Bayesian Analysis (2024) **TBA**, Number TBA, pp. 1

Forecasting Influenza Hospitalizations Using a Bayesian Hierarchical Nonlinear Model with Discrepancy

Spencer Wadsworth * and Jarad Niemi*

Abstract. The annual influenza outbreak leads to significant public health and economic burdens making it desirable to have prompt and accurate probabilistic forecasts of the disease spread. The United States Centers for Disease Control and Prevention (CDC) hosts annually a national flu forecasting competition which has led to the development of a variety of flu forecast modeling methods. For the first several years of the competition, the target to be forecast was weekly percentage of patients with an influenza-like illness (ILI), but in 2021 the target was changed to weekly hospitalization counts. Reliable state and national hospitalization data has only been available since 2021, but for ILI the data has been available since 2010 and has been successfully forecast for several seasons. In this manuscript, we introduce a two component modeling framework for forecasting weekly hospitalizations utilizing both hospitalization data and ILI data. The first component is for modeling ILI data using a dynamic nonlinear Bayesian hierarchical model. The second component is for modeling hospitalizations as a function of ILI. For hospitalization forecasts, ILI is first forecasted and then hospitalizations are forecast with ILI forecasts used as a linear or quadratic predictor. In a simulation study, two ILI forecast models, including one similar to the winning model for two seasons of the CDC forecast competition from [43] and a nonlinear Bayesian hierarchical model from [53] are compared. Also assessed is the usefulness of including a systematic model discrepancy term in the ILI model. Forecasts of state and national hospitalizations for the 2023-24 flu season are made, and different modeling decisions are compared. We found that including a discrepancy component in the ILI model tends to improve forecasts during certain weeks of the year. We also found that other modeling decisions such as the exact nonlinear function to use in the ILI model or the error distribution for hospitalization models may or may not be better than other decisions, depending on the season, location, or week of the forecast.

Keywords: Disease outbreak forecasting, Bayesian hierarchical modeling, Probabilistic forecasting, Model discrepancy.

1 Introduction

Every year the seasonal influenza outbreak burdens the public health system by infecting millions, causing an influx of primary care visits and hospitalizations and leading to between 290,000 and 650,000 deaths worldwide [57]. [42] estimated the United States' annual economic burden from medical costs, loss of income, and deaths to be over $87

arXiv: 2010.00000

*Department of Statistics, Iowa State University, Ames, IA 50011

billion. Accurate forecasting of infectious diseases can inform public decision making and ease the burden of an outbreak [52, 37]. There is a growing consensus that disease forecasts should be probabilistic in nature [20, 8], and it has been shown that reporting forecast uncertainty along with predictions may lead to better decision making [46, 28, 58].

To better inform public decision making regarding the flu epidemic, in 2013 the United States Centers for Disease Control and Prevention (CDC) organized a national flu forecasting competition, also known as FluSight [6, 39, 10]. Originally, over a dozen teams of researchers from academic and industry backgrounds participated in FluSight by contributing their own forecast models. Besides the 2020 season –or the flu season spanning the fall of 2020 and the winter of 2021– FluSight has been operated annually and researchers outside the CDC have been invited to participate. Initially the target data for forecasts was influenza-like illness (ILI) data. ILI is the proportion of patients who meet a healthcare provider and who display flu like symptoms, and ILI data has been available at the state and national level since the 2010 flu season [12, 11]. The collaborative ILI forecasting effort has led to a number of modeling developments in flu forecasting [40, 44, 43, 53, see references therein for more examples], and in their paper's introduction, [43] categorized the most commonly used flu forecasting models into four classes including mechanistic models based on differential equation compartmental models, agent based models based on population simulation, machine learning/regression models including data driven machine learning and statistical models, and data assimilation models which are constructed by assimilating mechanistic models into a probabilistic framework. An additional forecast model used in FluSight involves the combination of several forecasts into a single ensemble forecast, which has been shown to perform well relative to individual models [40, 47, 59].

The administration of FluSight saw few changes during the first seven seasons, but the onset of the COVID-19 pandemic and subsequent developments for COVID-19 forecasting led to major modifications. As a result of the COVID-19 pandemic which began during the 2019 flu season, the typical flu outbreak behavior was altered between the 2019 and 2022 seasons [39]. The COVID-19 pandemic led to the creation of the Health and Human Services (HHS) Patient Impact and Hospital Capacity Data System [22] which contains COVID-19 and flu hospitalization data, and the COVID-19 Forecast Hub was founded. The COVID-19 Forecast Hub was based on FluSight but with certain major adjustments including how the forecast uncertainty is represented and the addition of the weekly publication of a multi-model ensemble forecast as the official forecast of the CDC [8, 13]. Using estimated quantiles for representing forecast uncertainty and creating a multi-model ensemble are both aspects of the COVID-19 Forecast Hub which were adopted by the flu forecast competition. Additionally the target of the flu forecasts changed from being ILI data to being HHS hospitalization data, which reports the number of hospitalizations due to a laboratory confirmed flu infection [39, 22]. This is as a result of having COVID-19 cases in the population making ILI data, already only a proxy for flu behavior, more difficult to interpret.

The contribution of this manuscript is to introduce a two component framework for modeling HHS hospitalization forecasts where hospitalization data and years of ILI

data are used to inform forecast models. The first modeling component is a model of ILI data and the second is a model of hospitalization data with ILI as a predictive covariate. Herein we use ILI models similar to those in [43] and [53] for ILI forecasting. The model of [43] is a combined data assimilation and statistical regression model which involves a compartmental model in a probabilistic framework. Their model also includes an additional component for capturing a systematic discrepancy between the deterministic part of the model and the actual data, an idea which was first introduced by [30]. The model in [53] is a Bayesian hierarchical regression model with an underlying function intended to capture the trajectory of the seasonal ILI data. Herein, we provide a framework under which discrepancy modeling may be used along with a general function modeling ILI data, and we show the effectiveness of including discrepancy modeling during certain periods of the flu season.

In line with the newer FluSight standard of forecasting hospitalizations, we model hospitalizations as a linear function of ILI. Thus forecasts produced herein target flu hospitalizations and are a mapping of ILI forecasts to hospitalizations. This allows for ILI data from many seasons to be exploited and for ILI forecasts to assist in forecasting hospitalizations, which has fewer seasons of data than ILI. Several modeling schemes and their forecasts for the 2023 flu hospitalization season are compared, and it is shown that the modeling decisions produce good forecast results for different states or times during the flu season.

In section 2 we review the ILI and hospitalization data provided by the CDC and targeted by FluSight. In section 3 the modeling framework contributed by this manuscript is given. In the same section, functions similar to those used by [43] and [53] are defined. These functions are the susceptible-infectious-recovered (SIR) compartmental model and the asymmetric Gaussian (ASG) function respectively. Model fitting and implementation are described at the end of the section. Section 4 is a simulation study where four ILI forecast models and their use in forecasting hospitalizations are compared. Commonly used proper scoring rules [21], which are also introduced and defined in section 4, are used for comparing the forecasts. Forecasting of the 2023 flu outbreak along with assessment and comparison under several selected models is performed in section 5, with the analysis being done using the conventions of FluSight. Finally, the manuscript is concluded in section 6 with general observations and some discussion.

2 Flu outbreak data

In this section we introduce and define ILI and hospitalization data and evaluate the data visually. ILI and hospitalization data have been the object of forecasting for FluSight with ILI being the target for the first seven seasons and hospitalizations being the target since the 2022 season. Both of these data were collected at the state, territorial, and national level and were reported at least weekly. Overall the data is reported for 53 locations including the 50 US states, the District of Columbia (DC), Puerto Rico (PR), and at the US national level. We will refer to forecast targets throughout this manuscript. A target is the specific horizon, 1, 2, 3, or 4-weeks ahead, for a specific location and week during the season.

2.1 Influenza-like illness data

The US Outpatient Influenza-like Illness Surveillance Network (ILINet) collects information on respiratory illness from outpatient visits to health care providers. Over 3,400 outpatient health care providers in all 50 US states, PR, DC, and the US Virgin Islands report each week the total number of outpatient visits along with the number of ILI cases. An ILI case is defined as a "fever (temperature of 100°F[37.8°C] or greater) and a cough and/or a sore throat." Prior to the 2021 season, the definition included "without a known cause other than influenza" [12]. Because other illnesses such as COVID-19, RSV, and the common cold may induce similar respiratory symptoms, ILI may include patients infected with some disease other than influenza. To know whether or not a sick patient is infected with influenza would require a laboratory test.

In 2013, when FluSight began, the ILI data was the object of the forecasts. The data was released publicly at HHS region levels, and forecast teams were asked to provide forecasts of several ILI targets on the regional levels including season onset, 1-4 week ahead ILI levels, and the week of peak ILI activity [6, 41]. Currently, the ILI data is collected by the CDC and published on an online portal for viewing at the national, HHS region, census, and state levels [11]. To obtain ILI data, we used the `R` package `cdcfluview` which provides functions for downloading the data [48]. Weekly ILI data from the national, HHS region, and census levels are available from the 1997 flu season until the current season. At the state level, data is available from the 2010 flu season to the current season.

Figure 1 shows the ILI data at the national level for flu seasons 2010 to 2023. For most seasons there are 52 weeks listed, but for the 2010, 2015, and 2021 seasons there are 53 weeks because there were 53 Sundays during those seasons. To better align with the flu behavior, week 1 is set as the first week of August and week 52 or 53 is the last week in July of the following year. For example week 1 of the 2013 season corresponds to the first week of August 2013, and week 52 of the same season corresponds to the last week of July 2014. This convention is used for the remainder of this manuscript.

Notable from the plots in figure 1 is the regular trajectory of the ILI. With the exception of season 2020, the ILI begins low at week 1 and increases as the fall and winter progress until the ILI reaches a peak. As spring progresses to summer, the ILI decreases to low values. As [43] point out, there is nearly always either a global or local peak at week 22 which typically corresponds to the week between Christmas and New Year's day. Whether local or global, the ILI holiday peak is generally expected and thought to be due to widespread holiday travel, school closure, or other unique social behavior [14, 16]. The only seasons when there was not a peak at week 22 were season 2022 where the season peak occurred particularly early and season 2020 which was greatly influenced by the COVID-19 pandemic.

Figure 2 shows ILI data from five states and the District of Columbia, locations which received particular attention in [44]. The plots include the ILI data for all seasons from 2010 to 2023 in grey, and the black line is the per week ILI average over seasons. The patterns in the individual states are similar to the national level plots in figure 1 in that the ILI rises in the fall and winter until it peaks and descends as the spring and summer

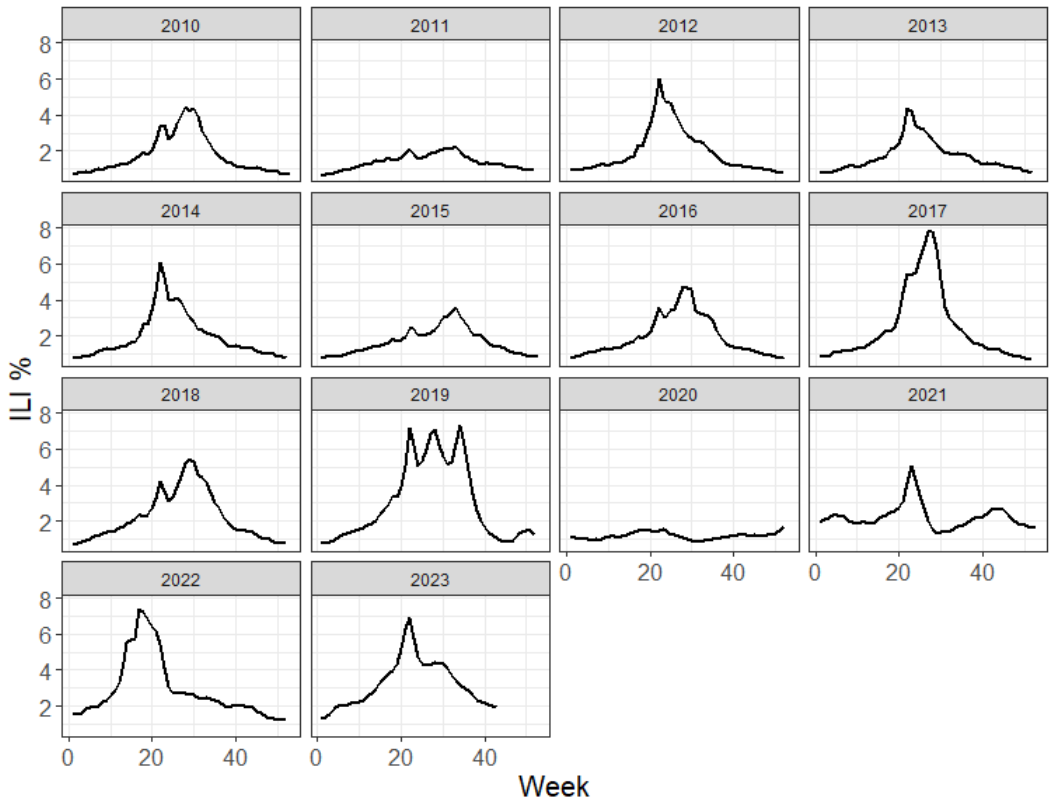

Figure 1: Percentage of outpatient visits with an influenza-like illness (ILI) in the US for seasons 2010 to 2023. Week 1 is the first week of August of the year the flu season begins and the last week of the season is the last week of July of the following year.

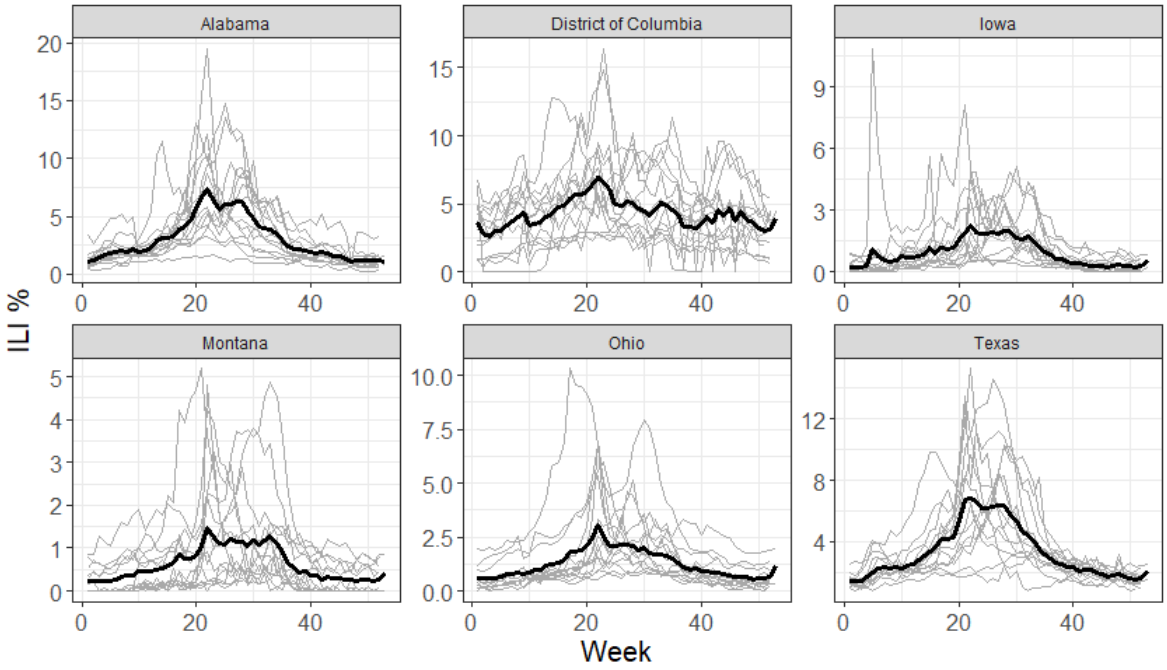

Figure 2: Percentage of outpatient visits with an influenza-like illness (ILI) in five different states and the District of Columbia for seasons 2010 to 2023. Week 1 is the first week of August of the year the flu season begins and the last week of the season is the last week of July of the following year. Plots include lines for ILI% from the 2010 flu season to 2023 (grey) and for the weekly ILI averaged over all seasons (black).

progress. For these locations ILI regularly peaks, either locally or globally, at or near week 22.

2.2 Hospitalization Data

Hospital admission data, used as the object of FluSight forecasting for the 2022 and 2023 seasons, is based on the CDC's National Healthcare Safety Network (NHSN) dataset entitled *HealthData.gov COVID-19 Reported Patient Impact and Hospital Capacity by State Timeseries*. Several targets of respiratory illnesses including COVID-19, RSV, and influenza are reported weekly by most hospitals in the US. In February 2022 it became mandatory for all hospitals to report the number of COVID-19 and influenza hospitalizations, and since then reporting of hospitalizations has become widespread. These data were updated every Wednesday and Friday according to NHSN guidelines [22].

Figure 3 shows the weekly national hospitalizations for the 2022 and 2023 flu seasons.

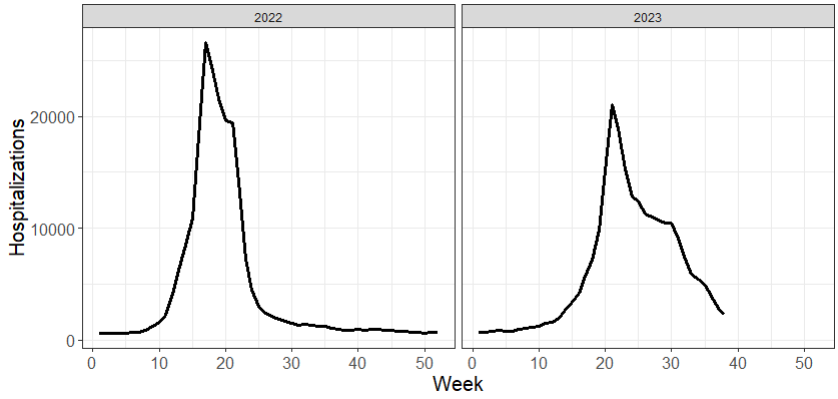

Figure 3: Weekly flu confirmed hospitalization counts at the national level for 2022 (left) and 2023 (right) flu seasons

These plots show similarities to the ILI plots in 1 in that at the early weeks of the season hospitalizations are low, but they increase in the fall to a peak after which they decrease until the flu outbreak ends. For both the 2022 and 2023 seasons, the hospitalizations peaked during the same week as ILI, and in 2023 that peak occurred during the holiday week 22. Figure 4 shows the 2022 and 2023 weekly hospitalizations for the same states from 2. Similar to the national data, the peak in 2022 came early compared to the peak of 2023.

Comparing figures 2 and 4 shows that ILI and hospitalizations share the similar pattern of increasing to a peak in the winter and decreasing thereafter. Figure 5 shows scatter plots with ILI% on the x-axes and hospitalizations on the y-axes, revealing a positive somewhat linear relationship between the two variables. This relationship motivates the forecast models outlined in the next section.

3 ILI and hospitalization forecast modeling

The typical behavior of the ILI data which starts at lower values in the late summer and fall but which increases to a peak, usually in December or January, followed by a decline motivates the use of a nonlinear function which follows of a similar trajectory for modeling ILI. Compartmental models are standard mathematical models used for disease outbreaks. One important compartmental model is the susceptible-infectious-recovered (SIR) model, which is used by some to model ILI data [43, 1]. [53] chose to use the asymmetric Gaussian (ASG) function to model ILI data. The SIR and ASG models may both be appropriate to describe ILI behavior over the course of a flu season, however there may also be systematic behavior not captured by either, necessitating an additional model component to capture the discrepancy. In the first part of this section, we present an ILI model similar to the model in [43]. With some generalization, the model may incorporate any appropriate nonlinear function, though the focus here is on

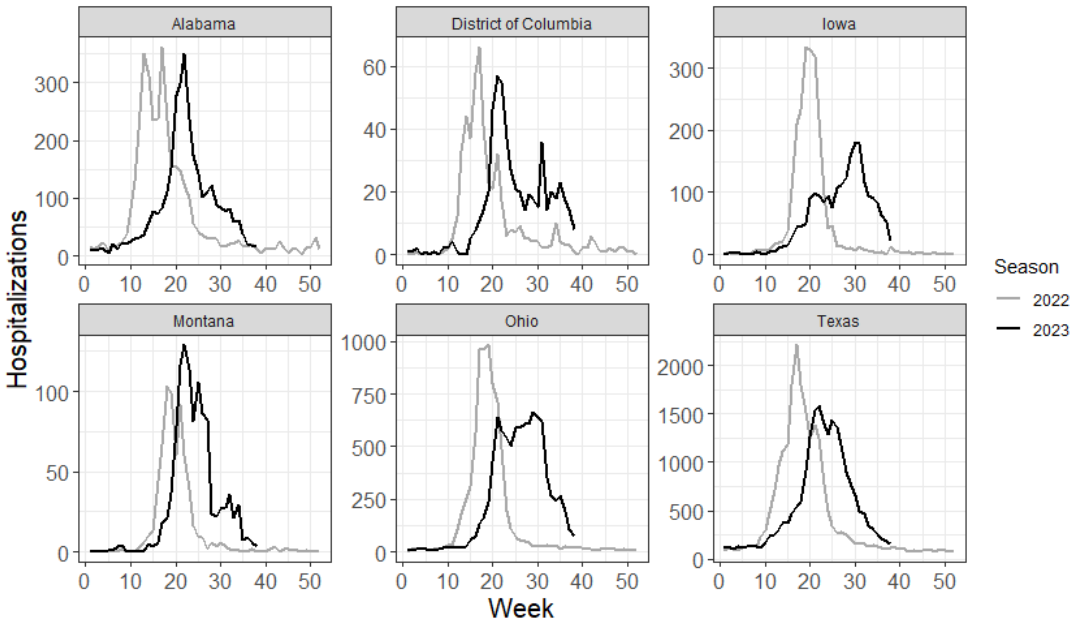

Figure 4: Weekly hospitalization counts for five states and DC for the 2022 (grey) and 2023 (black) flu seasons

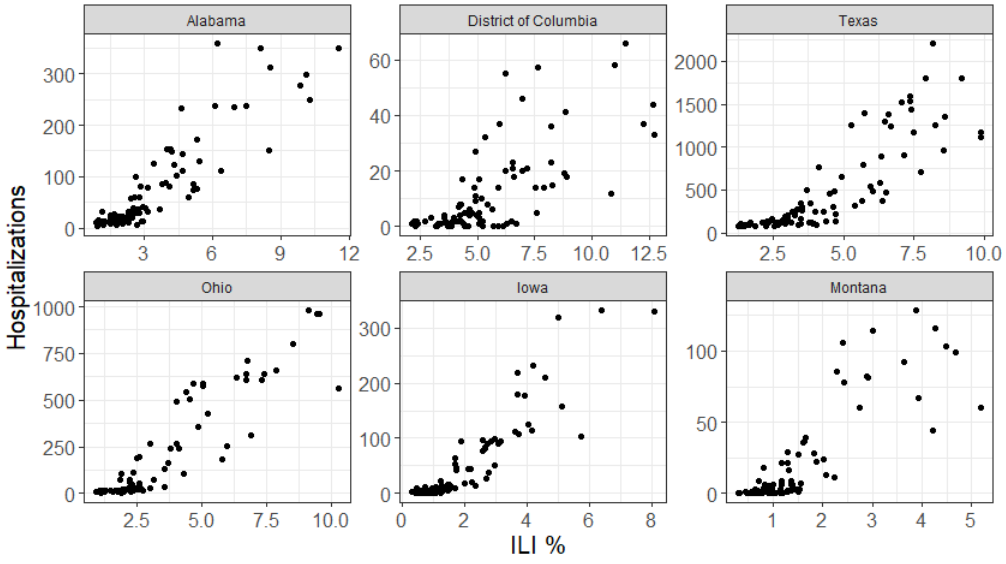

Figure 5: Weekly flu confirmed hospitalization counts (y-axis) for five states and the District of Columbia for the 2022 and 2023 flu seasons plotted against ILI% (x-axis)

the SIR and ASG functions which are defined.

With the aim of forecasting hospitalizations, we also introduce a linear model of hospitalizations data with ILI as a predictive covariate. To forecast hospitalizations, ILI data is first forecasted and the forecast is then plugged in as a covariate in the hospitalization model, thus producing hospitalization forecasts. The hospitalization model is also defined in this section, and the section is concluded with descriptions of selected prior distributions, model implementation, and posterior sampling.

3.1 ILI Model

The proposed model for ILI for any location is given in (1). Here $ILI_{s,w}$ is the ILI for flu season s and week $w = 1, 2, ..., W$, where $W = 52$ or $W = 53$, depending on how many Sundays there are in a given season. The ILI is a proportion, so the Beta random variable is a natural selection for modeling. Under the parameterization in of the Beta distribution used in (1) the expected value is $\pi_{s,w}$ and the variance is $\pi_{s,w}(1 - \pi_{s,w})/(1 + \kappa_s)$, making κ_s a scale parameter. The nonlinear function $f_{\theta_s}(w)$ captures the trajectory of the ILI, and γ_w is a discrepancy term for capturing the systematic patterns which $f_{\theta_s}(w)$ does not capture.

$$\begin{aligned} ILI_{s,w} &\overset{ind}{\sim} \text{Beta}(\pi_{s,w}\kappa_s,\ \kappa_s(1-\pi_{s,w})) \\ \text{logit}(\pi_{s,w}) &= f_{\theta_s}(w) + \gamma_w \end{aligned} \tag{1}$$

In [43] $f_{\theta_s}(w) = \text{logit}(I_{s,w})$ where $I_{s,w}$ is the infectious compartment of the SIR model from (2) in section 3.2. In [53] $f_{\theta_s}(w) = ASG_\theta(w)$ from (3) in section 3.3. In [53] modeling is done hierarchically over seasons.

3.2 Susceptible-Infectious-Recovered (SIR) compartmental model

The SIR compartmental model is a mathematical model used for modeling disease outbreaks and was introduced by [31]. Since then, compartmental models have become standard for modeling infectious diseases [2], and many extensions have been made and studied [49, 1, 54, for example]. The SIR mathematical model includes three compartments and assumes that at any time $t > 0$ every individual in a closed population belongs to exactly one compartment. The three compartments are susceptible (S), infectious (I), and recovered (R), and their interaction over the course of an outbreak is described by the differential equations in (2).

$$\frac{dS}{dt} = -\beta SI, \quad \frac{dI}{dt} = \beta SI - \delta I, \quad \frac{dR}{dt} = \delta I \tag{2}$$

Here S, I, and R represent the proportion of the population in each compartment such that $S + I + R = 1$ for all t. The trajectory is determined by the disease transmission rate $\beta > 0$ and the recovery rate $\delta > 0$. Respectively, these may be thought of as the expected proportion of susceptible individuals who will be infected by an infectious individual, and the expected rate of recovery to an immune state for a newly infected

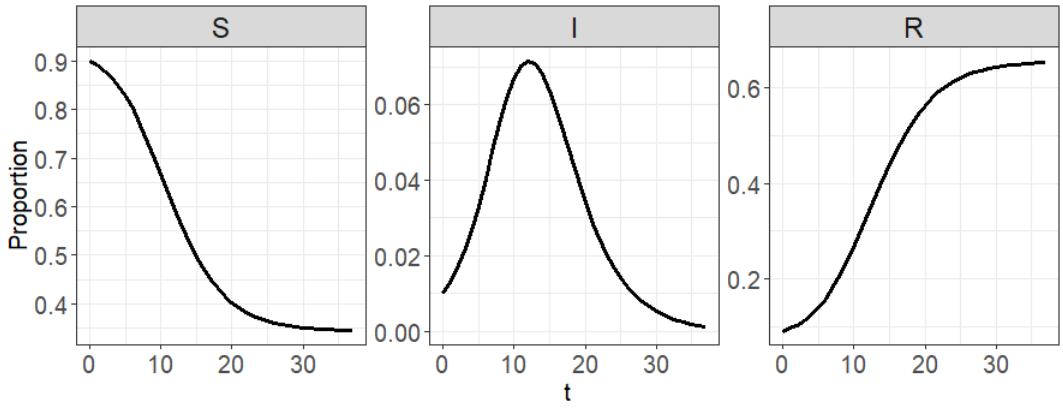

Figure 6: Susceptible-infectious-recovered (SIR) model separated by compartments. The three compartments are the susceptible compartment (left), infectious (center), and recovered (right). In this example, $S_0/\rho > 1$.

person. Whether or not a disease outbreak is classified as an epidemic is determined by the initial susceptible population S_0, or the susceptible population at time 0, and the parameter $\rho = \delta/\beta$. If $S_0/\rho > 1$, the outbreak is considered an epidemic. It is non-epidemic if $S_0/\rho \leq 1$ [43]. Figure 6 shows the trajectory of the three compartments of an SIR model where the S_0 and ρ were selected to match an outbreak that would classify as epidemic. In the case where $S_0 \leq \rho$, the trajectory for the I compartment would never be increasing. The increase to a peak and subsequent decrease in the I compartment of figure 6 suggest it is reasonable to model ILI by this compartment. Thus in modeling the ILI data, we consider the data to be analogous to the I proportion of the population.

3.3 Asymmetric Gaussian (ASG) function

The ASG function is another example of a nonlinear function which can approximate the trajectory of the flu outbreak. The ASG was previously used by Ulloa to model and forecast ILI [53], and it has been used to model vegetation growth and satellite sensor data [33, 26, 23, 5, 4]. The ASG is a modification of the asymmetric Gaussian distribution [56] and is characterized by its rise to a peak and fall from that peak which may not occur at the same rate, as shown in figure 7. The ASG function is denoted as $ASG_\theta(w)$ where $\theta = (\lambda, \nu, \mu, \sigma_1^2, \sigma_2^2)$, $\nu > 0$, $\lambda > 0$, $\mu \in (-\infty, \infty)$, $\sigma_1, \sigma_2 > 0$ and $w \in (1, ..., W)$ is week. The function is defined in (3).

$$ASG_\theta(w) = \begin{cases} \lambda + (\nu - \lambda)\exp[-(w-\mu)^2/2\sigma_1^2], & w < \mu \\ \lambda + (\nu - \lambda)\exp[-(w-\mu)^2/2\sigma_2^2], & w \geq \mu \end{cases} \tag{3}$$

For modeling in this manuscript, we use a slightly reparameterized version of the function in (4), where $\eta = \nu - \lambda > 0$. This constraint guarantees that the function has a

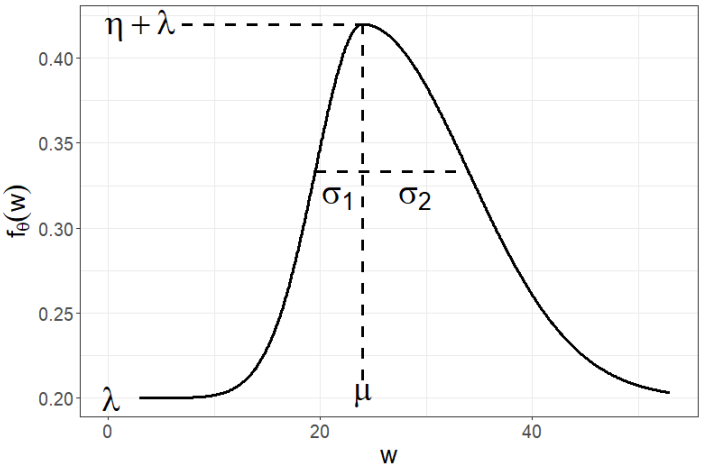

Figure 7: Example plot of asymmetric Gaussian (ASG) function showing the shape of the function in relation to the parameters λ, η, μ, σ_1, and σ_2

peak greater than λ.

$$ASG_\theta(w) = \begin{cases} \lambda + \eta \exp[-(w-\mu)^2/2\sigma_1^2], & w < \mu \\ \lambda + \eta \exp[-(w-\mu)^2/2\sigma_2^2], & w \geq \mu \end{cases} \tag{4}$$

3.4 Model discrepancy

The SIR and ASG functions are useful for capturing the main trend of the ILI data, but as [43] points out there may be systematic behavior that these or other possible functions may not capture. As noted in section 2, figures 1 and 2 show a regular peak at week 22 of the flu season. Figures 8 and 9 are used together to illustrate the systematic discrepancy from a fitted function. Figure 8 shows the US ILI percentage for all flu seasons from 2010 to 2022 excluding 2020 with a best fit ASG function plotted over the ILI. The fits for each season were made by obtaining the maximum likelihood estimate (MLE) of a model given the ILI data where we assume the ASG function is the mean parameter of a Beta distributed random variable. Figure 9 shows the discrepancy between the model fit and the data for the same seasons. The grey lines show the difference between the data and the functions from figure 8 for each season, and the black line is the average by week over all seasons. The lines show that the ASG function typically underpredicts week 22 and overpredicts week 23. Perhaps for other weeks, like week 30 for example, there also tends to be systematic behavior not captured by the ASG function.

The term γ_w, where w is the season week, is included in (1) to capture the per week discrepancy between ILI and the function. Modeling discrepancy has been used in uncertainty analysis of simulators to capture systematic differences between mathe-

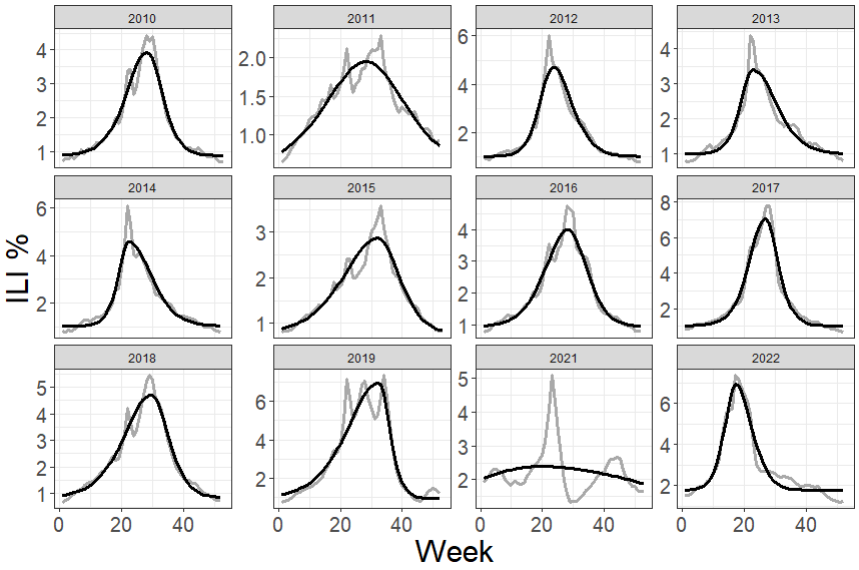

Figure 8: Observed US national influenza-like illness (ILI) percentage for seasons 2010 to 2022 excluding 2020 (grey) overlaid with MLE of an asymmetric Gaussian (ASG) model for the ILI data (black)

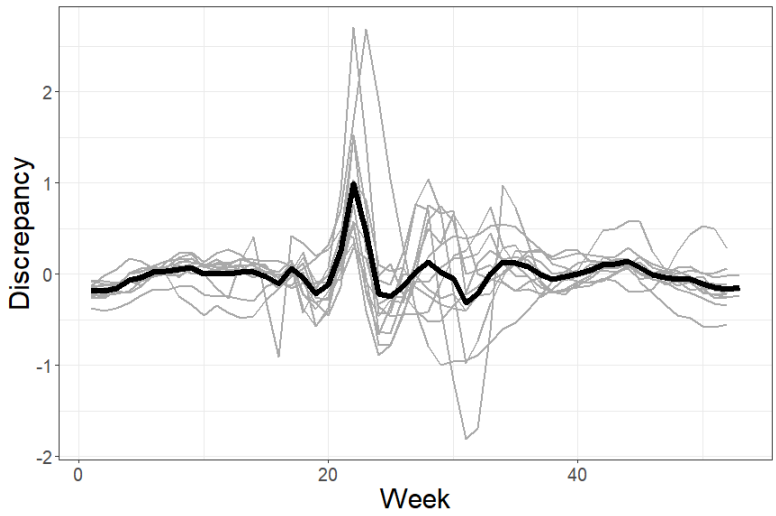

Figure 9: Difference between observed US national influenza-like illness (ILI) and MLE fits for an asymmetric Gaussian (ASG) model for each season 2010 to 2022 excluding 2020 (grey) and the average difference of all seasons (black)

matical models and reality [38, 9, 3, 30]. Modeling discrepancy can lead to overfitting, particularly in forecasting scenarios, and may also lead to identifiability issues. Thus, care must be taken in setting parameter constraints as well as in the selection of prior distributions [43, 9]. Modeling of the discrepancy for ILI was done by [43] during the 2015 and 2016 flu seasons where their model outperformed all others in the CDC flu forecasting challenge [43]. As in Osthus's model, γ_w is modeled as a reverse random walk, as shown in (5).

$$\gamma_w|\gamma_{w+1} \overset{ind}{\sim} N(\gamma_{t+1}, \sigma^2_\gamma), \quad \gamma_W \sim N(0, \sigma^2_{\gamma_W}) \tag{5}$$

The idea for using the reverse random walk is that there are several previous seasons of ILI data, and assuming the random walk captures systematic behavior, fitting it hierarchically over seasons can assist in predicting future behavior in the current season. Reverse random walks have also been used with success in election forecasting and other flu forecasting models [44, 43, 34]. The sum to zero constraint $-\gamma_1 = \sum_{w=2}^{W} \gamma_w$ is imposed on (5) to improve identifiability.

3.5 ILI forecasts

Model (1) is fit via Bayesian posterior updating. Future ILI forecasts are obtained via the posterior predictive distribution where for week w, the predictive distribution is obtained by integrating over the parameters π_s, κ_s, σ^2_γ, and $\sigma^2_{\gamma_W}$ as in (6) where $p(\boldsymbol{\pi}_s, \kappa_s, \sigma^2_\gamma, \sigma^2_{\gamma_W}|\mathbf{ILI})$ is the density function of the posterior distribution for the model parameters. If the current week is w^* then the desired forecasts are for weeks $w^* + i$ where i is a positive integer.

$$p(\widetilde{ILI}_{s,w^*}|\mathbf{ILI}) = \int\int\int\int p(\widetilde{ILI}_{s,w^*}|\boldsymbol{\pi}_s, \kappa_s, \sigma^2_\gamma, \sigma^2_{\gamma_W}) p(\boldsymbol{\pi}_s, \kappa_s, \sigma^2_\gamma, \sigma^2_{\gamma_W}|\mathbf{ILI}) d\boldsymbol{\pi}_s d\kappa_s d\sigma^2_\gamma d\sigma^2_{\gamma_W} \tag{6}$$

3.6 Hospitalization model

The second component for forecast modeling is the hospitalization model defined in (7). This is an example of an autoregressive model with exogenous variables where the autoregressive lag is one (ARX(1)) [45, 36]. Here $H_{s,w}$ is the number of hospitalizations for week w in season s, $\epsilon_{s,w}$ is an error term distributed according to some distribution D_s with mean parameter 0, scale parameter σ_{ϵ_s}, and the additional parameter ω_s is the degrees of freedom parameter when D_s belongs to the location-scale t (LST) family.

$$\begin{aligned} H_{s,w} &= \alpha_{0s} + \alpha_{1s}(ILI_{s,w} \times P) + \alpha_{2s}(ILI_{s,w} \times P)^2 + \phi H_{s,w-1} + \epsilon_{s,w} \\ \epsilon_{s,w} &\overset{iid}{\sim} D_s(0, \sigma^2_{\epsilon_s} \times P, \omega_s) \end{aligned} \tag{7}$$

This model is for any location, and for fitting purposes $ILI_{s,w}$ is always multiplied by P which is proportional to the population of the state or territory, in this case the

total population divided by 50,000. This is done as a means of scaling so that the prior distribution assigned to $\boldsymbol{\alpha}_s = (\alpha_{0s}, \alpha_{1s}, \alpha_{2s})$, σ_{ϵ_s}, and ω_s might reasonably be the same for all states.

Like the ILI model, the hospitalization model in (7) is also fit via Bayesian posterior updating. To obtain forecasts for H_{s,w^*+i}, the ILI posterior predictive distribution is used along with the posterior distribution for the parameters in (7). We considered three scenarios for model (7). A model where D_s belongs to the normal family (NORM), D_s belongs to the LST family, and one where $H_{s,w}$ is replaced with $\log(H_{s,w}+1)$ and D_s is from a normal family, or $H_{s,w}+1$ is lognormally distributed (LNORM). The population value P is excluded from the LNORM model, and we set $\boldsymbol{\alpha}_{22} = \boldsymbol{\alpha}_{23}$ to help with fitting. In the LNORM model, if the linear parameters are not set the same for seasons 22 and 23, the final variance was more prone to be extreme. Besides varying the distribution family of hospitalizations, we also considered $\alpha_{2s} = 0$ or there is no quadratic ILI term.

3.7 Prior selection

The priors selected for the ILI data model under both the SIR and ASG models largely follow the prior selections in [43] and [53] with a few exceptions where changes improved numerical stability and/or we felt the adjusted prior made more sense for the problem. For model (1), parameters which are common even when using different functions of $f_{\theta_s}(w)$ are κ_s, σ^2_γ, and $\sigma^2_{\gamma_W}$. For the SIR function $\theta_s = (S_{0s}, I_{0s}, R_{0s}, \alpha_s, \rho_s)$, and for the ASG function $\theta_s = (\alpha_s, \eta_s, \mu_s, \sigma^2_{1s}, \sigma^2_{2s})$. For the hospitalization model in (7) the parameter to be estimated is $\Psi = (\alpha_{0s}, \alpha_{1s}, \alpha_{2s}, \phi, \sigma_{\epsilon_s}, \omega_s)$.

The priors assigned were mostly noninformative, though in certain cases the prior distributions were selected for numerical stability as was the case for σ^2_γ and $\sigma^2_{\gamma_W}$. For these two scale parameters only, rather than assigning a half-normal prior to the standard deviation parameters, as recommended by [17], the priors were assigned to the variance parameters. Univariate parameters were assigned either a normal distribution prior if the support is on $\mathbb{R}$, a half-normal prior if the support is nonnegative, or a truncated-normal prior to match a more specific support. Under the ASG model, θ_s is modeled hierarchically over seasons so that for each season the transformed parameter $T(\theta_s) \sim N(\theta, \Sigma)$ and priors distributions are assigned to θ and Σ.

Additional prior constraints were made to improve parameter identifiability. In [43] the initial value of the susceptible population compartment of the SIR model was set to $S_0 = 0.9$. The parameters I_{0s}, β_s, and ρ_s were assigned informative priors. To improve identifiability when $f_{\theta_s}(w) = ASG_{\theta_s}(w)$ in (1) we followed a modular Bayesian approach for fitting the parameters. A modular Bayesian approach involves multiple steps of parameter fitting where some parameters may be estimated without priors via maximum likelihood estimation or other means. Fitting the rest of the model parameters involves assigning priors and conditioning on the previously fit parameters. This has been done in computer modeling to improve identifiability and other issues, though [35] warn this approach is not probabilistically sound if parameter inference is a priority [25, 3, 35]. We carried out the modular fit by first estimating the maximum likelihood estimate (MLE) for the parameter λ_s for each season in (4) and plugging in the MLE as a fixed value.

Table 1: Maximum $\hat{R}$ and minimum ESS over all parameters for four ILI models fitted on US data for week 14 of the 2023 season

	ASG	ASGD	SIR	SIRD
$\hat{R}$	< 1.001	< 1.001	< 1.001	1.001
ESS	72,057	9,980	17,745	6,259

Table 2: Maximum $\hat{R}$ and minimum ESS over all parameters for six hospitalization models fitted on US data for week 14 of the 2023 season

	NORM	NORM^2	LNORM	LNORM^2	LST	LST^2
$\hat{R}$	< 1.001	< 1.001	< 1.001	< 1.001	< 1.001	< 1.001
ESS	71,266	73,554	46,747	54,975	7,660	62,663

3.8 Parameter estimation and posterior predictive sampling

The models were fit via Markov chain Monte Carlo (MCMC) sampling using the `cmdstanr` package which was developed and is maintained by the [50] [15]. Stan implements Hamiltonian Monte Carlo (HMC) sampling with the No-U-turn sampler [24]. The `cmdstanr` package provides several diagnostic statistics for assessing the sampler. As mentioned, most model parameter prior distributions were intended to be uninformative. Plots of posterior distributions for select parameters from ILI and hospitalization models are shown in the supplementary material.

We assessed the model fit for four ILI models. These included the SIR and ASG models and models with and without discrepancy modeling. When discrepancy is included, the models are denoted as SIRD and ASGD. These models were fit using US national data from 2010 to 2023 flu seasons, where data from the 2020 season was excluded because of the unique behavior during that season. Assessment for hospitalization modeling was done for six different models. These include NORM, LNORM, and LST models, and models where a quadratic ILI term is included or excluded. To assess posterior sampling convergence, models were fit to data where ILI and hospitalization data up to week 14 of the 2023 season was included. Sampling was done with four chains where from each chain 60,000 posterior draws were sampled, and the first 10,000 draws were discarded as a burn-in. The $\hat{R}$ statistic [55] and the effective sample size (ESS) [18] were calculated for each parameter. Tables 1 and 2 summarize the maximum $\hat{R}$ and the minimum ESS over all parameters for ILI and hospitalization models respectively. Forecast models for all other weeks of the season were fit using one chain of 60,000 draws where the first 10,000 draws were discarded as a burn-in. For parameters of the ASG models that were prone to cause trouble in posterior sampling, the starting values were set to be the MLEs.

To obtain forecast distributions of hospitalizations, draws from the posterior predictive distribution from the ILI model were used in conjunction with the posterior distribution of the hospitalizations model. When fitting model (6), MCMC samples of

$\widetilde{ILI}_{s,w:(w+4)}$ were saved. Model (7) was fit and MCMC samples for the marginal distributions for the model parameters were saved. To obtain forecast distributions for $H_{s,w+i}$ where $i \in \{1,2,3,4\}$, the following steps are repeated K times where K is an integer for the number of desired samples. We set $K = 50,000$.

Step 1: Sample $\widetilde{ILI}^*_{s,w:(w+4)}$

Step 2: Sample α^*_{0s}, α^*_{1s}, α^*_{2s}, ϕ^*, $\sigma^*_{\epsilon_s}$, ω^*_s from respective marginal posterior distributions

Step 3: Sample $H^*_{s,w+i}$ from $D(\omega^*_s, \mu^*_{s,w+i}, \sigma^2_{\epsilon_s})$, where

$$\mu^*_{s,w+i} = \alpha^*_{0s} + \alpha^*_{1s}(ILI^*_{s,w+i} \times P) + \alpha^*_{2s}(ILI^*_{s,w+i} \times P)^2 + \phi^* H^*_{s,w+i-1}$$

Step 4: Repeat step 3 for $i \in \{1,2,3,4\}$ to obtain $H^*_{s,(w+1):(w+4)}$

Step 5: Repeat steps 1-4 K times

The sample $\{H^*_{s,w+i}\}^K$ was then used as the probabilistic forecast for hospitalizations at week $w+i$. For the forecast competition analysis in section 5, all negative values of $\{H^*_{s,w+i}\}^K$ were set to 0 to reflect realistic values of hospitalizations and comply with the FluSight forecasting rules.

4 Simulation Study

In this section, we present a simulation study conducted for comparing ILI models and further assessing the hospitalization forecast model. US ILI data is used, and hospitalization data is simulated. A leave-one-season-out (LOSO) approach was combined with a Monte Carlo simulation approach. For each replication, we simulated log-hospitalizations for all weeks during seasons 2010, ..., 2022, excluding 2020, using the existing ILI data as a predictive covariate. Each season was in turn "left-out" and treated as if it was the most recent season which we desired to forecast. Fitting and forecasting was then done for weeks 14, 20, 26, 32, and 38 of the left out season, giving two weeks that tend to occur as flu cases increase, two as cases decrease, and one that occurs when the cases may be increasing or decreasing. Week 20 is a week leading up to the holiday week 22 where ILI typically has a local peak. We were particularly interested in how important modeling discrepancy is for forecasting at week 20.

For the simulation of hospitalizations, the parameters $\boldsymbol{\alpha}_s = (\alpha_{0s}, \alpha_{1s}, \alpha_{2s})$ and $\sigma^2_{\epsilon_s}$ from the hospitalization model in (7) were considered the same across all seasons so that all $\boldsymbol{\alpha}_s = \boldsymbol{\alpha}$. The values for $\boldsymbol{\alpha}$, σ^2_ϵ, and ϕ were estimated by fitting model (7) using ILI and hospitalization data from the 2022 season. For fitting, the hospitalization data was first log-transformed. We took $\boldsymbol{\alpha}_\phi = (\boldsymbol{\alpha}, \phi)$ and assigned the noninformative prior $p(\boldsymbol{\alpha}_\phi, \sigma^2_\epsilon) \propto 1/\sigma^2_\epsilon$. The marginal posterior distribution $\boldsymbol{\alpha}_\phi | \sigma^2_\epsilon, \boldsymbol{H_{22}}$ was then the established posterior multivariate normal distribution and $\sigma^2_\epsilon | \boldsymbol{H_{22}}$ the inverse-χ^2 posterior distribution [18]. The posterior means of those parameters were used as the values from which log-hospitalizations were simulated. The number of replicates in the simulation was 500.

Model comparison was done by calculating the continuous ranked probability score (CRPS) and the logarithmic score (LogS) for each forecast. These scores are both proper

scoring rules which evaluate the forecast distribution and density functions respectively. Proper scoring rules are the current standard for comparing performance between probabilistic forecasts and selecting the best forecasts according to the notion of maximizing sharpness subject to (auto-)calibration [19, 51]. Proper scoring rules are commonly used in forecast comparison and have the property that a forecaster is incentivized to be honest in the reporting of their forecasts [21, 20]. The CRPS is defined in (8) and the LogS in (9). Here $F(\cdot)$ is the forecast distribution function, $f(\cdot)$ is the forecast density function, and y^* is the observed targeted value of the forecast. The orientation of both scores is negative, meaning the smaller the score the better.

$$\text{CRPS}(F, y^*) = \int_{-\infty}^{\infty} (F(x) - \mathbb{1}(y^* \leq x))^2 dx \tag{8}$$

$$\text{LogS}(f, y^*) = -\log(f(y^*)) \tag{9}$$

Both the CRPS and LogS are calculated using the `scoringRules` package in `R` [27]. To calculate the LogS when given MCMC samples from a posterior predictive distribution, a continuous density function was first estimated via kernel density estimation. The CRPS is calculated using a quantile decomposition from [32].

Figures 10 - 14 show boxplots of the CRPS and LogS for the four models. Figure 10 shows that the variation of overall CRPS is smallest for the ASG models and larger for the SIR models. The median scores for the SIR models also appears slightly higher than for the ASG models. When faceted by season in figure 11, the boxplots of the CRPS often show the same pattern for ASG and SIR models but not always. For example, the bulk of CRPS values for the SIRD model in 2019 appears to have smaller variation than the other models. Figure 12 shows CRPS boxplots faceted by week and horizon. Notable plots here are for week 20 where the two models accounting for the ILI discrepancy, ASGD and SIRD, have CRPS values with lower medians and lower variation than the two models not accounting for ILI discrepancy. This makes sense given the forecasts at week 20 are forecasting weeks, 22 and 23 where figure 9 shows a seasonal peak and trough.

Figures 13 and 14 show boxplots of the LogS for the simulated forecasts. Note that in these plots only values of less than 17 are included to improve visualization. The LogS plots show similar results to the CRPS, though there are some differences. For example, SIRD in figure 13 tends to show smaller LogS variation relative to the other three models than is seen by the CRPS of SIRD in figure 11. We also note the smaller relative LogS variation than CRPS variation for SIRD when comparing figure 14 with figure 12, particularly at week 14. Among the four models, ASG tends to have the lowest values in CRPS and LogS though this is not always the case. When forecasting weeks 21-23, the two models which model discrepancy in the ILI, ASGD and SIRD, tend to outperform the models not modeling the discrepancy. This confirms that there is value in modeling discrepancy at least during around the holiday weeks near week 22.

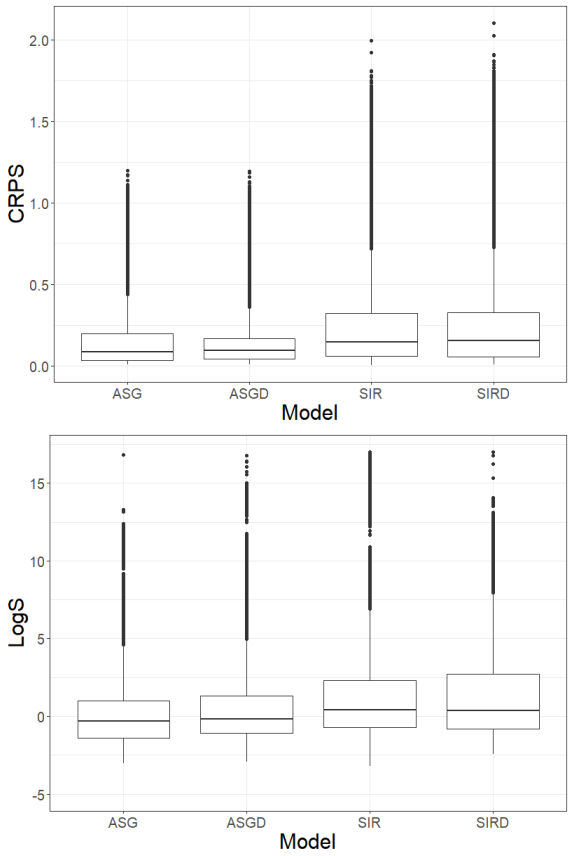

Figure 10: Boxplots of the continuous ranked probability score (CRPS) (left) and logarithmic score (LogS) (right) for the four ILI models over all seasons, weeks, and horizons in the simulation study

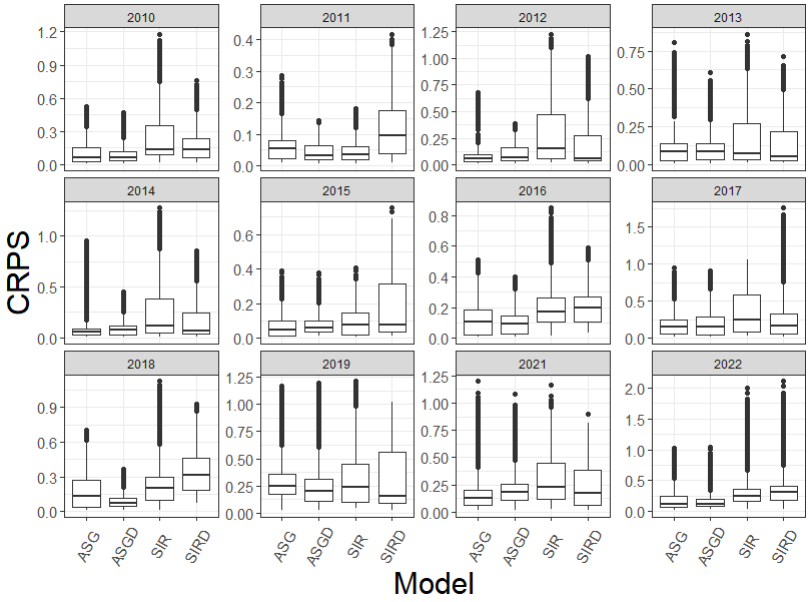

Figure 11: Boxplots of continuous ranked probability score (CRPS) for the four ILI models over all weeks and horizons in the simulation study faceted by season and including seasons 2010-2022, excluding 2020

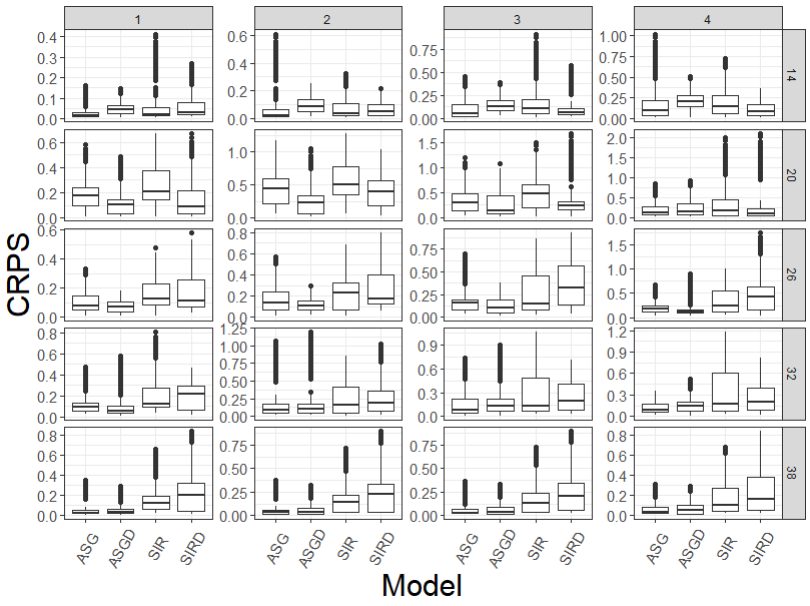

Figure 12: Boxplots of continuous ranked probability score (CRPS) for the four ILI models over all seasons in the simulation study faceted by horizon (x-axis) and week (y-axis). Horizons include 1-4 week ahead forecastss and weeks include weeks 14, 20, 26, 32, and 38 of the flu season

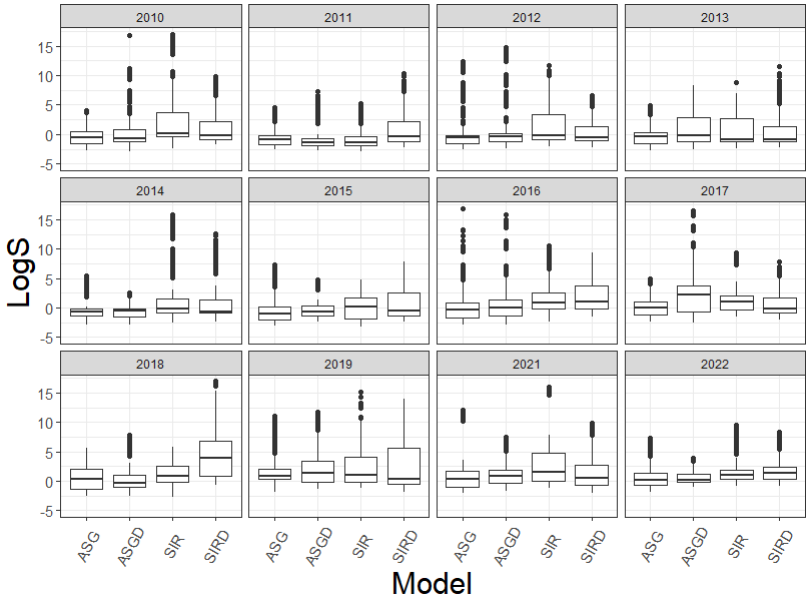

Figure 13: Boxplots of logarithmic score (LogS) for the four ILI models over all weeks and horizons in the simulation study faceted by season and including seasons 2010-2022, excluding 2020.

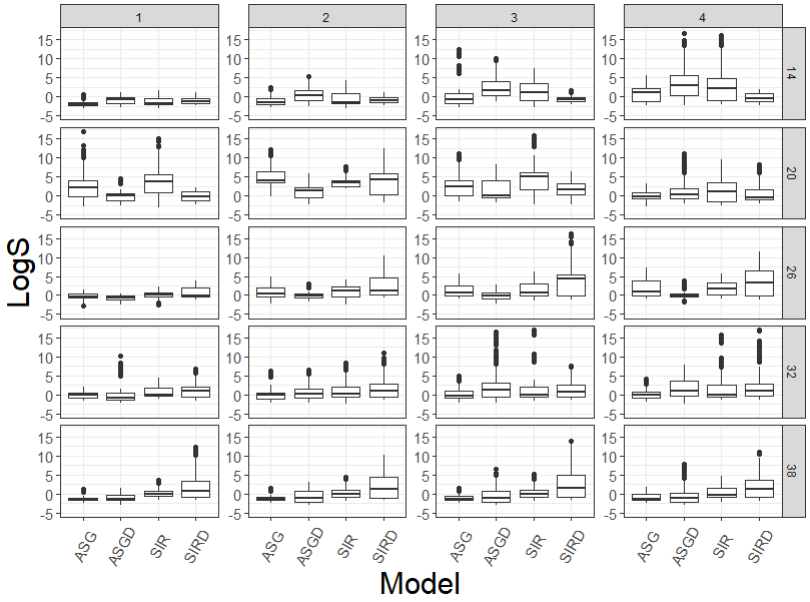

Figure 14: Boxplots of logarithmic (LogS) for the four ILI models over all seasons in the simulation study faceted by horizon (x-axis) and week (y-axis). Horizons include 1-4 week ahead forecasts and weeks include weeks 14, 20, 26, 32, and 38 of the flu season.

5 Analysis of forecasts for 2023 flu season

In this section we apply the forecast models to make forecasts of the 2023 flu season weekly hospitalizations. The scoring of the forecasts is in the context of the FluSight competition where each competing forecast was submitted as a set of 23 quantiles corresponding to the given probability levels $(0.010, 0.025, 0.050, 0.100, 0.150, \ldots, 0.950, 0.975, 0.990)$. A single forecast is thus comprised of 11 predictive intervals and a median. Forecasts of 1, 2, 3, and 4-week ahead hospitalization counts were requested, and forecasts were made at the state and national level. The first week of forecasting took place during the week of October 7, 2023, and the final week was the week of April 27, 2024 making 29 total weeks of forecasts. The same format was used during the 2021 and 2022 seasons and for the COVID-19 Forecast Hub [39, 8]. Primary scores for evaluating each forecast were the weighted interval score (WIS), the log-weighted interval score (LWIS), and the relative weighted interval score (RWIS). The focus in this section will be on the LWIS (see supplementary material for RWIS based results).

The WIS is a proper scoring rule used for scoring quantile or interval forecasts [21, 20, 8] and is defined in (10) where Q is a forecast represented by all included quantiles, B is the number of intervals, y^* is the observed value targeted by the forecast, $w_0 = 1/2$ and $w_b = \alpha_b/2$ are weights for each interval, and α_b is the nominal level of the b^{th} interval. IS_α is the interval score (IS), a proper scoring rule for a single interval as defined in (11). The goal of the forecaster is to minimize the WIS. The LWIS is the same as the WIS except that it is evaluated over the log of quantiles and the log of the observed value.

$$WIS_{0,B}(Q, y^*) = \frac{1}{B + 1/2} \times (w_0 \times |y^* - median| + \sum_{b=1}^{B} \{w_k \times IS_{\alpha_b}(Q, y^*)\}) \tag{10}$$

$$IS_\alpha(l, r; y^*) = (r - l) + \frac{2}{\alpha}(l - y^*)\mathbb{1}\{y^* < l\} + \frac{2}{\alpha}(y^* - r)\mathbb{1}\{y^* > r\} \tag{11}$$

We fit 24 forecast models for each location for all 29 weeks, and for each week forecast 1-4 week ahead hospitalizations. The 24 models included all combinations of ASG, ASGD, SIR, and SIRD ILI models, the NORM, LNORM, and LST hospitalization models, and both quadratic and linear hospitalization models. The prior distributions under the SIR model are in (12). Because we set $S_{0s} = 0.9$, prior distributions were assigned only to I_{0s}, β_s and ρ_s, recalling the parameter for the recovery rate $\delta_s = \rho_s\beta_s$. Here $\mathbb{1}_A$ represents the indicator function for values within the set A.

$$\begin{aligned} I_{0s} &\sim N(0.005, 0.03)\mathbb{1}_{(0,.1)} \\ \beta_s &\sim N^+(0.8, 0.3) \\ \rho_s &\sim N^+(0.68, 0.08) \end{aligned} \tag{12}$$

For the ASG model, the MLE $\hat{\theta}_s = (\hat{\lambda}_s, \hat{\eta}_s, \hat{\mu}_s, \hat{\sigma}^2_{1s}, \hat{\sigma}^2_{2s})$ was calculated and $\hat{\lambda}_s$ was accepted as a fixed value. The remaining estimates were used as starting values for posterior sampling. The transformation $T(\theta_s) = (\log(\eta_s), \mu_s, \log(\sigma^2_{1s}), \log(\sigma^2_{2s}))$ was made,

and a prior distribution was assigned to $T(\theta_s)$. The prior distributions for the ILI model under the ASG function are shown in (13). These are slightly informative priors because for most parameters we have an idea what reasonable values may be. Here $m = (0.3, 23, 3.69, 4.7)$ and $C = \text{diag}(0.2, 5, 2, 2)$ where $\text{diag}(\cdot)$ is the diagonal matrix for the given entries.

$$\begin{aligned} T(\theta_s) &\overset{ind}{\sim} MVN(\theta, \Sigma) \\ T(\theta) &\overset{ind}{\sim} MVN(m, C) \\ \Sigma &= \text{diag}(\zeta_1^2, ..., \zeta_4^2) \\ \zeta_i &\overset{ind}{\sim} N^+(0, 4^2) \end{aligned} \tag{13}$$

Parameters shared by both the SIR and ASG models are the scale parameter κ_s and the discrepancy parameters σ_γ^2 and $\sigma_{\gamma_W}^2$. These priors are in (14).

$$\begin{aligned} \kappa_s &\overset{ind}{\sim} N^+(0, 10,000^2) \\ \sigma_\gamma^2 &\sim N^+(0, .02^2) \\ \sigma_{\gamma_W}^2 &\sim N^+(\hat{\sigma}_W^2, 1^2) \end{aligned} \tag{14}$$

Because of the limited information for estimating $\sigma_{\gamma_W}^2$, we selected an informative prior distribution by first estimating $\hat{\sigma}_{\gamma_W}^2$. This was estimated for a given state by first calculating the MLE for θ_s. $\widehat{ILI}_{s,w}$ was then predicted such that $\text{logit}(\widehat{ILI}_{s,W}) = f_{\hat{\theta}_s}(W)$ for each season. Then $\hat{\sigma}_{\gamma_W}^2$ was calculated as the estimated variance over seasons of the differences $\text{logit}(\widehat{ILI}_{s,w}) - \text{logit}(ILI_{s,w})$. When $f_{\theta_s}(w)$ was the SIR function, $\widehat{\theta}_s$ was calculated using the `mle2` function in the `bblme` package [7] and the `ode` function in the `deSolve` package [29] function in `R`. Where the ASG function was used, $\hat{\theta}_s$ was calculated using the `optim` function. The prior for the hospitalization models are in (15).

$$\begin{aligned} \alpha_{0s} &\overset{ind}{\sim} N(0, 5^2) \\ \alpha_{1s} &\overset{ind}{\sim} N(0, 5^2) \\ \alpha_{2s} &\overset{ind}{\sim} N(0, 5^2) \\ \phi &\sim N(0, .4)\mathbb{1}_{(-1,1)} \\ \sigma_{\epsilon_s} &\overset{ind}{\sim} N^+(0, 4^2) \\ \omega_s &\overset{ind}{\sim} N^+(0, 15^2) \end{aligned} \tag{15}$$

Figure 15 shows 1-4-week ahead forecasts for US flu hospitalizations during the 2023 season under the NORM hospitalization model. The forecasts shown are for weeks 14, 20, 26, and 32. The predictive bands are the 50% and 95% predictive intervals. These plots show a tendency to often underpredict hospitalizations. The quadratic models appear to do a better job predicting hospitalizations at the season peak but a poorer

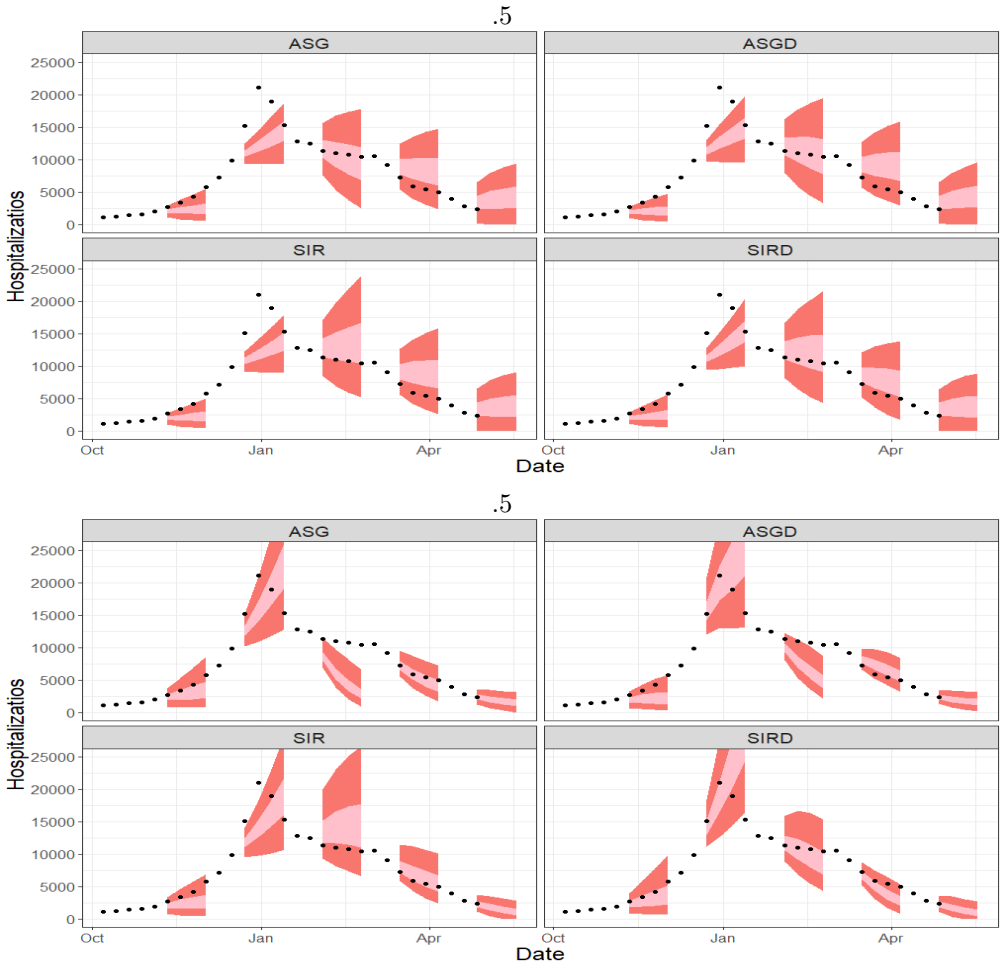

Figure 15: Forecasts 1-4 weeks ahead for US hospitalizations during the 2023 season for weeks 14, 20, 26, and 32. Forecasts are separated by ILI model, and the hospitalization models are all normally distribution. The figure includes hospitalization forecasts where ILI is a linear predictor (left) and where ILI is a quadratic predictor (right). 50% predictive intervals are pink and 95% predictive intervals are red.

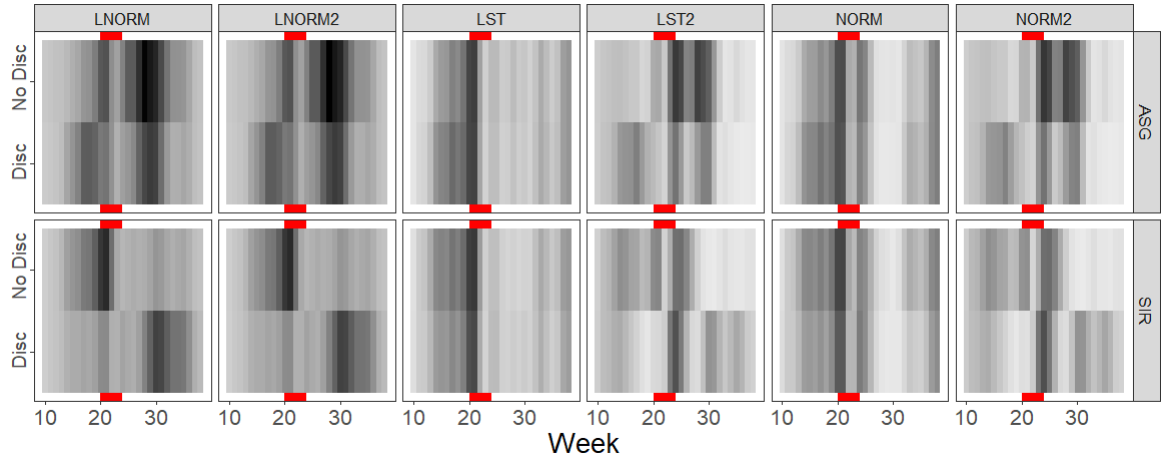

Figure 16: Each plot shows the log weighted interval score (LWIS) for every week of the 2023 flu season with scores for models including discrepancy in the ILI model (top) and excluding discrepancy (bottom). Scores are separated by hospitalization distribution family and by ILI as a linear or quadratic predictor. Scores for models with an ASG ILI model are above while those with an SIR model are below. The lighter the shade, the lower the LWIS with low LWIS being better.

job predicting after the peak, whereas the linear models seem to predict well or slightly overpredict after the peak.

Figure 16 shows model performance by LWIS for each week of the season for all 24 models of US hospitalizations. LWIS scores are grouped by ILI model, hospitalization model distribution, and by the linear or quadratic modeling. Here, a darker value represents a larger LWIS and worse forecast compared to lighter values. The weeks around the holiday week 22 are highlighted by a red band. In most cases, there appears to be a turning point in performance at or near week 22. The models which include discrepancy appear to forecast better around week 22, though they may not outperform the models without discrepancy through the whole season. Such is the case for the NORM and LST quadratic models and the linear LNORM models. It's also notable that for the SIR model, including discrepancy appears to make for poorer forecasts after week 22 whereas for ASG, including discrepancy appears to improve forecasts.

Figures 17 and 18 also show weekly forecast performance by LWIS, but all 53 locations are included. Overall, the weeks leading up to week 22 are the most difficult to forecast. In figure 17, it appears that both the SIR and ASG models which include discrepancy perform slightly better than the models which do not. Figure 18 shows the LWIS across the whole season faceted by hospitalization model distribution. The LNORM model appears to perform the best during the weeks leading up to week 22. The season overall score, calculated as the mean LWIS over all locations and weeks, for all 24 models is shown in table 3. The first main takeaway is that the top four

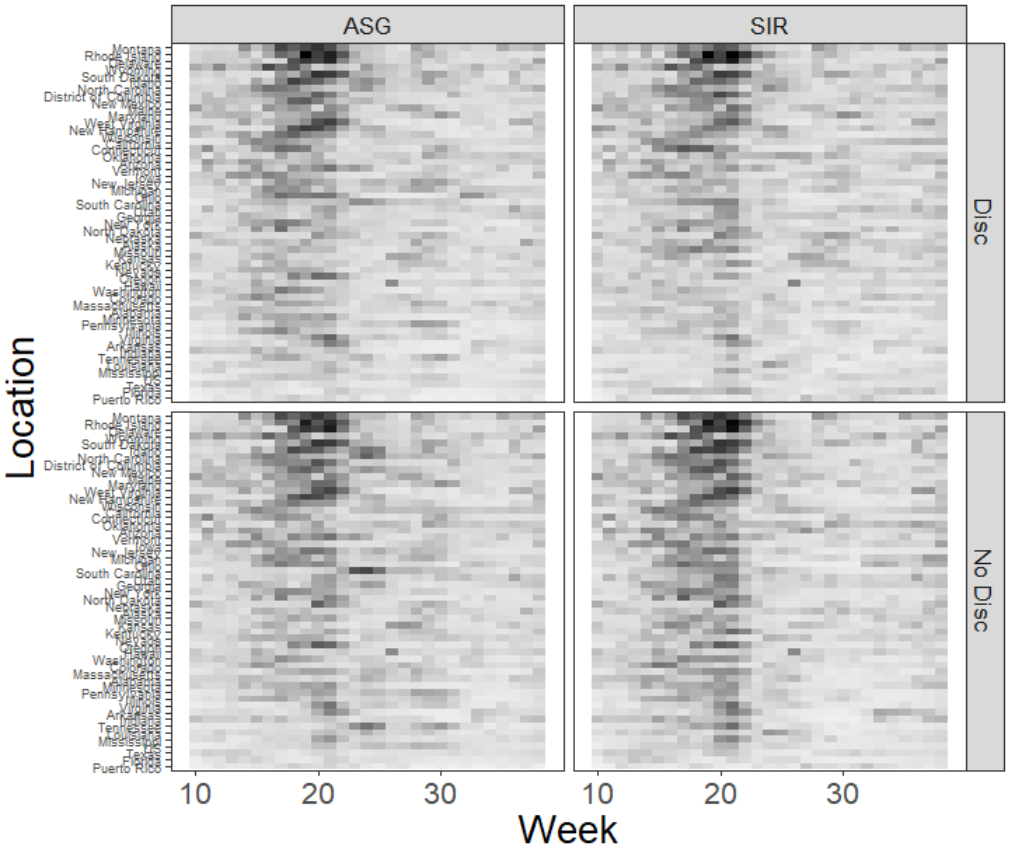

Figure 17: Each plot shows the log weighted interval scores (LWIS) for all 50 US states, PR, DC, and national level forecasts at each week during the 2023 flu season. Scores are averaged over all horizons 1-4 weeks ahead. Scores are faceted by ILI model function (columns) and by whether or not discrepancy modeling was included (rows). The lighter the shade, the lower the LWIS with low LWIS being better.

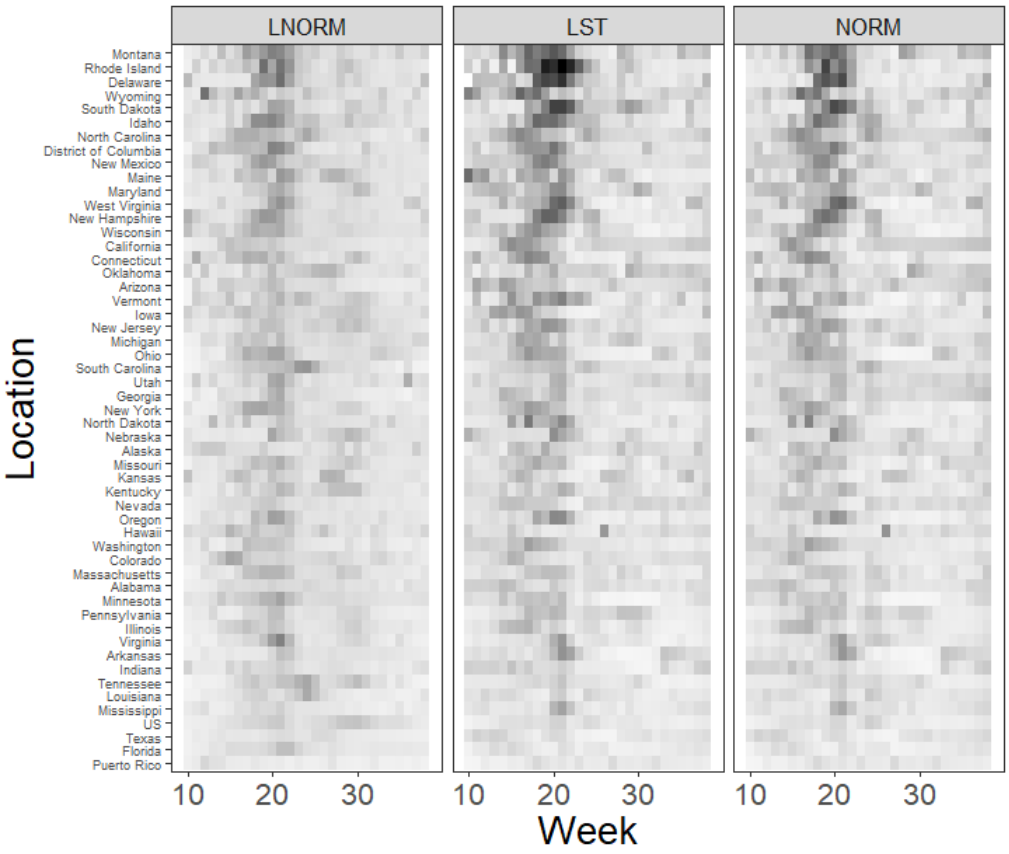

Figure 18: Each plot shows the log weighted interval scores (LWIS) for all 50 US states, PR, DC, and national level forecasts at each week during the 2023 flu season. Scores are averaged over all horizons 1-4 weeks ahead. Scores are faceted by hospitalization model distribution. The lighter the shade, the lower the LWIS with low LWIS being better.

Table 3: Overall scores for each of the 24 forecast models. The overall score is the log weighted interval score (LWIS) averaged over all locations, weeks, and horizons. The scores in the first two rows are for linear models, and the scores in the third and fourth rows are for quadratic models. The lowest WISs are bolded.

		ASG			SIR		
		LNORM	LST	NORM	LNORM	LST	NORM
Linear	No Disc	0.366	0.392	0.365	**0.355**	0.416	0.394
	Disc	0.358	0.391	0.365	**0.343**	0.382	0.354
Quadratic	No Disc	0.366	0.398	0.377	**0.355**	0.401	0.377
	Disc	0.358	0.390	0.369	**0.343**	0.378	0.356

performing models are SIR ILI models with the LNORM hospitalization models. The next main takeaway is that overall the models which include discrepancy outperformed the models without discrepancy.

It should not be assumed that these two takeaways will apply for future flu seasons. As suggested by the simulation study in the previous section, it may often be the case that the ASG ILI model is better for forecasting. Indeed, a close examination of figures 16, 17, and 18 shows that it is common for the ASG models to show better forecasting skill than the SIR models.

6 Conclusion

In this manuscript we introduce a statistical modeling framework which allows for the incorporation of several ILI forecast modeling methods. Specifically, we built upon [43] and introduced a framework for modeling ILI which includes the use of an arbitrary function for modeling the main trajectory of ILI along with modeling the discrepancy. We model flu hospitalizations by incorporating the ILI forecast model into a model forecasting hospitalizations where hospitalization predictions are a linear or quadratic function of ILI.

The simulation study in section 4 suggests the ASG function in ILI modeling may slightly outperform the SIR model according to the LogS and CRPS scoring rules, but in the analysis of the 2023 flu season forecasts, the SIR model was overall superior. The results from both the simulation study and the real data analysis suggest that the addition of a discrepancy component in ILI modeling may improve forecasts especially near the holiday week between Christmas and New Years day. It should not be assumed that these conclusions may be generalized for all locations in the US, weeks of a season, or for future flu seasons.

Forecasting the seasonal influenza outbreak remains a challenging task for forecasters. The general modeling framework in this manuscript is successful under diverse modeling conditions for all locations in the US and may contribute to future forecasting efforts. All forecast models were presented as separate from one another, but in the various locations and times of the flu season, different models perform better than others.

To build on the work done here, a natural step forward would be to combine all or a few selected forecasts into an ensemble forecast. Such an ensemble may work to cancel out certain model biases or highlight model strengths, leading to more robust forecasts.

Acknowledgments

This work is partially supported by the National Science Foundation under Grant No. 2152117. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

Supplementary Material

Posterior distribution plots for select parameters. Plots of posterior distributions for parameters in both ILI and hospitalization models

FluSight forecast competition scoring results. Flu forecast results with relative weighted interval scoring rule similar to that used in FluSight

References

[1] Allen, L. J. (2017). "A primer on stochastic epidemic models: Formulation, numerical simulation, and analysis." *Infectious Disease Modelling*, 2(2): 128–142. 7, 9

[2] Allen, L. J., Brauer, F., Van den Driessche, P., and Wu, J. (2008). *Mathematical Epidemiology*, volume 1945. Springer. 9

[3] Arendt, P. D., Apley, D. W., Chen, W., Lamb, D., and Gorsich, D. (2012). "Improving Identifiability in Model Calibration Using Multiple Responses." *Journal of Mechanical Design*, 134(10): 100909. 13, 14

[4] Atkinson, P. M., Jeganathan, C., Dash, J., and Atzberger, C. (2012). "Inter-comparison of four models for smoothing satellite sensor time-series data to estimate vegetation phenology." *Remote Sensing of Environment*, 123: 400–417. 10

[5] Beck, P. S., Atzberger, C., Høgda, K. A., Johansen, B., and Skidmore, A. K. (2006). "Improved monitoring of vegetation dynamics at very high latitudes: A new method using MODIS NDVI." *Remote Sensing of Environment*, 100(3): 321–334. 10

[6] Biggerstaff, M., Alper, D., Dredze, M., Fox, S., Fung, I. C.-H., Hickmann, K. S., Lewis, B., Rosenfeld, R., Shaman, J., Tsou, M.-H., et al. (2016). "Results from the Centers for Disease Control and Prevention's predict the 2013–2014 Influenza Season Challenge." *BMC Infectious Diseases*, 16(1): 1–10. 2, 4

[7] Bolker, B. and R Development Core Team (2023). *bbmle: Tools for General Maximum Likelihood Estimation*. R package version 1.0.25.1.
URL https://CRAN.R-project.org/package=bbmle 24

[8] Bracher, J., Ray, E. L., Gneiting, T., and Reich, N. G. (2021). "Evaluating epidemic

forecasts in an interval format." *PLOS Computational Biology*, 17(2): e1010592. 2, 23

[9] Brynjarsdóttir, J. and O'Hagan, A. (2014). "Learning about physical parameters: The importance of model discrepancy." *Inverse Problems*, 30(11): 114007. 13

[10] CDC (2024). "Centers for Disease Control and Prevention FluSight: About Flu Forecasting." https://www.cdc.gov/flu-forecasting/about/index.html?CDC_AAref_Val=https://www.cdc.gov/flu/weekly/flusight/how-flu-forecasting.htm. Accessed: 2024-10-22. 2

[11] — (2024). "Centers for Disease Control and Prevention FluView Portal." https://gis.cdc.gov/grasp/fluview/fluportaldashboard.html. Accessed: 2024-10-22. 2, 4

[12] — (2024). "Centers for Disease Control and Prevention FluView, U.S. Influenza Surveillance: Purpose and Methods." https://www.cdc.gov/fluview/overview/?CDC_AAref_Val=https://www.cdc.gov/flu/weekly/overview.htm. Accessed: 2024-10-22. 2, 4

[13] Cramer, E. Y., Huang, Y., Wang, Y., Ray, E. L., Cornell, M., Bracher, J., Brennen, A., Castro Rivadeneira, A. J., Gerding, A., House, K., Jayawardena, D., Kanji, A. H., Khandelwal, A., Le, K., Niemi, J., Stark, A., Shah, A., Wattanachit, N., Zorn, M. W., Reich, N. G., and Consortium, U. C.-. F. H. (2022). "The United States COVID-19 forecast hub dataset." *Scientific Data*, 9(1): 462. URL https://doi.org/10.1038/s41597-022-01517-w 2

[14] Ewing, A., Lee, E. C., Viboud, C., and Bansal, S. (2017). "Contact, travel, and transmission: the impact of winter holidays on influenza dynamics in the United States." *The Journal of Infectious Diseases*, 215(5): 732–739. 4

[15] Gabry, J., Češnovar, R., and Johnson, A. (2022). *cmdstanr: R Interface to 'CmdStan'*. Https://mc-stan.org/cmdstanr/, https://discourse.mc-stan.org. 15

[16] Garza, R. C., Basurto-Dávila, R., Ortega-Sanchez, I. R., Carlino, L. O., 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 Diseases*, 19(6): 938. 4

[17] Gelman, A. (2006). "Prior distributions for variance parameters in hierarchical models." *Bayesian Analysis*, 1(3): 515–533. 14

[18] Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., and Rubin, D. B. (2013). *Bayesian Data Analysis, Third Edition*. Chapman and Hall/CRC. 15, 16

[19] Gneiting, T., Balabdaoui, F., and Raftery, A. E. (2007). "Probabilistic forecasts, calibration and sharpness." *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 69(2): 243–268. 17

[20] Gneiting, T. and Katzfuss, M. (2014). "Probabilistic forecasting." *Annual Review of Statistics and Its Application*, 1: 125–151. 2, 17, 23

[21] Gneiting, T. and Raftery, A. E. (2007). "Strictly proper scoring rules, prediction, and estimation." *Journal of the American Statistical Association*, 102(477): 359–378. 3, 17, 23

[22] HealthData.gov (2024). "COVID-19 Reported Patient Impact and Hospital Capacity by State (RAW)." https://healthdata.gov/dataset/COVID-19-Reported-Patient-Impact-and-Hospital-Capa/6xf2-c3ie/about_data. Accessed: 2024-10-22. 2, 6

[23] Hird, J. N. and McDermid, G. J. (2009). "Noise reduction of NDVI time series: An empirical comparison of selected techniques." *Remote Sensing of Environment*, 113(1): 248–258. 10

[24] Hoffman, M. D., Gelman, A., et al. (2014). "The No-U-Turn sampler: adaptively setting path lengths in Hamiltonian Monte Carlo." *Journal of Machine Learning Research*, 15(1): 1593–1623. 15

[25] Jiang, Z., Apley, D. W., and Chen, W. (2015). "Surrogate preposterior analyses for predicting and enhancing identifiability in model calibration." *International Journal for Uncertainty Quantification*, 5(4). 14

[26] Jonsson, P. and Eklundh, L. (2002). "Seasonality extraction by function fitting to time-series of satellite sensor data." *IEEE ransactions on Geoscience and Remote Sensing*, 40(8): 1824–1832. 10

[27] Jordan, A., Krüger, F., and Lerch, S. (2019). "Evaluating Probabilistic Forecasts with scoringRules." *Journal of Statistical Software*, 90(12): 1–37. 17

[28] Joslyn, S. L. and LeClerc, J. E. (2012). "Uncertainty forecasts improve weather-related decisions and attenuate the effects of forecast error." *Journal of Experimental Psychology: Applied*, 18(1): 126. 2

[29] Karline Soetaert, Thomas Petzoldt, and R. Woodrow Setzer (2010). "Solving Differential Equations in R: Package deSolve." *Journal of Statistical Software*, 33(9): 1–25. 24

[30] 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. 3, 13

[31] Kermack, W. O. and McKendrick, A. G. (1927). "A contribution to the mathematical theory of epidemics." *Proceedings of the Royal Society of London. Series A, Containing papers of a mathematical and physical character*, 115(772): 700–721. 9

[32] Laio, F. and Tamea, S. (2007). "Verification tools for probabilistic forecasts of continuous hydrological variables." *Hydrology and Earth System Sciences*, 11(4): 1267–1277. 17

[33] Lewis-Beck, C., Walker, V. A., Niemi, J., Caragea, P., and Hornbuckle, B. K. (2020). "Extracting agronomic information from SMOS vegetation optical depth in the US Corn Belt using a nonlinear hierarchical model." *Remote Sensing*, 12(5): 827. 10

[34] Linzer, D. A. (2013). "Dynamic Bayesian forecasting of presidential elections in the states." *Journal of the American Statistical Association*, 108(501): 124–134. 13

[35] Liu, F., Bayarri, M., Berger, J., et al. (2009). "Modularization in Bayesian analysis, with emphasis on analysis of computer models." *Bayesian Analysis*, 4(1): 119–150. 14

[36] Ljung, L. (1987). *System Identification: Theory for the User*. Englewood Cliffs, NJ: Prentice-Hall. 13

[37] Lutz, C. S., Huynh, M. P., Schroeder, M., Anyatonwu, S., Dahlgren, F. S., Danyluk, G., Fernandez, D., Greene, S. K., Kipshidze, N., Liu, L., et al. (2019). "Applying infectious disease forecasting to public health: a path forward using influenza forecasting examples." *BMC Public Health*, 19(1): 1–12. 2

[38] Ma, P., Karagiannis, G., Konomi, B. A., Asher, T. G., Toro, G. R., and Cox, A. T. (2022). "Multifidelity computer model emulation with high-dimensional output: An application to storm surge." *Journal of the Royal Statistical Society Series C: Applied Statistics*, 71(4): 861–883. 13

[39] Mathis, S. M., Webber, A. E., León, T. M., Murray, E. L., Sun, M., White, L. A., Brooks, L. C., Green, A., Hu, A. J., Rosenfeld, R., et al. (2024). "Evaluation of FluSight influenza forecasting in the 2021–22 and 2022–23 seasons with a new target laboratory-confirmed influenza hospitalizations." *Nature Communications*, 15(1): 6289. 2, 23

[40] McAndrew, T. and Reich, N. G. (2021). "Adaptively stacking ensembles for influenza forecasting." *Statistics in Medicine*, 40(30): 6931–6952. 2

[41] McGowan, C. J., Biggerstaff, M., Johansson, M., Apfeldorf, K. M., Ben-Nun, M., Brooks, L., Convertino, M., Erraguntla, M., Farrow, D. C., Freeze, J., et al. (2019). "Collaborative efforts to forecast seasonal influenza in the United States, 2015–2016." *Scientific Reports*, 9(1): 683. 4

[42] 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. 1

[43] Osthus, D., Gattiker, J., Priedhorsky, R., and Del Valle, S. Y. (2019). "Dynamic Bayesian influenza forecasting in the United States with hierarchical discrepancy (with discussion)." *Bayesian Analysis*, 14(1): 261–312. 1, 2, 3, 4, 7, 9, 10, 11, 13, 14, 29

[44] Osthus, D. and Moran, K. R. (2021). "Multiscale influenza forecasting." *Nature Communications*, 12(1): 2991. 2, 4, 13

[45] Raftery, A. E., Kárnỳ, M., and Ettler, P. (2010). "Online prediction under model uncertainty via dynamic model averaging: Application to a cold rolling mill." *Technometrics*, 52(1): 52–66. 13

[46] Ramos, M. H., Van Andel, S. J., and Pappenberger, F. (2013). "Do probabilistic

forecasts lead to better decisions?" *Hydrology and Earth System Sciences*, 17(6): 2219–2232. 2

[47] Ray, E. L., Wattanachit, N., Niemi, J., Kanji, A. H., House, K., Cramer, E. Y., Bracher, J., Zheng, A., Yamana, T. K., Xiong, X., et al. (2020). "Ensemble forecasts of coronavirus disease 2019 (COVID-19) in the US." *medRxiv*, 2020–08. 2

[48] Rudis, B. (2021). *cdcfluview: Retrieve Flu Season Data from the United States Centers for Disease Control and Prevention ('CDC') 'FluView' Portal*. R package version 0.9.4.
URL https://CRAN.R-project.org/package=cdcfluview 4

[49] Simon, C. M. (2020). "The SIR dynamic model of infectious disease transmission and its analogy with chemical kinetics." *PeerJ Physical Chemistry*, 2: e14. 9

[50] Stan Development Team (2024). "Stan Modeling Language Users Guide and Reference Manual, 2.34." https://mc-stan.org. Accessed: 2024-10-22. 15

[51] Tsyplakov, A. (2013). "Evaluation of probabilistic forecasts: proper scoring rules and moments." *Available at SSRN 2236605*. 17

[52] Turtle, J., Riley, P., Ben-Nun, M., and Riley, S. (2021). "Accurate influenza forecasts using type-specific incidence data for small geographic units." *PLOS Computational Biology*, 17(7): e1009230. 2

[53] Ulloa, N. (2019). "Bayesian hierarchical modeling for disease outbreaks." Ph.D. thesis, Iowa State University Department of Statistics. 1, 2, 3, 7, 9, 10, 14

[54] Van den Driessche, P. (2008). "Deterministic compartmental models: extensions of basic models." In *Mathematical Epidemiology*, 147–157. Springer. 9

[55] Vehtari, A., Gelman, A., Simpson, D., Carpenter, B., and Bürkner, P.-C. (2021). "Rank-normalization, folding, and localization: An improved $\hat{R}$ for assessing convergence of MCMC (with discussion)." *Bayesian Analysis*, 16(2): 667–718. 15

[56] Wallis, K. F. (2014). "The two-piece normal, binormal, or double Gaussian distribution: its origin and rediscoveries." *Statistical Science*, 106–112. 10

[57] WHO (2024). "World Health Organization website Influenza (Seasonal) fact sheet." https://www.who.int/news-room/fact-sheets/detail/influenza-(seasonal). Accessed: 2024-10-22. 1

[58] Winkler, R. L. (1971). "Probabilistic prediction: Some experimental results." *Journal of the American Statistical Association*, 66(336): 675–685. 2

[59] Yamana, T. K., Kandula, S., and Shaman, J. (2016). "Superensemble forecasts of Dengue outbreaks." *Journal of The Royal Society Interface*, 13(123): 20160410. 2