Forecasting influenza by training on multiple data sources

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

February 22, 2024

Abstract

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.

2 Data Sources

Our model uses influenza data from four data sources, which are combined to produce three signals measuring the intensity of influenza activity over time and space:

- 'NHSN': Influenza hospital admissions as reported at the state level in a data set that is collected as part of the National Healthcare Safety Network (NHSN)
- 'ILI+': An approximate measure of
- 'FluSurv':

3 Model

Our model was constructed as an ensemble of statistical time series and machine learning models. We begin describe the component models in sections **??** and **??** and the ensemble methods in section

??. 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 **??**.

3.1 Component Model 1: SARIX

3.2 Component Model 2: GBM

3.3 Ensembling Methods: Quantile averaging and linear pools

3.4 Model adjustments used for real time forecasts

4 Evaluation metrics

In the sections below, we evaluate forecasts using two metrics: the weighted interval score (WIS) and quantile coverage rates.

We also compute relative versions of these metrics.

5 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.

5.1 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.

5.2 Evaluation results

TODO: finalize language here/check that statements made hold up at the end of the season.

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 1). Qualitatively, it appears that both the FluSight-ensemble and the Flusion model 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.

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 2 (b)).

Additionally, while its prediction interval coverage rates tended to be underconfident (i.e., prediction intervals were too wide on average), the 95% PI coverage rates for Flusion came closest to the nominal coverage rate among all models (Table 1). An examination of one-sided quantile coverage rates confirms that Flusion was unique among models contributed to the Hub in that its coverage rates for high quantile levels were too high (indicating that the upper tail of the predictive distributions tended to fall above the eventually observed data more often than expected), while nearly all other models missed in the upper tails more often than expected (Figure 2 (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?

6 Post hoc model exploration

In this section, we conduct some additional investigations designed to inform an understanding of why this model performed as well as it did. Specifically, the questions motivating these analyses are:

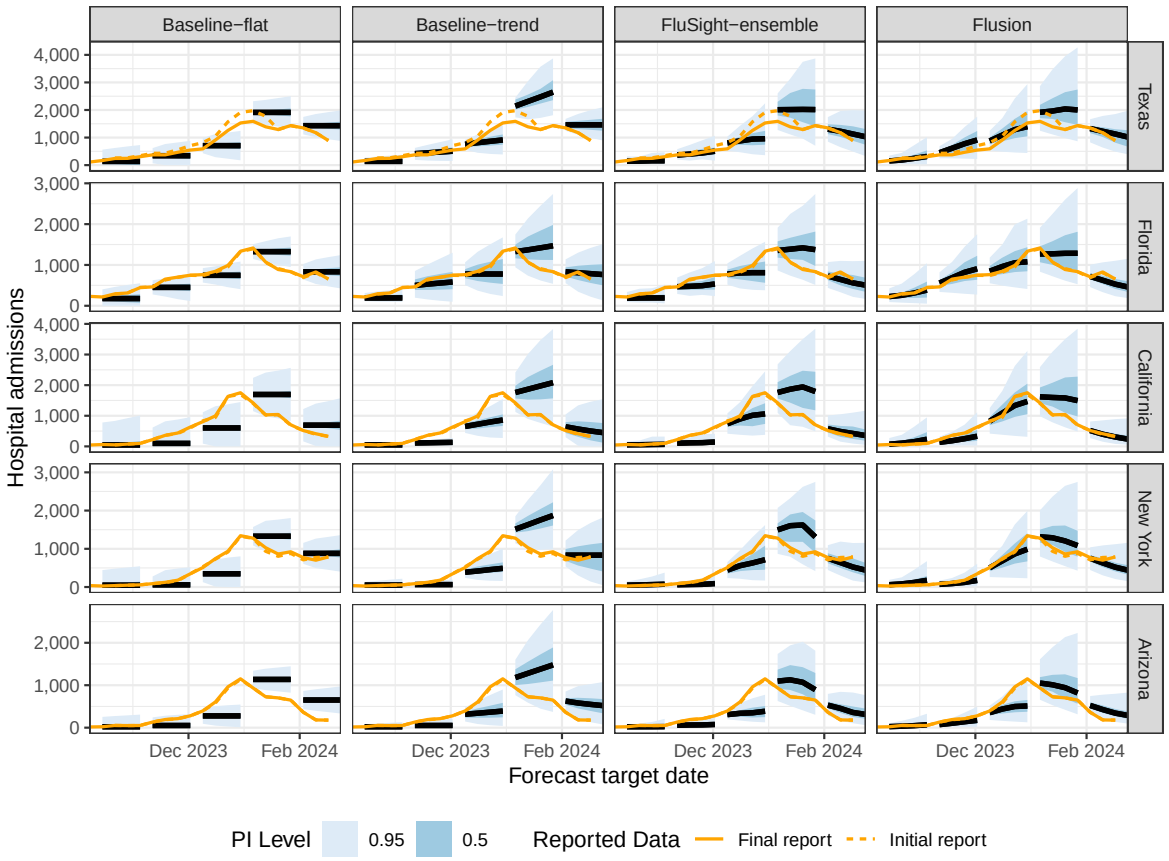

Figure 1: Influenza data and forecasts for the five states with the largest cumulative hospital admissions during the 2023/24 season. To avoid overplotting, in this figure forecasts from every fourth reference date are shown; evaluations include all reference dates. Forecasts are represented by the predictive median (black lines) and 50% and 95% prediction intervals (blue shaded regions). Solid orange lines show the finalized admission counts reported as of TODO, while dotted orange lines show the initial reported values that were available on the date predictions were generated.

1. are the individual GBM and SARIX models good, or did we get benefits from ensembling?
2. how helpful is training on other data streams?
3. this season was relatively early (only 1-2 weeks later than last season) and had peak incidence nearly identical to last season's peak. how much did we benefit from that? how much would performance degrade in larger/smaller/later seasons?
4. why does the model seem to respond so strongly to local level features and what can we do about it?

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.

Model	% Submitted	MWIS	RWIS	MAE	RAE	50% PI Cov.	95% PI Cov.
Flusion	99.9	**33.8**	**0.561**	**51.7**	**0.635**	0.595	**0.966**
Other Model #1	100.0	40.7	0.675	62.4	0.766	0.520	0.929
Other Model #2	97.8	44.6	0.729	68.3	0.827	0.382	0.827
FluSight-ensemble	100.0	44.3	0.738	69.2	0.851	0.431	0.886
Other Model #3	88.9	51.9	0.773	78.1	0.859	0.412	0.854
Other Model #4	100.0	47.0	0.779	68.0	0.834	0.479	0.836
Other Model #5	100.0	50.3	0.836	74.8	0.921	0.460	0.873
Other Model #6	100.0	53.0	0.876	79.8	0.980	0.443	0.928
Other Model #7	83.3	62.3	0.879	89.3	0.929	0.298	0.733
Baseline-trend	99.9	52.9	0.884	81.2	1.000	0.524	0.879
Other Model #8	95.7	54.1	0.887	78.2	0.949	**0.509**	0.815
Other Model #9	100.0	54.9	0.910	85.3	1.050	0.387	0.855
Other Model #10	100.0	55.1	0.912	77.1	0.945	0.312	0.716
Other Model #11	98.7	54.2	0.920	66.7	0.838	0.442	0.652
Other Model #12	94.4	53.3	0.983	79.5	1.090	0.350	0.836
Other Model #13	92.3	57.6	0.993	92.2	1.180	0.350	0.886
Baseline-flat	100.0	60.0	1.000	81.0	1.000	0.279	0.831
Other Model #14	94.3	63.8	1.010	77.0	0.902	0.225	0.644
Other Model #15	82.9	58.7	1.040	78.8	1.040	0.241	0.652
Other Model #16	90.3	66.3	1.050	96.5	1.130	0.360	0.729
Other Model #17	100.0	65.6	1.080	91.1	1.110	0.292	0.647
Other Model #18	96.8	72.4	1.180	106.0	1.280	0.347	0.789
Other Model #19	90.0	71.7	1.220	84.8	1.070	0.231	0.468
Other Model #20	88.9	75.9	1.240	95.8	1.160	0.223	0.508
Other Model #21	88.6	92.9	1.390	127.0	1.400	0.461	0.758
Other Model #22	88.8	179.0	2.790	214.0	2.470	0.137	0.354

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 75% 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.

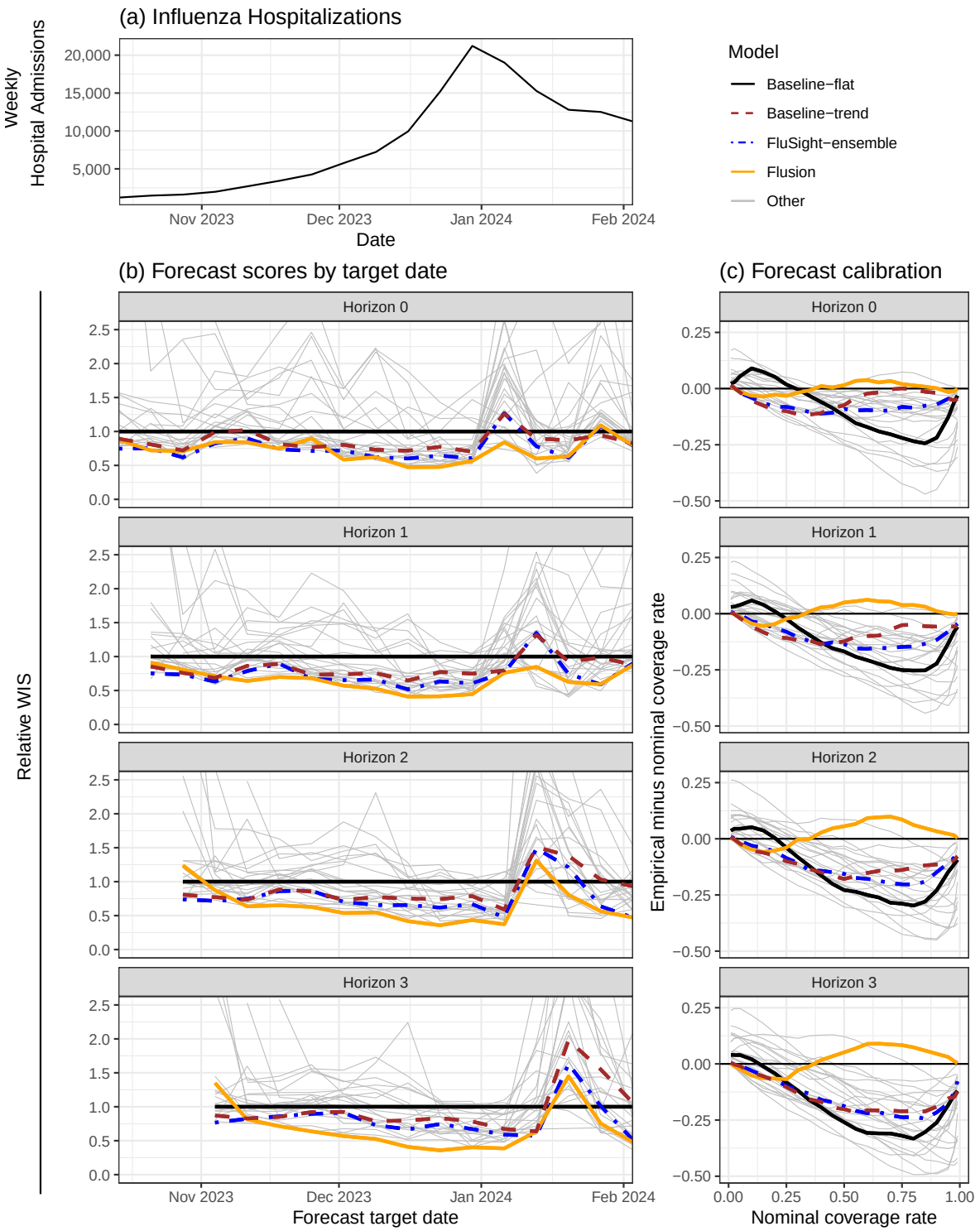

Figure 2: Influenza data and evaluation results. Panel (a): Target influenza hospital admissions data for the 2023/24 season, aggregated across all forecasted state-level geographic units. Panel (b): RWIS for models contributing to the FluSight Forecast Hub, stratified by forecast horizon (panels) and target date (horizontal axis). Lower relative WIS indicates better forecast performance. To focus on areas of interest, 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 that are less than 0.5 and a positive differential at nominal coverage rates greater than 0.5. In panels (b) and (c), we highlight performance for the Flusion model, two baselines, and the FluSight ensemble; results for other models contributed to the hub are shown in light grey.