduce a dynamic Bayesian (DB) flu forecasting model that exploits model discrepancy through a hierarchical model. The DB model allows forecasts of partially observed flu seasons to borrow discrepancy information from previously observed flu seasons. We compare the DB model to all models that competed in the CDC's 2015–2016 and 2016–2017 flu forecasting challenges. The DB model outperformed all models in both challenges, indicating the DB model is a leading influenza forecasting model.

Keywords: probabilistic forecasting, hierarchical modeling, discrepancy, influenza.

1 Introduction

Influenza is a respiratory illness caused by the influenza virus that hospitalizes hundreds of thousands of people and affects millions in the United States annually (Rolfes et al., 2016). Influenza also poses a significant burden on the U.S. economy through hospitalization costs and lost productivity from missing work (Molinari et al., 2007). Flu surveillance is a collaborative effort between the Centers for Disease Control and Prevention (CDC) and many state and local healthcare providers, clinics and emergency rooms (Centers for Disease Control and Prevention, 2016b). Monitoring the prevalence and geographic distribution of the flu is critical for targeted flu prevention strategies, such as vaccination campaigns and public education programs.

In addition to flu monitoring, the CDC is also interested in flu forecasting. To better understand flu forecasting capabilities and to improve their usefulness to public health decision-makers, the CDC organized the first national flu forecasting competition in 2013 (Biggerstaff et al., 2016). Participation in the challenge included over a dozen models. The CDC defined forecasting targets relevant to public health decision-maker.

[*] Statistical Sciences Group, Los Alamos National Laboratory, dosthus@lanl.gov
[†] Statistical Sciences Group, Los Alamos National Laboratory
[‡] High Performance Computing Environments, Los Alamos National Laboratory
[§] Information Systems and Modeling, Los Alamos National Laboratory

https://doi.org/10.1214/18-BA1117

These targets included the onset of the flu season, as well as the peak timing (PT) and peak intensity (PI) of the flu season. One-to-four week ahead forecasts (i.e., short term forecasts) were added as targets for the 2014–2015 challenge. From the 2013–2014 flu forecasting challenge, the CDC concluded that though flu forecasting is possible, much work remains. Flu forecasting is in its infancy and a concentrated effort to improve forecasting capabilities is needed in order for forecasts to be practically useful. The CDC has continued to organize an annual flu forecasting competition since the inaugural 2013–2014 challenge as a continuing effort to scope flu forecasting capabilities and provide an environment for collaboration and iterative improvement.

Nsoesie et al. (2014) and Chretien et al. (2014) provide reviews of the flu forecasting landscape.[1] Flu forecasting models can be broadly categorized into four groups: mechanistic models, agent-based models, machine learning/regression models, and data-assimilation/dynamic models.

- **Mechanistic models** are differential-equation model descriptions of the disease transmission mechanism. They include a class of models referred to as compartmental models that partition a population into compartments and mathematically describe how individuals in the population move between compartments (e.g., Towers and Feng, 2009).
- **Agent-based models** simulate a population that mimics a real population using, for example, U.S. Census data to match various aspects of the simulated population to a real population (e.g., demographic information). The disease is then propogated via simulation through the simulated population and used to approximate the transmission of disease through a real population (e.g., Mniszewski et al., 2008; Grefenstette et al., 2013).
- **Machine learning/regression models** are models that learn patterns in historical flu outbreaks and leverage those patterns for forecasting new flu seasons. This group includes such approaches as statistical time series (e.g., Soebiyanto et al., 2010), linear or regularized regression (e.g., Bardak and Tan, 2015), clustering (e.g., Viboud et al., 2003), and nonparametric approaches (e.g., Brooks et al., 2015). The machine learning/regression model approach to flu forecasting is often characterized by the absence of a mechanistic model.
- **Data-assimilation/dynamic models** usually involve embedding a mechanistic model into a probabilistic framework, allowing for the explicit modeling of the disease transmission process and observational noise (e.g., Osthus et al., 2017;

[1] Apart from the disease forecasting literature, much work has gone into parameter estimation in the context of embedding mechanistic compartmental models into a statistical framework (e.g., Boys et al., 2008; Pokharel and Deardon, 2016; Angulo et al., 2012; Jandarov et al., 2012, 2014). Pokharel and Deardon (2016) and Angulo et al. (2012) discuss approximations to the likelihood to improve computational efficiency, while Jandarov et al. (2012) use an approximate Bayesian computation approach in the context of parameter estimation. Jandarov et al. (2014) adds a discrepancy function in the form of a Gaussian process to improve parameter estimation interpretability for the purposes of answering scientifically meaningful questions. Our work deviates from all previously mentioned in that forecasting real-world data, not parameter estimation or interpretation, is our exclusive interest.

> Hickmann et al., 2015; Shaman et al., 2013; Dukic et al., 2012). That is, the dynamic modeling approach combines two sources of uncertainty in the modeling; parametric uncertainty in the mechanistic model and random uncertainty in the observations.

Our modeling approach extends the data assimilation/dynamic modeling approach and can be viewed as a combination of the machine learning/regression approach and the data assimilation/dynamic modeling approach. Our model, referred to as a *dynamic Bayesian (DB) model*, explicitly accounts for systematic deviations between the mechanistic model and the data that are unable to be explained by pure observational noise. This systematic deviation is referred to as *model discrepancy* and is modeled with a flexible, statistical model. Discrepancy modeling is an often used and effective modeling approach in the field of computer experiments, where systematic deviations between mechanistic models and data can be common (e.g., Kennedy and O'Hagan, 2001; Bayarri et al., 2007; Higdon et al., 2008; Brynjarsdóttir and O'Hagan, 2014).

Including a discrepancy model is an appealing way to account for the systematic inadequacy of the mechanistic model. The basic insight leading to the inclusion of a discrepancy model in our DB model is that the disease transmission model and the data-generating model are not equivalent. Disease transmission is merely a component of the data-generating process. Thus, even if a mechanistic model were able to accurately identify the disease transmission process, there may still be a systematic discrepancy between the disease transmission model and the data, introducing a source of uncertainty unable to be accounted for by observational noise alone.

Though effective for data fitting, discrepancy modeling can make extrapolation (i.e., forecasting) challenging due to potential overfitting (Bayarri et al., 2007). As opposed to previous Bayesian flu modeling approaches where flu tracking and parameter estimation were of interest (e.g., Coelho et al., 2011; Dukic et al., 2012), *our exclusive interest is forecasting*. Thus, discrepancy modeling must be done with care. We address the discrepancy/forecasting issue by modeling the discrepancy hierarchically across all flu seasons. This allows us to borrow common discrepancy structure observed in past seasons in the forecasting of the flu for the current season. The hierarchical discrepancy model thus serves as a balance between the flexibility needed to match the statistical model to data and the structure needed to make useful and valid forecasts.

This paper makes contributions and advances in the following ways. 1) We introduce and demonstrate the importance of discrepancy modeling to the growing and consequential field of flu forecasting. Discrepancy modeling is done hierarchically, allowing information to be shared across available flu seasons. 2) We demonstrate the superiority of our approach relative to all models that competed in the CDC's 2015–2016 and 2016–2017 flu forecasting challenges, providing yet another instance where discrepancy modeling is not only conceptually appealing but also practically effective. 3) In an effort to advance flu forecasting capabilities, much effort has been spent identifying possibly useful, nontraditional data sources such as Google (Ginsberg et al., 2009) and Wikipedia (Generous et al., 2014). Alternatively, as we demonstrate, flu forecasting can be improved through carefully made modeling choices, making use of the available traditional data hierarchically.

The paper is laid out as follows. In Section 2, we present the data. In Sections 3 and 4, we present the mechanistic model and statistical DB model, respectively. We investigate and assess the DB model in Sections 5.1 and 5.2, respectively. The DB model is compared to all participating models in the 2015–2016 and 2016–2017 flu forecasting challenges organized by the CDC in Section 5.3. We conclude with a discussion in Section 6.

2 Data

The CDC performs influenza surveillance in the United States via a multitude of surveillance efforts including virologic, outpatient, mortality, and hospitalization surveillance systems (Centers for Disease Control and Prevention, 2016b). In this paper, we focus exclusively on outpatient illness surveillance. Symptomatic information on patient visits to healthcare providers is collected through the United States outpatient influenza-like illness surveillance network (ILINet). ILINet is a collection of almost 3,000 healthcare providers across the United States. These participating healthcare providers supply information to the CDC regarding the number of patients seen for any reason and the number of patients seen with a diagnosed influenza-like illness on a weekly basis. An influenza-like illness is defined as a temperature greater than or equal to 100 degrees Fahrenheit and a cough or sore throat with no known cause other than influenza. It is important to note an influenza-like illness diagnosis and an influenza diagnosis are not equivalent. Many diseases have flu-like symptoms prior to fully developing such as measles, rubella, tuberculosis, food poisoning, dengue, and malaria. ILINet is ill-equipped and not designed to discriminate between the flu and diseases with flu-like symptoms. A flu diagnosis requires some form of laboratory test. We model and forecast influenza-like illness in this paper.[2]

The CDC aggregates, organizes, and ultimately releases influenza-like illness information to the public weekly throughout the year at both the national and health and human service region levels (Health and Human Services, 2015). In this paper, we focus exclusively on national level influenza-like illness surveillance. The proportion of the population with an influenza-like illness is estimated by the CDC with the quantity weighted influenza-like illness (wILI) where wILI is, “the percentage of patient visits to healthcare providers for influenza-like illness reported each week weighted on the basis of state population”, (Centers for Disease Control and Prevention, 2016b).

Figure 1 shows wILI for seasonal influenza from 1998 through 2015, excluding the pandemic H1N1 flu seasons 2008 and 2009. We focus on seasonal influenza rather than the more severe and substantially less frequent pandemic flu seasons as seasonal and pandemic flu transmission dynamics are appreciably different. Seasonal flu outbreaks

[2] In the description of the statistical model in Section 4, we are somewhat loose about the influenza-like illness versus influenza distinction. Shaman et al. (2013) proposed a scaling factor that multiplies ILI by the proportion of ILI patients testing positive for influenza as a better estimate of the proportion of influenza infectiousness in the population. We do not make use of this scaling factor in the statistical modeling that follows because our exclusive interest is forecasting, not parameter inference or scientific interpretation, and our statistical model has sufficient flexibility to account systematic differences between the mechanistic model outlined in Section 3 and the data. However, if model interpretation was of interest, such a scaling factor could be useful.

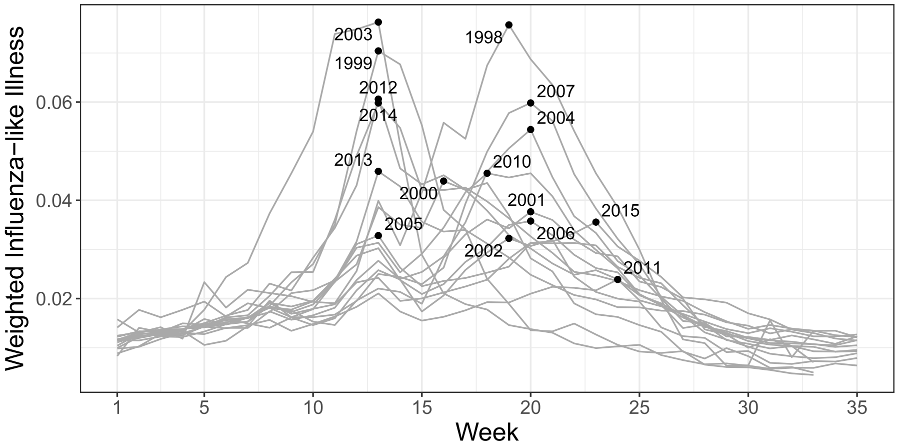


Figure 1: Weighted influenza-like illness for flu seasons 1998 through 2015, sans H1N1 seasons 2008 and 2009. Grey lines correspond to flu season trajectories. Week 1 is roughly the first week of October, while week 35 is roughly the last week of May. The black points are located at the peak timing (x-axis) and peak intensity (y-axis) of their respective flu season. Peak timing occurs between week 13 (roughly the end of December) and week 24 (roughly the middle of March).

have a relatively predictable profile, as can be seen in Figure 1. The population often has partial immunity to the circulating virus(es) of seasonal influenza and it occurs nearly ever year in the United States. Pandemic influenza follows much less predictable transmission patterns, due in part to the relatively low immunity in the population to the new, yet to be seen strain of influenza. As a result, pandemic influenza is typically non-recurring. There have been four instances of pandemic influenza since 1900: 1918, 1957, 1968, and 2009 (Centers for Disease Control and Prevention, 2016c).

A typical flu season begins in October and lasts until as late as May, thus, the 1998 flu season refers to the season starting in 1998 and ending in 1999. In this paper, flu season week 1, referred to as "week 1", corresponds to Morbidity and Mortality Weekly Report (MMWR) week 40. MMWR is a common epidemiological dating system used for reporting purposes (Centers for Disease Control and Prevention, 2016a). Week 1 roughly corresponds to the first week of October while week 35 roughly corresponds to the last week of May.

In Figure 1, the point corresponds to the peak timing (PT) and peak intensity (PI) of each flu season. We see most flu seasons either peak early (six flu seasons peaked on week 13 – roughly the end of December) or late (nine flu season peaked between week 18 and 24 – roughly the beginning of February through the middle of March), with the 2000 flu season peaking during week 16. All flu seasons in Figure 1 exhibit a similar pattern; wILI is low at the beginning of the flu season, increases to a maximum in the middle of the flu season, and reverts to low levels by the end of the flu season. Though each season shares this general pattern, heterogeneity exists between flu seasons. Some flu

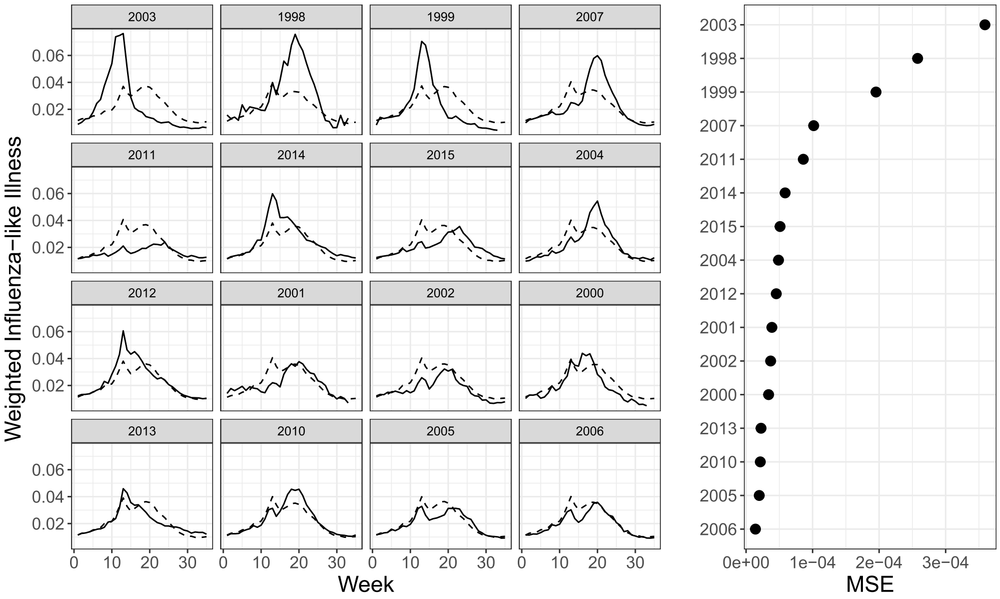


Figure 2: (Left) Weighted influenza-like illness (solid line) for each season and the average weighted influenza-like illness for all other flu seasons (dashed line). Seasons are ordered from top, left to bottom, right by decreasing mean-squared error (MSE). (Right) The MSE between each flu season and the average of the other flu seasons. Flu seasons 2003, 1998, and 1999 are the most "atypical" flu seasons.

seasons appear to deviate from "typical" flu seasons more than others. This observation is illustrated in Figure 2. Flu seasons 1998, 1999, and 2003 most significantly deviate from "typical" as calculated by the mean-squared error (MSE) between each flu season and the week-specific average of all other flu seasons. Other flu seasons, such as 2005, 2006, and 2010 are the most "typical" by this same measure.

3 Susceptible-Infectious-Recovered Model

The susceptible-infectious-recovered (SIR) model was introduced in the seminal work of Kermack and McKendrick (1927) and is considered the foundation for modern public health (Weiss, 2013). The SIR model is a mechanistic model that describes how an infectious disease spreads through a closed population via the following set of nonlinear, ordinary differential equations:

$$\frac{dS}{dt} = -\beta SI, \qquad \frac{dI}{dt} = \beta SI - \gamma I, \qquad \frac{dR}{dt} = \gamma I, \tag{1}$$

where $\beta > 0$ is the disease transmission rate, $\gamma > 0$ is the recovery rate, and $1/\gamma$ is the average infectious period. Under the SIR model, individuals in the population are partitioned into three mutually exclusive and exhaustive compartments: a susceptible, infectious, and recovered compartment. The SIR model describes the rate at which individuals move from being susceptible to infectious and finally recovered. S, I, and R in (1) represent proportions of the population, such that $S + I + R = 1$ for all times. An example of an SIR trajectory is shown in Figure 3.

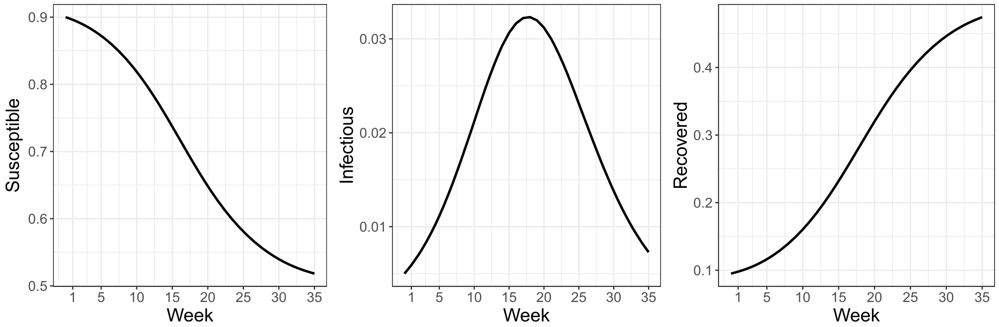


Figure 3: Solution to a susceptible-infectious-recovered model with $S_0 = 0.9$, $I_0 = 0.005$, $R_0 = 0.095$, $\gamma = 0.55$, and $\beta = 0.8$, where S_0, I_0, and R_0 are the proportions of the population susceptible, infectious, and recovered from the disease at time 0.

The trajectories of the susceptible and recovered SIR compartments are monotonically non-increasing and non-decreasing, respectively. All SIR infectious trajectories can be partitioned into two designations: epidemics and non-epidemics. These designations are completely determined by the relationship between S_0, the proportion of the population initially susceptible to the disease, and ρ, where $\rho = \gamma/\beta$ (Weiss, 2013). An epidemic designation occurs when $S_0/\rho > 1$, where S_0/ρ is the effective reproductive number. An epidemic graphically corresponds to an infectious trajectory that monotonically increases to a maximum followed by a monotonic decrease to zero as time goes to infinity. The infectious trajectory shown in Figure 3 is an epidemic. A non-epidemic designation occurs when $S_0/\rho \leq 1$, graphically meaning the infectious trajectory monotonically decreases from I_0, the proportion of the population initially infectious with the disease, to zero as time goes to infinity. As can be seen in Figure 1, every flu season we consider exhibits the general shape of an epidemic, where wILI is low at week 1 of the flu season, increases to a maximum between weeks 13 and 24, and declines to low levels by week 35.

Figure 4 plots the best fit SIR trajectory and wILI for each flu season constrained to $S_0 = 0.9$. Setting $S_0 = 0.9$ is an identifying constraint, as an SIR solution is effectively unidentifiable when only data related to the infectious trajectory is available (Capaldi et al., 2012). Estimating S_0, rather than fixing it, would require additional data related to non-infectious compartments of the SIR model. We see the SIR trajectories match the general shape of wILI. These trajectories do not, however, replicate some of the more nuanced structure of wILI. For instance, some seasons exhibit a double peak (e.g., 2002, 2005, 2006, and 2011). The SIR model, however, is incapable of capturing two peaks within a flu season; it can only capture one. Another interesting feature of wILI

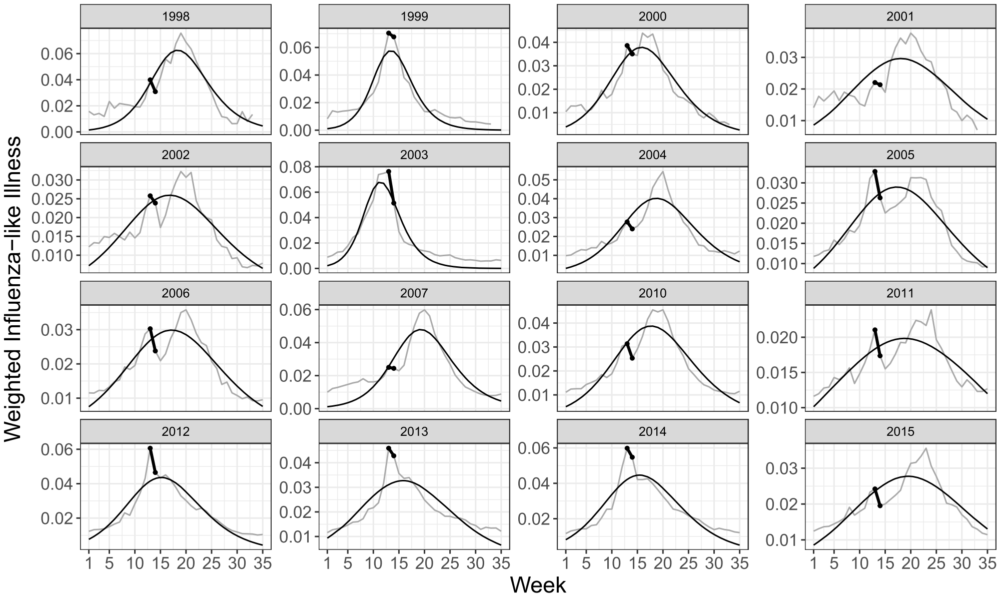


Figure 4: Fitted susceptible-infectious-recovered model (black curve) to weighted influenza-like illness (grey) by flu season. The black line segment denotes the systematic decline in weighted influenza-like illness from week 13 to week 14 in all flu seasons.

happens at weeks 13 and 14. In every flu season, wILI on week 13 is larger than wILI on week 14. For some flu seasons, this downturn in wILI from week 13 to 14 signifies the peak of the flu season; flu seasons 1999, 2003, 2005, 2012, 2013, 2014 all peak on week 13. For the other flu seasons, however, the decline in wILI from week 13 to week 14 does not signify the peak of the flu season as they all exhibit peaks later in the season. It is not known exactly why wILI reliably declines from week 13 to 14, though explanations have been posited, such as a change in disease transmission during winter holidays (Ewing et al., 2016; Garza et al., 2013; Huang et al., 2014). For the purposes of forecasting wILI, what is important is that the decline in wILI from week 13 to 14 is reliable and unable to be captured by the SIR model.

Figure 5 plots the residuals between wILI and the fitted SIR curves for all flu seasons in Figure 4, referred to as discrepancy trajectories, along with the average discrepancy trajectory. Figure 5 articulates the systematic deviations between best fit SIR trajectories and wILI. The SIR model tends to under estimate wILI near weeks at the start (week 1), end (week 35), and peak (weeks 13 and 20) of the flu season, while systematically over estimating wILI near weeks 10, 15, and 27.

Figure 4 suggests describing wILI with even the best fitting SIR model is inadequate. Furthermore, Figure 5 suggests the discrepancy between the best fit SIR model and wILI cannot plausibly be described by random error alone as there is structure across both seasons and time in the discrepancy. In the next section, we present the dynamic

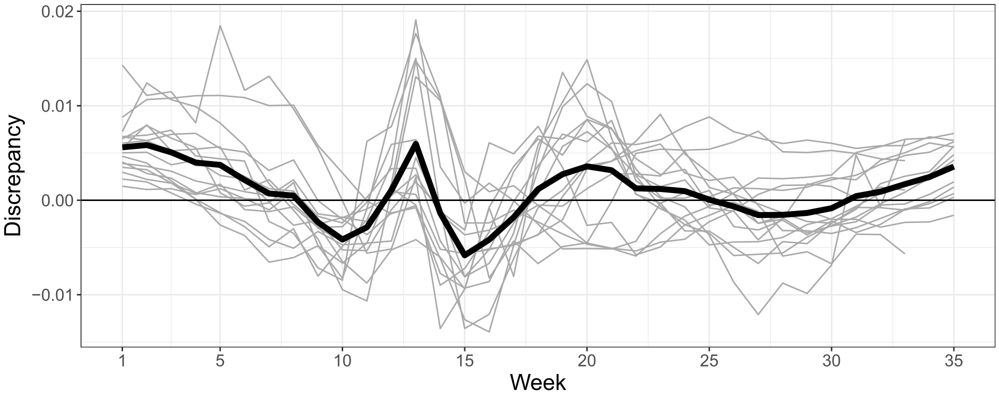


Figure 5: Discrepancy trajectories for all flu seasons are denoted by grey lines. The black line is the average discrepancy trajectory. Discrepancy tends to be greater than zero and the beginning (week 1), peak (weeks 13 and 20), and end of the season (week 35).

Bayesian model which incorporates both the SIR model and the structured discrepancy hierarchically for the ultimate purpose of forecasting future wILI.

4 Dynamic Bayesian Model

In this section, we describe the various components of the DB model, broadly partitioned into the data model (Section 4.1) and the process model (Section 4.2).

4.1 Data Model

Let $y_{j,t}$ be wILI for flu season $j = 1998, \ldots, 2007, 2010, \ldots, 2015$ during week $t = 1, 2, \ldots, T$ where $T = 35$. We model the proportion $y_{j,t}$ as

$$y_{j,t} \sim \text{Beta}(\lambda \pi_{j,t}, \lambda(1 - \pi_{j,t})), \tag{2}$$

where $\pi_{j,t} \in [0, 1]$ for all j and t is the true but unobservable proportion of the population with an influenza-like illness. The mean and standard deviation for $y_{j,t}$ are

$$\text{E}(y_{j,t}) = \pi_{j,t}, \tag{3}$$

$$\text{SD}(y_{j,t}) = \left(\frac{\pi_{j,t}(1 - \pi_{j,t})}{1 + \lambda} \right)^{0.5}. \tag{4}$$

Equation 2 assumes $y_{j,t}$ is unbiased for $\pi_{j,t}$. The concentration parameter λ governs the standard deviation of $y_{j,t}$. That is, λ governs the random variability of $y_{j,t}$ caused by such things as sampling variability, ILI diagnosis errors, and reporting variability. For a given λ, the standard deviation increases with increasing $\pi_{j,t}$ in the range 0 to 0.5. Increasing random variability with increasing levels of $\pi_{j,t}$ is a desired feature other models have attempted to mimic through ad hoc means (e.g., Shaman and Karspeck, 2012). The Beta distribution is able to capture this feature naturally. The random variability is not expected to vary across flu seasons and is poorly learned from the data. For these reasons, we set $\lambda = 4{,}500$, which we found to yield good predictions. Alternatively, λ could have been chosen via cross-validation. The choice of λ was motivated by the method of moments, where $\pi_{j,t}$ in (4) was set equal to the average of the training ILI on week T and the standard deviation of $y_{j,T}$ was computed using the method of moments for the training data. This choice of λ implies the standard deviation of $y_{j,t}$ is 0.0025 when $\pi_{j,t}$ is equal to 0.03 and $y_{j,t}$ is 0.0035 when $\pi_{j,t}$ is equal to 0.06.

4.2 Process Model

We model the logit of the true but unobservable proportion of influenza-like illness, $\pi_{j,t}$, as the sum of three components,

$$\text{logit}(\pi_{j,t}) = \text{logit}(I_{j,t}) + \mu_t + \delta_{j,t}. \tag{5}$$

Equation 5 decomposes $\text{logit}(\pi_{j,t})$ into an SIR model component $\text{logit}(I_{j,t})$, a discrepancy component common to all flu seasons μ_t, and a discrepancy component specific to each flu season $\delta_{j,t}$. The SIR model component represents the component of $\pi_{j,t}$ that can be described by the SIR model. Ideally, $\text{logit}(I_{j,t})$ would describe all of $\text{logit}(\pi_{j,t})$ implying the discrepancy terms are zero. On the basis of Figure 4, though, we know the SIR model cannot capture all the relevant features of $y_{j,t}$ and by extension, $\pi_{j,t}$. Thus, the common discrepancy component μ_t captures systematic differences between $\text{logit}(\pi_{j,t})$ and $\text{logit}(I_{j,t})$ shared by all flu seasons. We anticipate there is discrepancy structure common to all flu seasons on the basis of the non-zero, average discrepancy trajectory in Figure 5. The flu season-specific discrepancy term, $\delta_{j,t}$, captures the component of $\text{logit}(\pi_{j,t})$ unexplained by $\text{logit}(I_{j,t})$ and μ_t. Again, we anticipate season-specific discrepancy is needed on the basis of the season-specific discrepancy trajectories in Figure 5. In what follows, we specify the statistical models for each of the components of (5).

Model for logit($I_{j,t}$)

We model $I_{j,t}$, the infectious proportion of the population according to the SIR model for flu season j during week t, as the solution to (1). An explicit formula solution to (1), however, is unavailable. Thus, a numerical approximation method is used. We follow Osthus et al. (2017) and use the fourth order Runge–Kutta approximation method (RK4) to approximate the solution to the SIR model. The details of the RK4 method can be found in the accompanying supplementary material (Osthus et al., 2018). The RK4 approximation method is known to be more stable than the simpler Euler's method; a result we have also found to be true.

The quantity $I_{j,t}$ is completely determined once $S_{j,0}$, $I_{j,0}$, $R_{j,0}$, γ_j, and β_j are specified. We set $S_{j,0} = 0.9$ for all j following Osthus et al. (2017) as there is little information to learn about the susceptible and recovered trajectories of the SIR model from only

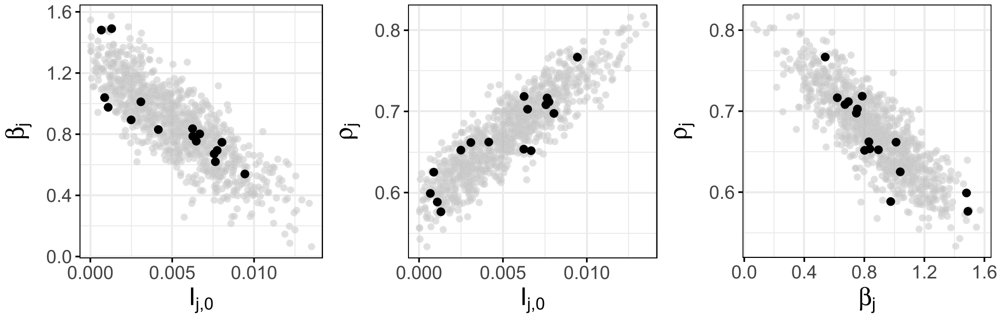


Figure 6: Parameter estimates for the best fit susceptible-infectious-recovered models corresponding to each flu season in Figure 4 (black) and 1,000 draws from the truncated Gaussian prior for $I_{j,0}$, β_j, and ρ_j (grey).

wILI (Capaldi et al., 2012). Setting $S_{j,0} = 0.9$ is, thus, an identifiability constraint. We assign an informative prior to $I_{j,0}$, β_j and ρ_j via empirical Bayes by fitting a multivariate Gaussian distribution to the parameter estimates of the fitted SIR models shown in Figure 4 and truncating to respect known and/or assumed boundary constraints. SIR parameter estimates and draws from the prior are shown in Figure 6. The truncation for $I_{j,0}$ is $(0, 0.1)$ ensuring that $I_{j,0}$ is positive and $S_{j,0} + I_{j,0} \leq 1$. To maintain mass balance, we set $R_{j,0}$ equal to $1 - S_{j,0} - I_{j,0}$. The truncation for β_j is $(0, \infty)$ ensuring it is positive. The truncation for ρ_j is $(0, .9)$, ensuring it is positive and less than $S_{j,0}$, thus restricting the SIR model to be an epidemic. We find the upper bound truncation on ρ_j helps with numerical stability in our posterior sampling in addition to aligning with our expectations for the shape of the SIR infectious trajectory.

We emphasize that the ultimate goal of the model is to predict future observations of wILI for a completely or partially unobserved flu season where data are not yet available. Data from partially observed flu seasons are not used in the prior specification of $I_{j,0}$, β_j, and ρ_j. For example, if the model is forecasting wILI for flu season 2015, then SIR parameter estimates from all non-2015 flu seasons are used to estimate the parameters for the prior on $I_{j,0}$, β_j, and ρ_j.

The prior for $I_{j,0}$, β_j, and ρ_j assumes that the best fitting SIR model for a completely or partially unobserved flu season comes from the same distribution as the best fitting SIR models for completely observed seasons. That is, we have an informative prior about the initial conditions and parameters of the best fitting SIR model prior to observing any data. We believe this is a reasonable assumption for the task of forecasting seasonal influenza. This, however, would be a questionable assumption for the task of forecasting a less predictable and non-recurrent disease, such as pandemic influenza.

Model for μ_t

The discrepancy process μ_t captures the systematic discrepancy common to all flu seasons and is what allows the forecasts of a partially observed season to borrow discrepancy information from other flu seasons. We specify the μ_t process as a reverse-random walk. Specifically,

$$\mu_T \sim \mathrm{N}(0, \sigma^2_{\mu_T}), \tag{6}$$
$$\mu_t | \mu_{t+1} \sim \mathrm{N}(\mu_{t+1}, \sigma^2_{\mu}). \tag{7}$$

The random-walk specification is a way to impose temporal structure to the common discrepancy model as μ_t depends on μ_{t+1}. The *reverse* random-walk specification is related to how wILI is released. Within a flu season, wILI is first available for week 1, then week 2, and so forth. As a result, we are always forecasting the end of the flu season and seldom the beginning of the flu season. From Figure 1, it is clear that the end of the flu season is relatively well-behaved and predictable as compared to the middle of the flu season. That is, though there is considerable uncertainty regarding the trajectory wILI will take, there is much less uncertainty regarding wILI's destination on week 35. A reverse-random walk helps bridge the gap between the last wILI observation and the end of the flu season. That is, the data observed at the beginning of the flu season imposes a constraint on the early part of the model while the reverse random-walk modeling assumption provides a constraint on the end of the flu season, restricting the spread of plausible trajectories. A reverse random-walk has been used with success in other forecasting contexts, such as presidential election forecasting (Linzer, 2013).

We assigned the following priors to the precisions of (6) and (7):

$$\sigma^{-2}_{\mu_T} \sim \mathrm{Gamma}(2, 2), \qquad \sigma^{-2}_{\mu} \sim \mathrm{Gamma}(2, 0.02). \tag{8}$$

The priors reflect a belief that σ_{μ_T} will be larger than σ_μ, as we expect the changes for μ_t at adjacent time points to be small relative to our uncertainty in the value of μ_T. The relative difference in these priors was motivated by these considerations. The specific hyperparameters, however, were chosen somewhat by trial and error based on the visual plausibility of realizations drawn from the prior predictive distribution. Though our choices yielded good forecasts (see Section 5.3), hyperparameter selection might be better optimized via cross-validation over the training data.

Model for $\delta_{j,t}$

We would like to explain $\mathrm{logit}(\pi_{j,t})$ with $\mathrm{logit}(I_{j,t})$ and μ_t if possible. Figure 5, however, suggests this is not possible and thus the model will likely benefit from a season-specific discrepancy term, $\delta_{j,t}$. The introduction of $\delta_{j,t}$ creates an issue with model identifiability. As an identifying constraint, we set

$$\delta_{j,T} = -\mathrm{logit}(I_{j,T}). \tag{9}$$

The constraint in (9) implies,

$$y_{j,T} \sim \mathrm{Beta}(\lambda \mathrm{logit}^{-1}(\mu_T), \lambda(1 - \mathrm{logit}^{-1}(\mu_T))), \tag{10}$$

as $\pi_{j,T} = \text{logit}^{-1}(\mu_T)$. That is, the data model for wILI on week T is a function of μ_T, the discrepancy component common to all flu seasons, and λ, the parameter governing the random variability of the data model.

We impose temporal structure on $\delta_{j,t}$ and encourage it to be close to zero for all $t \neq T$ by modeling $\delta_{j,t}$ with the following autoregressive, reverse random-walk:

$$\delta_{j,t}|\delta_{j,t+1} \sim \text{N}(\alpha_j \delta_{j,t+1}, \sigma^2_{\delta,j}). \tag{11}$$

The autoregressive parameter $\alpha_j \in [0,1]$ encourages $\delta_{j,t}$ to be close to zero. As with μ_t, the random-walk structure for $\delta_{j,t}$ imposes a temporal dependence, as $\delta_{j,t}$ depends on $\delta_{j,t+1}$. The *reverse* random-walk allows for easy incorporation of the identifying constraint in (9).

Both α_j and the variance $\sigma^2_{\delta,j}$ are modeled hierarchically. The model for α_j is,

$$\text{logit}(\alpha_j) \sim \text{TN}_{(\text{logit}(0.02),\text{logit}(0.98))}(\text{logit}(0.9), \sigma^2_\alpha), \tag{12}$$

$$\sigma_\alpha \sim \text{Gamma}(2,2), \tag{13}$$

where $\text{TN}_{(\text{logit}(0.02),\text{logit}(0.98))}$ is a truncated Gaussian distribution, truncated between logit(0.02) and logit(0.98). The truncation improved numerical stability of the posterior sampler. The mean in (12) reflects our prior belief that α_j is near 0.9 for all j. Note that if $\alpha_j = 0$, then (11) has no random-walk structure, as each $\delta_{j,t}$ is modeled as a mean 0 Gaussian distribution with a season-specific variance. Said another way, the closer α_j gets to zero, the less temporal structure exists in the season-specific discrepancy process.

Finally, we assign the following hierarchical prior to the precisions of the season-specific discrepancy model, $\sigma^{-2}_{\delta,j}$:

$$\sigma^{-2}_{\delta,j} \sim \text{Gamma}(a_\delta, b_\delta), \tag{14}$$

where $a_\delta \sim \text{Gamma}(5,1)$ and $b_\delta \sim \text{Gamma}(1,10)$. The parameter $\sigma^{-2}_{\delta,j}$ is a flexibility parameter. The smaller $\sigma^{-2}_{\delta,j}$ becomes (or equivalently, the larger the variance $\sigma^2_{\delta,j}$ becomes), the more flexible $\delta_{j,t}$ becomes. Thus, we expect flu seasons that more acutely deviate from "typical" wILI behavior to require larger variances, $\sigma^2_{\delta,j}$, than more "typical" flu seasons. The hierarchical specification for precisions $\sigma^{-2}_{\delta,j}$ allows the estimation of $\sigma^{-2}_{\delta,j}$ for a partially observed flu season to borrow information from fully observed flu seasons.

5 Results

In practice, the DB model is updated each week when new wILI data becomes available. We take a "leave-one-season-out" approach to model assessment, where we make use of all available data from the seasons *not* being forecasted as well as all of the *observed* data from the season being forecast. We refer to each model fit with a "Season.Week" naming convention. Model "Season.Week" refers to a model fit to all observations from flu seasons not equal to "Season" and all observations in "Season" from week 1 through week "Week". For example, model 2015.3 is a model fit to all wILI observations not in flu season 2015 plus weeks 1 through 3 of 2015 (recalling flu season 2015 means the 2015–2016 flu season). Forecasting model 2015.3 means forecasting the unobserved data for weeks 4 through 35 of flu season 2015.

The posterior is not known in closed form, thus we sample from it via Markov chain Monte Carlo (MCMC). The posterior sampling of the DB model was performed using the `rjags` package (Plummer, 2016) within the R programming language (R Core Team, 2016), which calls the software "Just Another Gibbs Sampler", or JAGS (Plummer, 2003). JAGS queries a set of internal samplers based on the specification of the model. The internal samplers can be highly catered to specific models or highly generic (e.g., slice sampling) for any model specified as a directed acyclic graph. For more details, we direct the reader to the JAGS user manual (Plummer, 2017).

Model parameter convergence was checked for the 2015.3 model by running four chains for 100,000 iterations, throwing away the first half as burn-in and thinning every 20th iteration, resulting in four chains each of length 2,500. We assessed MCMC convergence for all latent quantities of the model with the Gelman–Rubin diagnostic, $\hat{R}$ (Gelman and Rubin, 1992). $\hat{R}$ was computed using the `gelman.diag()` function in the `coda` package (Plummer et al., 2006). All $\hat{R}$s were less than 1.1, suggesting no evidence for lack of convergence. For all other models, we ran one chain for 50,000 iterations, throwing away the first half as burn-in and thinning every 10th iteration, resulting in a chain of 2,500. Running the chain for 50,000 iterations takes approximately 10 to 15 minutes on a MacBook Pro with a 2.8 GHz Intel Core i7 processor.

In Section 5.1, we illustrate how the model is updated each week for the 2015 flu season and discuss the different model components. We assess the model for all seasons in the context of predictive empirical coverage in Section 5.2. Finally, in Section 5.3, we compare the DB model's forecasting accuracy to the 14 and 30 flu forecasting models that participated in the CDC's 2015–2016 and 2016–2017 flu forecasting challenges, respectively.

5.1 DB Model Fit to the 2015–2016 Flu Season

Models 2015.3 through 2015.30, inclusively, were fit mimicking the sequential model fitting for an entire flu season. Weeks 3 and 30 roughly correspond to the forecasting window used in the CDC flu forecasting challenge. Figure 7 shows the posterior predictive mean and 95% point-wise posterior predictive intervals for select model fits. Predictive uncertainty is largest when forecasts are made early in the flu season, but gradually diminishes throughout the flu season as more data are observed and incorporated into the model fitting. Forecasts early in the flu season reflect the bimodal nature of peak timing in the non-2015 flu seasons. That is, the forecasts for 2015 suggests there could either be an early peak to the flu season at week 13, or the peak could occur later around week 20. The average forecast exhibits a sharp decline from week 13 to week 14, reflecting the decline in wILI from week 13 to week 14 in all non-2015 seasons (recall Figure 4). The reason the forecast exhibits bimodality is because of the hierarchical discrepancy model. Importantly, the empirical coverage for the 95% nominal predictive intervals for all 2015.3 through 2015.30 model forecasts was 95.2%. Empirical coverage will be discussed more in Section 5.2.

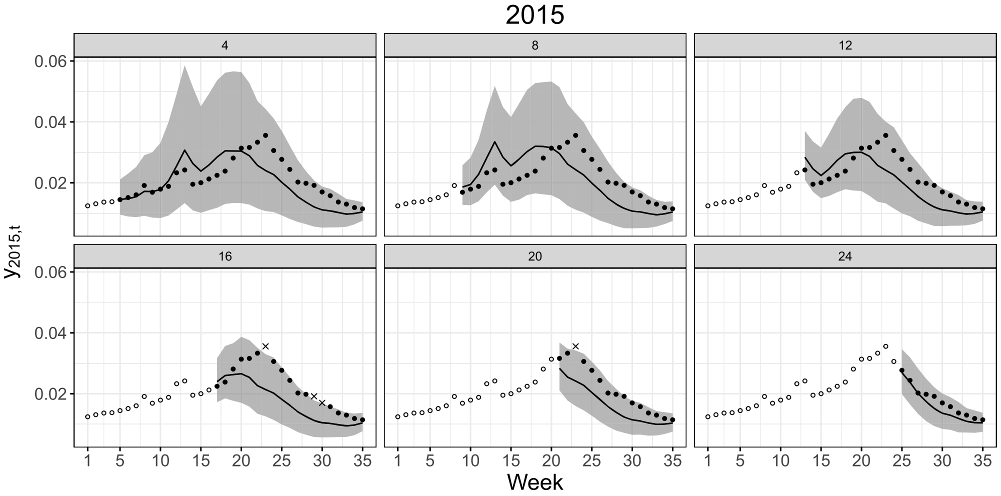


Figure 7: Forecasts for the 2015 flu season. The panel label number denotes the "Week" of the "2015.Week" model fit. Black, hollow circles denote observed data. Grey bands denote 95% point-wise posterior predictive intervals for unobserved data while the black line is the posterior predictive mean. Black, solid circles denote unobserved data that fell within the 95% point-wise posterior predictive intervals. Black 'X's denote unobserved data that fell outside the 95% point-wise posterior intervals. The empirical 95% coverage for all forecasts in 2015 was 95.2%.

The 95% point-wise posterior predictive intervals corresponding to the model components of (5) are presented in Figure 8 for various weeks of the 2015 flu season. In the second row of Figure 8, we see the SIR component of the model approximates wILI. As more observations are incorporated into the analysis, the infectious trajectory better approximates the data. Though the infectious trajectory approximates the data, it is unable to match it exactly (as expected). Thus, there is a non-zero discrepancy, captured by $\mu_t + \delta_{2015,t}$ and plotted in the third row of Figure 8. When the average infectious trajectory underestimates the data, the discrepancy $\mu_t + \delta_{2015,t}$ is greater than zero and vice versa. The discrepancy, thus, compensates for the inadequacies of the infectious trajectory. The point-wise 95% posterior interval for $\mu_t + \delta_{2015,t}$ is typically larger for weeks corresponding to unobserved data than weeks for observed data.

The discrepancy $\mu_t + \delta_{2015,t}$ is further decomposed into the common discrepancy μ_t and season-specific discrepancy $\delta_{2015,t}$ in the fourth and fifth rows of Figure 8, respectively. The posterior mean and 95% point-wise posterior interval for μ_t is relatively unchanged for all model fits in Figure 8. This is because μ_t is common to all seasons, meaning the complete data from all 15 non-2015 flu seasons are informing the estimate of μ_t. The incremental increase in data added to the analysis for season 2015 is a relatively small proportion of the total data informing μ_t. We also see the estimate of μ_t captures the discrepancy bump between weeks 10 and 15. Finally, note that the posterior mean and 95% posterior interval for μ_T is roughly -4.56 (-4.64, -4.47) for all

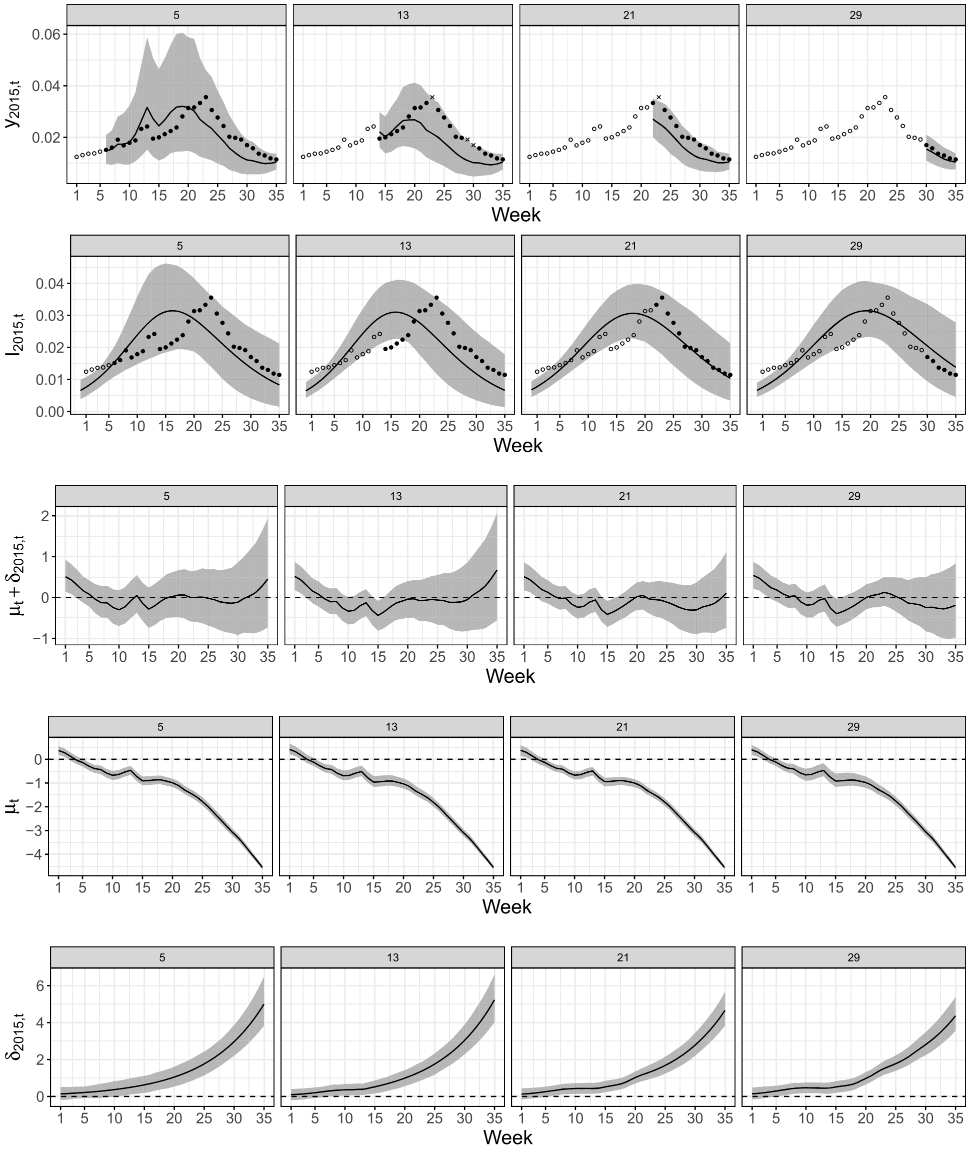


Figure 8: Posterior predictive distribution for future weighted influenza-like illness (wILI) (top row) and posterior distributions for the various components of (5) (rows 2 through 5) for models 2015.5, 2015.10, 2015.15, 2015.20, 2015.25, and 2015.30 (columns from left to right). In the top row, hollow, black circles are observed wILI. Solid, black circles are unobserved wILI that fell within the 95% predictive band. Black 'X's are unobserved wILI that fell outside the 95% predictive band. In the second row, hollow, black circles are observed wILI and solid, black circles are unobserved wILI. In all rows, the black line and grey bands represent the posterior mean and 95% point-wise posterior interval for the row-specific quantity, respectively.

models displayed in Figure 8. The mean of logit($y_{j,T}$) for all non-2015 seasons is -4.59. The identifying constraint of (9) effectively sets μ_T equal to the average of logit($y_{j,T}$), as supported by these results.

The season-specific discrepancy term, $\delta_{2015,t}$, does change throughout the season, as it is capturing the season-specific discrepancy unaccounted for by the common discrepancy and infectious trajectory. Also by the identifying constraint of (9), $\delta_{2015,T}$ is set to -logit($I_{2015,T}$). $\delta_{2015,t}$ gradually reverts to near zero from week T to week 1, as was encouraged by (11) and (12).

Figure 9 displays the posterior credible intervals for $\sigma_{\delta,j}$, the evolution standard deviation of the season-specific discrepancy trajectory. The larger the evolution standard deviation, the more flexible $\delta_{j,t}$ is. Greater flexibility for $\delta_{j,t}$ is needed for seasons whose wILI deviates more acutely from "typical" wILI flu seasons. From Figure 9, we see the three largest $\sigma_{\delta,j}$s as measured by posterior means correspond to flu seasons 1998, 1999, and 2003. These were also the three most "atypical" flu seasons as measured by MSE in Figure 2, supporting the interpretation that $\sigma_{\delta,j}$ captures season-specific discrepancy flexibility.

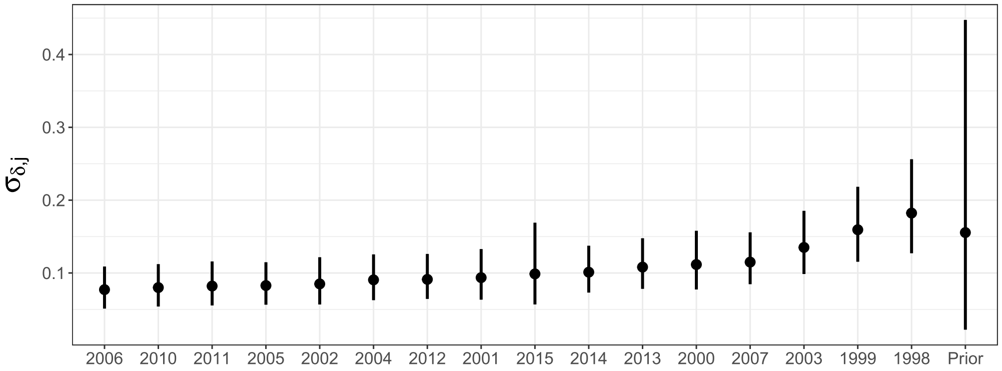
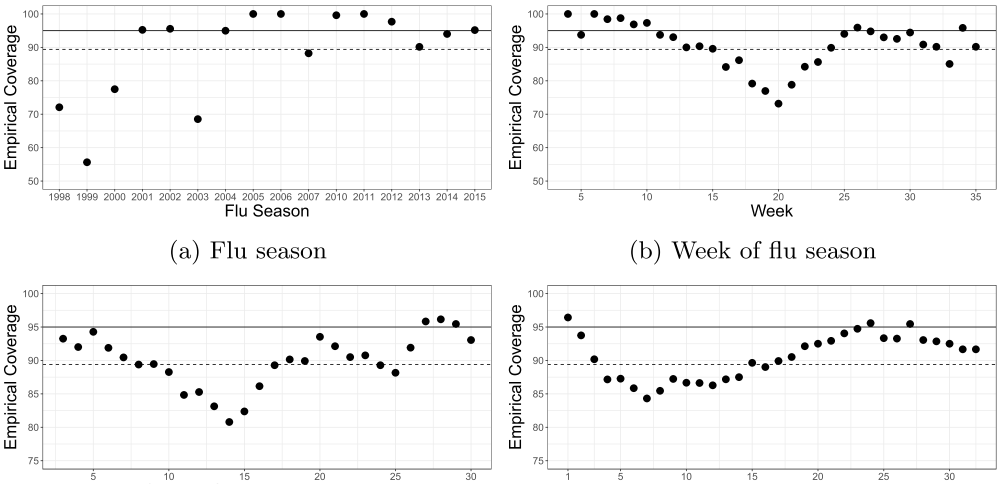


Figure 9: 95% credible intervals (line segments) along with means (points) for the prior and posterior distributions for $\sigma_{\delta,j}$ for model 2015.3, with seasons ordered by ascending posterior means.

5.2 DB Model Assessment

For each of 16 flu seasons shown in Figure 4, we fit and forecast models Season.3 through Season.30. Recall that when we fit, for example, model 2015.3, we forecast weeks 4 through 35 (32 forecasts) of season 2015. When we fit model 2015.30, we forecast weeks 31 through 35 (5 forecasts) of season 2015. For each complete season, we make 518 forecasts. Weighted ILI for weeks 34 and 35 are unavailable for seasons 1998 through 2001, thus 462 forecasts were made for those seasons. The totality of all forecasts across all "leave-one-season-out" model fits was 8,064 forecasts. Each forecast is a 95% point-wise posterior predictive interval for a future observation of wILI. The overall



(a) Flu season

(b) Week of flu season

(c) "Week" of "Season.Week" model fit

(d) Week-ahead forecast.

Figure 10: Empirical forecast coverage (points) by various partitionings of forecasts. The overall empirical coverage was 89.4% (dashed line). The solid line represents the nominal 95% coverage.

empirical coverage was 89.4%, suggesting the DB model forecasts, on balance, exhibit undercoverage.

Figure 7 suggests that empirical coverage is correlated, with future wILI often either falling within or outside predictive intervals on consecutive weeks. We can interrogate the DB model's forecasting accuracy by subsetting the 8,064 forecasts. Figure 10 plots empirical coverage versus various partitionings of the forecasts, revealing both areas the DB model's forecasts perform well and need improvement as well as times in the forecasting process that are inherently more difficult to forecast than others.

Figure 10a plots empirical coverage by flu season. We see empirical coverage ranges from 100% during seasons 2005, 2006, and 2011 to 55.6% in 1999. The flu seasons whose empirical coverage deviated most significantly below the nominal 95% coverage were also the most "atypical" flu seasons as determined by MSE in Figure 2: 1998, 1999, and 2003. Hierarchical models are powerful models for borrowing strength across a collection of similar units (e.g., flu seasons). An underlying assumption of hierarchical models, however, is that a new flu season is "similar" to the flu seasons that have already been seen (i.e., a new flu season comes from the same superpopulation of flu seasons the observed flu seasons came from). When a new flu season deviates from "typical", forecasts struggle. If seasons 1998, 1999, and 2003 were removed from the coverage assessment, the overall coverage would be 94.6%. The forecast undercoverage appears to be driven by a few flu seasons, suggesting more variability should be accounted for in the model.

Figure 10b plots empirical coverage by the week of the flu season. Empirical coverage ranged from 100% on weeks 4 and 6 to 73.2% on week 20. We see a general trend of declining empirical coverage from week 4 to week 20, followed by a recovery in coverage to week 35. Weeks near week 20 represent the most challenging period of the flu season to forecast, as wILI for some of the later peaking flu seasons are still ascending to their peak, while the early peaking flu seasons have been reverting to low levels for multiple weeks. Though the range in wILI at week 20 is the same as at week 13 (both 0.055, see Figure 1 for reference), all wILI trajectories are ascending to week 13 while some are ascending and some are descending to week 20. This ascending/descending distinction is an added source of uncertainty for week 20 not shared with week 13, making it so challenging to forecast.

Figure 10c plots the "Week" of the Season.Week model fit. For example, the plotted empirical coverage for "Week" 5 is the average empirical coverage over all forecasts made when only the first 5 weeks of the flu season have been observed. The empirical coverage ranges from 96.2% for Season.28 models to 80.8% for Season.14 models. There is a general decline in empirical coverage in Season.5 through Season.14, with an upturn in empirical coverage from Season.14 to Season.30. Forecasts for model Season.14 represent a fork in the forecasting process. Recall all flu seasons exhibit a downturn in wILI from week 13 to week 14. That downturn either signifies the worst of the flu season has occurred or a temporary decline in an otherwise still ascending flu season. Forecasts corresponding to model Season.14, thus, exhibit appreciable uncertainty. The information in wILI for weeks 15, 16, and 17 provide much information about whether the flu season will continue to ascend or descend.

Finally, Figure 10d plots the empirical forecast versus the week-ahead-forecast. Empirical coverage ranges from 96.4% for one-week-ahead forecasts to 84.3% for seven-week-ahead forecasts. Empirical coverage generally declines from one to seven-week-ahead forecasts, and then increases with increasing week-ahead-forecasts. The average empirical coverage for all one to four-week ahead forecasts is 91.9%, representing an improvement when compared to the overall empirical coverage of 89.4%. Seasonal flu forecasting represents an example of forecasting where forecast accuracy does not decline with increasing week-ahead-forecasts. Figure 1 illustrates why this is. Again, wILI is relatively well-behaved and predictable at the beginning and the end of the flu season. The bulk of the uncertainty occurs in the middle of the flu season. In October, it is easier to predict wILI in May (roughly 30 weeks into the future) than it is to predict wILI in December (roughly 10 weeks into the future).

5.3 DB Model Comparisons to Other Flu Models

Forecasting challenges are effective ways to both identify and improve predictive capabilities in a myriad of fields, including influenza forecasting (Tetlock et al., 2017). The CDC has hosted an influenza forecasting challenge, open to the public, since the 2013–2014 flu season. In the inaugural 2013–2014 challenge, over a dozen forecasting models participated in the challenge (Biggerstaff et al., 2016). Since then, the challenge has grown, with 30 forecasting models participating in the 2016–2017 challenge

(Epidemic Prediction Initiative, 2016). The CDC's flu forecasting challenge is an opportunity for the CDC to scope flu forecasting capabilities. It is also an opportunity for teams to compare their forecasting models against the leading forecasting models in the field. Competition drives innovation and incentives iterative improvement.

To see how the DB model compares to cutting-edge forecasting competition, we compare the DB model to the 14 models and 30 models that participated in the 2015–2016 and 2016–2017 flu forecasting challenges, respectively. These models represent a diverse collection of mechanistic, machine learning, ensemble, and statistical models, making use of numerous data sources including Internet based sources, such as Wikipedia and Twitter, as well as non-Internet based sources, such as weather attributes and school vacation schedules. The weekly submissions for all models are publicly available (FluSight Influenza Forecasting Challenge, 2016). We stress that we did not use 2015–2016 data when constructing priors for the DB model and did not even have access to the 2016–2017 data when the DB model was developed, ensuring that our forecasts were true out-of-sample forecasts.

Model comparison follows the evaluation criteria of the flu forecasting challenge (Epidemic Prediction Initiative, 2015a), comparing each model's ability to accurately predict seven targets throughout the flu season:

- **Peak intensity (PI)**: the maximum value of wILI for the flu season.
- **Timing of peak intensity (PT)**: the week the PI occurs.
- **Onset**: the start of the flu season, defined as the first of three consecutive weeks of wILI equal to or above the national baseline (Epidemic Prediction Initiative, 2015b). For the 2015–2016 and 2016–2017 flu seasons, the national baselines were 0.021 and 0.022, respectively.
- **One, two, three, and four week ahead forecasts**: short-term forecasts.

Each week of the flu season, a submission for all targets is made in the form of a probabilistic forecast. For each target and submission week, probabilities are assigned to mutually exclusive and exhaustive bins such that the probabilities sum to one. For PI and the short-term forecasts, bins range from 0 to 0.13, with bin widths of 0.005 in 2015–2016 and 0.001 in 2016–2017. For completion, there is a catch-all bin from 0.13 to 1. For PT and onset, each bin corresponds to a week. There is an additional bin of "no onset" for onset, as there is no guarantee a flu season will have three consecutive weeks of wILI at or above baseline.

The evaluation of each target is done by computing a logarithmic score. Let

$$\boldsymbol{p}_{\text{mod},\text{wk},\text{tgt}} = (p_{\text{mod},\text{wk},\text{tgt},1}, p_{\text{mod},\text{wk},\text{tgt},2}, \dots, p_{\text{mod},\text{wk},\text{tgt},n_{\text{tgt}}})' \tag{15}$$

represent the vector of binned probabilities for model "mod" submitted on week "wk" corresponding to target "tgt", where $p_{\text{mod},\text{wk},\text{tgt},i}$ corresponds to the probability assigned to bin $i = 1, 2, \dots, n_{\text{tgt}}$.

Assume the target falls in bin i^*. Then, the log score corresponding to $\boldsymbol{p}_{\text{mod,wk,tgt}}$ is defined as,

$$S(\boldsymbol{p}_{\text{mod,wk,tgt}}, i^*) = ln(p_{\text{mod,wk,tgt},i^*-1} + p_{\text{mod,wk,tgt},i^*} + p_{\text{mod,wk,tgt},i^*+1}). \tag{16}$$

That is, the log score is the natural log of the sum of the probabilities assigned to the correct bin and the immediately preceding and proceeding bins.[3] For example, if the forecasted probability of the PT occurring on weeks 19, 20, and 21 are 0.1, 0.3, and 0.2, respectively, and the true PT is week 20, then the log score is $ln(0.1+0.3+0.2) = -0.51$. A perfect log score is 0 and is achieved if all the probability is assigned to the correct or immediately adjacent bins. A score of -10 is assigned to all undefined natural log scores (e.g., $ln(0)$ is undefined), all late submissions, and all submissions where the sum of $\boldsymbol{p}_{\text{mod,wk,tgt}} > 1.1$. Good forecasts, as determined by (16), are forecasts that concentrate appreciable probability tightly around the bin of the true target value.

For the CDC challenge participating models, the binned probabilities are publicly available. For the DB model, bins were populated by simulating from the posterior predictive distribution for each target and submission week.

Weighted ILI estimates are revised every week. For example, when wILI on week 3 of the 2015–2016 flu season was first publicly released, it was 0.0135. The next week, when week 4 was first publicly released, the estimate for wILI on week 3 was revised from 0.0135 to 0.0141. This process of revision can occur every week. These weekly revisions can cause wrinkles when retrospectively comparing a new model (e.g., the DB model) to models that participated in past forecasting challenges, as the retrospective model fitting is often based on wILI estimates that were unavailable on the date real-time submissions were made. For the comparison of the DB model to the other models, we used the wILI estimates that were available on each submission week, allowing us to faithfully replicate the forecasting conditions. In general, though, the weekly wILI revisions can cause problems with retrospective model comparisons.

For every target, model, and season we computed the average log score over all submission weeks. The results are plotted in Figure 11. The DB model compared favorably to all models with respect to all targets. The DB model beat all models with respect to onset and two week ahead forecasts in 2015–2016 and beat all models with respect to three and four week ahead forecasts both seasons. Furthermore, the DB model ranked no worse than fourth for all targets in 2015–2016 and no worse than sixth in 2016–2017. Averaging over the log scores for all targets provides an estimate of a model's forecasting ability. The DB model beat all models with respect to overall average log score in both seasons.

As shown in Table 1, the DB model was the only model to rank first for more than one target in 2015–2016 and one of two models to rank first for more than one target in 2016–2017, indicating that it was the best forecasting model with respect to multiple targets. The DB model was also the only model to rank no worse than fourth with respect to all forecasting targets in 2015–2016 and no worse than sixth for all targets

[3] For 2016–2017, the five immediately preceding and proceeding bins were summed for PI and all short term forecasts, as was specified by the CDC.

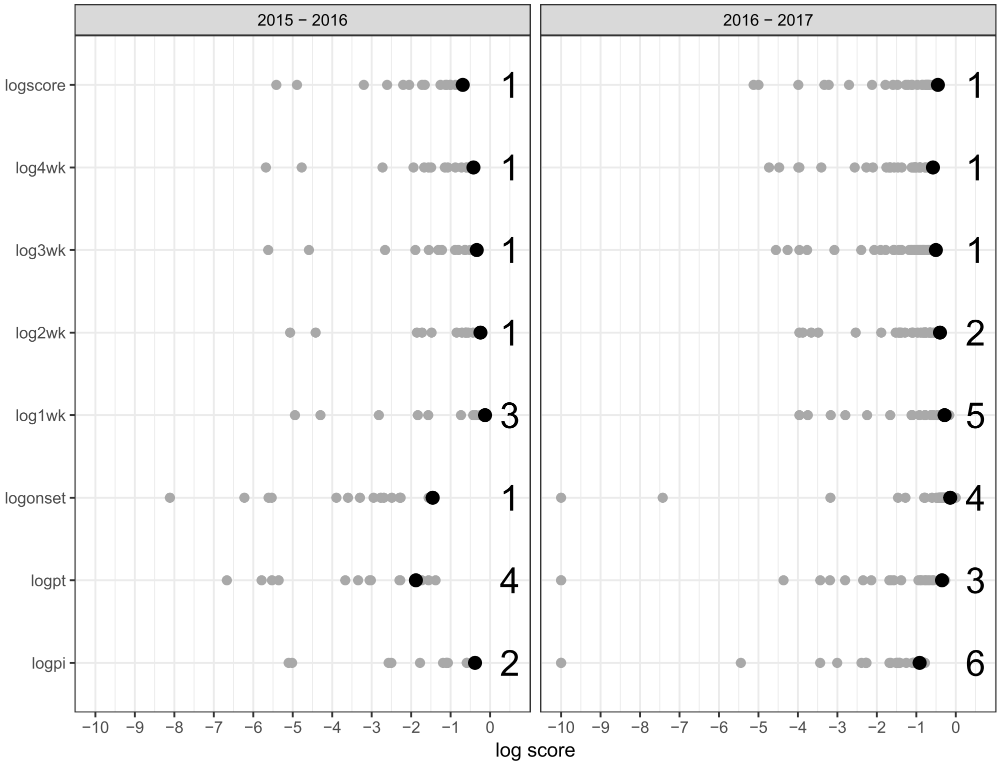


Figure 11: Average log score for every target and model in the CDC's 2015–2016 (left) and 2016–2017 (right) flu forecasting challenge (grey) and the DB model (black). The overall average log score is shown as 'logscore'. The numbers on the right represent the rank of the DB model relative to all other models where 1 is the best log score. The DB model had the best overall average log score for both seasons.

in 2016–2017, suggesting the DB model was not deficient at forecasting any flu season target, in a relative sense.

Though the DB model had the best overall log score for both seasons and appears to be one of the best models for all considered targets, there is still room for improvement. Figure 12 illustrates one such area and articulates why accounting for the weekly wILI revisions is important for faithful retrospective comparisons between new models and flu challenge participating models. The top of Figure 12 shows that on week 13 of the 2015–2016 season, the DB model's log score for onset was -10, indicating zero probability was assigned to the correct or neighboring bins of the true onset (week 16). The bottom of Figure 12 reveals why this occurred. Based on the data available through the first 13 weeks of the 2015–2016 flu season, week 11 was, by definition, the onset (i.e., week 11 was the first of three consecutive weeks equal to or above baseline). Thus, on week 13, the DB model forecasted week 11 to be the onset with probability one. Declaring week

Season	**Model**	**Top 1**	**Top 2**	**Top 3**	**Top 4**	**Top 5**	**Top 6**
2015 – 2016	DB	4	5	6	7	7	7
	Model 1	1	4	4	5	5	6
	Model 2	1	1	1	2	2	5
	Model 3	1	1	1	1	4	5
	Model 4	0	2	2	2	2	2
	Model 5	0	1	3	6	7	7
	Model 6	0	0	2	3	5	5
	Model 7	0	0	1	1	2	2
	Model 8	0	0	1	1	1	1
	Models 9 – 10	0	0	0	0	0	1
	Models 11 – 14	0	0	0	0	0	0
2016 – 2017	DB	2	3	4	5	6	7
	Model 1	2	2	2	3	3	4
	Model 2	1	1	3	3	3	4
	Model 3	1	1	1	1	2	2
	Model 4	1	1	1	1	1	1
	Model 5	0	2	3	5	5	6
	Model 6	0	2	3	4	5	5
	Model 7	0	1	2	2	3	3
	Models 8 – 9	0	0	1	1	1	1
	Model 10	0	0	0	2	3	3
	Models 11 – 12	0	0	0	0	1	1
	Models 13 – 15	0	0	0	0	0	1
	Models 16 – 30	0	0	0	0	0	0

Table 1: The number of top 1 through top 6 rankings by each model. The DB model was the only model to have more than one top 1 ranking in 2015–2016 and one of two models with more than one top 1 ranking in 2016–2017. The DB model was the only model to rank no worse than fourth for all seven targets in 2015–2016 and no worse than sixth for all seven targets in 2016–2017. Models 1 through 30 represent the anonymized comparison models. Model X in 2015–2016 does not correspond to Model X in 2016–2017.

11 the onset, however, ignores the weekly wILI revisions. The very next week when 14 weeks of wILI estimates were available, week 11's wILI estimate was revised and fell below the national baseline, indicating it was not the onset. The scenario displayed in Figure 12 articulates that the forecasts of the DB model are missing a source of uncertainty caused by wILI revisions. It also articulates that using revised versions of wILI estimates to retrospectively compare a model to forecasts based on currently available wILI estimates gives an unfair advantage to the new model. The log onset score of -10 would not have occurred for the DB model were forecasts based on wILI estimates available at the end of the flu season, biasing the log scores up.

Even without accounting for the uncertainty caused by weekly wILI revisions in the forecasts, the DB model outperformed all models it was compared against, suggesting that it is one of the leading flu forecasting models with room for improvement.

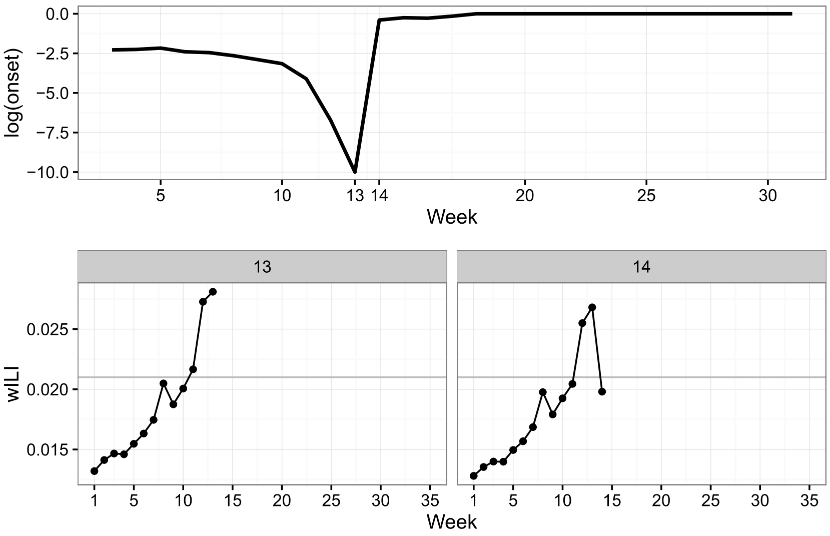


Figure 12: (Top) The weekly onset log score for the DB model in 2015–2016. (Bottom) Weighted influenza-like illness estimates available through the first 13 and 14 weeks of the 2015–2016 flu season, respectively. The grey, horizontal line is the national baseline equal to 0.021.

6 Discussion

In this paper, we introduced a novel dynamic Bayesian influenza forecasting model that exploits discrepancy structure. The basic insight and motivation leading to the development of the DB model is that the disease transmission model (e.g., the SIR model) and the data-generating model are not equivalent; disease transmission is a component of but not equal to the data-generating process. The data-generating model is non-exhaustively comprised of a disease transmission process, a healthcare provider visitation process, an influenza-like illness determination process, and a reporting process. Thus, even if a disease transmission model more sophisticated than the SIR model were used, of which there are numerous (e.g., the SIRS model, the SEIR model), there might still be a disagreement between the best version of the disease transmission model and the data. Rather than attempt to model each component of the data-generating model, we acknowledge there will likely be a systematic disagreement between the best version of the disease transmission model and the data. We then model the commonalities of the discrepancy across flu seasons with a flexible, hierarchical model. The hierarchical discrepancy model allows us to leverage patterns in the data the disease transmission model is incapable of capturing and not simply model the discrepancy as white noise.

The DB model assumes future flu seasons will exhibit similar trajectories to past flu seasons. We showed that the more dissimilar a flu season was as compared to the other considered flu seasons, the worse forecasts were. Because of this underlying assumption, the DB model would be inappropriate for forecasting pandemic influenza. A companion model to the DB model that tracks how “dissimilar” a new flu season is compared to the collection of observed seasonal flu seasons could be useful to gauge when the DB model should be trusted and when it should not. Alternatively, λ in (2) could be modeled as a function of how similar/dissimilar the current flu season is to past flu seasons and forecast uncertainty could be modified accordingly.

When compared to the forecasting models participating in the CDC’s 2015–2016 and 2016–2017 flu forecasting challenges, the DB model had the best overall scores. Comparisons were facilitated by the CDC coordinating the flu forecasting challenge and making the submissions publicly available. These submissions provide an excellent test case for future models to be compared against.

The work of Ginsberg et al. (2009) demonstrated the potential value of monitoring flu outbreaks with Google search queries. The basic idea being, when individuals experience symptoms of the flu, they may go to their web browser to search for more information. Thus, an increase in searches for flu related terms may indicate an increase in flu incidence in the population. The work of Ginsberg et al. (2009) sparked a large research effort to investigate other digital surveillance sources and their possible connection to disease surveillance (e.g., Generous et al., 2014; Wilson and Brownstein, 2009; Polgreen et al., 2008). Many forecasting models have augmented wILI with digital surveillance data (e.g., Hickmann et al., 2015; Brooks et al., 2015; Shaman et al., 2013), including those that participated in the 2013–2014 CDC flu forecasting competition (Biggerstaff et al., 2016). Recently, the value of digital surveillance data with respect to flu forecasting has been curbed (e.g., Lazer et al., 2014; Priedhorsky et al., 2017). In fact, as Biggerstaff et al. (2016) conclude from the 2013–2014 flu forecasting competition, “not all digital data are equally accurate, and the algorithms and methodologies underpinning these data require constant upkeep to maintain their accuracy. . . . Influenza forecasting models informed by digital data are subject to the biases and errors of their underlying source data”. The DB model does not use digital surveillance data. Incorporating digital surveillance data may or may not improve forecasts; investigation into this might serve as a next iteration of the DB model. It is worth noting, however, that even without digital surveillance data, the DB model compared favorably to all comparison models, some of which did make use of digital surveillance data. The results presented in Section 5.3 suggest influenza forecasting can be improved without augmenting wILI with digital surveillance data but rather focusing on statistical model development.

Supplementary Material

Dynamic Bayesian Influenza Forecasting in the United States with Hierarchical Discrepancy: Supplementary Material (DOI: 10.1214/18-BA1117SUPP; .pdf).

References

Angulo, J. M., Yu, H.-L., Langousis, A., Madrid, A. E., and Christakos, G. (2012). “Modeling of space–time infectious disease spread under conditions of uncertainty." *International Journal of Geographical Information Science*, 26(10): 1751–1772. 262

Bardak, B. and Tan, M. (2015). "Prediction of influenza outbreaks by integrating Wikipedia article access logs and Google flu trend data." In *Bioinformatics and Bioengineering (BIBE), 2015 IEEE 15th International Conference on*, 1–6. IEEE. 262

Bayarri, M. J., Berger, J. O., Paulo, R., Sacks, J., Cafeo, J. A., Cavendish, J., Lin, C.-H., and Tu, J. (2007). "A framework for validation of computer models." *Technometrics*, 49(2): 138–154. MR2380530. doi: https://doi.org/10.1198/004017007000000092. 263

Biggerstaff, M., Alper, D., Dredze, M., Fox, S., Fung, I. C.-H., Hickmann, K. S., Lewis, B., Rosenfeld, R., Shaman, J., Tsou, M.-H., Velardi, P., Vespignani, A., Finelli, L., and the Influenza Forecasting Contest Working Group (2016). "Results from the centers for disease control and prevention's predict the 2013–2014 Influenza Season Challenge." *BMC Infectious Diseases*, 16(1): 357. 261, 279, 285

Boys, R. J., Wilkinson, D. J., and Kirkwood, T. B. (2008). "Bayesian inference for a discretely observed stochastic kinetic model." *Statistics and Computing*, 18(2): 125–135. MR2390814. doi: https://doi.org/10.1007/s11222-007-9043-x. 262

Brooks, L. C., Farrow, D. C., Hyun, S., Tibshirani, R. J., and Rosenfeld, R. (2015). "Flexible modeling of epidemics with an empirical Bayes framework." *PLoS Computational Biology*, 11(8): e1004382. 262, 285

Brynjarsdóttir, J. and O'Hagan, A. (2014). "Learning about physical parameters: the importance of model discrepancy." *Inverse Problems*, 30(11): 114007. MR3274591. doi: https://doi.org/10.1088/0266-5611/30/11/114007. 263

Capaldi, A., Behrend, S., Berman, B., Smith, J., Wright, J., and Lloyd, A. L. (2012). "Parameter Estimation and Uncertainty Quantification for an Epidemic Model." *Mathematical Biosciences and Engineering*, 553. MR2957535. 267, 270

Centers for Disease Control and Prevention (2016a). "MMWR Weeks." Accessed: 03-18-2017. URL https://wwwn.cdc.gov/nndss/document/MMWR_week_overview.pdf 265

Centers for Disease Control and Prevention (2016b). "Overview of influenza surveillance in the United States." Accessed: 03-18-2017. URL https://www.cdc.gov/flu/weekly/overview.htm 261, 264

Centers for Disease Control and Prevention (2016c). "Past Pandemics." Accessed: 03-18-2017. URL https://www.cdc.gov/flu/pandemic-resources/basics/past-pandemics.html 265

Chretien, J.-P., George, D., Shaman, J., Chitale, R. A., and McKenzie, F. E. (2014). "Influenza forecasting in human populations: a scoping review." *PloS One*, 9(4): e94130. 262

Coelho, F. C., Codeço, C. T., and Gomes, M. G. M. (2011). "A Bayesian framework for parameter estimation in dynamical models." *PloS One*, 6(5): e19616. 263

Dukic, V., Lopes, H. F., and Polson, N. G. (2012). "Tracking epidemics with Google flu trends data and a state-space SEIR model." *Journal of the American Statistical Association*, 107(500): 1410–1426. MR3036404. doi: https://doi.org/10.1080/01621459.2012.713876. 262, 263

Epidemic Prediction Initiative (2015a). "Forecast Evaluation." Accessed: 03-21-2017. URL https://predict.phiresearchlab.org/legacy/flu/evaluation.html 280

Epidemic Prediction Initiative (2015b). "Forecast Targets." Accessed: 03-21-2017. URL https://predict.phiresearchlab.org/legacy/flu/targets.html 280

Epidemic Prediction Initiative (2016). "FluSight 2016-17: Seasonal Influenza Forecasting." Accessed: 03-18-2017. URL https://predict.phiresearchlab.org/post/57f3f440123b0f563ece2576 279

Ewing, A., Lee, E. C., Viboud, C., and Bansal, S. (2016). "Contact, travel, and transmission: the impact of winter holidays on influenza dynamics in the United States." *Journal of Infectious Diseases*, jiw642. 268

FluSight Influenza Forecasting Challenge (2016). "Ensemble forecasts for the CDC's 2015–2016 Flu Forecasting Challenge." Accessed: 04-18-2017. URL https://github.com/cdcepi/FluSight_ensemble/tree/master/Data/2015-2016 280

Garza, R. C., Basurto-Dávila, R., Ortega-Sanchez, I. R., Oreste Carlino, L., Meltzer, M. I., Albalak, R., Balbuena, K., Orellano, P., Widdowson, M.-A., and Averhoff, F. (2013). "Effect of Winter School Breaks on Influenza-like Illness, Argentina, 2005–2008." *Emerging Infectious Disease*, 19(6). 268

Gelman, A. and Rubin, D. B. (1992). "Inference from iterative simulation using multiple sequences." *Statistical Science*, 457–472. 274

Generous, N., Fairchild, G., Deshpande, A., Del Valle, S. Y., and Priedhorsky, R. (2014). "Global disease monitoring and forecasting with Wikipedia." *PLoS Computational Biology*, 10(11): e1003892. 263, 285

Ginsberg, J., Mohebbi, M. H., Patel, R. S., Brammer, L., Smolinski, M. S., and Brilliant, L. (2009). "Detecting influenza epidemics using search engine query data." *Nature*, 457(7232): 1012–1014. 263, 285

Grefenstette, J. J., Brown, S. T., Rosenfeld, R., DePasse, J., Stone, N. T., Cooley, P. C., Wheaton, W. D., Fyshe, A., Galloway, D. D., Sriram, A., Guclu, H., Abraham, T., and Burke, D. S. (2013). "FRED (A Framework for Reconstructing Epidemic Dynamics): an open-source software system for modeling infectious diseases and control strategies using census-based populations." *BMC Public Health*, 13(1): 940. 262

Health and Human Services (2015). "Regional Offices." Accessed: 03-29-2017. URL https://www.hhs.gov/about/agencies/regional-offices/index.html?language=es 264

Hickmann, K. S., Fairchild, G., Priedhorsky, R., Generous, N., Hyman, J. M., Deshpande, A., and Del Valle, S. Y. (2015). "Forecasting the 2013–2014 influenza season using Wikipedia." *PLoS Computational Biology*, 11(5): e1004239. 262, 285

Higdon, D., Gattiker, J., Williams, B., and Rightley, M. (2008). "Computer model calibration using high-dimensional output." *Journal of the American Statistical Association*, 103(482): 570–583. MR2523994. doi: https://doi.org/10.1198/016214507000000888. 263

Huang, K. E., Lipsitch, M., Shaman, J., and Goldstein, E. (2014). "The US 2009 A/H1N1 influenza epidemic: quantifying the impact of school openings on the reproductive number." *Epidemiology (Cambridge, Mass.)*, 25(2): 203. MR2572908. doi: https://doi.org/10.1016/j.mbs.2009.06.002. 268

Jandarov, R., Haran, M., Bjørnstad, O., and Grenfell, B. (2014). "Emulating a gravity model to infer the spatiotemporal dynamics of an infectious disease." *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 63(3): 423–444. MR3238160. doi: https://doi.org/10.1111/rssc.12042. 262

Jandarov, R., Haran, M., and Ferrari, M. (2012). "A compartmental model for meningitis: separating transmission from climate effects on disease incidence." *Journal of Agricultural, Biological, and Environmental Statistics*, 1–22. MR2993273. doi: https://doi.org/10.1007/s13253-012-0101-2. 262

Kennedy, M. C. and O'Hagan, A. (2001). "Bayesian calibration of computer models." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 63(3): 425–464. MR1858398. doi: https://doi.org/10.1111/1467-9868.00294. 263

Kermack, W. O. and McKendrick, A. G. (1927). "A contribution to the mathematical theory of epidemics." In *Proceedings of the Royal Society of London A: mathematical, physical and engineering sciences*, 700–721. The Royal Society. 266

Lazer, D., Kennedy, R., King, G., and Vespignani, A. (2014). "The parable of Google Flu: traps in big data analysis." *Science*, 343(6176): 1203–1205. 285

Linzer, D. A. (2013). "Dynamic Bayesian forecasting of presidential elections in the States." *Journal of the American Statistical Association*, 108(501): 124–134. MR3174607. doi: https://doi.org/10.1080/01621459.2012.737735. 272

Mniszewski, S. M., Del Valle, S. Y., Stroud, P. D., Riese, J. M., and Sydoriak, S. J. (2008). "EpiSimS simulation of a multi-component strategy for pandemic influenza." In *Proceedings of the 2008 Spring simulation multiconference*, 556–563. Society for Computer Simulation International. 262

Molinari, N.-A. M., Ortega-Sanchez, I. R., Messonnier, M. L., Thompson, W. W., Wortley, P. M., Weintraub, E., and Bridges, C. B. (2007). "The annual impact of seasonal influenza in the US: measuring disease burden and costs." *Vaccine*, 25(27): 5086–5096. 261

Nsoesie, E. O., Brownstein, J. S., Ramakrishnan, N., and Marathe, M. V. (2014). "A systematic review of studies on forecasting the dynamics of influenza outbreaks." *Influenza and Other Respiratory Viruses*, 8(3): 309–316. 262

Osthus, D., Gattiker, J., Priedhorsky, R., and Del Valle, S. Y. (2018). "Dynamic Bayesian Influenza Forecasting in the United States with Hierarchical Discrepancy: Supplementary Material." *Bayesian Analysis*. doi: https://doi.org/10.1214/18-BA1117SUPP. 270

Osthus, D., Hickmann, K. S., Caragea, P. C., Higdon, D., and Del Valle, S. Y. (2017). "Forecasting seasonal influenza with a state-space SIR model." *Annals of Applied Statistics*, 11(1): 202–224. MR3634321. doi: https://doi.org/10.1214/16-AOAS1000. 262, 270

Plummer, M. (2003). "JAGS: A program for analysis of Bayesian graphical models using Gibbs sampling." In *Proceedings of the 3rd international workshop on distributed statistical computing*, volume 124, 125. Vienna. 274

Plummer, M. (2016). *rjags: Bayesian Graphical Models using MCMC*. R package version 4-6. URL https://CRAN.R-project.org/package=rjags 274

Plummer, M. (2017). "JAGS version 4.3.0 User Manual." Accessed: 11-11-2017. URL http://people.stat.sc.edu/hansont/stat740/jags_user_manual.pdf 274

Plummer, M., Best, N., Cowles, K., and Vines, K. (2006). "CODA: Convergence Diagnosis and Output Analysis for MCMC." *R News*, 6(1): 7–11. URL http://CRAN.R-project.org/doc/Rnews/ 274

Pokharel, G. and Deardon, R. (2016). "Gaussian process emulators for spatial individual-level models of infectious disease." *Canadian Journal of Statistics*, 44(4): 480–501. MR3574133. doi: https://doi.org/10.1002/cjs.11304. 262

Polgreen, P. M., Chen, Y., Pennock, D. M., Nelson, F. D., and Weinstein, R. A. (2008). "Using Internet searches for influenza surveillance." *Clinical Infectious Diseases*, 47(11): 1443–1448. 285

Priedhorsky, R., Osthus, D., Daughton, A. R., Moran, K. R., Generous, N., Fairchild, G., Deshpande, A., and Del Valle, S. Y. (2017). "Measuring Global Disease with Wikipedia: Success, Failure, and a Research Agenda." In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing*, 1812–1834. ACM. 285

R Core Team (2016). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/ 274

Rolfes, M. A., Foppa, I. M., Garg, S., Flannery, B., Brammer, L., Singleton, J. A., Burns, E., Jernigan, D., Reed, C., Olsen, S. J., and Bresee, J. (2016). "Estimated Influenza Illnesses, Medical Visits, Hospitalizations, and Deaths Averted by Vaccination in the United States." Accessed: 04-01-2017. URL https://www.cdc.gov/flu/about/disease/2015-16.htm 261

Shaman, J. and Karspeck, A. (2012). "Forecasting seasonal outbreaks of influenza." *Proceedings of the National Academy of Sciences*, 109(50): 20425–20430. 270

Shaman, J., Karspeck, A., Yang, W., Tamerius, J., and Lipsitch, M. (2013). "Real-time influenza forecasts during the 2012–2013 season." *Nature Communications*, 4. 262, 264, 285

Soebiyanto, R. P., Adimi, F., and Kiang, R. K. (2010). “Modeling and predicting seasonal influenza transmission in warm regions using climatological parameters.” *PloS One*, 5(3): e9450. 262

Tetlock, P. E., Mellers, B. A., and Scoblic, J. P. (2017). “Bringing probability judgments into policy debates via forecasting tournaments.” *Science*, 355(6324): 481–483. 279

Towers, S. and Feng, Z. (2009). “Pandemic H1N1 influenza: Predicting the course of vaccination programme in the United States.” *Euro Surveillance*, 14. 262

Viboud, C., Boëlle, P.-Y., Carrat, F., Valleron, A.-J., and Flahault, A. (2003). “Prediction of the spread of influenza epidemics by the method of analogues.” *American Journal of Epidemiology*, 158(10): 996–1006. MR0618792. 262

Weiss, H. H. (2013). “The SIR model and the Foundations of Public Health.” *Materials Matemàtics*, 1–17. 266, 267

Wilson, K. and Brownstein, J. S. (2009). “Early detection of disease outbreaks using the Internet.” *Canadian Medical Association Journal*, 180(8): 829–831. 285

Acknowledgments

We thank Matthew Biggerstaff and Craig McGowan for their help facilitating CDC forecasting challenge model comparisons, C.C. Essix for her encouragement and support, and three anonymous referees, the associate editor, and the editor for their insightful comments that greatly improved the manuscript. This work is supported in part by NIH/NIGMS/MIDAS under grant U01-GM097658-01 and by the U.S. Department of Energy, Office of Science, Advanced Scientific Computing Research SciDAC program. LANL is operated by Los Alamos National Security, LLC for the Department of Energy under contract DE-AC52-06NA25396. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript. Approved for public release: LA-UR-17-22749.

Invited Discussion

Lance A. Waller[*][†]

1 On the interface of statistical and mathematical modeling

Many thanks to Osthus et al. (2019) for a thoughtful addition to the literature operating at the interface of statistical and mathematical modeling of dynamic systems with particular application to inference regarding the annual influenza season. Having served as a judge in the 2013-2014 influenza modeling competition sponsored by the US Centers for Disease Control and Prevention (CDC), I am grateful to see continuing developments in this area. As noted by the authors, past competitions featured a variety of approaches applied to the weekly case reports of influenza-like illness (ILI) from the CDC. Some approaches link machine learning, search engine results, and social media posts to the ILI reports and seek to identify predictive patterns over the annual flu season, others utilize dynamic models via systems of differential equations or large-scale agent-based models to describe the cycle of infection across local, regional, and national populations. The authors' approach blends these via a hierarchical framework linking a smooth underlying dynamic systems model with stochastic patterns of observational discrepancies both across and within individual flu seasons.

The authors' modeling approach builds on a general hierarchical combination of a deterministic mathematical *process model* defining the general dynamics of the system, and a statistical *data model* given the process model. The data model defines stochastic variation from the dynamic process resulting in the observed data values. A hierarchical framework of a process model and a data model given the process was introduced in the context of time series modeling by Berliner (1996) and is a popular framework in the environmental sciences, especially within climate science and ecology (Wikle (2003)). The data model (given the process) can be simple (e.g., Gaussian measurement error) or can capture multiple levels of structured uncertainty and covariation, as in the case of the authors' approach.

Generally speaking, the hierarchical process and data model approach is attractive for the study of the dynamics of disease outbreaks particularly because the framework requires modelers to understand, appreciate, and apply both advanced mathematical and statistical modeling in an integrated inferential framework. This integration often requires statisticians and mathematical modelers to step outside of their traditional comfort zones, training systems, and worldviews. I previously summarized some aspects of this contrast in a discussion of a model of chronic wasting disease by Heisey

[*] This work supported in part by grant number R01HD09258, US Eunice Kennedy Shriver National Institute of Child Health and Human Development (NICHD). The opinions expressed here reflect those of the author and do not necessarily represent those of the funding agency.

[†] Department of Biostatistics and Bioinformatics, Rollins School of Public Health, Emory University, 1518 Clifton Road NE, Atlanta GA 30322, lwaller@emory.edu

et al. (2010) (Waller (2010)) and a few points raised there also merit mention here. At one extreme, a mathematical modeling-only approach seeks to define the underlying dynamics of a system via the process model and often only includes a very simple data model involving measurement error (e.g., using least squares or another error measure to assess goodness-of-fit via differences between observed and model-predicted values). At the other extreme, a statistical modeling-only approach focuses on defining inference on associations between observed outcomes and potential covariates, often without an underlying model of how these variables interact (e.g., various regressions). Machine-learning approaches can extend the statistical approach further to build sophisticated predictive algorithms based on patterns between variables but again with little attention to any underlying biologic or epidemiologic process. Such approaches can provide improved prediction but may ignore known features of disease epidemiology and can go astray for periods of time, as was the case with Google Flu Trends (Lazar et al. (2014)). While our current training systems often concentrate on one aspect or the other (mathematical or statistical modeling), there is substantial value and insight in incorporating both perspectives at a high level. In the present paper, the authors (Osthus et al. (2019)) provide a framework for drawing from the strengths of both worlds connected through a hierarchical structure allowing Bayesian inference on model parameters.

While the authors use the process and data model terminology in the titles of Sections 4.1 and 4.2, I feel it can be helpful to remind readers of the general framework, which is simple to state but widely extendable and general. Additional focused discussion of the roles of the process and data models also allows discussion of whether the hierarchical terms μ_t and $\delta_{j,t}$ represent components of the underlying susceptible-infected-recovered (SIR) *process model* (where they are defined currently in Section 4.2) or whether, conceptually, they are better framed as components of the *data model* by providing more nuanced statistical modeling of discrepancies between the underlying mathematical model and the observed data.

I frame my thinking as follows. Typically, one can partition the full set of model parameters into process parameters (relating to dynamics) and data parameters (relating to variation and covariation in observations) and it is helpful to think of each of these in turn. In disease outbreak applications of the process and data model framework, process parameters typically focus on biological/epidemiologic quantities linked to the dynamics of disease transmission. In the case of a SIR process model, process parameters include the rate (for deterministic implementations) or probability (for stochastic implementations) of transmission per contact between infected (assuming, for simplicity, all infected individuals are infectious) and susceptible individuals and the rate (or probability) of susceptible-infected contacts. Such parameters provide epidemiologic understanding of the dynamics and potential control of the outbreak as they help define the basic and effective reproductive numbers that, in turn, define whether and how an outbreak will grow or shrink. However, these process parameters are defined at the individual infection level and these parameters can be notoriously difficult to estimate due to identifiability concerns, if the only data available are aggregate counts of cases over a given period of time (e.g., weekly ILI counts). For example, it can be difficult to assess whether an observed number of cases is more likely due to a low rate of transmission per contact and a high rate of contact, or a high rate of transmission per contact and a low rate of contact. While epidemiologically meaningful, the process parameters may only weakly be identifiable (or, in some cases, not identifiable at all) from the data at hand.

The authors make use of three tools to help address the identifiability challenges. First, pooling data across multiple outbreaks of the same disease or adding data with additional information can help for weakly identifiable parameters, but, as the authors carefully illustrate, the fundamental mismatch between the biologic/epidemiologic process parameters of interest and the parameters identifiable in the data can continue to present a challenge in model fitting. Second, setting constraints for some parameters also provides a tool to aid identifiability and proves useful within the authors' model. The Bayesian formulation of the hierarchical process and data model system provides a third tool to address aspects of the identifiability challenge, namely the use of prior distributions for model parameters and random effects. In a sense, the definition of prior distributions provides a formal inferential mechanism for the first two tools by (1) defining a structure for borrowing information across data sets through the definition of random effects and (2) implementing constraints via subjective prior distributions (providing a potential way of weakening dependence upon fixed constraints). A particularly attractive feature of the authors' approach is the creative use of the discrepancy random effect parameters (and their prior distributions) to frame borrowing of information across outbreaks via the μ_t parameters and within each outbreak via the autoregressive $\delta_{j,t}$ parameters for each year t. The authors' prior specifications define how the full model will combine information across annual outbreaks and across months within a given outbreak in order to model systematic deviations from the smooth SIR process model. While the authors define the μ_t and $\delta_{j,t}$ parameters as part of the process model in Section 4.2 (I suspect primarily due to the convenience of defining these parameters as part of equation 5), I suggest that, conceptually, they are better considered part of an expanded data model to define how and to what extent the model borrows information across observations in order to allow statistical estimation of the underlying process parameters. The distinction of whether to include these parameters as part of the process model or the data model may be a bit of a quibble on my part, but I feel some additional discussion by the authors on this topic may be helpful in placing the authors' model within the broader context of hierarchical process and data models for complex systems.

2 On modeling discrepancy

The discrepancy parameters μ_t and $\delta_{j,t}$ are key to the authors' approach and allow statistical modeling of how and when we see consistent deviation between the smooth underlying SIR model of transmission dynamics and individual quirks in observations each year. Importantly, the discrepancy parameters allow the authors' model to adjust between annual peaks that tend to occur either early or late in the season, and to model the unusual but consistent deviations observed in week 13 of the season each year. That is, while the Runge-Kutta algorithm minimizes the overall deviation between the dynamic SIR process and the observed data within each year, the added hierarchical elements identify that there is a pattern to these deviations both across influenza seasons and within each individual season.

The discrepancy modeling within the authors' model provides improved fit over the individual models in the CDC competition as illustrated in their performance measures. The authors' model clearly benefits from information across repeated outbreaks but it is not clear whether, in the competition setting of predicting within an ongoing outbreak, the other models would have access to the same level of data completeness used by the authors. The current comparisons clearly show the added value of borrowing information across seasons and months, but it was not entirely clear to me whether direct comparison to the competition model performances is entirely fair with respect to the data assumed available at any particular point during an ongoing flu season.

The authors' comments regarding the anomaly during week 13 reflect an advantage of their discrepancy modeling approach. As noted, week 13 of any flu season typically overlaps the December holidays of Christmas and New Year's Eve/Day and this week is clearly different than others with respect to ILI reports. With holidays from work and school, behaviors are different, locations are different, interaction patterns are different, health seeking behaviors are different, healthcare staffing may be different, and record reporting may be different at the end of the calendar year. All of these changes can result in different overall transmission dynamics, even if the individual probability of transmission per contact remains unchanged. Such changes can be very difficult to incorporate within a process model (either SIR or agent-based), and the authors' discrepancy approach offers a mechanism to summarize the annual pattern of this impact through the data model without seeking to detail such changes within the process model to capture a single, anomalous week of the season.

The modeling of discrepancies across and within seasons has appeal but, as noted in general above, also pushes the limits of identifiability within the model. As the authors' Figure 8 illustrates, the combined discrepancies defined by $\mu_t + \delta_{j,t}$ are well identified and the separate identifiability between μ_t and $\delta_{j,t}$ is aided by "anchoring" of the endpoints through the implemented identifiability constraints coupled with the temporal random walk priors. The annual weekly pattern in μ_t captures the consistent week 13 effect, as one might expect. However, the (general) pattern of negative μ_t and a similar-in-magnitude positive $\delta_{j,t}$ for any particular week t suggests potential identifiability challenges in applying the approach to other data sets. Additional comments/recommendations from the authors directed toward others hoping to apply or extend the models in other settings would be welcome in this regard.

3 On the value of multiple models

In closing, I again express my thanks to the authors for their thoughtful work on an inferential framework for assessing both the process and patterns within a given disease outbreak. Their hierarchical structure links the underlying dynamic system to systematic and stochastic data discrepancies and provides posterior inference on all model parameters. As noted in Biggerstaff et al. (2016) and Biggerstaff et al. (2018) in reviewing the 2013–2014 and 2014–2015 flu season CDC modeling challenges, model-based influenza forecasting is still in its early stages of development. Linking model outcomes to measurable public health goals remains a priority and requires expansion of the collaboration between mathematical and statistical modeling mindsets above with public health practice to ensure accurate, reliable, reproducible, and actionable forecasts. As a judge for one of the competitions, I found value in reviewing multiple models of the same outbreak to see where they agreed, and, almost as importantly, where they differed in order to explore potential features of a particular outbreak that differed from the "average" outbreak. The authors' discrepancy parameters may offer a general mechanism for summarizing and exploring these differences and may be extendable to model comparisons in future analyses (and competitions!), perhaps through application of similar data models on top of different process models. Identifying which discrepancies represent regular features of outbreaks and which represent one-time anomalies could be particularly powerful in assessing a current outbreak and customizing potential intervention strategies while formally learning from historical outbreaks of the same disease.

References

Berliner, L. M. (1996). "Hierarchical Bayesian time series models." In Hanson, K. and Silver, R. (eds.), *Maximum Entropy and Bayesian Methods*, volume 79 of *Fundamental Theories of Physics (An International Book Series on The Fundamental Theories of Physics: Their Clarification, Development and Application)*, 15–22. Dordrecht: Springer. MR1446712. doi: `https://doi.org/10.1007/978-94-011-5430-7`. 291

Biggerstaff, M., Alper, D., Dredze, M., Fox, S., Fung, I. C.-H., Hickman, K. S., Lewis, B., Rosenfeld, R., Shaman, J., Tsou, M.-H., Velardi, P., Vespignani, A., Finelli, L., and the Influenza Forcasting Contest Working Group (2016). "Results from the Centers for Disease Control and Prevention's Predict the 2013–2014 Influenza Season Challenge." *BMC Infectious Diseases*, 16(357). 294

Biggerstaff, M., Johansson, M., Alper, D., Brooks, L. C., Chakraborty, P., Farrow, D. C., Hyun, S., Kandula, S., McGowan, C., Ramakrishnan, N., Rosenfeld, R., Shaman, J., Tibshirani, R., Tibshirani, R. J., Vespignani, A., Yang, W., Zhang, Q., and Reed, C. (2018). "Results from the second year of a collaborative effort to forecast influenza seasons in the United States." *Epidemics*, 24: 26–33. 294

Heisey, D. M., Osnas, E. E., Cross, P. C., Joly, D. O., Langenberg, J. A., and Miller, M. W. (2010). "Linking process to pattern: Estimating spatiotemporal dynamics of a wildlife epidemic from cross-sectional data." *Ecological Monographs*, 80(2): 221–240. 291

Lazar, D., Kennedy, R., King, G., and Vespignani, A. (2014). "The parable of Google Flu: Traps in big data analysis." *Science*, 343: 1203–1205. 292

Osthus, D., Gattiker, J., Priedhorsky, R., and Del Valle, S. Y. (2019). "Dynamic Bayesian influenza forecasting in the United States with hierarchical discrepancy." *Bayesian Analysis*. 291, 292

Waller, L. A. (2010). "Bridging gaps between statistical and mathematical modeling in ecology." *Ecology*, 9(12): 3500–3502. 292

Wikle, C. K. (2003). "Hierarchical models in environmental science." *International Statistical Review*, 71(2): 181–199. 291

Invited Discussion

Leonhard Held* and Johannes Bracher†

We congratulate Osthus et al. (2018) for their interesting work which proposes new methodology for modelling and prediction of influenza epidemics. The repeated success of their methods in flu forecasting competitions is impressive. We separate our discussion into comments on modelling and forecasting.

1 Modelling

Osthus et al. (2018) analyse weekly influenza-like illness (ILI) surveillance data from the Centers for Disease Control (CDC) on the US national level. As the authors note, the data are actually available at a finer resolution stratified by ten surveillance regions, see the map in Figure 1. An interesting question is whether an analysis on this finer geographical resolution would give improved national ILI forecasts. In a study of norovirus gastroenteritis incidence in the twelve regions of Berlin, Germany, multivariate modelling has generally led to better predictions, even of aggregated forecast targets (Held et al., 2017). A simple approach would be to apply the Osthus et al. (2018) methodology separately to each ILI surveillance region with subsequent aggregation of the forecasts. It would then also be of interest to investigate whether region-specific discrepancy trajectories show similar structures as on the national level (Osthus et al., 2018, Figure 5). However, this simple approach ignores dependencies between regions and could be improved by a joint space-time model for all regions.

Indeed, models at the national level without accounting for differences in geography may lead to suboptimal forecasts if ILI characteristics are different across regions (Chakraborty et al., 2018). Such differences in peak time and incidence are clearly visible in Figure 1 and are not surprising given that the territory in question spans four time zones with very diverse climatic conditions. It has been suggested that spatial patterns in peak timing are relatively stable across seasons among older adults, with the Western United States peaking earlier (Wenger and Naumova, 2010). In a recent analysis of insurance claim data for ILI, Charu et al. (2017) find pronounced spatial patterns in onset times. Their results point to a predominantly localized mode of transmission which suggests that an even finer resolution than the crude ten-region resolution will be required to capture spatio-temporal spread. Long-range transmission events are less common as the distribution of human travel distances approximately decays as a power law (Brockmann et al., 2006). In this context, air travel information has been successfully incorporated into analyses of the CDC ILI database at the ten-region resolution (Brownstein et al., 2006; Paul et al., 2008). In the absence of such information, a power law formulation (Meyer and Held, 2014), possibly combined with a gravity model (Xia

*Epidemiology, Biostatistics and Prevention Institute, University of Zurich, leonhard.held@uzh.ch
†Epidemiology, Biostatistics and Prevention Institute, University of Zurich, johannes.bracher@uzh.ch

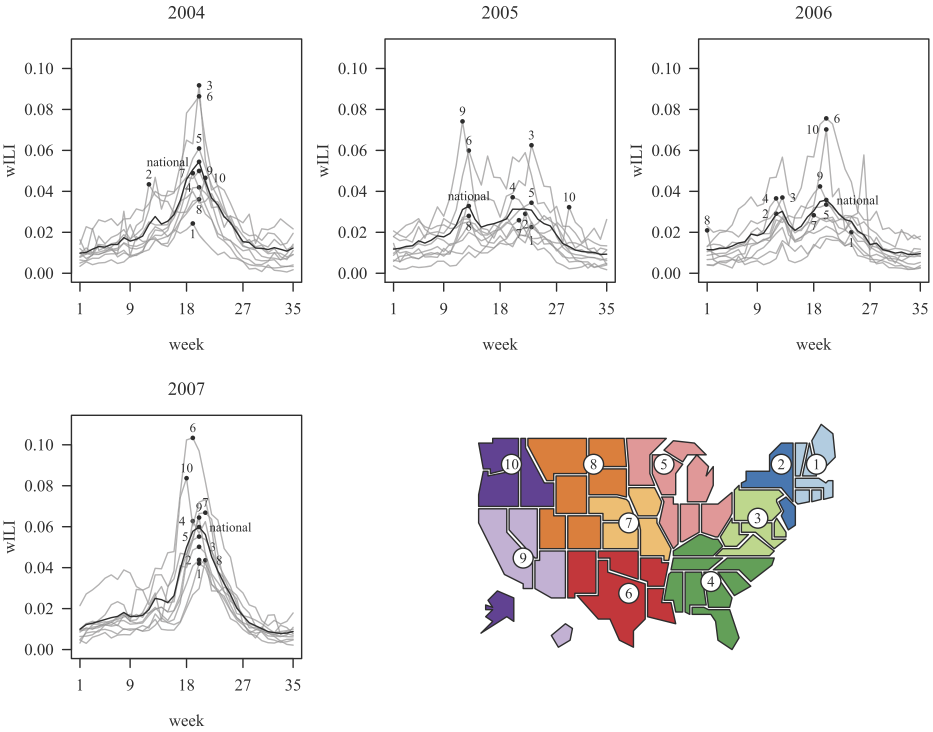


Figure 1: Weighted influenza-like illness for flu seasons 2004–2007 in the US: ten surveillance regions (grey lines) and aggregate national level (black line). Dots indicate the respective peak weeks and peak intensities. The seasons 2004 and 2007 show only limited variation in peak timing, while pronounced differences between regions can be seen in 2005 and 2006. In all four seasons substantial differences in peak intensity can be observed among the ten regions. The map shows the partition of the US into the ten surveillance regions.

et al., 2004) can serve as a useful proxy in a suitable statistical model. See Wakefield et al. (2019) for a comprehensive summary of different statistical modelling approaches for space-time infectious disease surveillance data.

There is also a large body of literature on how to extend the deterministic SIR (susceptible-infectious-recovered) model used by Osthus et al. (2018) to a stratified population (Daley and Gani, 1999, Section 2.4). A recent application to ILI can be found in Pei et al. (2018) who integrated commuter data in a metapopulation SIRS model. They found this approach to outperform isolated region-specific forecasts for 35 US states. It would be interesting to see whether the proposed hierarchical discrepancy approach can also improve such multivariate SIR analyses. For example, region-specific discrepancy components could be modelled using correlated random walks, see Riebler et al. (2012) for an application in mortality forecasting.

It is worth emphasizing that the basis of the stratification is not necessarily only geographical. Other potentially useful stratification variables include age group (Meyer and Held, 2017), virus strain (Goldstein et al., 2011) and degree of urbanization (Dalziel et al., 2018). Indeed, surveillance data stratified by region and age group are often available and have been used for multivariate forecasting (Held et al., 2017).

Turning to the proposed hierarchical discrepancy model, Osthus et al. (2018, Section 4.2) use a reverse random walk formulation for the discrepancy component μ_t common to all seasons and a reverse autoregressive model for the season-specific discrepancy components $\delta_{j,t}$ during week $t = 1, \ldots, 35$. The reverse-time formulation is suggested because it imposes a constraint on the late rather than the early part of the season. A possible alternative is an intrinsic autoregression (Rue and Held, 2005, Chapter 3), where the proper normal prior on μ_T is replaced by a flat improper prior. This would lead to a less informative prior distribution, constraining neither the early nor the late part of the season. The deterministic constraint on $\delta_{j,T}$ (equation (9) in Osthus et al. (2018)) could be replaced by a sum-to-zero constraint $\sum_j \delta_{j,t} = 0$ for all t, see Knorr-Held and Besag (1998) for a similar formulation applied to age-group specific random walks in a space-time disease mapping model.

2 Forecasting

Osthus et al. (2018) provide a thorough comparison with other flu prediction models based on proper scoring rules. We welcome this as often improper measures are used for the evaluation of probabilistic forecasts. Specifically, Osthus et al. (2018) consider the onset of the flu season, peak timing (PT) and peak intensity (PI) as forecast targets, as well as one-to-four week ahead forecasts. This reflects the aims of the CDC forecasting competitions, but long-term flu forecasts may also be of interest (Ray et al., 2017; Held and Meyer, 2019).

Empirical coverage of the dynamic Bayesian forecasting model is investigated for nominal 95% prediction intervals, which could be complemented with probability integral transform (PIT) histograms (Gneiting and Katzfuss, 2014). Empirical coverage turns out to be correlated (Figure 10) which suggests to assess calibration in a multivariate fashion (Gneiting et al., 2008). It may also be of interest to compute proper scoring rules for multivariate forecasts. This could be applied to path forecasts of the epidemic curve in one season (Held et al., 2017) or to the joint distribution of onset, peak timing and peak incidence. As noted by Wenger and Naumova (2010), these quantities are dependent with early seasons often showing higher intensity.

References

Brockmann, D., Hufnagel, L., and Geisel, T. (2006). “The scaling laws of human travel.” *Nature*, 439: 462–465. URL `https://www.nature.com/articles/nature04292`. 296

Brownstein, J. S., Wolfe, C. J., and Mandl, K. D. (2006). “Empirical evidence for the effect of airline travel on inter-regional influenza spread in the United States.” *PLoS*

Medicine, 3(10): e401. URL https://doi.org/10.1371/journal.pmed.0030401. 296

Chakraborty, P., Lewis, B., Eubank, S., Brownstein, J. S., Marathe, M., and Ramakrishnan, N. (2018). "What to know before forecasting the flu." *PLOS Computational Biology*, 14(10): 1–7. URL https://doi.org/10.1371/journal.pcbi.1005964. 296

Charu, V., Zeger, S., Gog, J., Bjørnstad, O. N., Kissler, S., Simonsen, L., Grenfell, B. T., and Viboud, C. (2017). "Human mobility and the spatial transmission of influenza in the United States." *PLOS Computational Biology*, 13(2): 1–23. URL https://doi.org/10.1371/journal.pcbi.1005382. 296

Daley, D. J. and Gani, J. (1999). *Epidemic Modelling: An Introduction*. Cambridge: Cambridge University Press. MR1688203. doi: https://doi.org/10.1017/CBO9780511608834. 297

Dalziel, B. D., Kissler, S., Gog, J. R., Viboud, C., Bjørnstad, O. N., Metcalf, C. J. E., and Grenfell, B. T. (2018). "Urbanization and humidity shape the intensity of influenza epidemics in U.S. cities." *Science*, 362(6410): 75–79. URL http://science.sciencemag.org/content/362/6410/75. 298

Gneiting, T. and Katzfuss, M. (2014). "Probabilistic forecasting." *Annual Review of Statistics and Its Application*, 1(1): 125–151. URL https://doi.org/10.1146/annurev-statistics-062713-085831. 298

Gneiting, T., Stanberry, L. I., Grimit, E. P., Held, L., and Johnson, N. A. (2008). "Assessing probabilistic forecasts of multivariate quantities, with an application to ensemble predictions of surface winds." *TEST*, 17(2): 211. MR2434318. doi: https://doi.org/10.1007/s11749-008-0114-x. 298

Goldstein, E., Cobey, S., Takahashi, S., Miller, J. C., and Lipsitch, M. (2011). "Predicting the epidemic sizes of influenza A/H1N1, A/H3N2, and B: A statistical method." *PLOS Medicine*, 8(7): 1–12. URL https://doi.org/10.1371/journal.pmed.1001051. 298

Held, L. and Meyer, S. (2019). "Forecasting based on surveillance data." In Held, L., Hens, N., O'Neill, P., and Wallinga, J. (eds.), *Handbook of Infectious Disease Data Analysis*. Chapman & Hall/CRC Press. URL https://arxiv.org/abs/1809.03735. 298

Held, L., Meyer, S., and Bracher, J. (2017). "Probabilistic forecasting in infectious disease epidemiology: the 13th Armitage lecture." *Statistics in Medicine*, 36(22): 3443–3460. MR3696502. doi: https://doi.org/10.1002/sim.7363. 296, 298

Knorr-Held, L. and Besag, J. (1998). "Modelling risk from a disease in time and space." *Statistics in Medicine*, 17(18): 2045–2060. URL https://doi.org/10.1002/(SICI)1097-0258(19980930)17:18<2045::AID-SIM943>3.0.CO;2-P. 298

Meyer, S. and Held, L. (2014). "Power-law models for infectious disease spread." *Annals of Applied Statistics*, 8(3): 1612–1639. MR3271346. doi: https://doi.org/10.1214/14-AOAS743. 296

Meyer, S. and Held, L. (2017). "Incorporating social contact data in spatio-temporal models for infectious disease spread." *Biostatistics*, 18(2): 338–351. MR3824756. doi: https://doi.org/10.1093/biostatistics/kxw051. 298

Osthus, D., Gattiker, J., Priedhorsky, R., and Del Valle, S. Y. (2018). "Dynamic Bayesian influenza forecasting in the United States with hierarchical discrepancy." *Bayesian Analysis*. Advance publication. URL https://doi.org/10.1214/18-BA1117. 296, 297, 298

Paul, M., Held, L., and Toschke, A. M. (2008). "Multivariate modelling of infectious disease surveillance data." *Statistics in Medicine*, 27(29): 6250–6267. MR2522320. doi: https://doi.org/10.1002/sim.3440. 296

Pei, S., Kandula, S., Yang, W., and Shaman, J. (2018). "Forecasting the spatial transmission of influenza in the United States." *Proceedings of the National Academy of Sciences*, 115(11): 2752–2757. URL http://www.pnas.org/content/115/11/2752. 297

Ray, E. L., Sakrejda, K., Lauer, S. A., Johansson, M. A., and Reich, N. G. (2017). "Infectious disease prediction with kernel conditional density estimation." *Statistics in Medicine*, 36(30): 4908–4929. MR3734482. doi: https://doi.org/10.1002/sim.7488. 298

Riebler, A., Held, L., and Rue, H. (2012). "Estimation and extrapolation of time trends in registry data – Borrowing strength from related populations." *Annals of Applied Statistics*, 6(1): 304–333. MR2951539. doi: https://doi.org/10.1214/11-AOAS498. 297

Rue, H. and Held, L. (2005). *Gaussian Markov Random Fields: Theory and Applications*. Boca Raton, FL: Chapman & Hall/CRC. MR2130347. doi: https://doi.org/10.1201/9780203492024. 298

Wakefield, J., Dong, T. Q., and Minin, V. N. (2019). "Spatio-Temporal Analysis of Surveillance Data." In Held, L., Hens, N., O'Neill, P., and Wallinga, J. (eds.), *Handbook of Infectious Disease Data Analysis*. Chapman & Hall/CRC. URL https://arxiv.org/abs/1711.00555. 297

Wenger, J. B. and Naumova, E. N. (2010). "Seasonal synchronization of influenza in the United States older adult population." *PLOS ONE*, 5(4): 1–11. URL https://doi.org/10.1371/journal.pone.0010187. 296, 298

Xia, Y., Bjørnstad, O., and Grenfell, B. (2004). "Measles metapopulation dynamics: A gravity model for epidemiological coupling and dynamics." *The American Naturalist*, 164(2): 267–281. PMID: 15278849. URL https://doi.org/10.1086/422341. 296

Invited Discussion

Vladimir N. Minin[*], Jonathan Fintzi[†], Luis J. Martinez Lomeli[‡], and Jon Wakefield[†,§]

The authors present an elegant method for accurate prediction of influenza-like-illness (ILI) incidence during an ongoing flu season. Their method combines ordinary differential equation-based (ODE-based) mechanistic modeling of ILI spread with flexible modeling of discrepancies between the ODE trajectories and observed incidence. The key idea is that these discrepancies behave similarly across flu seasons. Capturing these similarities in a Bayesian hierarchical model, the authors arrive at a predictive semi-parametric model of ILI spread. The authors conjecture that there is room for improving their approach and discuss some enhancements to the nonparametric component of their model. Below we argue that more careful handling of the parametric model component may also be a fruitful strategy to pursue in parallel with nonparametric model enhancements.

Flexible modeling and forecast sharpness

The authors motivate their discrepancy model component by correctly pointing out that certain consistently repeated features of ILI incidence time series cannot be predicted using deterministic mechanistic epidemic models. The authors' results show that the new Bayesian hierarchical model can indeed capture these features. For example, Figure 7 in the Osthus et al. manuscript shows that a consistent, but mysterious drop in ILI incidence from week 13 to week 14 can be seen in the authors' short term forecasts. However, the same figure shows that using the first 4 and 8 weeks of ILI data produces weeks 10-25 predictive intervals that are so large that they cover almost the entire plausible range of weighted ILI (wILI) counts. This suggests that the authors' model may be a little too flexible. There are multiple ways to tighten the authors' model, but from our perspective, the most intriguing avenue to pursue is to try to improve the parametric model component. Specifically, we first concentrate modeling efforts on improving the mean model, to reduce bias. Second, we finesse the wILI variance model, in particular paying attention to how the variance depends on the mean, so that we obtain an appropriate measure of uncertainty.

SIR-only model

Incidence ODEs with unknown initial conditions To establish a baseline, we wanted to see how SIR-only predictions compare to the authors' much more advanced modeling. Following the authors, we model the transmission dynamics of wILI in the population

[*] Department of Statistics, University of California, Irvine, CA, vminin@uci.edu

[†] Department of Biostatistics, University of Washington, Seattle, WA

[‡] Center for Complex Biological Systems, University of California, Irvine, CA

[§] Department of Statistics, University of Washington, Seattle, WA using a Susceptible-Infected-Recovered (SIR) model, represented as a system of ODEs. Let $\mathbf{X}^{(j)}(t) = (S^{(j)}(t), I^{(j)}(t), R^{(j)}(t)),\ S^{(j)}(t) + I^{(j)}(t) + R^{(j)}(t) = K$, denote the vector of compartment counts at time t in season $j \in \{1998, \dots, 2014\}$, where K is the population size that we set to 3×10^8 to approximate the size of U.S. population. We also let $\mathbf{X}_0^{(j)} = (S_0^{(j)}, I_0^{(j)}, R_0^{(j)})$ be the initial compartment counts.

The standard ODE representation of the SIR model expresses the time-evolution of the compartment counts as the solution to the following system of ODEs:

$$\begin{aligned} &\frac{\mathrm{d}S^{(j)}(t)}{\mathrm{d}t} = -\beta_j S^{(j)}(t) I^{(j)}(t),\ \frac{\mathrm{d}I^{(j)}(t)}{\mathrm{d}t} = \beta_j S^{(j)}(t) I^{(j)}(t) - \gamma_j I^{(j)}(t), \\ &\frac{\mathrm{d}R^{(j)}(t)}{\mathrm{d}t} = \gamma_j I^{(j)}(t),\ \text{such that, } \mathbf{X}^{(j)}(0) = \mathbf{X}_0^{(j)}, \end{aligned} \tag{1}$$

where β_j is the per–contact infection rate in season j and γ_j is the recovery rate. This is the same model that the authors use as their parametric component.

We modify the authors' SIR model in two ways. First, we are skeptical of the authors' claim that the initial number of susceptible individuals in each season is not identifiable. This claim may be true if only one season/outbreak is observed, but availability of multiple season onsets can make the initial number of susceptibles identifiable. To explore this issue, we introduce an additional parameter, C_j, for the number of susceptibles who are effectively removed at the start of season j, e.g., due to pre–existing immunity or geographic isolation. Second, to make the SIR model more appropriate for the incidence data, we follow Bretó and Ionides (2011) and Ho et al. (2018) and reparameterize the SIR ODEs in terms of cumulative incidence. Let $\mathbf{N}^{(j)}(t) = (N_{SI}^{(j)}(t), N_{IR}^{(j)}(t))$ denote the cumulative numbers of infections and recoveries and $\mathbf{N}^{(j)}(0)$ be the initial numbers of these events. The SIR ODEs for cumulative incidence and recoveries are given by

$$\begin{aligned} &\frac{\mathrm{d}N_{SI}^{(j)}(t)}{\mathrm{d}t} = \beta_j \left(S_0^{(j)} - C_j - N_{SI}^{(j)}(t)\right)\left(I_0^{(j)} + N_{SI}^{(j)}(t) - N_{IR}^{(j)}(t)\right), \\ &\frac{\mathrm{d}N_{IR}^{(j)}(t)}{\mathrm{d}t} = \gamma_j \left(I_0^{(j)} + N_{SI}^{(j)}(t) - N_{IR}^{(j)}(t)\right), \mathbf{N}^{(j)}(0) = (0, 0). \end{aligned} \tag{2}$$

Notice that we need the initial compartment counts $\mathbf{X}_0^{(j)}$ in the above system. Technically, we do not need to have both C_j and R_0^j in our model, because they represent the same number of initially removed individuals. We set $R_0^j = 0$ and estimate C_j due to constraints of our pre-baked implementation of the SIR model.

We fit two versions of our modified model to 15 seasons corresponding to years 1998–2007 and 2010–2014. In the first model A we assume that $C_j = C$, for j, with C being an unknown parameter that we estimate together with season-specific infection and recovery rates. We use this model primarily to test whether C is identifiable. The second model B is more realistic and assumes that each season j can have its own number of initially removed individuals, C_j. The model is hierarchical in that it assumes that *a priori* C_j's are drawn independently from the same distribution. More specifically, $\text{logit}(C_j/K) \sim \mathcal{N}(\mu_C, \sigma_C^2)$, with unknown parameters μ_C and σ_C^2 that we estimate.

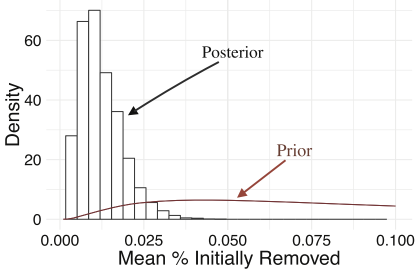


Figure 1: Prior density and posterior histogram of the proportion of initially removed individuals C/K. The prior and posterior are for the simple model in which all seasons start with the same number of initially removed individuals.

Data model Let $P_{SI}^{(j)}(t) = N_{SI}^{(j)}(t)/K$ be the attack rate (% of the population infected) up to time t in season j. Let $\Delta P_{SI}^{(j)}(t_\ell) = P_{SI}^{(j)}(t_\ell) - P_{SI}^{(j)}(t_{\ell-1})$ denote the attack rate in week ℓ. We model the observed wILI in week ℓ of season j, denoted $Y_\ell^{(j)}$, as

$$\operatorname{logit}\left(Y_\ell^{(j)}\right) \sim \mathcal{N}\left(\operatorname{logit}\left(\Delta P_{SI}^{(j)}(t_\ell)\right), \frac{\omega_0 + \omega_1 \Delta P_{SI}^{(j)}(t_\ell)}{\Delta P_{SI}^{(j)}(t_\ell)\left(1 - \Delta P_{SI}^{(j)}(t_\ell)\right)^2}\right), \tag{3}$$

where ω_0 and ω_1 control the variance of the emission distribution. This measurement model derives from an application of the delta method to a normal approximation of an overdispersed binomial distribution for detected wILI cases under the assumption that the rate of patient visits is not changing across time. The main motivation for this fairly complicated data model is our desire to model the dependence of wILI count variance on the latent/unobserved population incidence.

Priors and posterior inference We assign informative, scientifically meaningful priors, detailed in Table 1, for the parameters of models A and B. Note that we assign Dirichlet-Multinomial prior to the initial state $\mathbf{X}_0$ in such a way that there are no removed individuals at time 0, because we have a separate parameter to the number of initially removed individuals, C_j. For model A, where all parameters but the number of removed individuals C, are decoupled across all the seasons, we used our custom Markov chain Monte Carlo (MCMC) algorithm to approximate the posterior distribution of $(\beta_j, \gamma_j, I_0^j, S_0^j, \omega_0, \omega_1)$ for each season j and C that is common to all seasons. We show the prior and posterior distributions of the number of removed individuals C in Figure 1. The apparent differences between the prior and posterior distributions suggests that parameter C is identifiable. Moreover, our proportion of initially removed individuals C/K is much lower than 0.1 — the number used by Osthus et al. In Model B, we used MCMC to target the posterior distribution of $(\beta_j, \gamma_j, I_0^j, S_0^j, C_j, \omega_0, \omega_1)$ for each season j and (μ_C, σ_C). We found that the season-specific C_j's and their overall prior mean μ_C and standard deviation σ_C were also identifiable. We omit most of posterior summaries for the sake of brevity.

Model	**Parameter**	**Interpretation**	**Prior**	**Prior Median (90% Interval)**
A	$R0^{(j)} = \beta_j(K - C)/\mu_j - 1$	Basic reproduction #-1	LogNormal(log(0.4), 1.25)	$R0^{(j)} = 1.4$ (1.05, 4.10)
B	$R0^{(j)} = \beta_j(K - C_j)/\mu_j - 1$	Basic reproduction #-1	LogNormal(log(0.4), 1.25)	$R0^{(j)} = 1.4$ (1.05, 4.10)
A,B	$7/\mu_j - 1$	Mean infectious period (days-1)	LogNormal(log(7), 0.843)	$7/\mu_j = 7$ (1.75, 28)
A	C/K	% initially removed	LogitNormal(logit(0.1), 1)	$C/K = 0.1$ (0.02, 0.37)
B	μ_C	Mean logit % initially removed	LogitNormal(logit(0.1), 0.63)	expit(μ_C) = 0.1 (0.04, 0.24)
B	σ_C	Std.dev. logit % initially removed	Exponential(4)	$\sigma_C = 0.17$ (0.013, 0.75)
B	C_j/K	% initially removed	LogitNormal(μ_c, σ_C^2)	—
A,B	ω_0	Variance parameter	Exponential(3×10^8)	$\omega_0 = 2.3 \times 10^{-9}$ (1.7×10^{-10}, 1.0×10^{-8})
A,B	ω_1	Variance parameter	Exponential(5)	$\omega_1 = 0.14$ (0.01, 0.60)
A,B	$\mathbf{X}_0$	Initial compartment counts	Dirichlet–Multinom.(150,2,0)	

Table 1: Parameters and priors used in fitting SIR models to wILI data. Under Model A, the initial depletion of susceptibles is common to all seasons, whereas Model B hierarchically allows for season–specific initial depletion of susceptibles.

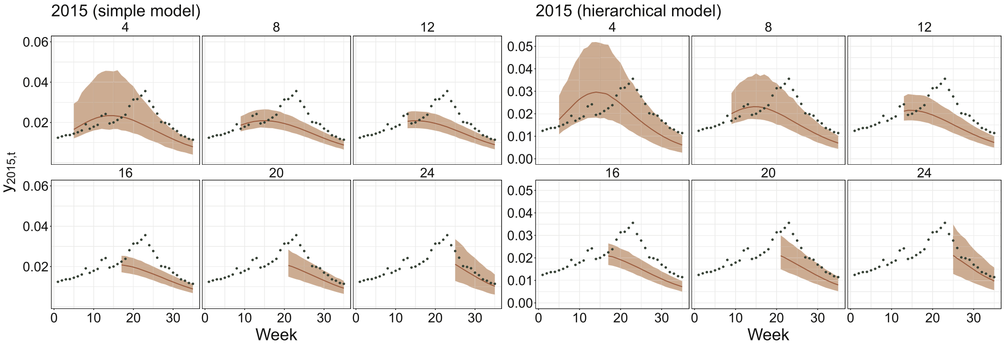


Figure 2: Forecasts for 2015 season under models A and B. The left six plots show forecasts produced by model A, where all seasons share the same number of initially removed individuals. The right six plots show forecasts produced by model B, where this number of initially removed individuals is season-specific. Each plot has the first z weeks/points used as training data, with the rest of the data being withheld during model fitting. This number z is shown above each plot (e.g., $z = 4$ in the top left plot). The solid red lines show the medians of the predictive distributions on which the forecasts are based. The shaded areas designate 95% predictive intervals.

SIR-only predictions

Now we use our SIR models A and B to make predictions about wILI incidence in season 2015. During this forecasting exercise, we use the estimated posterior distributions of SIR model parameters for seasons 1998–2007 and 2010–2014 in the following way. We pool MCMC samples of season-specific parameters and fit a multivariate Gaussian mixture model to these samples. For model A, separately from the mixture model fitting, we approximate the posterior distribution of initially removed individuals C with a univariate log-normal distribution. We use these approximations to the posterior model parameters as priors in our analysis of partial data from the 2015 season. As in the authors' paper, we fit our SIR model A to the first z weeks of data and use this model to predict the rest of the season for $z = 4, 8, 12, 16, 20, 24$. Prediction results are shown in Figure 2. Both sets of priors result in reasonable short term forecasts in weeks 4, 8, 12, and 24, but the timing of the epidemic peak is not predicted well. We see that a mixture model-based prior distribution of the initial number of removed individuals, obtained from the posterior samples under model B, produces better forecasts than predictions based on a prior distribution obtained from the posterior of model A. However, similarly to the Osthus et al. hierarchical model, this improvement comes at the expense of wider predictive intervals. Still, our experiments with the initial states of flu seasons demonstrate that careful modeling of initially removed individuals may be a fruitful forecasting strategy, at least in the context of time homogeneous infectious disease dynamics.

Speculative remarks

Although our SIR-only predictions are not competitive with the state-of-the-art ILI forecasting methods, they establish a parametric modeling starting point, which is different from the starting point of Osthus et al. Combining parametric modeling similar to ours with the authors' hierarchical discrepancy model may improve ILI forecasting even further. More specifically, it would be interesting to see if including the initial number of removed individuals as a free parameter and/or a data model with a mean/variance relationship into Osthus et al. model would lead to better forecasts.

Another way to improve SIR-only predictions is to use stochastic SIR modeling and to move to a nonparametric modeling of the infection rate, as was recently proposed by Xu et al. (2016) in a wider context of stochastic epidemic modeling. For example, we can assume that for season each j the time-varying infection rate has the form $\beta_j(t) = \alpha_j \times \beta(t)$, where α_j's are season-specific multipliers and $\beta(t)$ captures commonalties in infection rate changes across seasons. *A priori* modeling of $\beta(t)$ as a Gaussian process or another suitable functional prior would result in nonparametric estimation of $\beta(t)$. In summary, we are excited about the successes of Osthus et al. forecasting method based on semi-parametric modeling of infectious disease dynamics and looking forward to future modeling and forecasting improvements in this area.

References

Bretó, C. and Ionides, E. (2011). "Compound Markov counting processes and their applications to modeling infinitesimally over–dispersed systems." *Stochastic Processes and their Applications*, 121: 2571–2591. MR2832414. doi: https://doi.org/10.1016/j.spa.2011.07.005. 302

Ho, L., Crawford, F., and Suchard, M. (2018). "Direct likelihood-based inference for discretely observed stochastic compartmental models of infectious disease." *The Annals of Applied Statistics*, 12: 1993–2021. MR3852706. doi: https://doi.org/10.1214/18-AOAS1141. 302

Xu, X., Kypraios, T., and O'Neill, P. (2016). "Bayesian non-parametric inference for stochastic epidemic models using Gaussian processes." *Biostatistics*, 17(4): 619–633. MR3604269. doi: https://doi.org/10.1093/biostatistics/kxw011. 306

Contributed Discussion

David Conesa[*][†], Rubén Amorós[*][‡],
Antonio López-Quílez[*][†], and Miguel-Angel Martinez-Beneito[§]

We would like to start by congratulating the authors for this great piece of work that provides a novel and competitive model in the complicated field of influenza forecasting.

Influenza surveillance has become a challenging issue in public health practice due to its propensity to cause large scale seasonal epidemics and even pandemics. Many surveillance information sources (real-time internet surveys, over the counter sales, absenteeism registers, syndromic/sentinel surveillance, hospital admissions, influenza mortality rates, etc.) have been used to forecast its behaviour. As a result, always a new model is proposed to perform this forecast, it is important to describe how the model could be extended to deal with other kind of data. The beta modelling proposed by the authors clearly fits for the weighted influenza-like illness (wILI) data provided by the Centers for Disease Control and Prevention, but if the outcome of interest was collected in terms of counts (or even rates) as in Conesa et al. (2015), a Poisson or a Binomial model would more adequate. Could the proposal by Osthus et al. (2018) be adapted to these alternative information sources that provide other outcomes?

As stated by the authors, the key issue when trying to forecast influenza is how to model the discrepancy between seasons. In our experience, the behaviour of some epidemic seasons can be very different to what the authors appropriately named a "typical" season. Indeed, we have found seasons in which the influenza is not present (Martinez-Beneito et al., 2008). Is the proposal of Osthus et al. (2018) robust enough as for dealing with data of this kind? The causes of this discrepancy and the way it can be included in the final model will guide the remaining comments of Osthus et al. (2018) paper.

The extension of the region analysed is one of these possible causes. It is not clearly the same to perform global forecasting (like in the USA) than to do it in states or even counties. As stated by the authors, the behaviour of wILI in the whole USA is low at the beginning of the influenza season, increases to a maximum in the middle, and reverts to low levels by the end of the season. But, this behaviour could not be appropriate if the interest is to forecast in smaller parts of the country.

A possible first option to deal with this would be to use the Dynamic Bayesian Influenza Forecasting model proposed by the authors to analyse each region separately,

[*] This work has been partially supported by grants MTM2016-77501-P and TEC2016-81900-REDT from the Spanish Ministerio de Ciencia, Innovación y Universidades – Agencia Estatal de Investigación (jointly financed by the European Regional Development Fund, FEDER).

[†] Dept. Estadística e Inv. Operativa, Universitat de València, C/Dr. Moliner 50, 46100 Burjassot (Valencia) Spain, david.v.conesa@uv.es; antonio.lopez@uv.es

[‡] School of Mathematics, University of Edinburgh, The King's Buildings, Peter Guthrie Tait Road, EH9 3FD, Scotland, UK, ruben.amoros@ed.ac.uk

[§] Fundación para el Fomento de la Investigación Sanitaria y Biomédica de la Comunitat Valenciana (FISABIO), Valencia, Spain, martinez_mig@gva.es

although an option taking into account the spatial relationship among counties or states would be more helpful in order to understand the usual behaviour with diseases (Banks et al., 2012; Zou et al., 2012). In this line, including a spatio-temporal term in formulae (5) of Osthus et al. (2018) could take into account both temporal and spatial structures of the data and so it could provide a better performance of the forecasting system.

Another possible way to describe the discrepancies between seasons could be the environmental and/or climatic effects. In line with our previous comment, these effects could also be incorporated in the forecasting model as covariates in formulae (5) of Osthus et al. (2018). These covariates would also be helpful to describe above mentioned differences between states or counties.

Finally, the nowcasting or, even better, the forecasting of the onset of influenza epidemics is sometimes as important as the forecasting of the time series on its own. Predicting that particular feature of the epidemics has important consequences in real terms as it allows to prepare health services for the starting of the outbreak, which is the more critic time point for the health system. Do the authors find that their proposal could be useful in some sense for this particular purpose?

References

Banks, D., Datta, G., Karr, A., Lynch, J., Niemi, J., and Vera, F. (2012). "Bayesian CAR models for syndromic surveillance on multiple data streams: Theory and practice." *Information Fusion*, 13: 105–116. 308

Conesa, D., Martinez-Beneito, M. A., Amorós, R., and López-Quílez, A. (2015). "Bayesian hierarchical Poisson models with a hidden Markov structure for the detection of influenza epidemic outbreaks." *Statistical Methods in Medical Research*, 24: 206–223. MR3336291. doi: https://doi.org/10.1177/0962280211414853. 307

Martinez-Beneito, M. A., Conesa, D., López-Quílez, A., and López-Maside, A. (2008). "Bayesian Markov switching models for the early detection of influenza epidemics." *Statistics in Medicine*, 27: 4455–4468. MR2528524. doi: https://doi.org/10.1002/sim.3320. 307

Osthus, D., Gattiker, J., Priedhorsky, R., and Del Valle, S. Y. (2018). "Dynamic Bayesian influenza forecasting in the United States with hierarchical discrepancy." *Bayesian Analysis*. In press. 307, 308

Zou, J., Karr, A. F., Banks, D., Heaton, M. J., Datta, G., Lynch, J., and Vera, F. (2012). "Bayesian methodology for the analysis of spatial-temporal surveillance data." *Statistical Analysis and Data Mining*, 5: 194–204. MR2929962. doi: https://doi.org/10.1002/sam.10142. 308

Rejoinder

Dave Osthus[*], James Gattiker[*], Reid Priedhorsky[†], and Sara Y. Del Valle[‡]

We thank the *Bayesian Analysis* editorial team for organizing this discussion and all the discussants for their thoughtful and encouraging comments. In the rejoinder, we provide needed clarification on the Dynamic Bayesian (DB) model and address issues raised regarding model interpretation and future model development directions.

Response to Waller

We appreciate the discussion of data and process models. We chose to include discrepancy modeling components μ_t and $\delta_{j,t}$ in the description of the process model for the convenience of defining equation 5, as correctly conjectured by the discussant. Rather than include discrepancy in either the data or process models, we think partitioning the model into data, discrepancy, and process is a better representation, as is often done in the computer experiments modeling literature (e.g., Kennedy and O'Hagan, 2001; Higdon et al., 2004). That is, we prefer the representation that hierarchically decomposes the DB model into

$$[\text{data} \mid \text{discrepancy, process}] \times [\text{discrepancy} \mid \text{process}] \times [\text{process}].$$

This formulation does not change the DB model, but rather makes more explicit the three different modeling components. Namely, section "Model for logit($I_{j,t}$)" describes the disease transmission dynamics of the susceptible-infectious-recovered (SIR) process model, sections "Model for μ_t" and "Model for $\delta_{j,t}$" collectively describe the discrepancy model conditional on the process model, as $\delta_{j,T}$ is defined in equation 9 conditional on the process model, and section "Data Model" is defined conditional on both the process and discrepancy model, as $\pi_{j,t}$ is a function of both the process and discrepancy model. This description preserves the process model interpretation advocated for by the discussant while making clear discrepancy is a significant modeling component deserving of thoughtful consideration separate from the data model. The process, discrepancy, and data model description can be viewed as either an extension to the process/data model decomposition described by Berliner (1996) or as the hierarchical, rather than additive, decomposition of the model presented by Kennedy and O'Hagan (2001).

We agree that in some scenarios discrepancy unidentifiability could be problematic, but since the focus of this work is exclusively on forecasting, this does not directly apply. It is a known problem that adding in a flexible discrepancy model to account for systematic deviations between the process model and the data comes at the cost of model interpretation (Brynjarsdóttir and O'Hagan, 2014). However, not including a discrepancy model may lead to biased inference and inappropriate uncertainties. Building

[*] Statistical Sciences, Los Alamos National Laboratory, dosthus@lanl.gov
[†] High Performance Computing Environments, Los Alamos National Laboratory
[‡] Information Systems and Modeling, Los Alamos National Laboratory

a model that has both identifiable model components *and* appropriate uncertainties is a challenging problem. Formulating constraints on the discrepancy model, either through setting parameters to constant values, using informative/regularizing priors, or imposing model structure, are mechanisms that can improve model identifiability. The discrepancy constraints we chose (e.g., equation 9) were highly catered to this application and we only argue that they are pragmatically useful, not optimal in any sense. Determining the best way to impose constraints on the discrepancy model so that it is sufficiently flexible to capture residual structure but not too flexible to produce unreasonably large predictive uncertainty intervals is an open area of ongoing research.

Regarding the comparisons to other challenge participants in the paper, we affirm that direct comparisons are fair. We worked with challenge organizers from the Centers for Disease Control and Prevention (CDC) when writing this manuscript to ensure this, hence our discussion of using only data that would have been available at the time a forecast is rendered. For additional context, the DB model participated in the CDC's 2017–2018 flu forecasting challenge (not retrospectively) and, at the national scale, placed 3rd while our next iteration of the DB model, one that incorporates Google search volume data, placed 1st.

Response to Held and Bracher

We agree that extending the DB model to finer stratifications, such as age groups, flu strains, and/or geographic scales, is a really promising direction for future model development for improved forecast accuracy at more actionable levels for public services and funding. In fact, we have already implemented one of these promising directions. In 2017, the CDC piloted a state-level flu forecasting challenge, accompanied with the public release of state-level influenza-like illness (ILI) data. Since then, we have developed a spatio-temporal variation of the DB model that is currently participating in the CDC's 2018–2019 flu forecasting challenge (Centers for Disease Control and Prevention, 2019). Regional and national forecasts are directly derived as upscaled quantities of the state-level model, facilitating a unified view of forecasting not requiring different models at different scales. We are excited to see evidence provided by the discussants that suggests improved coarse-scale forecasting can be accomplished by modeling at finer scales.

The suggested alternative approaches to hierarchical discrepancy modeling, such as an intrinsic autoregression or a deterministic sum-to-zero constraint on $\sum_j \delta_{j,t} = 0$ for all t, provide additional options for constraining the discrepancy model and potentially improving identifiability. These are directly applicable to the question of identifiability raised by Waller.

The suggestion to consider probability integral transform histograms and proper scoring rules for multivariate forecasts is well-received. Model assessments for flu forecasting are less sophisticated and less popular than model comparisons. We think there is substantial potential for flu forecasting model improvements based on improved model assessments. Work to develop and introduce appropriate model assessment techniques to the flu forecasting community are admittedly needed to better diagnose areas for future model improvement. Model assessments can further be complemented with post-challenge analyses to identify areas for improvement (Anderson-Cook et al., 2019).

Response to Minin, Fintzi, Martinez Lomeli, and Wakefield

The discussants highlight that in addition to nonparametric improvements, parametric model improvements should also be considered. We agree improvements to the parametric SIR model is a potentially fruitful avenue for model improvement worth pursuing, especially for applications where the primary focus goes beyond prediction.

The discussants comment that the DB predictive intervals cover nearly the entire plausible weighted influenza-like illness (wILI) range when early season forecasts are made, suggesting our model may be too flexible. We agree the predictive intervals are large early in the season, but that is because there is little information in the early season wILI data to discriminate between seasons (e.g., mild vs. intense, early vs. late peaking), not because the model is too flexible. Furthermore, Figure 10(c) presents empirical coverage results that do not suggest the early season predictive intervals are too wide.

As the discussants correctly asserted and helpfully demonstrated, the initial number of susceptible individuals can be identified when multiple seasons are considered and model assumptions linking the initial susceptibility across seasons are made. Our comment about lack of identifiability of the initially susceptible population was in the context of fitting an SIR model to a single season of wILI data. We like the suggestion by the discussants to model $S_{j,0}$ hierarchically across seasons, as we did with other SIR parameters. This, in fact, might help improve our general under coverage problem (Figure 10) by enlarging our predictive intervals, a result that was illustrated by the discussant's Figure 2 and appears to be a consequence of weakening the constraint on $S_{j,0}$ across seasons.

The discussant's equation 3 allows for modeling the relationship between the variance of the wILI data and the latent population incidence. We point out that the Beta distribution used in our data model (equation 2) also allows for mean-variance modeling. Specifically, equation 4 shows that the variance of $y_{j,t}$ grows as $\mathrm{E}(y_{j,t}|\pi_{j,t}) = \pi_{j,t}$ increases from 0 to 0.5 for a fixed, positive λ. We had difficulties learning λ, which is partially why we set it. It would be interesting to see if the discussants were able to learn ω_0 and ω_1 well in their model.

Response to Conesa, Amorós, López-Quílez, and Martinez-Beneito

In response to the question about how the DB model can be extended to incorporate different types of surveillance information sources, such as internet surveys and hospital admissions, we note the answer will depend on the information source. For instance, internet surveys can be piped through a machine learning short-term forecasting model and appended to the end of available wILI data; this is what was done in the next iteration of the DB model that placed first nationally in the 2017–2018 challenge. Hospital admissions data might be related to the variability of wILI data. In equation 4, λ could be modeled as a function of hospital admissions. Whether it is useful or not to include additional information sources should be measured by improved predictive performance.

We note the DB model can be adapted to accommodate other forms of response data, such as counts or rates. The distribution of the data model should correspond to the support of the data being modeled. If, for instance, the response data were counts, the Beta distribution could be replaced with a Poisson(θ) distribution where θ would be modeled as a function of the process and discrepancy models.

The discussants also wonder about the effectiveness of the model when little or no influenza is present. The DB model is able to handle these scenarios straightforwardly, as little influenza corresponds to $y_{j,t}$ near zero. This is an important case from a practical perspective as different models are not needed for different degrees of circulating influenza; they are all handled by the same unified model.

Finally, regarding the question about forecasting features of the flu season, such as the onset of the flu season, we note that this is in fact what the DB model does for the flu forecasting challenge. The peak intensity, the timing of the peak intensity, and the flu season onset are examples of public health-relevant flu season features of interest.

References

Anderson-Cook, C. M., Myers, K. L., Lu, L., Fugate, M. L., Quinlan, K. R., and Pawley, N. (2019). "How to Host a Data Competition: Statistical Advice for Design and Analysis of a Data Competition." *arXiv preprint arXiv:1901.05356*. 311

Berliner, L. M. (1996). "Hierarchical Bayesian time series models." In *Maximum entropy and Bayesian methods*, 15–22. Springer. MR1446713. 309

Brynjarsdóttir, J. and O'Hagan, A. (2014). "Learning about physical parameters: The importance of model discrepancy." *Inverse Problems*, 30(11): 114007. MR3274591. doi: https://doi.org/10.1088/0266-5611/30/11/114007. 309

Centers for Disease Control and Prevention (2019). "FluSight: Flu Forecasting." Accessed: 01-28-2019. URL https://www.cdc.gov/flu/weekly/flusight/index.html 310

Higdon, D., Kennedy, M., Cavendish, J. C., Cafeo, J. A., and Ryne, R. D. (2004). "Combining field data and computer simulations for calibration and prediction." *SIAM Journal on Scientific Computing*, 26(2): 448–466. MR2116355. doi: https://doi.org/10.1137/S1064827503426693. 309

Kennedy, M. C. and O'Hagan, A. (2001). "Bayesian calibration of computer models." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 63(3): 425–464. MR1858398. doi: https://doi.org/10.1111/1467-9868.00294. 309