Flusion: Integrating multiple data sources for accurate influenza predictions

Evan L. Ray[1], Yijin Wang[1], Russell D. Wolfinger[2], and Nicholas G. Reich[1]

[1]Department of Biostatistics and Epidemiology, University of Massachusetts, Amherst, MA, United States

[2]JMP Statistical Discovery, Cary, NC, United States

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 influenza hospital admissions through NHSN began within the last few years, and as such only a limited amount of historical data are available for this target signal. 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 selected set of 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 of influenza hospital admissions in each state for 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 adds to the pool of available training data.

Keywords: infectious disease, forecasting, gradient boosting, transfer learning

1 Introduction

Since the early 2010s, short-term forecasting for infectious diseases has become an increasingly common activity, often through collaborations between governmental and industry or academic partners. Starting during the 2013/2014 influenza season, the FluSight collaborative forecasting exercises organized by the US Centers for Disease Control and Prevention (CDC) have brought together teams of academic, industry and governmental

researchers to produce probabilistic forecasts of influenza activity in the United States [5]. A primary motivation for the FluSight challenges is to carefully evaluate forecasts against real data and to use them to improve situational awareness and yield more effective public health actions in realms such as the distribution of limited medical resources or the implementation of non-pharmaceutical interventions [58, 32, 34].

Starting with FluSight seasons just following the COVID pandemic, new high-resolution and low-latency data streams that came online during the pandemic have been used as the "ground truth" target for forecasting exercises. Specifically, from 2022 through 2024, the forecasting targets for FluSight were based on hospital admissions reported in the CDC's National Healthcare Safety Network (NHSN) surveillance system [17]. Because reporting of influenza hospital admissions through NHSN began during the COVID pandemic, the NHSN data have only a limited amount of historical information about influenza hospitalizations. While the NHSN data surpassed previous surveillance systems in terms of providing a specific measure of influenza activity at fine spatial and temporal scales, the lack of an extensive history of data for model training introduced challenges to learning about seasonal patterns in influenza burden in the US. To address this challenge, for the 2023/2024 FluSight season we developed a new model, called *Flusion*, which pulled in data from external data sources with a longer history of observations. This model was the top-performing model submitted to FluSight in the 2023/2024 season, and it even outperformed the ensemble forecast that the FluSight organizers produced by combining predictions from all contributing models; these ensemble forecasts have typically had some of the most accurate performance overall in past influenza seasons [38, 49]. The purpose of this paper is to describe the Flusion model and provide insights into what aspects of its design were associated with its strong performance.

Stimulated in large part by outbreak forecasting challenges organized by governmental agencies [e.g., 56, 38, 26], there has been substantial methodological development in the area of infectious disease forecasting over the last 10 years, both for stand-alone and ensemble modeling strategies [63, 49, 50, 47]. Most successful stand-alone approaches can be characterized broadly as using statistical or machine learning methods in conjunction with conceptual models that are informed by infectious disease transmission dynamics, either explicitly through a compartmental transmission model or implicitly through inclusion of seasonality or observations from recent time-points [48, 33]. There has simultaneously been a large increase in worldwide participation in online data science competition sites like Kaggle, including many competitions with time series forecasting and even a series of five weeks of real time forecasting of COVID-19 in 2020 [1, 23, 59, 60, 61]. The top performers in these competitions make heavy use of advanced feature engineering in conjunction with gradient boosted tree methods [19], using the popular packages XGBoost, LightGBM, and CatBoost. For example, the majority of the top-performing competitors in the M5 forecast accuracy competition used gradient boosting [35].

The Flusion model fits into the machine learning model category, although it uses features that emphasize both seasonality as well as recent observations and recent trends as a way of encoding underlying transmission dynamics of influenza. In terms of underlying methodology, our primary focus in this work is on forecasting using gradient boosting. In particular, our work closely follows the approach of [29], which was the leading approach in several time series forecasting competitions including strong performance for forecasting COVID-19 [33]. Our methods adapt that work to forecasting seasonal influenza by including features intended to help the model identify seasonal trends of disease activity and by incorporating multiple data sources measuring influenza activity.

A key area where the Flusion model provides a methodological innovation is in how it leverages long historical time-series data that are closely related to the target data source that has little history. There is existing work that focuses on situations where auxiliary data signals related to a primary signal of interest are used to improve short-term predictions. For example, in settings where a primary signal has delayed or unreliable recent reporting, research has shown that other signals that are available contemporaneously or with a shorter lag can help inform estimation and prediction of trends [64, 18, 42, 30, 37, 25]. Some of these modeling efforts, among others, also highlight how sharing contemporaneous information across different locations can be helpful

[18, 37, 43].

In contrast, the Flusion model, in addition to borrowing information across locations, augments a primary data signal with a very short history with other closely related signals that measure the same general system but for a longer history. This approach bears a close resemblance to multi-task learning and transfer learning, where models are trained on related tasks with a large number of available observations either simultaneously with training on more limited data for the primary task, or in a first stage that is followed by fine-tuning on data for the target task [44]. In a forecasting context, these methods have been used in domains such as energy use and climate control [51, 22]. However, to our knowledge the literature contains only a few applications of this general approach to disease outbreak forecasting. We are aware of two studies that investigated forecasting a pathogen of interest using data from other pathogens to inform model estimation: [53] considered using historical data on dengue to inform Zika predictions as well as the use of past influenza data to inform COVID-19 forecasts, while [10] used models trained on data measuring dengue incidence to generate predictions for Chikungunya and Zika. In another line of work, Zou et al. used multitask learning to forecast influenza-like illness by training models jointly on data for multiple geographic locations, and demonstrated that this was helpful in settings where some locations had only limited data but others had a longer history of observations [65]. Unlike this existing literature, our work augments a signal measuring seasonal influenza in the US that has a limited history with other signals that measure the same disease system over a longer time span.

We note that contemporaneously with our development of Flusion, two other groups also explored methods for forecasting influenza hospitalizations as reported in NHSN using other data streams to extend the amount of available training data [40, 4]. Their approaches differed from ours in that they used these other data streams to impute additional historical observations of hospitalizations, rather than training a model directly on data from multiple surveillance signals. Although those methods did not outperform Flusion or the FluSight ensemble, they were ranked 2nd and 13th among the 29 individual models included in our evaluation of FluSight forecasts for the 2023/24 season (see Section 6.2). The success of the Flusion model and these other methods that incorporated multiple data sources in the 2023/2024 FluSight season suggests that this approach may be a promising direction, especially in the context of public health data modernization initiatives that may replace long-standing surveillance data systems with new data streams.

The remainder of this article is structured as follows. We begin with some context for our work, including a discussion of the data sources we use in section 2 and an overview of the FluSight forecasting exercise in 3. We establish notation in section 4 and describe our modeling approaches in section 5, and then we discuss the performance of our real-time forecasts relative to other contributors to FluSight in the 2023/24 season in section 6. Section 7 details a series of ablation studies that provide insight about which aspects of our modeling approach were important to its success. We conclude in section 8.

2 Data Sources

Our model used three measures of influenza activity (Figure 1). The first of these was weekly influenza hospital admissions reported to the National Healthcare Safety Network (NHSN, [17]), which was the target signal for FluSight (see section 3 for more detail). NHSN reports counts of hospital admissions where the patient tested positive for influenza at the time of admission or within the previous two weeks. Requirements for facilities reporting to NHSN have changed over time, but for the time period included in our analyses reporting was generally mandatory for all hospitals registered with the Centers for Medicare and Medicaid Services. We worked with reported data that had been aggregated to the state level for each of the 50 US states, the District of Columbia, Puerto Rico and the national level. Reporting of influenza hospitalizations through NHSN began in 2020, and the first two years of reporting for this signal showed 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.

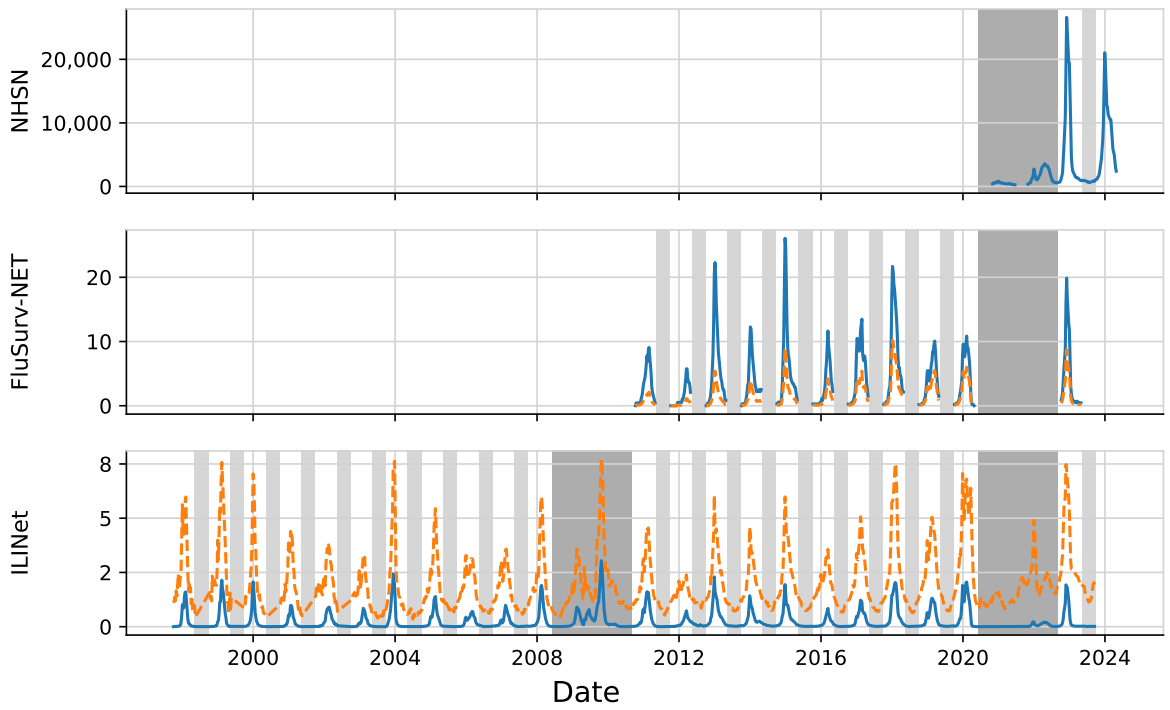

Figure 1: Influenza data at the national level in the US. The top panel shows weekly hospital admissions from NHSN, including the 2023/24 season that was the target season for predictions described in this article. In the second panel, a dashed orange line shows raw data reported from FluSurv-NET; the modeled data, in blue, are obtained by scaling up the raw data using per-season inflation factors designed to account for varying testing rates and test sensitivity. 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. Dark 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 H1N1 influenza, and the 2020/21 and 2021/22 seasons which were impacted by low influenza activity during the COVID pandemic. Light grey shaded regions indicate the off-season, which was not used for training the GBQR and GBQR-no-level models. Additionally, FluSurv-NET and ILI+ data for the 2023/24 season were not used for model training in this work.

To address this limitation of the target surveillance signal, our model used two other measures of weekly influenza activity that have a longer reporting history. The first of these was 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 hospital facilities [15]. Our models used data from 14 sites in 13 states (including two sites in New York, which we aggregated to the state level), as well the "Entire Network", which we use as an estimate of influenza activity at the national level. The other signal was 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, the World Health Organization (WHO), and the National Respiratory Enteric Virus Surveillance System (NREVSS) [21, 17]. Our models use ILI+ data at the state level and national level, as well as at the regional level for the ten Health and Human Services regions, which are groups of between 4 and 10 states and/or territories of the US. As will be described further in section 5, in this work our goal in using these data was to provide models with additional historical data to use for learning about trends in seasonal influenza activity. Because contemporaneous use of these signals was not our primary goal, and in order to avoid complications with integrating multiple data sources with different reporting lags and backfill behaviors, in this work we only used historical data prior to the 2023/24 season for these alternative

data signals.

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 from season to season. The CDC produces annual estimates of hospital burden due to influenza that adjust for these factors [16]. We used these total burden estimates to estimate season-specific inflation factors that scale up reported rates from FluSurv-NET, with the intent of producing a more consistent measure of influenza activity over time; see supplemental section 2 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 a fever with a temperature of at least 100°F (37.8°C) and a cough and/or a sore throat [17]. 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 computed ILI+ as the product of this ILI signal and influenza test positivity rates from laboratory testing sites reporting to WHO and 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. 21, 54].

Our models were not trained on the 2008/09 and 2009/10 seasons, which were impacted by pandemic H1N1 influenza, and the 2020/21 and 2021/22 seasons, which were impacted by low influenza activity during the COVID pandemic. Additionally, some of our models were not trained on the influenza off-season (see section 5.2 for more detail on our models). Designating US Epidemic week 31 as season week 1 (generally falling in early August, a time of consistently low influenza activity), we define the on-season as season weeks 10 through 40 and the off-season to be the remaining weeks, which fall during the summer months [13]. Omission of the off-season from the training set forms a natural implementation of purging/embargoing, a method that is popular in econometric time series forecasting in which the time series is split into blocks for cross-validation with a buffer between the time blocks to prevent leakage of information across the cross-validation folds. [12, 29].

3 The FluSight collaborative forecasting exercise

For the 2023/24 season, the primary target for forecasts collected in the FluSight forecasting exercise organized by CDC was weekly hospital admissions with confirmed influenza as reported in the NHSN data set for each of the 50 US states, the District of Columbia, Puerto Rico, and in total at the national level (Figure 1, top panel).

Predictions were submitted to FluSight 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. 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 the 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 week 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 $z_{s,l,t}$ 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 refer to 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 the coverage rates of central 50% and 95% prediction intervals. In some figures, we will also examine one-sided quantile coverage rates. Our evaluations of forecast skill cover only predictions of NHSN admissions ($s = 1$). All of these metrics have been used by FluSight for forecast evaluation. The absolute error and weighted interval score are proper scoring rules for the predictive median and the full collection of predictive quantiles collected by FluSight, respectively [20, 6]. This means that in expectation under a forecaster's predictive distribution, the scores are optimized by submitting the predictive median or the collection of elicited predictive quantiles of that distribution.

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 [6]:

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

$$QS_{\alpha_k}(q_{i,k}, z_i) = \alpha_k \max(z_i - q_{i,k}, 0) + (1 - \alpha_k) \max(q_{i,k} - z_i, 0)$$

The quantile score for a single quantile prediction assigns an asymmetric penalty to the distance between the prediction and the observation. The magnitudes of the penalties for underprediction and overprediction are set so that the expected value of the quantile score with respect to the forecast distribution is minimized by the quantile of the predictive distribution at the specified quantile level [20]. 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 [11]. In a comparison of forecast accuracy among M models, computation of rMAE (rMWIS) 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 rMWIS) then normalizes this geometric mean relative to the value of θ^m for a

baseline model:

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

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

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. In this situation, comparing MAE or MWIS could be misleading because a forecaster could achieve good MAE or MWIS by omitting forecasts for difficult locations or time points, such as near a peak in disease incidence. The relative metrics account for this by evaluating each pair of forecasters on the prediction tasks they shared in common, enabling a direct comparison of their scores. However, ultimately the rMAE and rMWIS metrics may also be susceptible to gaming by a forecaster who selectively omits forecasts that may have poor performance. Of note, considerations about missingness are relevant to the evaluation of submissions to FluSight in section 6, but in the experimental results in section 7 all forecasts were provided. A secondary goal of the relative scores 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 rMWIS indicate better performance relative to the other models in the comparison pool, and in particular values less than 1 indicate performance that is better than the baseline.

Finally, we examine probabilistic calibration through the empirical coverage rates of 50% and 95% prediction intervals and individual quantiles. Let $l, u \in \{1, \ldots, K\}$ denote the indices for the lower and upper quantile levels defining the central prediction interval at a specified nominal level. For example, with our notation 95% intervals correspond to $l = 2$ and $u = 22$, with $\alpha_l = 0.025$ and $\alpha_u = 0.975$. The empirical prediction interval coverage rate across prediction tasks in the index set $\mathcal{I}$ is defined as

$$\frac{1}{|\mathcal{I}|} \sum_{i \in \mathcal{I}} \mathbb{1}(q_{i,l} \leq z_i \leq q_{i,u}),$$

where $|\mathcal{I}|$ is the size of the index set $\mathcal{I}$ and $\mathbb{1}(q_{i,l} \leq z_i \leq q_{i,u})$ is 1 if $q_{i,l} \leq z_i \leq q_{i,u}$ and 0 otherwise.

One-sided coverage rates at individual quantile levels can provide more detail than interval coverage rates. In some figures below, we show the difference between the empirical one-sided coverage rate at each quantile level and the corresponding nominal coverage rate:

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

A well-calibrated forecaster will have $\delta_k \approx 0$, 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

Flusion was constructed as an ensemble of statistical and machine learning time series models. All models were fitted to data that were 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 were fitted 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 applied the following operations to compute the transformed version of the signal, $\tilde{z}_{s,l,t}$:

1. For NHSN admissions, we divided 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. We used population counts reported by the US Census Bureau [7, 8]. 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 took a fourth root transformation to stabilize the variance of the signal across times of low and high influenza activity. The fourth root was selected through visual inspection of data plots with various power transforms, using the criteria that the amount of variability around a local trend should be fairly consistent throughout all phases of the influenza season.
3. We scaled by dividing by the 95th percentile of all observations for each location and data source, and centered by subtracting the mean for each location and data source. These transformations adjusted 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

The first two models used gradient boosting for quantile regression, which we abbreviate as GBQR; this methodological choice was motivated by the success of gradient boosting in several recent forecasting exercises as described in the introduction. Specifically, we used 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 learned a mapping $f_{\alpha_k}(x_i)$ 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 $h(i)$. In gradient boosting, the estimated function $f_{\alpha_k}(x_i)$ takes the form of a sum of regression trees which are estimated sequentially. In each step of estimation, a new regression tree is added by fitting to pseudo-observations which are the gradient of the objective function (here, average quantile loss) with respect to the predictions for each data point after the previous iteration. At each quantile level, the final prediction was obtained using bagging, a procedure in which predictions are formed by ensembling models fit to different random samples drawn from the full data set. In our implementation, the predictions were computed as the median of the predictive quantiles from 100 separate fits that were each based on a different randomly selected 70% of the seasons in the training set (including partial data for the current season). The models were fitted using the LightGBM package in Python [28] with default settings for all hyperparameters.

The models were trained jointly on data for all data sources, locations, dates, and forecast horizons. However, the features x_i for a particular prediction task indexed by i contained 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 (corresponding to different prediction tasks with different indices i) allowed 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 were 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. Of note, the inclusion of data for the same past season from multiple data sources and the inclusion of data that have been aggregated to the regional and national

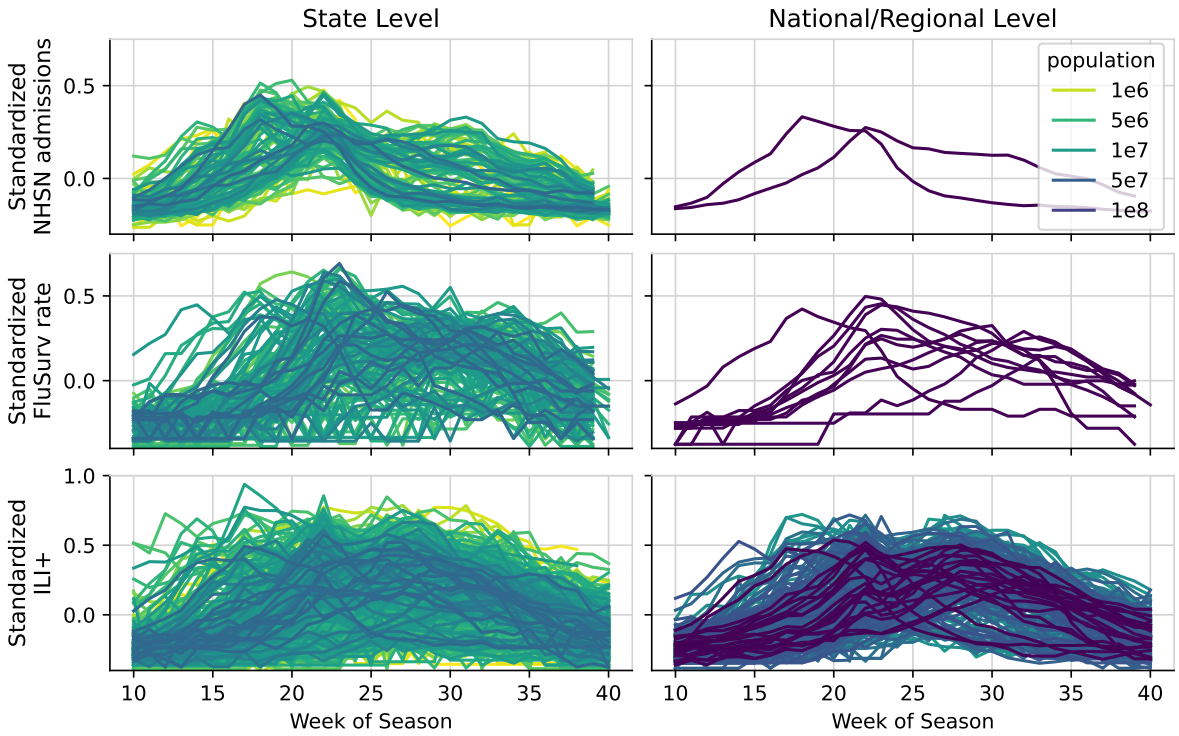

Figure 2: Influenza data for all surveillance signals and all locations available for each data source after standardizing transformations have been applied. The top row shows weekly hospital admissions from NHSN, the second row shows data from FluSurv-NET, and the third row shows ILI+. The left column shows all state-level locations, while the right column shows national level data, as well as data at the level of HHS regions for the ILI+ signal. The horizontal axis is the season week. We define the season to begin on US Epidemic Week 31, which generally falls in 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.

levels means that the model sees data related to any given time period and location multiple times during training.

Both the features x_i and the prediction targets y_i were calculated based on the standardized version of the original surveillance signal, $\tilde{z}$ (see section 5.1). 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 defined 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 was trained to predict the change in influenza activity over the next $h(i)+1$ time steps. Predictions of this target were 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 5.1.

For the primary GBQR model, the feature vector x_i contained 114 features as outlined in table 1. These features included information about the data source and location being forecasted, the time of season when the forecast was generated, the difference between the time of season when the forecast was generated and Christmas week (intended to help the model adjust for a consistent peak in reported influenza activity near

the holidays), the forecast horizon, and measures of the local level, trend, and curvature of the surveillance signal in the weeks leading up to the time $d(i)$. Measures of local level of the signal included the signal value itself, rolling means of the signal in a trailing window, and the intercepts of Taylor polynomials fit to a trailing window of observations; these Taylor polynomial fits were also used to obtain the features measuring local trend and curvature. More information about the calculation of these features is given in supplemental section 3.

Group	Description	Count
1	A one-hot encoding of the data source.	3
2	A one-hot encoding of the location.	65
3	A one-hot encoding of the spatial scale of the location ("state", "region", or "national").	3
4	The population of the location.	1
5	The week of the season with the most recent reported data, $d(i) - 1$.	1
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.	1
7	The forecast horizon.	1
8	The most recent reported value of the surveillance signal, for the time $d(i) - 1$.	1
9	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 $d(i) - 1$. These coefficients are estimates of the local level, first derivative, and second derivative of the signal at the time $d(i) - 1$.	6
10	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 $d(i) - 1$.	4
11	The rolling mean of the signal over the last w weeks, where $w \in \{2, 4\}$.	2
12	The values of all features from groups 8 through 11 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.	26

Table 1: Features used in the GBQR and GBQR-no-level models. There were 12 groups of features. The "Count" column gives the number of features in each group; there were 114 features in total across all groups. The GBQR-no-level model did not use the features from groups 8 through 12 that measured the local level of the signal. See supplemental section 3 for details of rolling mean and Taylor polynomial calculations.

Measures of the local level of the surveillance signal (i.e., the reported signal value, rolling means and the intercepts of Taylor polynomial fits) had a high feature importance in the primary GBQR model; see section 7.3 below and supplemental section 4 for more detail. Starting in the eighth week of the season, on the reference date of December 2, 2023, we included a second variation on the GBQR 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 GBQR 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. Although this model variation did not have access to any local level features (including the most recent data report), the formulation of the prediction target y_i as the difference between the most recent data value and the data value on the target date meant that predictions from this model were still placed in the correct location relative to recent observations.

In experimental results below, we refer to this model variation as GBQR-no-level.

5.3 Component Model 3: ARX

Our third component model was a Bayesian auto-regressive time series model with covariates (ARX). This model is simpler than the models based on GBQR, but we included it in the Flusion ensemble for two reasons. First, simple AR models have a history of strong performance for infectious disease forecasting [e.g., 48]. Second, past theoretical and applied investigations have shown that even relatively weak individual models can be valuable ensemble members if their prediction errors are not highly correlated with the other ensemble members, and in particular including methodologically diverse forecasters can be beneficial [e.g., 3].

The ARX model included 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 was intended to help the model account for the consistent peak in influenza activity that is observed near the holiday.

The ARX model was trained only on NHSN admissions, but as with the GBQR models, it was trained jointly on data for all locations. The autoregressive coefficients were shared across locations, while a separate variance parameter for the innovations was estimated for each location. Specifically, the model had 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 $\sigma_{\nu,l}$, $l = 1, \ldots, L$, were assigned independent Half-Cauchy priors, and all of the α_j, β_j, and γ_j parameters were given the shared hierarchical normal prior $\alpha_j, \beta_j, \gamma_j \sim \text{Normal}(0, \xi)$ with a scale ξ that followed a Half-Cauchy prior.

Note that unlike many ARX model specifications, our model did not take future values of the covariate $x_{l,t}$ as known, but rather it predicted the covariate alongside the primary prediction target. In our modeling setting, where $x_{l,t}$ was a deterministic function of the season week, this behavior was 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, though we did not make this update prior to the 2023/24 season due to time constraints. 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 were 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 is similar to the strategy of joint estimation of shared parameters across multiple observed series that has been employed elsewhere [e.g., 41]. 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).

In this model, forecasts more than one step after the most recent observed data were obtained by iterating

one-step-ahead predictions. As with the GBQR models, the model is specified in terms of the transformed data $\tilde{z}$, so predictions on the original scale of the data were obtained by inverting the initial data transformations described in section 5.1.

We used the the No-U-Turn Sampler (NUTS, [24]) algorithm for model estimation in NumPyro [45].

5.4 Ensemble Methods: Quantile averaging

For prediction task i, each of the three models above produced predictive quantiles $q^m_{i,k}$, $k = 1, \ldots, 23$. At each quantile level, the prediction for the Flusion model was the mean of the component predictive quantiles: $q_{i,k} = (1/M) \sum_m q^m_{i,k}$. This method has been referred to as quantile averaging [e.g., 31] or as Vincent ensembling after [57].

5.5 Model adjustments used for real time forecasts

Here we describe a few minor variations on the methodology outlined above that we introduced over the course of the season.

As we described above, the GBQR and ARX component models were used throughout the full season, but the GBQR-no-level model was introduced starting in the eighth week. In the first week of the season only, we used an additional model that obtained a predictive median using the same method as GBQR, but obtained predictions at other quantile levels by bootstrapping out-of-sample residuals.

In the first week of the season, as a hedge against possible reporting revisions we formed our submitted predictions by combining forecasts based on all available data and forecasts based on data up to the second-to-last observation (i.e., omitting the final reported value).

Finally, in the submission for reference date of December 16, 2023, we dropped our forecasts for Massachusetts. In that week, the most recent NHSN data release did not include full reporting for Massachusetts, and the FluSight organizers at CDC announced that they would not score or publicly communicate forecasts for Massachusetts that were submitted that week.

The experimental results below indicate that these changes had only a minor impact on forecast performance. For example, averaging over the full season, the MWIS of ensembles that included GBQR-no-level or omitted it differed by only 0.3 (Section 7, Table 3).

5.6 Software availability and reproducibility

Code for fitting models and generating predictions for real-time submissions and retrospective experiments is available at `https://github.com/reichlab/flusion`. Feature preprocessing functionality is implemented in a Python module at `https://github.com/reichlab/timeseriesutils/`, and the ARX model is implemented in a Python module at `https://github.com/elray1/sarix`. Code for analyses in the manuscript and supplement is available at `https://github.com/reichlab/flusion-manuscript`. The manuscript is generated with a reproducible workflow using Docker [39] and knitr [62]. Analyses were conducted using R version 4.4.0 [46] and Python version 3.10.12 [52].

6 Real-time performance: the 2023/24 FluSight season

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

6.1 Evaluation setup

We only evaluated forecasts that were made at the state level; MWIS and MAE can be dominated by locations with large populations, and we have found that averages that include scores for national level predictions often only reflect predictive skill at the national level. 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 two thirds of the combinations of state-level locations, reference dates, and non-negative horizons for which FluSight collected forecasts over the course of the season. Because a comprehensive evaluation of all FluSight contributors is not the aim of this manuscript, we anonymized the names of other individually-contributed models in these results to focus attention on the comparisons that are of interest for our purposes. However, we included these other models in the analysis to give a more complete sense of how Flusion compared to all other contributing forecasters. We also included the ensemble forecasts that were produced by FluSight and were used by CDC as the source of official communications throughout the season. These ensemble forecasts were calculated by taking the median of the predictions at each quantile level across all contributing models.

We also included results from two baseline methods. The first baseline, which we refer to as Baseline-flat, was a random walk model produced by FluSight (labeled as FluSight-baseline in submissions), which produced forecasts that extended 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 $\delta_{i,t} = y_{i,t} - y_{i,t-1}$ were 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 were generated by iteratively sampling from this collection of symmetrized weekly differences.

The second baseline method, Baseline-trend, was produced by our team. It followed a similar process to the Baseline-flat model, with a few modifications that were designed so that the resulting forecasts tended to follow the trend of recent observations. It was a quantile averaging ensemble of 16 variations on the baseline method. Most importantly, it incorporated variations that did not symmetrize the past differences, and rather than using all available history, it collected differences in a rolling window of the past few weeks. The 16 variations were 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 was applied (no transformation or square root), and whether or not symmetrization was used. We emphasize that although this baseline was more methodologically involved than Baseline-flat, it produced an epidemiologically naive forecast that pushed forward a local estimate of the trend observed over the few most recent weeks.

Scores were calculated based on the version of the NHSN target data signal that was released by FluSight on May 1, 2024. Reporting of influenza admissions to NHSN was not mandatory beyond that date, and so this data release represents the latest available complete data release to use for forecast evaluation.

6.2 Evaluation results

Forecasts from Flusion 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, Supplemental Figures 3 through 8). Qualitatively, both the FluSight-ensemble and Flusion 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 Flusion (as well as the FluSight ensemble) typically produced incorrect predictions of

continued decreases after the first peak heading into the second peak. There were also several instances, such as predictions made in late December in Florida, California, Pennsylvania, and Michigan, where predictions did not accurately capture the decline after the holiday peak. We return to these challenges in the discussion. These general trends in forecast performance also held in other states (Supplemental Figures 3 through 8).

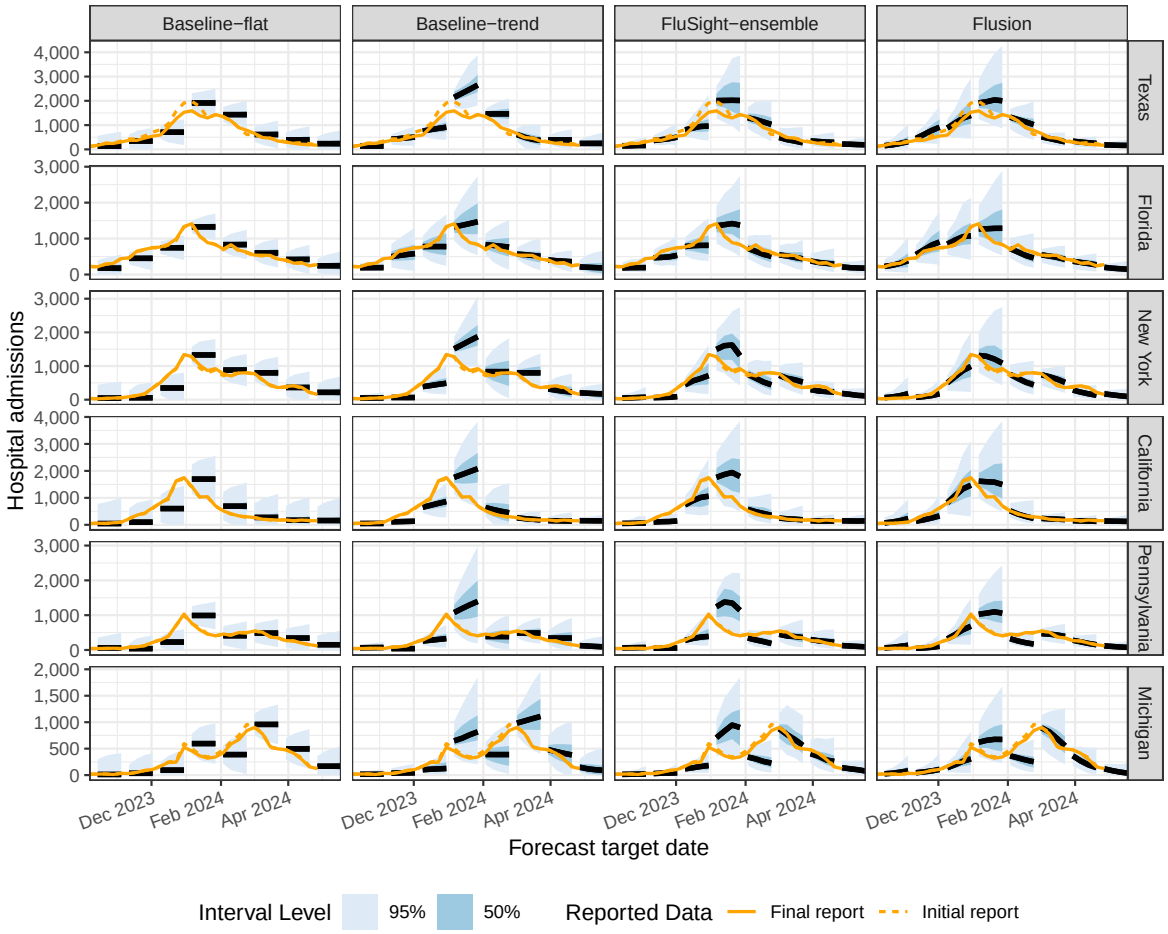

Figure 3: Influenza data and forecasts for the six states with the largest cumulative hospital admissions during the 2023/24 season. To avoid overplotting, in this figure forecasts from every fourth reference date are shown; evaluations include all reference dates. Forecasts are represented by the predictive median (black lines) and 50% and 95% prediction intervals (blue shaded regions). Solid orange lines show the finalized admission counts reported as of May 1, 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, Flusion had the best performance as measured by rMWIS and rMAE among all models that contributed to FluSight for the 2023/24 season (Table 2). In a sensitivity analysis, we found that these results still held when the evaluation was conducted on the subset of forecasts generated for locations and reference dates for which the latest available data at the time of the forecast were not subsequently revised by 10 or more admissions (supplemental section 7). Flusion was consistently among the top-ranking models contributing to FluSight across individual forecast reference dates and forecast horizons (Figure 4 (b)). Forecasts had decreasing accuracy at larger forecast horizons (Supplemental Figure 9).

Prediction intervals from Flusion tended to be underconfident, i.e., prediction intervals were too wide on average (Table 2). An examination of one-sided quantile coverage rates indicates that marginally, the predictive

quantiles in the upper tail of the forecast distribution were fairly well calibrated, while predictions for the lower quantile levels were too small on average (Figure 4 (c)). Overall, the probabilistic calibration of Flusion was comparable to or better than that of other models contributed to FluSight, and it was superior to the calibration of the baseline and ensemble models.

Model	% Submitted	MWIS	rMWIS	MAE	rMAE	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 2: Overall evaluation results for forecasts submitted to FluSight. 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. Models are ordered by rMWIS, with better-performing models at the top of the table. Results for the model with the best MWIS, rMWIS, MAE, and rMAE 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 Flusion 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 FluSurv-NET data, the use of a fourth root data transform, and the importance of the features that were used by the GBQR model.

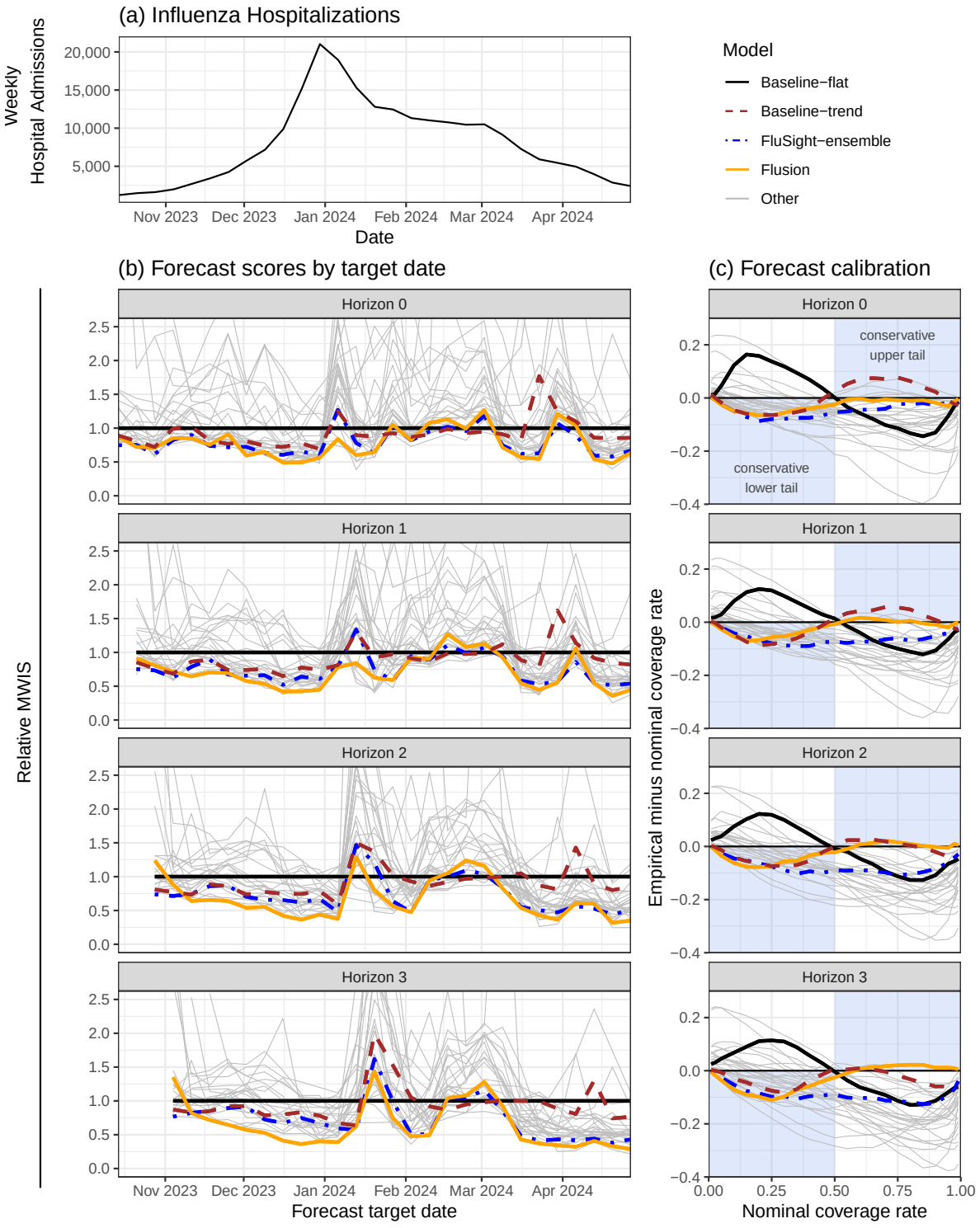

Figure 4: Influenza data and evaluation results. Panel (a): Weekly influenza hospital admissions reported in NHSN for the 2023/24 season, aggregated across all forecasted state-level locations. Panel (b): rMWIS for models contributing to FluSight, by forecast horizon (panels) and target date (horizontal axis). Lower rMWIS indicates better forecast performance. rMWIS 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, indicated with blue shading.

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 GBQR, GBQR-no-level, and ARX models and the quantile averaging ensemble method that were used for Flusion 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.

In this comparison, the most important determinant of performance was whether or not an ensemble included the GBQR model (Table 3, Experiment A). Score differences among the top four model variations were small, and all of those model variations included the GBQR model either alone or in combination with GBQR-no-level and/or ARX. There was a drop-off in performance for other variations that did not include GBQR. Thus, the fact that Flusion was constructed as an ensemble of methodologically distinct models was not a key driver of its performance. Indeed, GBQR alone and GBQR-no-level alone would each have placed first among all FluSight submissions if they had been submitted instead of Flusion, while ARX would have placed third among all contributing models. (We would note that gradient boosting methods are often considered to be ensemble methods, and remind the reader that we used bagging, another ensembling approach, within the GBQR method. Our observation here is that GBQR alone performed about as well as ensembles including GBQR alongside other distinct components.)

The GBQR-no-level model, which was trained without access to features measuring the local level of the time series, had worse performance than the primary GBQR 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 GBQR model to serve as a useful ensemble member. In contrast, although ARX was the worst of our individual models, 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 GBQR model was trained jointly on data from all three surveillance signals (NHSN, FluSurv-NET, and ILI+) and on data for all locations at the state, HHS regional, and US national levels. To investigate the value of this joint model training approach, we considered two alternative methods:

1. The GBQR-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 GBQR-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 GBQR model in terms of MWIS, MAE, and interval coverage rates (Table 3, Experiment B). GBQR-by-location would have been among the top three contributing models to FluSight, and GBQR-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.

Experiment A: Component model performance

Model	% Submitted	MWIS	rMWIS	MAE	rMAE	50% Cov.	95% Cov.
GBQR, 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
GBQR	100.0	30.3	0.625	46.3	0.682	0.529	**0.947**
GBQR, GBQR-no-level	100.0	30.4	0.628	47.1	0.694	0.546	0.958
GBQR-no-level, ARX	100.0	33.2	0.685	52.2	0.769	0.528	0.958
GBQR-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	rMWIS	MAE	rMAE	50% Cov.	95% Cov.
Flusion	100.0	**30.2**	**0.622**	46.6	0.686	0.558	0.963
GBQR	100.0	30.3	0.625	**46.3**	**0.682**	**0.529**	**0.947**
GBQR-by-location	100.0	37.8	0.780	57.9	0.854	0.327	0.891
GBQR-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	rMWIS	MAE	rMAE	50% Cov.	95% Cov.
GBQR-no-reporting-adj	100.0	**29.1**	**0.600**	**44.4**	**0.654**	0.510	0.940
Flusion	100.0	30.2	0.622	46.6	0.686	0.558	0.963
GBQR	100.0	30.3	0.625	46.3	0.682	0.529	0.947
GBQR-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 3: 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 GBQR model, which is trained jointly on data for all locations and data sources, and variations trained separately for each location (GBQR-by-location) and trained only on hospital admissions from NHSN (GBQR-only-NHSN). Experiment C gives results for a variation on the GBQR model that does not incorporate reporting adjustments designed to improve the degree to which ILINet and FluSurv-NET data reflect influenza activity (GBQR-no-reporting-adj) and a variation that does not use a fourth-root transform (GBQR-no-transform), along with the original GBQR model which uses the reporting adjustments and the fourth-root transform. In Experiments B and C, we did not reconstruct the Flusion ensemble with the alternative GBQR formulations; we include Flusion as a reference, but appropriate comparisons are among variations on the GBQR model. 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, rMWIS, MAE, and rMAE are highlighted. Results for the models where empirical PI coverage rates are closest to the nominal levels are highlighted. Models are ordered by rMWIS, with better-performing models at the top of each table.

7.3 Data preprocessing

In a third experiment, we fitted two model variations to investigate the value of some of the data preprocessing steps we used. The GBQR-no-reporting-adj model omitted the adjustments described in Section 2 that were intended to address reporting inconsistencies in the ILINet and FluSurv-NET 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 FluSurv-NET rather than attempting to account for time-varying case capture rates in the FluSurv-NET system. In our evaluations, GBQR-no-reporting-adj outperformed the original GBQR model by a small amount (Table 3, Experiment C). In our model formulation, 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 GBQR-no-transform model was fit to data without using a power transform, though other preprocessing steps described in section 2 were used, including converting hospital admissions 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 GBQR-no-transform model had slightly worse performance than the original GBQR model, indicating that the power transform was helpful (Table 3, 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 GBQR model (supplemental section 4). 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 indicator 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

The Flusion model documented in this manuscript was the top-ranked model in the FluSight forecasting exercise for the 2023/24 season as measured by MAE and MWIS, and its probabilistic calibration was comparable to or better than that of other participating models. The experimental results presented here indicate that this strong performance was primarily driven by the use of a gradient boosting model that was trained jointly on data from multiple surveillance signals and locations. In contrast, other modeling decisions we made had a more minor impact on forecast accuracy. For example, forming predictions as an ensemble of the GBQR model and a more classical ARX model yielded only a small gain in performance, and our attempts to compensate for irregularities in reporting for FluSurv-NET and ILI data were counterproductive.

Not all modelers that contributed to FluSight provided a description of their methods, but we reviewed the available model metadata [14] and identified at least 3 other models that in some way used data for multiple surveillace signals and trained jointly on data for multiple locations, along with an additional 13 that used multiple surveillance signals and another 8 that were trained jointly on data for multiple locations. As with previous forecast evaluation efforts that have not found consistent associations between individual model design decisions and forecast skill [e.g., 36], these methods had a range of performance rankings. This indicates the challenge of identifying important modeling decisions by comparing scores from models that differ in many aspects of their model formulations, and highlights a strength of our study: the use of ablation studies to investigate the value-added from specific aspects of our model setup while controlling for other parts of the

model design. That said, the second-best individual FluSight model overall (denoted as "Other Model #1" in Table 2) also used gradient boosting and multiple surveillance signals [40], which lends additional support to our finding that these design choices were helpful.

A limitation of the results presented in this manuscript is that they report on performance only for a single season in the United States. Ongoing evaluation of the methods we have outlined will be necessary to ensure that the strong performance we documented here generalizes across multiple influenza seasons. Additionally, it would be valuable to understand how our methods would perform in the face of an influenza pandemic. The strong performance of a similar model using gradient boosting to forecast COVID offers some reassurance on this front [33], but our reliance on a long history of training data representing seasonal influenza activity may impact model performance in other settings.

There are numerous avenues for improving on our methods, and we are pursuing some of these in future work. Our intuition is that it would be valuable to use contemporaneous observations of multiple signals, such as NHSN admissions, FluSurv-NET, and ILI together, to inform predictions of trends in disease incidence. Care will need to be taken with this since the relative magnitudes of these signals can vary across geographies and over time. Along these lines, it may also be possible to use other signals, such as insurance claims or internet activity, to improve forecast accuracy [e.g., 42, 37]. A challenge with any approach integrating multiple data streams in real time is that they may have different reporting lags and different revision behaviors.

Although our models were trained jointly on data from multiple spatial units, the forecasts they produced were not directly informed by the spatial structure. We anticipate that improvements in accuracy could be achieved by including features that measure trends in flu activity in neighboring locations [55]. Alternatively, it may be possible to improve forecast accuracy by reconciling predictions made at multiple hierarchical levels, for example generating probabilistic forecasts at the state, regional, and national levels and subsequently adjusting the predictions so that the forecasts at those scales are consistent with each other [2]. These methods would likely be particularly useful for locations with small populations, which typically have a relatively low signal-to-noise ratio in reported surveillance data.

Forecasting the turning point of an epidemic wave is a challenging problem [9], and as we saw in Section 6.2 both Flusion and the FluSight ensemble often failed to accurately predict the turaround after a local peak in disease activity or the increase leading into a second peak within the same season. We conjecture that it may be possible to improve forecasting performance in these instances through better handling of holiday effects or by extending our model to take into account epidemiological understanding of disease transmission. Improved handling of holiday effects, e.g. through features designed to measure trends accounting for the typical magnitude of holiday effects, could help models to predict a decline in activity immediately after the holidays and might allow the model to better see a second wave coming.

Features derived from mechanistic insights, such as measures of vaccine uptake and efficacy or the circulation of multiple strains of the influenza virus at different times over the course of the season may also be helpful [27]. For instance, a second wave of influenza later in the season is often at least partially driven by a different influenza type or subtype than was circulating early in the season. A model that saw data on disease incidence broken down by flu strain might be better able to predict a second wave in these situations. Information about vaccine uptake and efficacy might help a model predict season severity more accurately. A challenge with these approaches is that timely data may not be available for use in a real-time forecasting exercise.

We expect that the main insights presented in this work regarding the value of using data from multiple surveillance signals and locations could be useful in multiple modeling frameworks and for forecasting infectious diseases other than seasonal influenza. These results are of particular importance to the infectious disease forecasting community since they indicate a path forward in settings where new public health surveillance systems may come online and shut down in a span of a few years. In the absence of a long history of data for a signal of interest, borrowing information from similar data sets can provide important context for models to

learn about patterns of disease transmission.

Acknowledgements

We are grateful to Ryan Tibshirani and Logan Brooks for constructive comments on an initial version of this manuscript, and to Rebecca Sweger for code review.

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] Ben Hamner Addison Howard. *COVID19 Global Forecasting (Week 1)*. 2020. URL: https://www.kaggle.com/competitions/covid19-global-forecasting-week-1/ (visited on 07/24/2024).

[2] George Athanasopoulos et al. "Forecast reconciliation: A review". In: *International Journal of Forecasting* 40.2 (2024), pp. 430–456. ISSN: 0169-2070. DOI: https://doi.org/10.1016/j.ijforecast.2023.10.010.

[3] Roy Batchelor and Pami Dua. "Forecaster diversity and the benefits of combining forecasts". In: *Management Science* 41.1 (1995), pp. 68–75. DOI: 10.1287/mnsc.41.1.68.

[4] Amy E. Benefield, Desiree Williams, and VP Nagraj. "An Imputation-Based Approach for Augmenting Sparse Epidemiological Signals". In: *medRxiv* (2024). DOI: 10.1101/2024.07.31.24311314.

[5] Matthew Biggerstaff et al. "Results from the Centers for Disease Control and Prevention's predict the 2013–2014 Influenza Season Challenge". In: *BMC Infectious Diseases* 16.1 (July 2016), p. 357. ISSN: 1471-2334. DOI: 10.1186/s12879-016-1669-x.

[6] Johannes Bracher et al. "Evaluating epidemic forecasts in an interval format". In: *PLOS Computational Biology* 17.2 (2021), e1008618.

[7] US Census Bureau. *National Population Totals and Components of Change: 2010-2019*. Oct. 2024. URL: https://www.census.gov/data/tables/time-series/demo/popest/2010s-national-total.html (visited on 07/11/2024).

[8] US Census Bureau. *National Population Totals and Components of Change: 2020-2023*. June 2024. URL: https://www.census.gov/data/tables/time-series/demo/popest/2020s-national-total.html (visited on 07/11/2024).

[9] Mario Castro et al. "The turning point and end of an expanding epidemic cannot be precisely forecast". In: *Proceedings of the National Academy of Sciences* 117.42 (2020), pp. 26190–26196. DOI: 10.1073/pnas.2007868117.

[10] Flàvio Codeço Coelho, Nicolaus Linneu de Holanda, and Beatriz Coimbra. "Transfer learning applied to the forecast of mosquito-borne diseases". In: *medRxiv* (2020). DOI: 10.1101/2020.02.03.20020164.

[11] 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.

[12] Marcos Lopez De Prado. *Advances in financial machine learning*. John Wiley & Sons, 2018.

[13] Centers for Disease Control and Prevention. *Event Codes & Other Surveillance Resources*. Apr. 2024. URL: https://ndc.services.cdc.gov/event-codes-other-surveillance-resources/ (visited on 07/09/2024).

[14] Centers for Disease Control and Prevention. *FluSight-forecast-hub*. URL: https://github.com/cdcepi/FluSight-forecast-hub/ (visited on 09/12/2024).

[15] Centers for Disease Control and Prevention. *Influenza Hospitalization Surveillance Network (FluSurv-NET)*. Oct. 2023. URL: https://www.cdc.gov/flu/weekly/influenza-hospitalization-surveillance.htm (visited on 07/03/2024).

[16] Centers for Disease Control and Prevention. *Past Seasons Estimated Influenza Disease Burden*. Feb. 2024. URL: https://www.cdc.gov/flu/about/burden/past-seasons.html (visited on 07/03/2024).

[17] Centers for Disease Control and Prevention. *U.S. Influenza Surveillance: Purpose and Methods*. Oct. 2023. URL: https://www.cdc.gov/flu/weekly/overview.htm (visited on 07/03/2024).

[18] David Farrow. "Modeling the Past, Present, and Future of Influenza". PhD Thesis. Pittsburgh, PA: Carnegie Mellon University, 2016. URL: http://reports-archive.adm.cs.cmu.edu/anon/cbd/CMU-CB-16-101.pdf (visited on 07/09/2024).

[19] Jerome H. Friedman. "Greedy function approximation: a gradient boosting machine". In: *Annals of Statistics* 29.5 (2001), pp. 1189–1232.

[20] Tilmann Gneiting. "Quantiles as optimal point forecasts". In: *International Journal of Forecasting* 27.2 (2011), pp. 197–207. ISSN: 0169-2070. DOI: 10.1016/j.ijforecast.2009.12.015.

[21] 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.

[22] Thomas Grubinger, Georgios C. Chasparis, and Thomas Natschläger. "Generalized online transfer learning for climate control in residential buildings". In: *Energy and Buildings* 139 (2017), pp. 63–71. ISSN: 0378-7788. DOI: 10.1016/j.enbuild.2016.12.074.

[23] Ben Hamner. *COVID19 Global Forecasting (Week 2)*. 2020. URL: https://www.kaggle.com/competitions/covid19-global-forecasting-week-2/ (visited on 07/24/2024).

[24] Matthew D. Hoffman, Andrew Gelman, et al. "The No-U-Turn sampler: adaptively setting path lengths in Hamiltonian Monte Carlo." In: *J. Mach. Learn. Res.* 15.1 (2014), pp. 1593–1623.

[25] Maria Jahja, Andrew Chin, and Ryan J. Tibshirani. "Real-Time Estimation of COVID-19 Infections: Deconvolution and Sensor Fusion". In: *Statistical Science* 37.2 (May 2022). Publisher: Institute of Mathematical Statistics, pp. 207–228. ISSN: 0883-4237, 2168-8745. DOI: 10.1214/22-STS856.

[26] Michael A. Johansson et al. "An open challenge to advance probabilistic forecasting for dengue epidemics". In: *Proceedings of the National Academy of Sciences of the United States of America* 116.48 (Nov. 2019), pp. 24268–24274. ISSN: 0027-8424. DOI: 10.1073/pnas.1909865116.

[27] Sasikiran Kandula, Wan Yang, and Jeffrey Shaman. "Type- and Subtype-Specific Influenza Forecast". In: *American Journal of Epidemiology* 185.5 (Feb. 2017), pp. 395–402. ISSN: 0002-9262. DOI: 10.1093/aje/kww211.

[28] Guolin Ke et al. "LightGBM: a highly efficient gradient boosting decision tree". In: *Proceedings of the 31st International Conference on Neural Information Processing Systems*. NIPS'17. Long Beach, California, USA: Curran Associates Inc., 2017, pp. 3149–3157. ISBN: 9781510860964.

[29] A. David Lainder and Russell D. Wolfinger. "Forecasting with gradient boosted trees: augmentation, tuning, and cross-validation strategies: Winning solution to the M5 Uncertainty competition". In: *International Journal of Forecasting* 38.4 (2022), pp. 1426–1433.

[30] Sequoia I. Leuba et al. "Tracking and predicting U.S. influenza activity with a real-time surveillance network". In: *PLOS Computational Biology* 16.11 (Nov. 2020). Publisher: Public Library of Science, e1008180. ISSN: 1553-7358. DOI: 10.1371/journal.pcbi.1008180.

[31] Kenneth C. Lichtendahl Jr, Yael Grushka-Cockayne, and Robert L. Winkler. "Is it better to average probabilities or quantiles?" In: *Management Science* 59.7 (2013), pp. 1594–1611.

[32] Marc Lipsitch et al. "Improving the Evidence Base for Decision Making During a Pandemic: The Example of 2009 Influenza A/H1N1". In: *Biosecurity and Bioterrorism: Biodefense Strategy, Practice, and Science* 9.2 (June 1, 2011), pp. 89–115. ISSN: 1538-7135. DOI: 10.1089/bsp.2011.0007. URL: https://www.liebertpub.com/doi/full/10.1089/bsp.2011.0007 (visited on 11/30/2021).

[33] Velma K. Lopez et al. "Challenges of COVID-19 Case Forecasting in the US, 2020–2021". In: *PLOS Computational Biology* 20.5 (2024), e1011200. DOI: 10.1371/journal.pcbi.1011200.

[34] Chelsea S. Lutz et al. "Applying infectious disease forecasting to public health: a path forward using influenza forecasting examples". In: *BMC Public Health* 19.1 (Dec. 2019), p. 1659. ISSN: 1471-2458. DOI: 10.1186/s12889-019-7966-8.

[35] Spyros Makridakis, Evangelos Spiliotis, and Vassilios Assimakopoulos. "M5 accuracy competition: Results, findings, and conclusions". In: *International Journal of Forecasting* 38.4 (2022), pp. 1346–1364. DOI: 10.1016/j.ijforecast.2021.11.013.

[36] Sarabeth M Mathis et al. "Evaluation of FluSight influenza forecasting in the 2021–22 and 2022–23 seasons with a new target laboratory-confirmed influenza hospitalizations". In: *Nature communications* 15.1 (2024), p. 6289. DOI: 10.1038/s41467-024-50601-9.

[37] Daniel J. McDonald et al. "Can auxiliary indicators improve COVID-19 forecasting and hotspot prediction?" In: *Proceedings of the National Academy of Sciences of the United States of America* 118.51 (Dec. 2021), e2111453118. ISSN: 0027-8424. DOI: 10.1073/pnas.2111453118.

[38] Craig J. McGowan et al. "Collaborative efforts to forecast seasonal influenza in the United States, 2015–2016". In: *Scientific Reports* 9.1 (Jan. 2019), p. 683. ISSN: 2045-2322. DOI: 10.1038/s41598-018-36361-9.

[39] Dirk Merkel. "Docker: lightweight linux containers for consistent development and deployment". In: *Linux journal* (2014). URL: https://www.linuxjournal.com/content/docker-lightweight-linux-containers-consistent-development-and-deployment.

[40] Austin G Meyer et al. "A prospective real-time transfer learning approach to estimate Influenza hospitalizations with limited data". In: *medRxiv* (2024). DOI: 10.1101/2024.07.17.24310565.

[41] Pablo Montero-Manso and Rob J. Hyndman. "Principles and algorithms for forecasting groups of time series: Locality and globality". In: *International Journal of Forecasting* 37.4 (2021), pp. 1632–1653. ISSN: 0169-2070. DOI: 10.1016/j.ijforecast.2021.03.004.

[42] Dave Osthus, Ashlynn R. Daughton, and Reid Priedhorsky. "Even a good influenza forecasting model can benefit from internet-based nowcasts, but those benefits are limited". In: *PLOS Computational Biology* 15.2 (Feb. 2019), e1006599. ISSN: 1553-7358. DOI: 10.1371/journal.pcbi.1006599.

[43] Dave Osthus and Kelly R. Moran. "Multiscale influenza forecasting". In: *Nature Communications* 12.1 (May 2021), p. 2991. ISSN: 2041-1723. DOI: 10.1038/s41467-021-23234-5.

[44] Sinno Jialin Pan and Qiang Yang. "A survey on transfer learning". In: *IEEE Transactions on knowledge and data engineering* 22.10 (2009), pp. 1345–1359. DOI: 10.1109/TKDE.2009.191.

[45] Du Phan, Neeraj Pradhan, and Martin Jankowiak. *Composable Effects for Flexible and Accelerated Probabilistic Programming in NumPyro*. 2019. DOI: 10.48550/arXiv.1912.11554.

[46] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. Vienna, Austria, 2024. URL: https://www.R-project.org/.

[47] Evan L. Ray et al. "Comparing trained and untrained probabilistic ensemble forecasts of COVID-19 cases and deaths in the United States". In: *International Journal of Forecasting* 39.3 (July 2023), pp. 1366–1383. ISSN: 0169-2070. DOI: 10.1016/j.ijforecast.2022.06.005.

[48] Nicholas G. Reich et al. "A collaborative multiyear, multimodel assessment of seasonal influenza forecasting in the United States". In: *Proceedings of the National Academy of Sciences of the United States of America* 116.8 (Feb. 2019), pp. 3146–3154. ISSN: 1091-6490. DOI: 10.1073/pnas.1812594116.

[49] Nicholas G. Reich et al. "Accuracy of real-time multi-model ensemble forecasts for seasonal influenza in the U.S". In: *PLOS Computational Biology* 15.11 (Nov. 2019), e1007486. ISSN: 1553-7358. DOI: 10.1371/journal.pcbi.1007486.

[50] Julia Reis et al. "Superensemble forecast of respiratory syncytial virus outbreaks at national, regional, and state levels in the United States". In: *Epidemics* 26 (Mar. 2019), pp. 1–8. ISSN: 1755-4365. DOI: 10.1016/j.epidem.2018.07.001.

[51] Mauro Ribeiro et al. "Transfer learning with seasonal and trend adjustment for cross-building energy forecasting". In: *Energy and Buildings* 165 (2018), pp. 352–363. ISSN: 0378-7788. DOI: 10.1016/j.enbuild.2018.01.034.

[52] G. van Rossum. *Python tutorial*. Tech. rep. CS-R9526. Amsterdam: Centrum voor Wiskunde en Informatica (CWI), May 1995.

[53] Kirstin Roster, Colm Connaughton, and Francisco A. Rodrigues. "Forecasting new diseases in low-data settings using transfer learning". In: *Chaos, Solitons & Fractals* 161 (2022), p. 112306. ISSN: 0960-0779. DOI: 10.1016/j.chaos.2022.112306.

[54] Jeffrey Shaman et al. "Real-time influenza forecasts during the 2012–2013 season". In: *Nature Communications* 4.1 (2013). DOI: 10.1038/ncomms3837.

[55] Gabrielle Thivierge, Aaron Rumack, and F. William Townes. "Does Spatial Information Improve Influenza Forecasting?" In: *arXiv* (2024). DOI: 10.48550/arXiv.2408.12722.

[56] Cécile Viboud et al. "The RAPIDD Ebola forecasting challenge: Synthesis and lessons learnt". In: *Epidemics* 22 (Mar. 2018), pp. 13–21. ISSN: 1878-0067. DOI: 10.1016/j.epidem.2017.08.002.

[57] Stella Burnham Vincent. *The Functions of the Vibrissae in the Behavior of the White Rat*. Vol. 1. 5. University of Chicago, 1912.

[58] Jacco Wallinga, Michiel van Boven, and Marc Lipsitch. "Optimizing infectious disease interventions during an emerging epidemic". In: *Proceedings of the National Academy of Sciences* 107.2 (Jan. 12, 2010), pp. 923–928. ISSN: 0027-8424, 1091-6490. DOI: 10.1073/pnas.0908491107. URL: https://www.pnas.org/content/107/2/923 (visited on 11/30/2021).

[59] Addison Howard Walter Reade. *COVID19 Global Forecasting (Week 3)*. 2020. URL: https://www.kaggle.com/competitions/covid19-global-forecasting-week-3/ (visited on 07/24/2024).

[60] Addison Howard Walter Reade. *COVID19 Global Forecasting (Week 4)*. 2020. URL: https://www.kaggle.com/competitions/covid19-global-forecasting-week-4/ (visited on 07/24/2024).

[61] Addison Howard Walter Reade. *COVID19 Global Forecasting (Week 5)*. 2020. URL: https://www.kaggle.com/competitions/covid19-global-forecasting-week-5/ (visited on 07/24/2024).

[62] Yihui Xie. "knitr: A Comprehensive Tool for Reproducible Research in R". In: *Implementing Reproducible Computational Research*. Ed. by Victoria Stodden, Friedrich Leisch, and Roger D. Peng. ISBN 978-1466561595. Chapman and Hall/CRC, 2014.

[63] Teresa K. Yamana, Sasikiran Kandula, and Jeffrey Shaman. "Superensemble forecasts of dengue outbreaks". In: *Journal of The Royal Society Interface* 13.123 (Oct. 2016). Publisher: Royal Society, p. 20160410. DOI: 10.1098/rsif.2016.0410.

[64] Shihao Yang, Mauricio Santillana, and Samuel C. Kou. "Accurate estimation of influenza epidemics using Google search data via ARGO". In: *Proceedings of the National Academy of Sciences* 112.47 (2015). Publisher: National Acad Sciences, pp. 14473–14478. DOI: 10.1073/pnas.1515373112.

[65] Bin Zou, Vasileios Lampos, and Ingemar Cox. "Multi-task learning improves disease models from web search". In: *Proceedings of the 2018 world wide web conference*. 2018, pp. 87–96. DOI: 10.1145/3178876.3186050.