Flusion: Integrating multiple data sources for accurate influenza predictions

Evan L. Ray, possible additions include Nicholas G. Reich, Russ Wolfinger, Yijin Wang

July 2, 2024

Abstract

Over the last ten years, the US Centers for Disease Control and Prevention (CDC) has organized an annual influenza forecasting challenge with the motivation that accurate probabilistic forecasts could improve situational awareness and yield more effective public health actions. Starting with the 2021/22 influenza season, the forecasting targets for this challenge have been based on hospital admissions reported in the CDC's National Healthcare Safety Network (NHSN) surveillance system. Reporting of hospital admissions through NHSN began during the COVID pandemic, and as such NHSN has only a limited amount of historical data about influenza hospitalizations. To produce forecasts in the presence of limited data for the target surveillance system, we augmented these data with two signals that have a longer historical record: 1) ILI+, which estimates the proportion of outpatient doctor visits where the patient has influenza; and 2) rates of laboratory-confirmed influenza hospitalizations at a set of sentinel healthcare facilities. Our model, Flusion, is an ensemble model that combines two machine learning models using gradient boosting for quantile regression based on different feature sets with a Bayesian autoregressive model. The gradient boosting models were trained on all three data signals, while the autoregressive model was trained on only data for the target surveillance signal, NHSN admissions; all three models were trained jointly on data for multiple locations. In each week of the influenza season, these models produced quantiles of a predictive distribution for influenza hospital admissions in each state in the current week and the following three weeks; the ensemble prediction was computed by averaging these quantile predictions. Flusion emerged as the top-performing model in the CDC's influenza prediction challenge for the 2023/24 season. In this article we investigate the factors contributing to flusion's success, and we find that its strong performance was primarily driven by the use of a gradient boosting model that was trained jointly on data from multiple surveillance signals and multiple locations. These results indicate the value of sharing information across multiple locations and surveillance signals, especially when doing so enlargens the pool of available training data.

1 Introduction

There is a long history of forecasting seasonal influenza, often through formal collaborative forecasting exercises organized by the US Centers for Disease Control and Prevention (CDC).

- Long history going back 10 years
- After a pause during the first two years of the COVID-19 pandemic, this forecasting exercise restarted in the 2022/23 flu season.

A challenge in recent forecasting exercises is that new data streams have come online which are being used as the "ground truth" target for these forecasting exercises. These data sources have been great in many ways, but the lack of an extensive history of data for model training introduces difficulty for learning about seasonal patterns in flu. To address this challenge, this season we developed a new model, called 'flusion', which pulls in data from external data sources with a long history of observations. This model was the top-performing model that was contributed to the FluSight Forecast Hub in the 2023/24 influenza season. The purpose of this paper is to describe that model and investigate what aspects of its design were associated with its strong performance.

Relevant literature to review, 1 paragraph each:

- methods for infectious disease forecasting, including methods that integrate multiple data sources.
  - David Farrow cmu thesis on data fusion
  - maybe papers using surveillance signals like Quidel?
  - maybe some covid papers using things like mobility data
- forecasting with gradient boosting.
  - Russ's paper
  - M5 results paper

The remainder of this article is structured as follows...

2 Data Sources

Our model uses three measures of influenza activity (Figure 1). The first of these is weekly influenza hospital admissions reported to the National Healthcare Safety Network (NHSN), which is the target signal for the FluSight forecasting exercise (see section 3 for more detail). Reporting of hospital admissions through NHSN began in 2020, and the first two years of reporting for this signal show very low influenza activity during the COVID-19 pandemic. As a result, at the start of the 2023/24 season, this signal had only one season's worth of data showing patterns typical of seasonal influenza.

To address this limitation of the target surveillance signal, our model uses two other measures of influenza activity that have a longer reporting history. The first of these is a measure of hospital admissions where the patient has a positive influenza test as reported by FluSurv-NET, expressed as a rate per 100,000 population in the catchment areas of selected sentinel hospital facilities (TODO: cite). The other signal is ILI+, which is an approximate measure of the proportion of outpatient doctor visits where the patient has influenza that is derived by combining data from ILINet and WHO/NREVSS (TODO: cite, cite).

For both of these additional signals, we included adjustments designed to correct for known challenges with interpreting the data collected by FluSurv-NET and ILINet as consistent measures of influenza activity over time. The FluSurv-NET data report on patients with a positive influenza test, and as such are subject to varying levels of underreporting depending on changing testing rates and test sensitivity and specificity from season to season. The CDC produces annual estimates of hospital burden due to influenza that adjust for these factors (TODO: cite). We used these total burden estimates to estimate season-specific inflation factors that scale up reported rates from FluSurv-NET, with the intent of

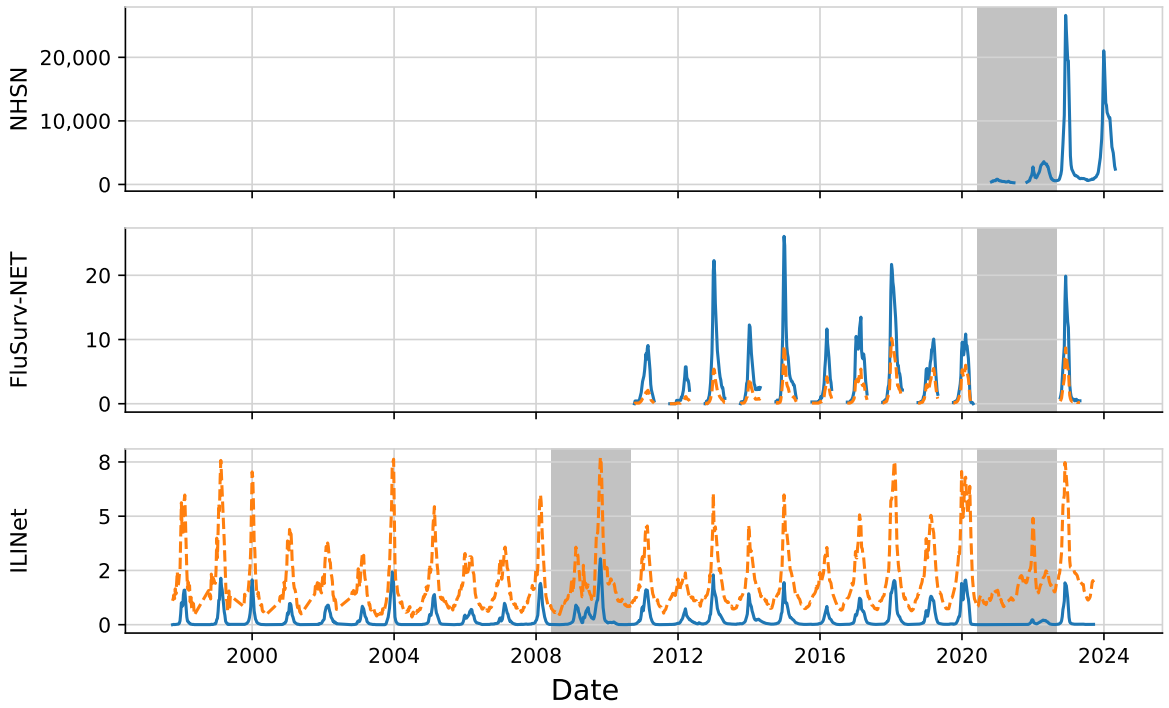

Figure 1: Influenza data at the national level in the US. The top panel shows weekly hospital admissions from NHSN, including the 2023/24 season that was the target season for predictions described in this article. In the second panel, a dashed orange line shows raw data reported from FluSurv-NET; the modeled data, in blue, are obtained by scaling up the raw data using per-season inflation factors designed to account for varing testing rates and test sensitivity/specificity. In the third panel, a dashed orange line shows raw ILI data from ILINet; the modeled data in blue are ILI+ values obtained by combining ILI with test positivity rates. Grey shaded regions indicate pandemic seasons that were not used for model training; these include the 2008/09 and 2009/10 seasons which were impacted by pandemic swine influenza, and the 2020/21 and 2021/22 seasons which were impacted by low influenza activity during the COVID pandemic. Additionally, FluSurv-NET and ILI+ data for the 2023/24 season were not used for model training in this work.

producing a more consistent measure of influenza activity over time; see supplemental section TODO for more details. These inflation factors were generally larger in earlier seasons than in later seasons, indicating that FluSurv-NET undercounted influenza activity more in early seasons than it did in later seasons.

ILINet reports a measure of influenza-like illness (ILI), as the percent of outpatient doctor visits where the patient has symptoms consistent with influenza without another known cause. Because ILI is defined symptomatically, this signal generally includes some patients who have respiratory diseases other than influenza such as RSV and COVID. To address this, we compute ILI+ as the product of this ILI signal and influenza test positivity rates from laboratory testing sites reporting to World Health Organization (WHO) and National Respiratory Enteric Virus Surveillance System (NREVSS) systems. After converting to a proportion scale, ILI+ can be interpreted as an estimate of the proportion of outpatient doctor visits where the patient has influenza, and has been used in previous forecasting work as a more specific measure of influenza activity than ILI [e.g. 2, 3].

3 The FluSight collaborative forecasting exercise

For the 2023/24 season, the primary forecast target collected in the FluSight forecasting exercise was weekly hospital admissions with confirmed influenza as reported in the NHSN data set for the 50 US states, the District of Columbia, Puerto Rico, and in total at the national level (Figure 1, top panel). Predictions for a second target representing a categorical measure of the direction and magnitude of change in admissions were also collected by the hub, but we did not make predictions for that target with the Flusion model.

Predictions were submitted to the FluSight hub on Wednesday each week. The week of submission is anchored relative to the Saturday after the submission date, which corresponds to the final day of that US epidemic week and is denoted as the *reference date* for the predictions. Predictions were made for hospital admissions from Sunday to Saturday in the current week and each of the three following weeks, corresponding to forecast horizons of 0, 1, 2, and 3 weeks ahead relative to the week of submission. Initially, the hub also collected predictions at a horizon of -1, representing a "hindcast" of admissions in the week before submission, but these hindcasts were discontinued a few weeks into the season and we do not analyze them here. A new data release from NHSN was made public at approximately noon on Wednesday each week including reported admissions up through the previous Saturday, and predictions were based on models fit to that data release. This data release was delayed by one day on the week ending on April 13, 2024, and for that week the forecast submission due date was extended so that the latest data could be used.

For this hospital admissions target, probabilistic predictions were represented with a set of predictive quantiles at $K = 23$ quantile levels α_k, corresponding to a predictive median and the endpoints of 11 central prediction intervals at the nominal 10%, ..., 90%, 95%, and 98% levels. For example, the predictive median corresponds to $\alpha_{12} = 0.5$ and the 98% interval corresponds to predictive quantiles at the levels $\alpha_1 = 0.01$ and $\alpha_{23} = 0.99$.

4 Notation and evaluation metrics

We use $z_{s,l,t}$ to denote the observed value of surveillance signal s (with $s = 1$ for NHSN data, $s = 2$ for FluSurv data, and $s = 3$ for ILI+ data) in location l in the week ending on date t. We denote the predictive quantile at level α_k for the value of the signal s in location l, generated on reference date d at forecast horizon h by $q_{s,l,d,h,k}$. Note that the observation at time t is the prediction target across the four combinations of reference date d and forecast horizon h with $d + h = t$. For brevity, we will often use the index i to denote a forecast task consisting of a combination of values of $s = s(i)$, $l = l(i)$, $d = d(i)$, and $h = h(i)$, with $q_{i,k}$ denoting the prediction at quantile level α_k for that task and z_i denoting the corresponding observed value.

In this manuscript, we evaluate forecasts using three metrics: the mean absolute error (MAE) of the predictive median, the mean weighted interval score (MWIS), and one-sided quantile coverage rates. Our evaluations of forecast skill cover only predictions of NHSN admissions ($s = 1$).

MAE measures the average distance between the predictive median and the eventual observation, with smaller values indicating better forecast accuracy. Recalling that the predictive median corresponds to the quantile level $\alpha_{12} = 0.5$ (i.e., $k = 12$), the absolute error of the prediction for task i is $|q_{i,12} - z_i|$. In our evaluations, the mean absolute error averages the absolute error across predictions for different

locations, dates, and horizons.

WIS can be viewed as a generalization of the absolute error to a set of quantile predictions, and is equivalent to an average of quantile scores (sometimes referred to as pinball losses) computed for each quantile prediction:

$$WIS(\{q_{i,k} : k = 1, \ldots, K\}, z_i) = \frac{1}{K} \sum_k 2 \cdot QS_{\alpha_k}(q_{i,k}, z_i)$$
$$QS_{\alpha_k}(q_{i,k}, z_i) = \alpha_k \max(z_i - q_{i,k}, 0) + (1 - \alpha_k) \max(q_{i,k} - z_i, 0)$$

The quantile score for a single quantile prediction assigns an asymmetric penalty to the distance between the prediction and the observation, with the magnitudes of the penalties for underprediction and overprediction set so that in expectation, the quantile score is minimized by the quantile of the predictive distribution at a specified quantile level. The quantile score for the quantile level 0.5 is equal to one half of the absolute error of the median. Again, lower values of WIS indicate a better alignment of forecasts with the observed data. MWIS is the average WIS across multiple locations, dates, and horizons.

We also compute relative versions of MAE and MWIS, denoted rMAE and rMWIS, using the "pairwise tournament" approach outlined in [1]. The primary purpose of this procedure is to correct for the varying level of difficulty of the predictions submitted by different forecasters in settings where some forecasters did not provide predictions for all locations or time points. This is relevant to the evaluation of submissions to the FluSight hub in section 6, but in the experimental results in section 7 all forecasts were provided. A secondary goal is to standarize scores relative to a baseline model with known behavior, in our case the flat baseline described in section 6. Smaller values of rMAE or rWIS value indicate better performance relative to the other models in the comparison pool, and in particular values less than 1 indicate performance better than the baseline.

In a comparison of forecast accuracy among M models, computation of rMAE (rWIS) has two steps. First, we obtain a summary of the average performance for model m relative to each other model m', denoted θ^m. This summary is computed as the geometric mean of the ratio of the MAE (MWIS) for model m to the MAE (MWIS) for each other model m', where for each model pair the MAE averages across the set $\mathcal{I}_{m,m'}$ of locations, dates, and forecast horizons for which both models submitted predictions. The rMAE (or rWIS) then normalizes this geometric mean relative to the value of θ^m for a baseline model:

$$rMAE^m = \frac{\theta^m}{\theta^{baseline}}, \qquad \text{where } \theta^m = \left(\prod_{m' \neq m} \frac{MAE^m_{\mathcal{I}_{m,m'}}}{MAE^{m'}_{\mathcal{I}_{m,m'}}} \right)^{1/(M-1)}$$

Here, $MAE^m_{\mathcal{I}}$ denotes the MAE for model m across all predictions for tasks i in the index set $\mathcal{I}$.

Finally, we examine probabilistic calibration through the difference between the empirical one-sided coverage rate at each quantile level and the nominal coverage rate:

$$\delta_k = \frac{1}{|\mathcal{I}|} \sum_{i \in \mathcal{I}} \mathbb{1}(z_i \leq q_i) - \alpha_k,$$

where $|\mathcal{I}|$ is the size of the index set $\mathcal{I}$ and $\mathbb{1}(z_i \leq q_i)$ is 1 if $z_i \leq q_i$ and 0 otherwise. A well-calibrated forecaster will have $\delta_i \approx 0$, while a conservative forecaster with wide prediction intervals will have

$\delta_k > 0$ for $\alpha_k > 0.5$ and $\delta_k < 0$ for $\alpha_k < 0.5$.

5 Model

Our model was constructed as an ensemble of statistical and machine learning time series models. All models are fit to data that have been preprocessed to standardize across different signals, locations, and time points, and we describe these preprocessing steps in section 5.1. We describe the component models in sections 5.2 and 5.3 and the ensemble methods in section 5.4. The precise formulation of the models included in the ensemble and the ensembling methods varied slightly over the course of the season, and we describe these aspects of our setup for generating real-time predictions in section 5.5.

5.1 Data standarization

Our models are fit to preprocessed versions of the surveillance data, with transformations designed to put the data on a similar scale for different surveillance signals and across different locations. Starting with the observed value $z_{s,l,t}$ of a particular surveillance signal s in location l at time t, we apply the following operations to compute the transformed version of the signal, $\tilde{z}_{s,l,t}$:

1. For NHSN admissions, we divide by the population of the location l in units of 100,000 people to convert to a hospital admissions rate per 100,000 population, which is comparable across locations of different sizes. Note that the ILI+ and FluSurv signals are naturally expressed as rates or percents and so the magnitudes of those signals do not depend on population size.

2. We take a fourth root transformation to stabilize the variance of the signal across times of low and high influenza activity.

3. We scale by dividing by the 95th percentile of all observations for each location and data source, and center by subtracting the mean for each location and data source. These transformations adjust for varying magnitudes of the surveillance signals for different data sources and locations.

The resulting transformed data used as an input to the models are shown in Figure 2.

5.2 Component Models 1 and 2: GBQR and GBQR-no_level

TODO: update GBM to GBQR everywhere.

The first two models use gradient boosting for quantile regression, which we abbreviate as GBQR. Specifically, we use GBQR to obtain separate quantile regression fits for each required quantile level α_k, using the average quantile loss QS_{α_k} as the learning objective. These models learn a mapping from features x_i to the specified quantile of a predictive distribution for the target y_i, where i indexes combinations of data source $s(i)$, location $l(i)$, reference date $d(i)$, and forecast horizon $h(i)$. At each quantile level, the final prediction is obtained using bagging, by taking the median of predictive quantiles from 100 separate fits that are each based on a randomly selected 70% of the seasons in the training set (including partial data for the current season). The models are fit using the LightGBM package in Python (TODO: cite) with default settings for all hyperparameters.

The models are trained jointly on data for all data sources, locations, dates, and forecast horizons. However, the features x_i contain information only about influenza activity for the particular data source $s(i)$ and location $l(i)$. Inclusion of multiple locations and data sources in the training data set

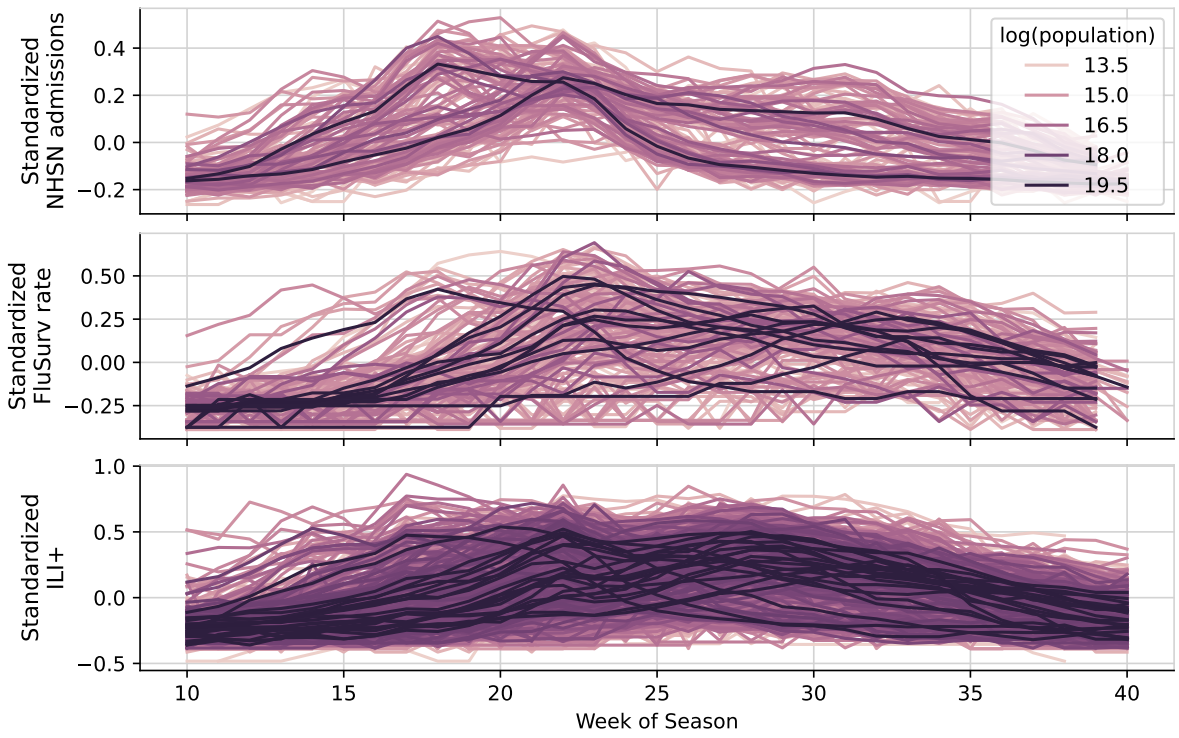

Figure 2: Influenza data for all surveillance signals and all locations available for each data source after standardizing transformations have been applied. The top panel shows weekly hospital admissions from NHSN, the second panel shows data from FluSurv-NET, and the third panel shows ILI+. The horizontal axis is the season week. We define the season to begin on US Epidemic Week 31, which generally corresponds to early August; the range of season weeks shown corresponds approximately to the active flu season. Within each panel, there is one line for each combination of season and location for all seasons and locations that are available for the given surveillance system at the state, regional, and national levels. Line color corresponds to the population size of the location; the darkest lines are for the national level while the lightest lines are for states with small populations.

allows the model to use past examples from multiple locations and data sources to learn a mapping from x to y. However, in our model setup, predictions of NHSN admissions in a particular location are not informed by contemporaneous observations of NHSN admissions in other locations or by contemporaneous observations of other surveillance signals in that same location. The use of contemporaneous observations from other locations or signals to inform predictions remains a topic for future work.

Both the features x_i and the prediction targets y_i are calculated based on the standardized version of the original surveillance signal, $\tilde{z}$ (see section 2). As was described above, on reference date d the most recent available data report on influenza activity in the previous week, $d-1$. For forecast task i with reference date $d(i)$ and forecast horizon $h(i)$, we define the prediction target for this model to be $y_i := \tilde{z}_{l(i),s(i),d(i)+h(i)} - \tilde{z}_{l(i),s(i),d(i)-1}$, the difference between the transformed signal value on the target date and on the date with most recent available reported data. Thus, the model is trained to predict the change in influenza activity over the next $h(i)+1$ time steps. Predictions of this target are converted to the original scale by adding the last observed value ($\tilde{z}_{l(i),s(i),d(i)-1}$) and inverting the initial data transformation operations described in section 2.

For the primary GBM model, the feature vector x_i contains the following 114 features:

1. A one-hot encoding of the data source.
2. A one-hot encoding of the location.
3. A one-hot encoding of the spatial scale of the location ("state", "region", or "national").
4. The population of the location.
5. The week of the season with the most recent reported data
6. The difference between the week of the season with the most recent reported data and Christmas week; for instance, a value of 3 means that the most recent data report is for the week three weeks after Christmas.
7. The coefficients of a degree 2 Taylor polynomial fit to the trailing w weeks of data, where $w \in \{4, 6\}$, with the reference point for the polynomial set to the time $d(i) - 1$ with the most recent reported data. These coefficients are estimates of the local level, first derivative, and second derivative of the signal at the time $d(i) - 1$.
8. The coefficients of a degree 1 Taylor polynomial fit to the trailing w weeks of data, where $w \in \{3, 5\}$. These coefficients are estimates of the local level and first derivative of the signal at the time $d(i) - 1$.
9. The rolling mean of the signal over the last w weeks, where $w \in \{2, 4\}$.
10. The values all features from points 6, 7, and 8 at lags 1 and 2, representing estimates of the local level and first and second derivatives of the signal in each of the previous two weeks.
11. The forecast horizon.

The computation of features based on Taylor polynomials is described in more detail in Supplemental section TODO.

Measures of the local level of the surveillance signal (i.e., rolling means and the intercepts of Taylor polynomial fits) had a high feature importance in the primary GBM model; see section 7.3 below for more detail. Starting the week of TODO, we fit a second variation on the GBM model that was not allowed to see these "local level" features. This was motivated by two considerations: (1) a model fit without features that had high importance in the primary GBM model might introduce more model diversity to the flusion ensemble; and (2) in seasons with particularly high or low incidence, measures of local level might not be a reliable indicator of the magnitude and direction of changes in future values of influenza activity.

5.3 Component Model 3: ARX

Our third component model is a Bayesian auto-regressive time series model with covariates (ARX). This model includes only one covariate: a spike function indicating proximity to the week of Christmas, taking the value 3 on Christmas week, 2 in the week before and the week after Christmas, 1 two weeks before and two weeks after Christmas, and 0 otherwise. This covariate is intended to help the model account for the consistent peak that is observed near the holiday.

The ARX model is trained only on NHSN admissions, but as with the GBQR models, it is trained jointly on data for all locations. The autoregressive coefficients are shared across locations, while a

separate variance parameter for the innovations is estimated for each location. Specifically, the model has the following structure, where we suppress the index s and use J to denote the autoregressive order (which we set to 8 in this work):

$$\begin{aligned}
\tilde{Z}_{l,t} \mid \tilde{z}_{l,t-1}, \ldots, \tilde{z}_{l,t-J}, x_{l,t-1}, \ldots, x_{l,t-J}, \varepsilon_{l,t} &= \sum_{j=1}^{J} \alpha_j \tilde{z}_{l,t-j} + \sum_{j=1}^{J} \beta_j x_{l,t-j} + \varepsilon_{l,t} \\
X_{l,t} \mid x_{l,t-1}, \ldots, x_{l,t-J}, \nu_{l,t} &= \sum_{j=1}^{J} \gamma_j x_{l,t-j} + \nu_{l,t} \\
\varepsilon_{l,t} &\sim \text{Normal}(0, \sigma_{\varepsilon,l}) \\
\nu_{l,t} &\sim \text{Normal}(0, \sigma_{\nu,l})
\end{aligned}$$

The parameters $\sigma_{\varepsilon,l}$ and $\sigma_{\nu,l}$ are assigned independent Half-Cauchy priors, and all of the α_j, β_j, and γ_j parameters are given a shared hierarchical normal prior, $\alpha_j, \beta_j, \gamma_j \sim \text{Normal}(0, \xi)$ with a scale ξ that follows a Half-Cauchy prior.

Note that unlike many ARX model specifications, our model does not take future values of the covariate $x_{l,t}$ as known, but rather it predicts the covariate alongside the primary prediction target. In our current modeling setting, where $x_{l,t}$ is a deterministic function of the season week, this behavior is likely not ideal. Remedying this to allow for the provision of known future values of covariates is on a short list of model improvements to make. Anecdotally, we noted that near Christmas week the model's predictions of the covariate x captured the spike function's behavior.

The motivation for sharing the auto-regressive coefficients across locations was that this might help prevent the model from overfitting to a limited amount of training data. With an autoregressive order of $J = 8$, there are a total of 24 α, β, and γ parameters to estimate (and even if the values of the covariate were held fixed, there would be 16 α and β parameters to estimate). Estimating these parameters based on only a single past season of data for one location would likely be infeasible. This motivation is similar to the strategy of joint estimation of shared parameters across multiple observed series that has been employed elsewhere, e.g. in the PatchTST model (TODO Cite). On the other hand, we felt that it was important to estimate separate variance parameters for different locations because the amount of noise in the observed data varies substantially across locations with small and large populations (Figure 2).

5.4 Ensembling Methods: Quantile averaging and linear pools

TODO, we mostly did quantile averaging. sometimes i used a linear pool when i didn't trust the latest data report.

5.5 Model adjustments used for real time forecasts

TODO, we tried a few different model variations over the course of the season. To pull from `https://github.com/cdcepi/FluSight-forecast-hub/blob/main/model-metadata/UMass-flusion.yml`

5.6 Software

TODO: Describe the software used: R, python, timeseriesutils, flusion, flusion-manuscript, sarix, reproducible workflow for rendering manuscript using Docker and Rnw.

6 Real-time performance: the 2023/24 season

In this section, we summarize model performance results for real-time submissions to the FluSight Forecast Hub in the 2023/24 season.

6.1 Evaluation setup

To avoid distortion of WIS and AE results, we did not evaluate forecasts that were made at the national level. Although the FluSight forecast hub originally allowed for collection of predictions at a "horizon" of -1 week, these were discontinued; our analysis includes predictions made at horizons of 0 weeks ("nowcasts"), and predictions at horizons of 1, 2, and 3 weeks ahead relative to the reference date.

We included all models that contributed forecasts for at least 75% of the combinations of state-level locations, reference dates, and non-negative horizons for which the Hub collected forecasts over the course of the season. Because a comprehensive evaluation of all Hub contributors is not the aim of this manuscript, we have anonymized the names of other individually-contributed models in these results to focus attention on the comparisons that are of interest for our purposes.

The FluSight hub produced two ensemble forecasts during the season: one using a quantile averaging approach and one using a linear pool. These two ensembles had similar performance, though the linear pool had slightly better calibration. However, the quantile averaging ensemble was used by CDC as the source of official communications throughout the season, and so we include results from only that ensemble here.

We also included results from two baseline methods. **Baseline-flat** is a random walk model produced by the Hub (labeled as **FluSight-baseline** in Hub submissions), which produces forecasts that extend from the most recent observation in a flat line, with expanding uncertainty based on historical differences in weekly hospital admissions. In this method, for each location i the historical differences $\delta_{i,t} = y_{i,t} - y_{i,t-1}$ are collected, along with their negative values $-\delta_{i,t}$ (a process which we refer to as "symmetrizing" the differences). Forecasts at multiple step-ahead horizons are generated by iteratively sampling from this collection of symmetrized weekly differences.

The second baseline method, **Baseline-trend**, follows a similar process with a few modifications that are designed so that the resulting forecasts tend to follow the trend of recent observations. It is a quantile averaging ensemble of 16 variations on the baseline method. Most importantly, it incorporates variations that do not symmetrize the past differences, and rather than using all available history, it collects differences in a rolling window of the past few weeks. The 16 variations are obtained by using different options for the rolling window size, the temporal resolution of data used as an input (daily or weekly), a data transformation that is applied (no transformation or square root), and whether or not symmetrization is used. We emphasize that although this baseline is more methodologically involved than **Baseline-flat**, it produces an epidemiologically naive forecast that pushes forward a local estimate of the trend observed over the few most recent weeks. This model is named **UMass-trends_ensemble** in real-time Hub submissions.

6.2 Evaluation results

Forecasts from the Flusion model often appeared similar to forecasts from the FluSight-ensemble, though in several states (e.g. Florida, California, and New York) Flusion did a better job of capturing the increase in weekly hospital admissions in the early part of the season and a slightly better job

of predicting the turnaround after the peak in late December (Figure 3). Qualitatively, both the FluSight-ensemble and the Flusion model generally captured trends in hospital admissions better than the baseline models during all phases of the season – during the rise in the early part of the season, near the peak, and on the way down. An exception to this can be seen in forecasts for New York, Pennsylvania, and Michigan that were produced with a reference date of February 3, 2024. Those states had two local peaks: one near Christmas, and a second in late February or early March. This was a common pattern in many states in the northern US in the 2023/24 season, and in these instances the Flusion model (as well as the FluSight ensemble) typically produced incorrect predictions of continued decreases after the first peak heading into the second peak.

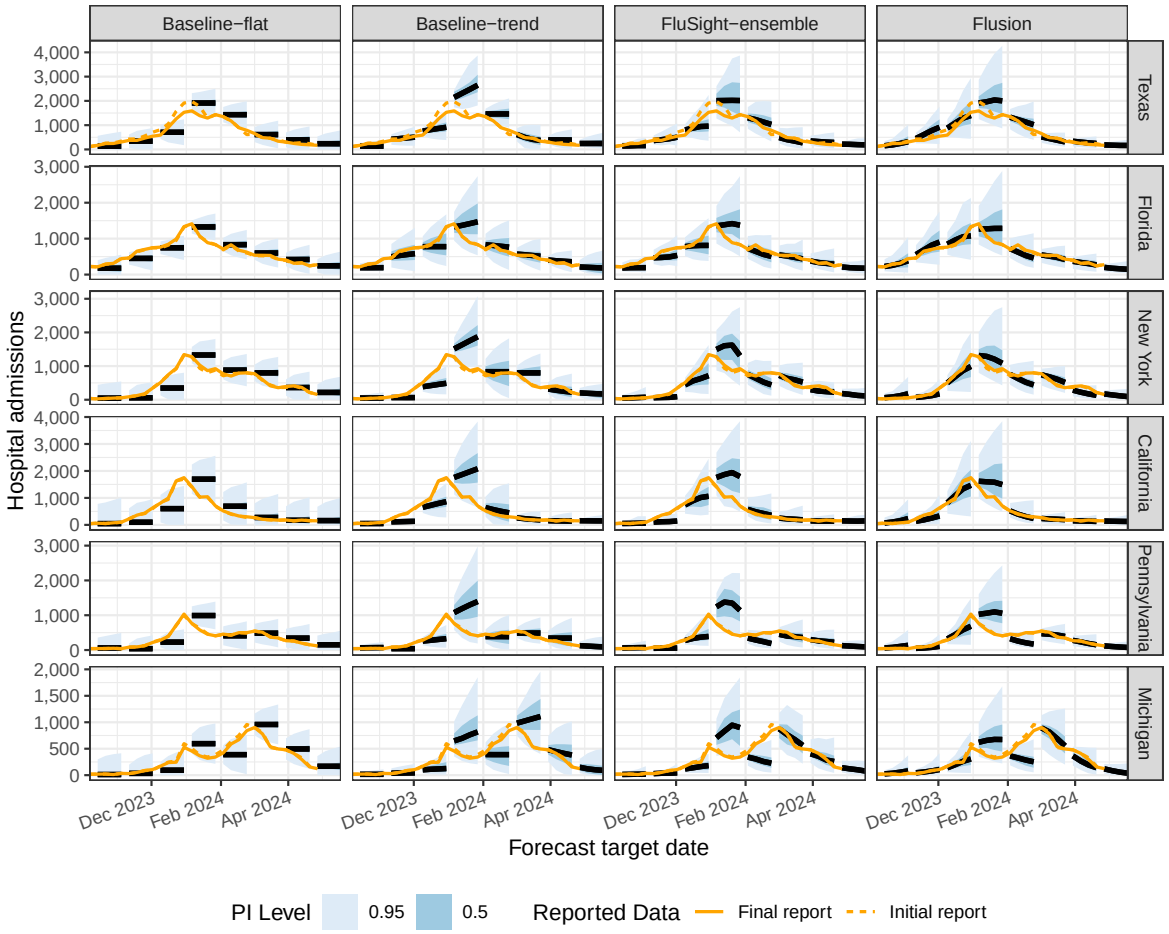

Figure 3: Influenza data and forecasts for the six states with the largest cumulative hospital admissions during the 2023/24 season. To avoid overplotting, in this figure forecasts from every fourth reference date are shown; evaluations include all reference dates. Forecasts are represented by the predictive median (black lines) and 50% and 95% prediction intervals (blue shaded regions). Solid orange lines show the finalized admission counts reported as of May 17, 2024, while dotted orange lines show the initial reported values that were available on the date predictions were generated.

Aggregating across all forecast dates and forecast horizons, the Flusion model had the best performance as measured by RWIS and RAE among all models that contributed to the FluSight Forecast Hub (Table 1). The Flusion model was consistently among the top-ranking models contributing to the forecast hub for individual forecast reference dates and forecast horizons (Figure 4 (b)).

Prediction intervals from the Flusion model tended to be underconfident, i.e., prediction intervals were too wide on average (Table 1). An examination of one-sided quantile coverage rates indicates that marginally, the predictive quantiles in the upper tail of the forecast distribution are fairly well calibrated, while predictions for the lower quantile levels were too small on average (Figure 4 (c)). Overall, the probabilistic calibration of the Flusion model was comparable to or better than that of other models contributed to the Hub, and it was superior to the calibration of the baseline and ensemble models.

TODO: Supplemental analysis with similar results, omitting forecasts affected by data revisions. Rankings don't change much.

Model	% Submitted	MWIS	RWIS	MAE	RAE	50% Cov.	95% Cov.
Flusion	99.9	**29.6**	**0.610**	**45.6**	**0.670**	0.583	0.967
FluSight-ensemble	100.0	35.5	0.731	55.4	0.814	0.516	0.926
Other Model #1	100.0	35.6	0.731	54.0	0.792	0.558	**0.940**
Other Model #2	89.1	40.4	0.773	61.5	0.840	0.479	0.908
Other Model #3	97.8	39.9	0.806	59.3	0.857	0.363	0.793
Other Model #4	100.0	40.0	0.823	60.5	0.890	**0.497**	0.884
Other Model #5	67.3	45.0	0.827	68.7	0.899	0.487	0.866
Other Model #6	100.0	41.5	0.851	64.4	0.945	0.466	0.903
Other Model #7	85.5	45.7	0.852	66.1	0.878	0.418	0.824
Other Model #8	100.0	41.6	0.856	60.7	0.893	0.460	0.855
Other Model #9	100.0	42.1	0.865	60.9	0.894	0.442	0.827
Other Model #10	98.8	44.3	0.901	67.7	0.986	0.456	0.939
Baseline-trend	99.9	43.9	0.906	67.0	0.990	0.618	0.922
Other Model #11	95.7	45.0	0.908	66.2	0.956	0.554	0.870
Other Model #12	87.0	45.0	0.936	70.7	1.050	0.449	0.929
Other Model #13	96.4	42.4	0.948	64.2	1.030	0.429	0.896
Other Model #14	93.6	48.7	0.980	70.8	1.020	0.473	0.838
Other Model #15	99.2	47.3	0.993	58.1	0.870	0.596	0.793
Baseline-flat	100.0	48.5	1.000	67.9	1.000	0.282	0.888
Other Model #16	72.2	59.1	1.010	87.1	1.060	0.416	0.823
Other Model #17	96.3	51.8	1.040	65.3	0.934	0.242	0.751
Other Model #18	98.2	51.3	1.040	73.0	1.060	0.395	0.773
Other Model #19	76.7	61.9	1.090	87.8	1.100	0.288	0.717
Other Model #20	84.2	52.6	1.150	72.0	1.130	0.368	0.768
Other Model #21	85.1	57.8	1.180	73.2	1.070	0.316	0.615
Other Model #22	88.3	61.3	1.180	89.1	1.230	0.377	0.802
Other Model #23	69.0	42.6	1.280	59.0	1.270	0.386	0.772
Other Model #24	85.5	65.2	1.300	83.5	1.200	0.219	0.494
Other Model #25	92.5	80.2	1.550	110.0	1.520	0.389	0.821
Other Model #26	92.6	126.0	2.540	154.0	2.220	0.174	0.429

Table 1: Overall evaluation results for forecasts submitted to the FluSight Forecast Hub. Model names other than **Flusion**, **FluSight-ensemble**, **Baseline-flat**, and **Baseline-trend** are anonymized. The percent of all combinations of location, reference date, and horizon for which the given model submitted forecasts is shown in the "% Submitted" column; only models submitting at least 2/3 of forecasts were included. Results for the model with the best MWIS, RWIS, MAE, and RAE are highlighted. Results for the models where empirical PI coverage rates are closest to the nominal levels are highlighted.

7 Post hoc model exploration

In this section, we investigate the degree to which the following aspects of the Flusion model contibuted to its strong performance: 1. the formulation of Flusion as an ensemble of three individual

models; 2. joint training on multiple data sets and multiple locations; 3. data preprocessing, including corrections for reporting inconsistencies in the ILINet and FluSurvNET data, the use of a fourth root data transform, and the importance of the features that were used by the GBQ model.

7.1 Component models and ensembling

To investigate the skill of our individual component models and the added value of ensembling, we computed scores for each of the three component models that were members of the Flusion ensemble and for ensembles formed using two of the three components. As documented in section 5, the component models and ensembling method that we used in real time changed over the course of the season. To enable a clearer understanding of the contributions of these models, the results we present here are based on the specifications of the individual GBQ, GBQ-no-level, and ARX models and the quantile averaging ensemble method that were used for the Flusion model starting the week of December 2, 2023. In instances where predictions from one or more component models were not created in real time, we created post hoc model fits and predictions using the data that would have been available in real time.

Although two of the ensembles outperformed the GBQ model alone, the most important determinant of performance was whether or or not the GBQ model was included (Table 2, Experiment A). Score differences among the top four model variations were small, and all of those model variations included the GBQ model. There was a drop-off in performance for other variations that did not include GBQ. Thus, the fact that Flusion was constructed as an ensemble of three models was not a key driver of its performance. Indeed, GBQ alone and GBQ-no-level alone would each have placed first among all FluSight submissions, while ARX would have placed third among all contributing models.

The GBQ-no-level model, which was trained without access to features measuring the local level of the time series, had worse performance than the primary GBQ model, and ensembles that included it generally performed slightly worse than ensembles that did not include it. We view this as evidence that the approach of omitting local level features was harmful to individual model performance without introducing enough differentiation from the primary GBQ model to serve as a useful ensemble member. In contrast, although ARX was the worst individual model, ensemble variations that included ARX were slightly better than ensemble variations that did not include it. Including models with more structural differences can be helpful in an ensemble, although in these results the gains in performance from including ARX were generally small.

7.2 Joint training on multiple data sets and multiple locations

The primary GBQ model was trained jointly on data from all three surveillance signals (NHSN, ILINet, and FluSurv) and on data for all locations at the state, HHS regional, and US national level. To investigate the value of this joint model training approach, we considered two alternative methods:

1. The GBQ-only-NHSN model was trained on data for all locations, but using only hospital admissions from NHSN, the surveillance signal used as the prediction target.

2. The GBQ-by-location model was trained separately for each state-level jurisdiction using data for that location from all three data sources.

Both of these alternatives underperformed relative to the GBQ model (Table 2, Experiment B). GBQ-by-location would have been among the top three contributing models to the forecast hub, and GBQ-

only-NHSN would have been among the top ten contributing models; however, both would have underperformed relative to the FluSight ensemble. The decisions to train on multiple data sources and to train jointly on data for all locations were critical for achieving strong model performance.

7.3 Data preprocessing

In a third experiment, we fit two model variations to investigate the value of some of the data preprocessing steps we used. The GBQ-no-reporting-adj model omitted the adjustments described in Section 2 that were intended to address reporting inconsistencies in the ILINet and FluSurvNET data. Specifically, this model used the ILI signal directly rather than using test positivity rates to convert to ILI+, and it used the raw rates reported by FluSurvNET rather than attempting to account for time-varying case capture rates in the FluSurvNET system. In our evaluations, GBQ-no-reporting-adj outperformed the original GBQ model by a small amount (Table 2, Experiment C). These reporting adjustments were not helpful to model performance, and indeed the evidence suggests that they were counterproductive.

In a second model variation, we investigated whether or not the use of a fourth root data transform was helpful. The GBQ-no-transform model was fit to data without using a power transform, though other transformations described in section 2 were used, including converting hospital admission to a rate per 100,000 population and applying centering and scaling operations to make the data more comparable across different locations and data sources. The GBQ-no-transform model had slightly worse performance than the original GBQ model, indicating that the power transform was helpful (Table 2, Experiment C).

We also investigated feature importance as measured by the number of times each feature was used for the splitting criterion in a tree node in the gradient boosting fits for the GBQ model (Supplemental Figure TODO). For this investigation, we used a representative fit from the reference date of January 6, 2024. We averaged the importance score across the gradient boosting fits from all 100 bags and all 23 quantile levels. The top five features were the current season week, the population of the target location, the most recent observation of the surveillance signal (after preprocessing transformations), the forecast horizon, and the difference between the current season week and Christmas week. These were followed by a group of features that also had fairly high importance, primarily consisting of features measuring the local level, trend, and curvature of disease incidence, as well as an indicators of whether the location was Puerto Rico (which sees substantively different trends in influenza activity than other locations), and indicators of what the data source was. A final group of features with lower importance included indicators for all other locations and indicators of the aggregation level for the location (state, regional, or national).

8 Discussion

summary

- we had top-ranked performance as measured by MAE and MWIS
- thing that mattered most was training jointly on data from multiple surveillance signals and locations
- other decisions had a more minor impact or were unhelpful: data transformations, attempting to

correct for reporting artifacts, adding in ensemble members other than our main GBQR model.

next steps

- incorporating spatial dependence
- contemporaneous use of multiple signals
- other possible signals – quidel, insurance claims
- vaccine uptake and efficacy
- better handling of holiday effects and the possibility of a second peak
- flu strain data – how to use, laggy?
- improvements to ARX model

Acknowledgements

This work has been supported by the National Institutes of General Medical Sciences (R35GM119582) and the U.S. CDC(1U01IP001122). The content is solely the responsibility of the authors and does not necessarily represent the official views of NIGMS, the National Institutes of Health, or CDC.

References

[1] Estee Y Cramer et al. "Evaluation of individual and ensemble probabilistic forecasts of COVID-19 mortality in the United States". In: *Proceedings of the National Academy of Sciences* 119.15 (2022), e2113561119.

[2] Edward Goldstein et al. "Predicting the epidemic sizes of influenza A/H1N1, A/H3N2, and B: a statistical method". In: *PLoS medicine* 8.7 (2011), e1001051.

[3] Jeffrey Shaman et al. "Real-time influenza forecasts during the 2012–2013 season". In: *Nature communications* 4.1 (2013), p. 2837.

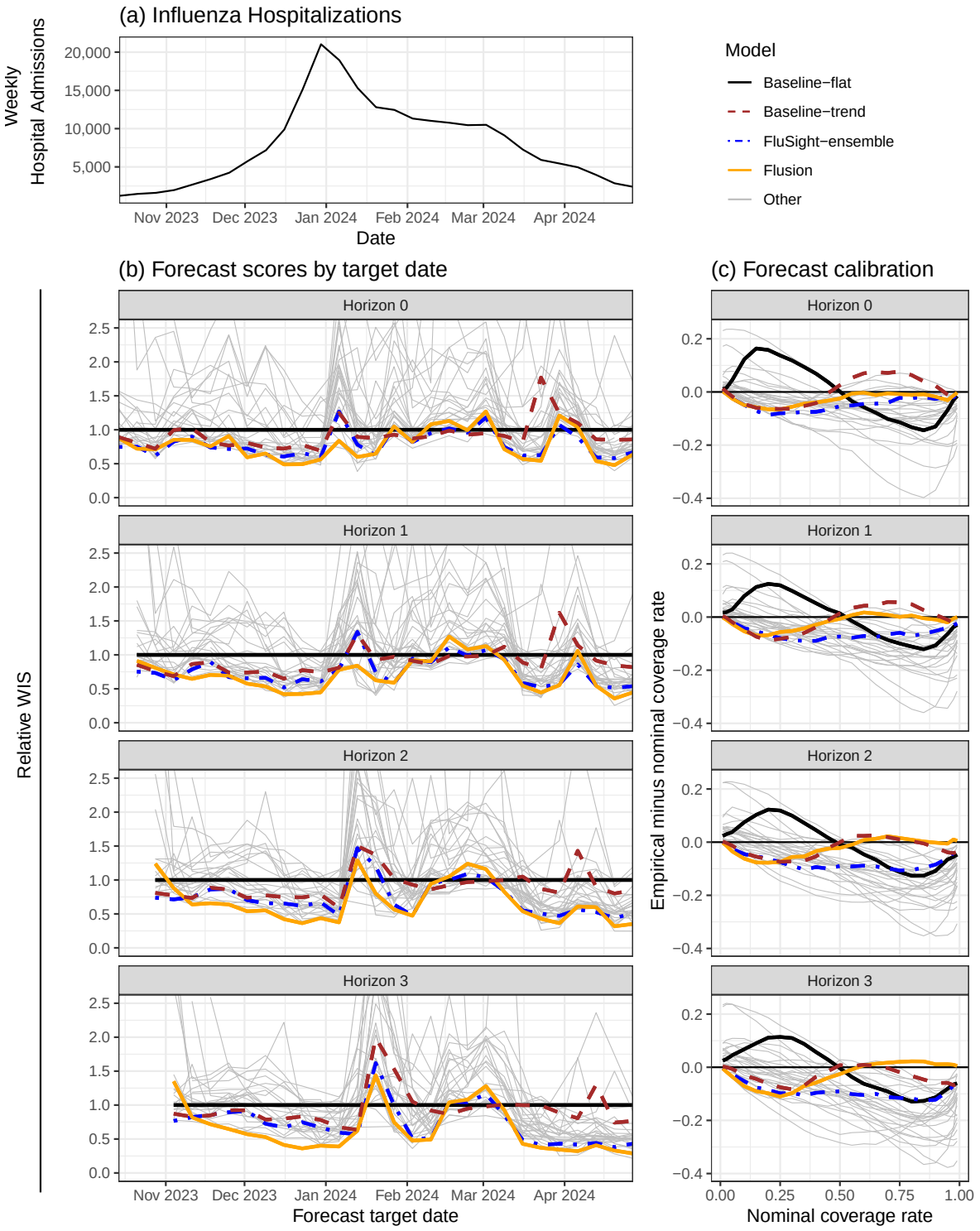

Figure 4: Influenza data and evaluation results. Panel (a): Weekly influenza hospital admissions reported in NHSN for the 2023/24 season, aggregated across all forecasted state-level locations. Panel (b): RWIS for models contributing to the FluSight hub, by forecast horizon (panels) and target date (horizontal axis). Lower relative WIS indicates better forecast performance. RWIS values greater than 2.5 are not displayed. Panel (c): One-sided quantile coverage differential, computed as empirical coverage rate minus nominal coverage rate. A well-calibrated model has a differential of 0, while a conservative method (with wide prediction intervals) has a negative differential at nominal coverage rates less than 0.5 and a positive differential at nominal coverage rates greater than 0.5.

Experiment A: Component model performance

Model	% Submitted	MWIS	RWIS	MAE	RAE	50% Cov.	95% Cov.
GBQ, ARX	100.0	**29.9**	**0.618**	**45.3**	**0.668**	0.570	0.958
Flusion	100.0	30.2	0.622	46.6	0.686	0.558	0.963
GBQ	100.0	30.3	0.625	46.3	0.682	0.529	**0.947**
GBQ, GBQ-no-level	100.0	30.4	0.628	47.1	0.694	0.546	0.958
GBQ-no-level, ARX	100.0	33.2	0.685	52.2	0.769	0.528	0.958
GBQ-no-level	100.0	33.9	0.698	52.6	0.775	0.523	0.944
ARX	100.0	39.5	0.815	60.0	0.884	**0.485**	0.917
Baseline-flat	100.0	48.5	1.000	67.9	1.000	0.282	0.888

Experiment B: Reduced training data

Model	% Submitted	MWIS	RWIS	MAE	RAE	50% Cov.	95% Cov.
GBQ	100.0	**30.3**	**0.625**	**46.3**	**0.682**	**0.529**	**0.947**
GBQ-by-location	100.0	37.8	0.780	57.9	0.854	0.327	0.891
GBQ-only-NHSN	100.0	41.5	0.857	63.7	0.939	0.361	0.838
Baseline-flat	100.0	48.5	1.000	67.9	1.000	0.282	0.888

Experiment C: Data preprocessing

Model	% Submitted	MWIS	RWIS	MAE	RAE	50% Cov.	95% Cov.
GBQ-no-reporting-adj	100.0	**29.1**	**0.600**	**44.4**	**0.654**	0.510	0.940
GBQ	100.0	30.3	0.625	46.3	0.682	0.529	0.947
GBQ-no-transform	100.0	31.1	0.642	48.0	0.708	**0.497**	**0.948**
Baseline-flat	100.0	48.5	1.000	67.9	1.000	0.282	0.888

Table 2: Evaluation results for post hoc experiments investigating determinants of model performance. Experiment A gives results for individual component models in the Flusion ensemble, ensembles of pairs of components, and the full Flusion ensemble including all three components. Experiment B gives results for the GBQ model, which is trained jointly on data for all locations and data sources, and variations trained separately for each location (GBQ-by-location) and trained only on hospital admissions from NHSN (GBQ-only-NHSN). Experiment C gives results for a variation on the GBQ model that does not incorporate reporting adjustments designed to improve the degree to which ILINet and FluSurvNET data reflect influenza activity (GBQ-no-reporting-adj) and a variation that does not use a fourth-root transform (GBQ-no-transform), along with the original GBQ model which uses the reporting adjustments and the fourth-root transform. The percent of all combinations of location, reference date, and horizon for which the given model submitted forecasts is shown in the "% Submitted" column; in these retrospective experiments, we produced forecasts for all locations and time points. Within each experiment group, results for the model with the best MWIS, RWIS, MAE, and RAE are highlighted. Results for the models where empirical PI coverage rates are closest to the nominal levels are highlighted.