

Flusion:

Integrating multiple data sources for
accurate influenza predictions

Evan L. Ray, Yijin Wang, Russel D. Wolfinger, Nicholas G. Reich
University of Massachusetts, Amherst

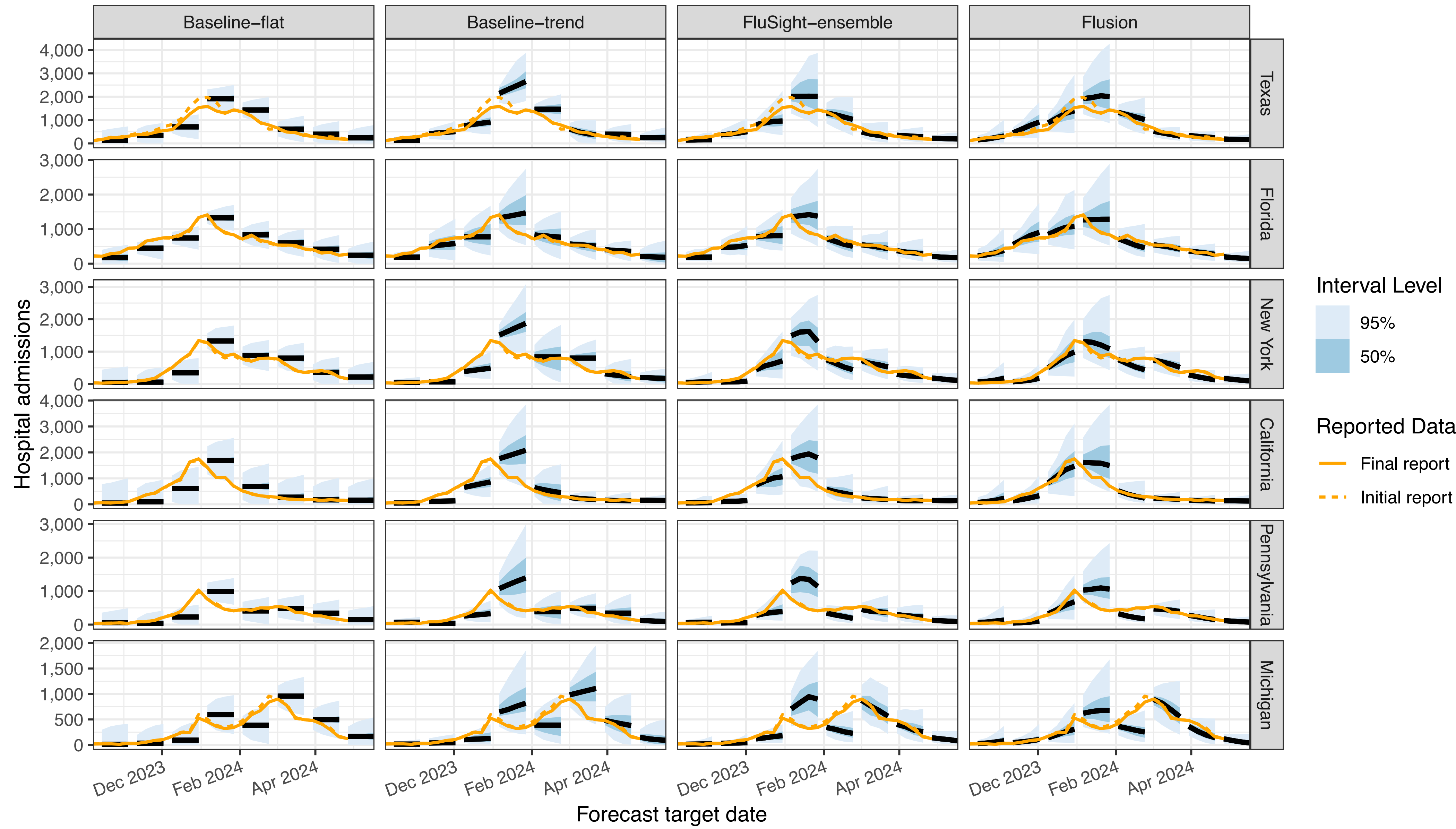
Joint Statistical Meetings
August 6, 2024



Overview of this talk

- **Preview of results**
- Our model
 - Data
 - Model Setup
- Ablation Experiments
- Conclusions

First look: FluSight forecasts, 2023/24 season



Overall Results: FluSight

		Model	% Submitted	MWIS	rMWIS	MAE	rMAE	50% Cov.	95% Cov.
Higher rank Better Performance ↑		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
Lower rank Worse Performance ↓		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

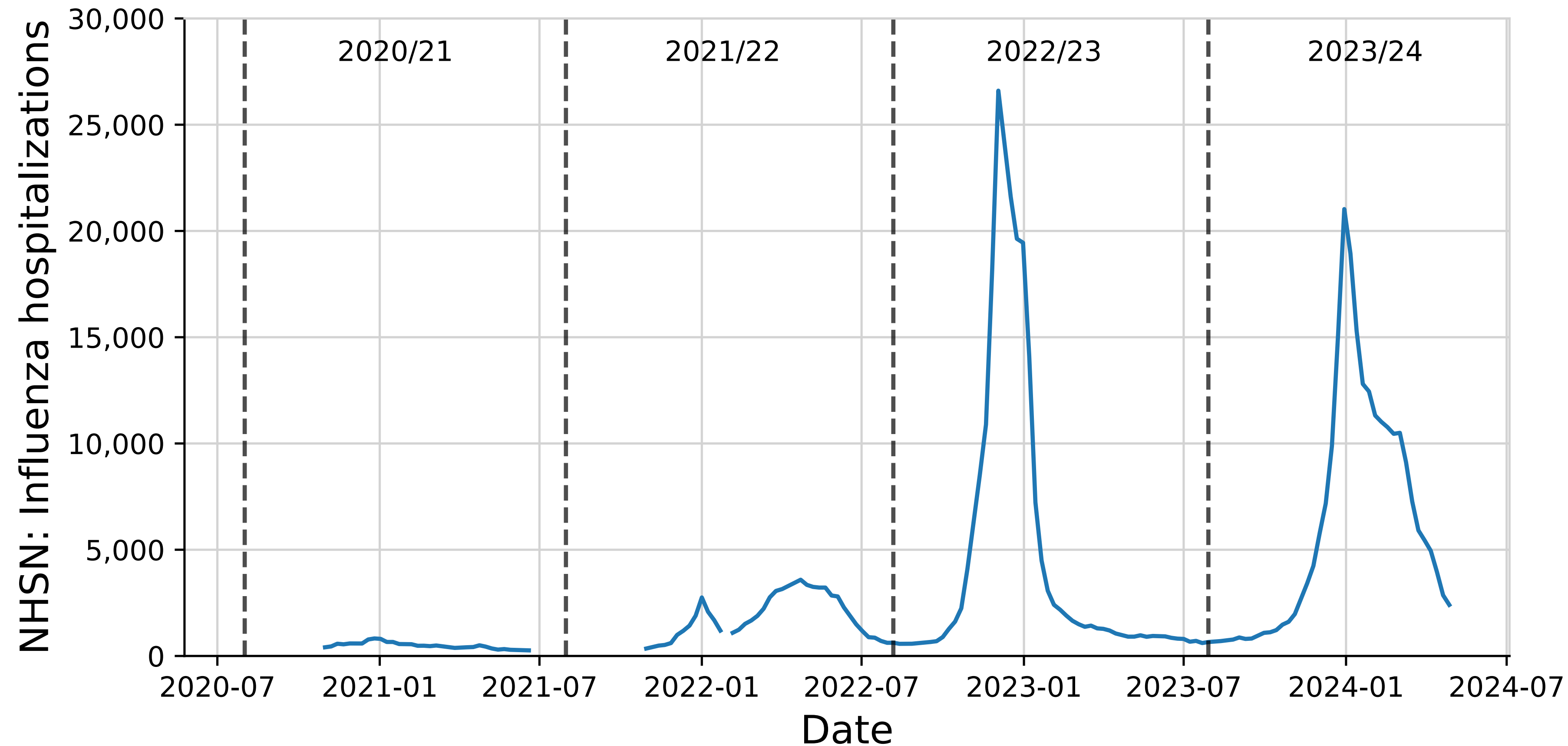
(Results for 11 lower-ranked models are suppressed for brevity)

Overview of this talk

- Preview of results
- **Our model**
 - **Data**
 - Model Setup
- Ablation Experiments
- Conclusions

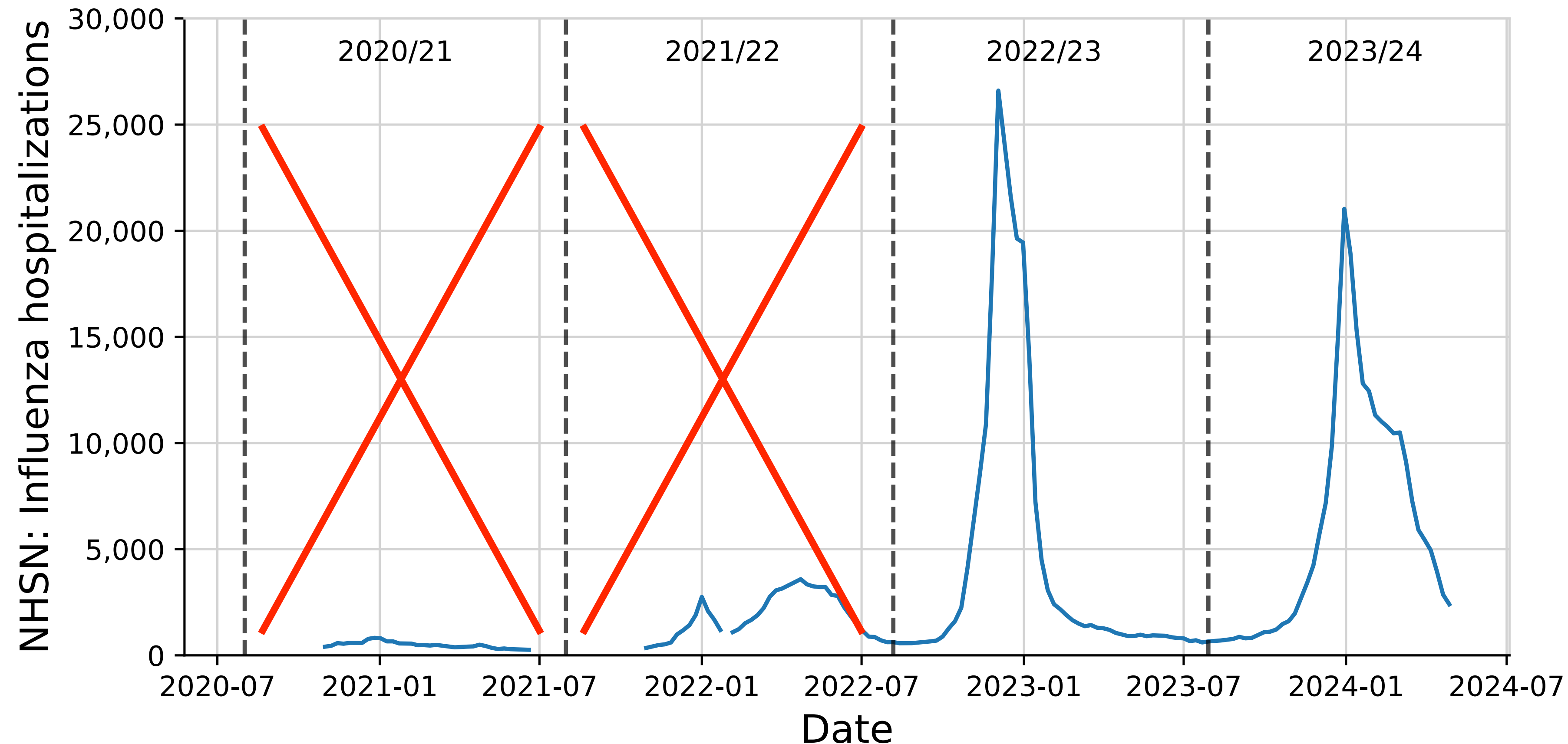
A central data challenge

- Since the COVID-19 pandemic, FluSight is based on a new data stream:
- Hospitalizations with influenza as reported in National Healthcare Safety Network (NHSN)



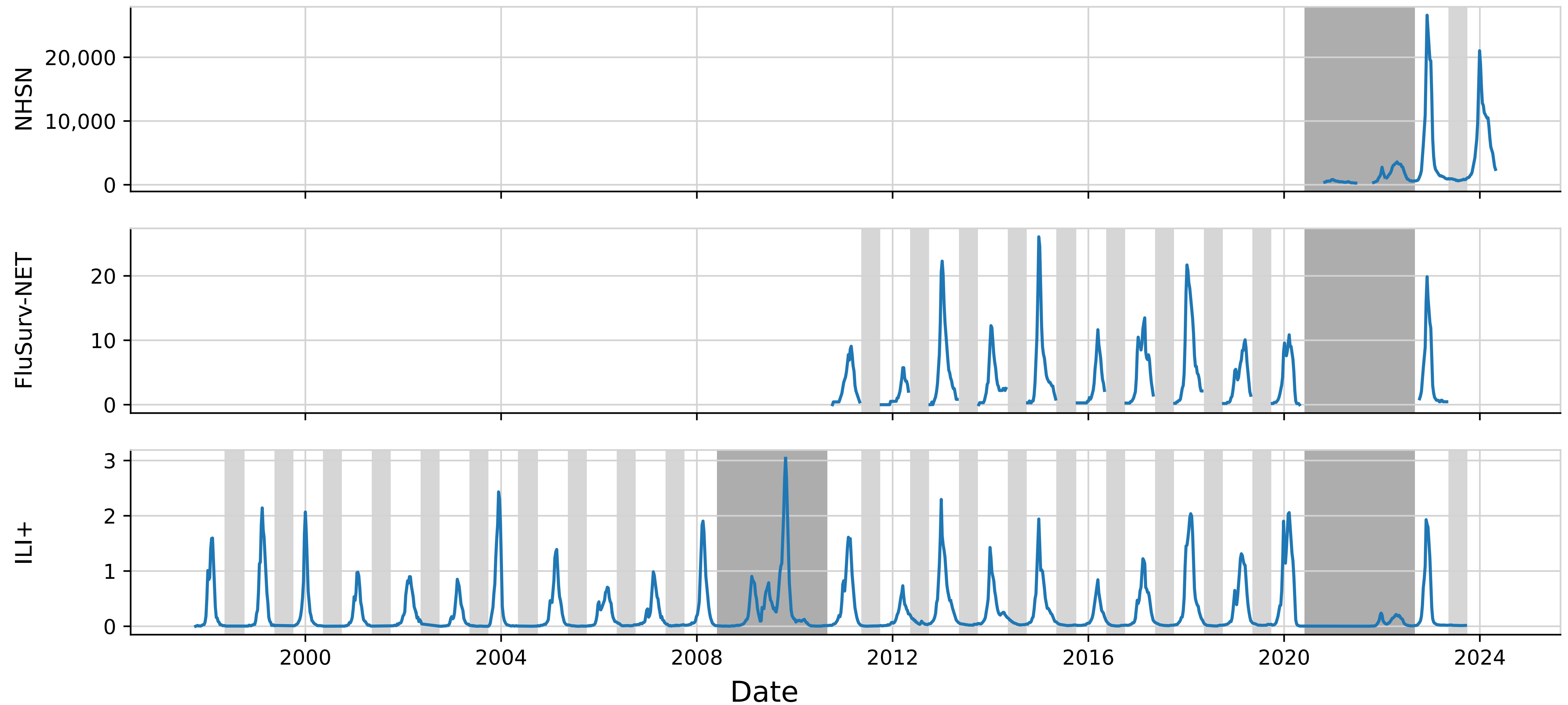
A central data challenge

- Since the COVID-19 pandemic, FluSight is based on a new data stream:
- Hospitalizations with influenza as reported in National Healthcare Safety Network (NHSN)
- This surveillance signal came online during the COVID-19 pandemic
- At 2023/24 season start, only 1 past season of data with typical patterns of flu transmission



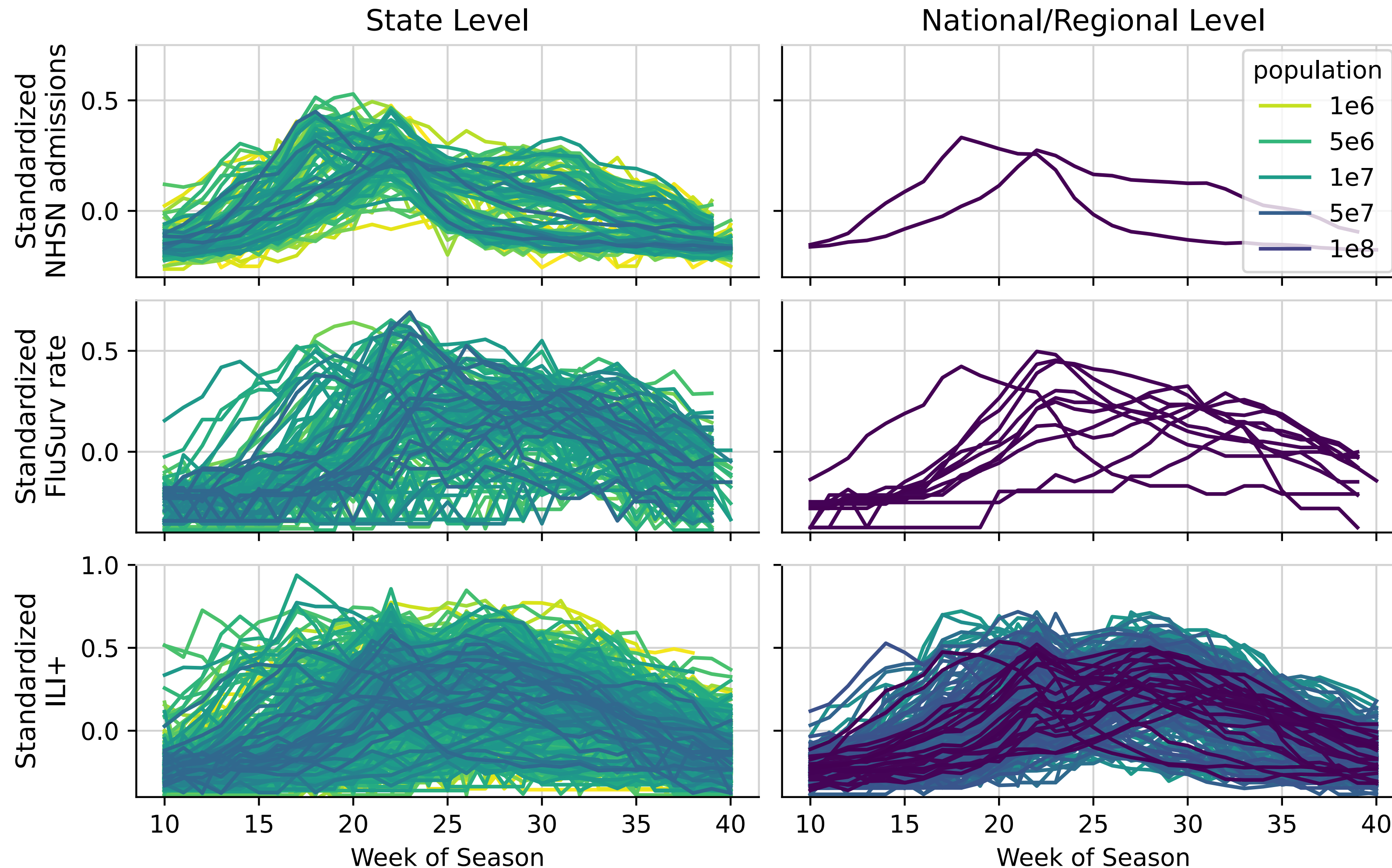
Our solution

- We augment the target NHSN data with 2 other signals with a longer history
 - FluSurv-NET: influenza hospitalizations in selected hospitals
 - ILI+: estimated percent of outpatient doctor visits where patient has influenza



Data preprocessing

- Initial transforms with two goals:
 - Center and scale: put the data on a similar scale for different locations, signals
 - Fourth root: stabilize variance across different times



Overview of this talk

- Preview of results
- **Our model**
 - Data
 - **Model Setup**
- Ablation Experiments
- Conclusions

Flusion: an ensemble of 3 models

- Three component models:
 1. GBQR: A gradient boosting quantile regression model
 - Learns a mapping $f_{\alpha}(x)$ from features x to a predictive quantile at each quantile level α
 - Used 114 features:
 - Measures of local level, trend, curvature in the signal
 - One-hot encoding of location
 - Week of season, week relative to Christmas
 - ...
 - Note: when predicting a target signal and location, features measure information only about that signal and location
 2. GBQR-no-level: Same as GBQR, but not allowed to see measures of local level of signal
 3. ARX: Bayesian autoregressive model with 1 covariate, a spike function peaking at Christmas
- Each model produces a set of predictive quantiles at 23 quantile levels from 0.01 to 0.99
- Flusion takes the average of these quantiles

Overview of this talk

- Preview of results
- Our model
 - Data
 - Model Setup
- **Ablation Experiments**
- Conclusions

Experiment A: Component models


- **Question:** Which component model(s) drove Flusion's performance?
- **Experiment:** Scored individual components and ensembles of component model pairs
- **Results:**

Experiment A: Component model performance								} Top 4 include GBQR
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	

- Primary driver was whether or not GBQR was included
- Both GBQR and GBQR-no-level would have placed first among all FluSight submissions
- ARX was our worst individual component, but ensembles including it were better

Experiment B: Reduced training data

- **Question:** was training jointly on multiple signals and locations helpful?
- **Experiment:** Fit 2 model variations:
 - GBQR-by-location: fit to each location separately, all 3 data sources
 - GBQR-only-NHSN: fit to all locations jointly, only data from NHSN
- **Results:**

Higher rank
Better Performance


Experiment B: Reduced training data							
Model	% Submitted	MWIS	rMWIS	MAE	rMAE	50% Cov.	95% Cov.
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

- Training jointly on data from all locations and data sources was key to strong performance

Summary & conclusions

- Flusion had the top rank among all contributors to FluSight in the 2023/24 season
- Key drivers of its performance were:
 - The use of a gradient boosting model for forecasting
 - Joint training on all locations
 - Joint training on data for the target system and 2 other signals with a longer history
- This approach indicates a way forward in a setting where public health data modernization initiatives may bring new surveillance systems online

Thanks for your attention!

- Funding acknowledgment:

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.

- Thanks to co-authors Yijin Wang, Russell D. Wolfinger, Nicholas G. Reich
- Thanks to Ryan Tibshirani and Logan Brooks for comments on an initial version of this work
- Preprint: <https://arxiv.org/abs/2407.19054>



Forecast evaluation

We use 6 metrics to evaluate forecast accuracy and calibration

- Mean absolute error (MAE)
 - $|m - y|$, where m is the predictive median and y is the observed value
- Mean weighted interval score (MWIS)
 - Let $\{q_k : k = 1, \dots, K\}$ denote a set of predictive quantiles at levels $\alpha_1, \dots, \alpha_K$.

$$WIS(\{q_k : k = 1, \dots, K\}, y) = \frac{1}{K} \sum_k 2 \cdot QS_{\alpha_k}(q_k, y)$$

- $$QS_{\alpha_k}(q_k, z_i) = \alpha_k \max(y - q_k, 0) + (1 - \alpha_k) \max(q_k - y, 0)$$
- Relative MAE (rMAE), Relative MWIS (rMWIS), see next slide
- 50% Interval Coverage, 95% Interval Coverage
 - What proportion of the time did central prediction intervals include the eventually observed value?

Relative score metrics

Challenge:

- different forecasters submit predictions for different locations and dates
- MAE and WIS are sensitive to the scale of the prediction target
- MAE and WIS values for forecasts in different locations and dates are not comparable

Our approach has 3 steps:

1. For each pair of models m and m' , compute the MAE (or MWIS) on the subset of location/dates they have in common, denoted by $MAE_{\mathcal{J}_{m,m'}}^m$ and $MAE_{\mathcal{J}_{m,m'}}^{m'}$

2. For model m , compute the geometric mean of ratios of MAEs for m compared to all other models

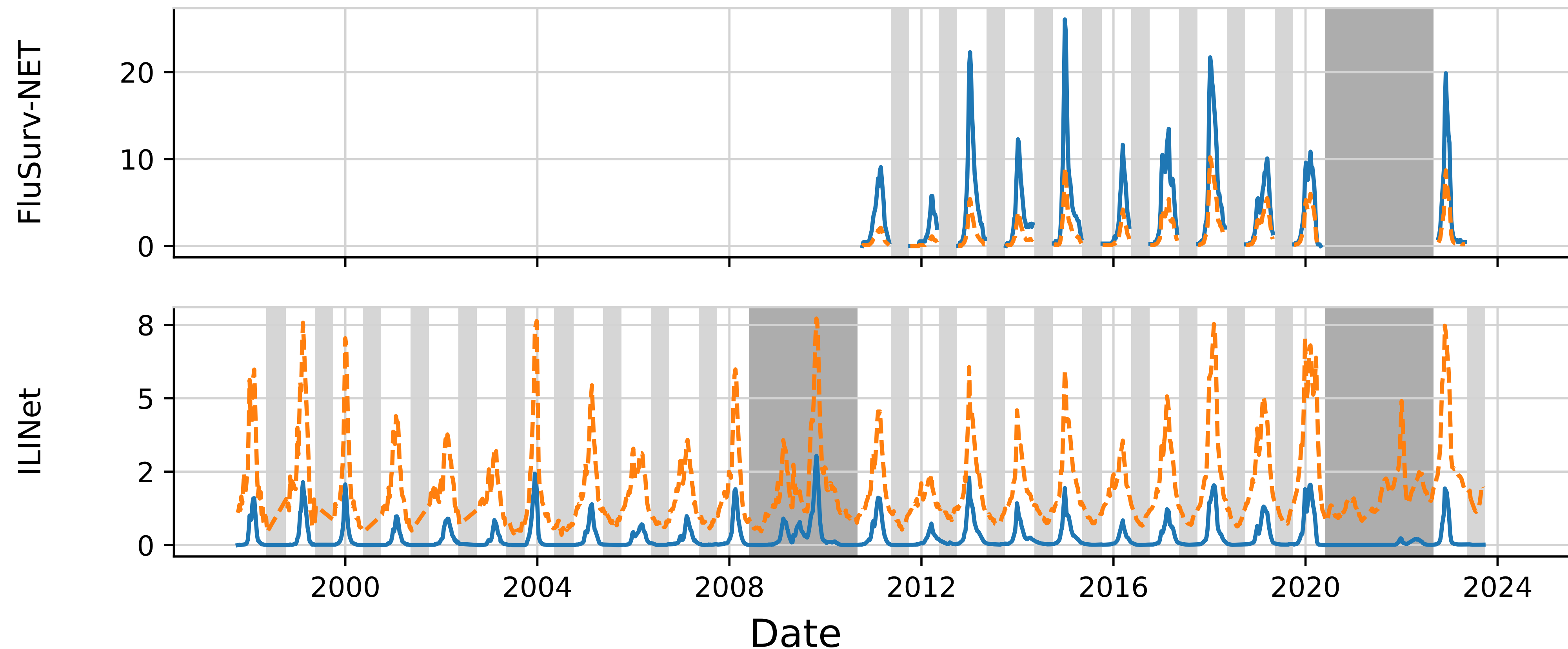
$$\theta^m = \left(\prod_{m' \neq m} \frac{MAE_{\mathcal{J}_{m,m'}}^m}{MAE_{\mathcal{J}_{m,m'}}^{m'}} \right)^{1/(M-1)}$$

3. Standardize relative to a baseline (in our case, Baseline-flat)

$$rMAE^m = \frac{\theta^m}{\theta^{baseline}}$$

Data adjustments

- For both other signals, we employ adjustments (original in **orange**, adjusted in **blue**)
 - FluSurv-NET: adjust for different case capture rates from season to season due to changing testing rates and test sensitivity
 - ILI+: combine a measure of influenza-like illness (ILI) with influenza test positivity rates to get a more specific measure of flu activity



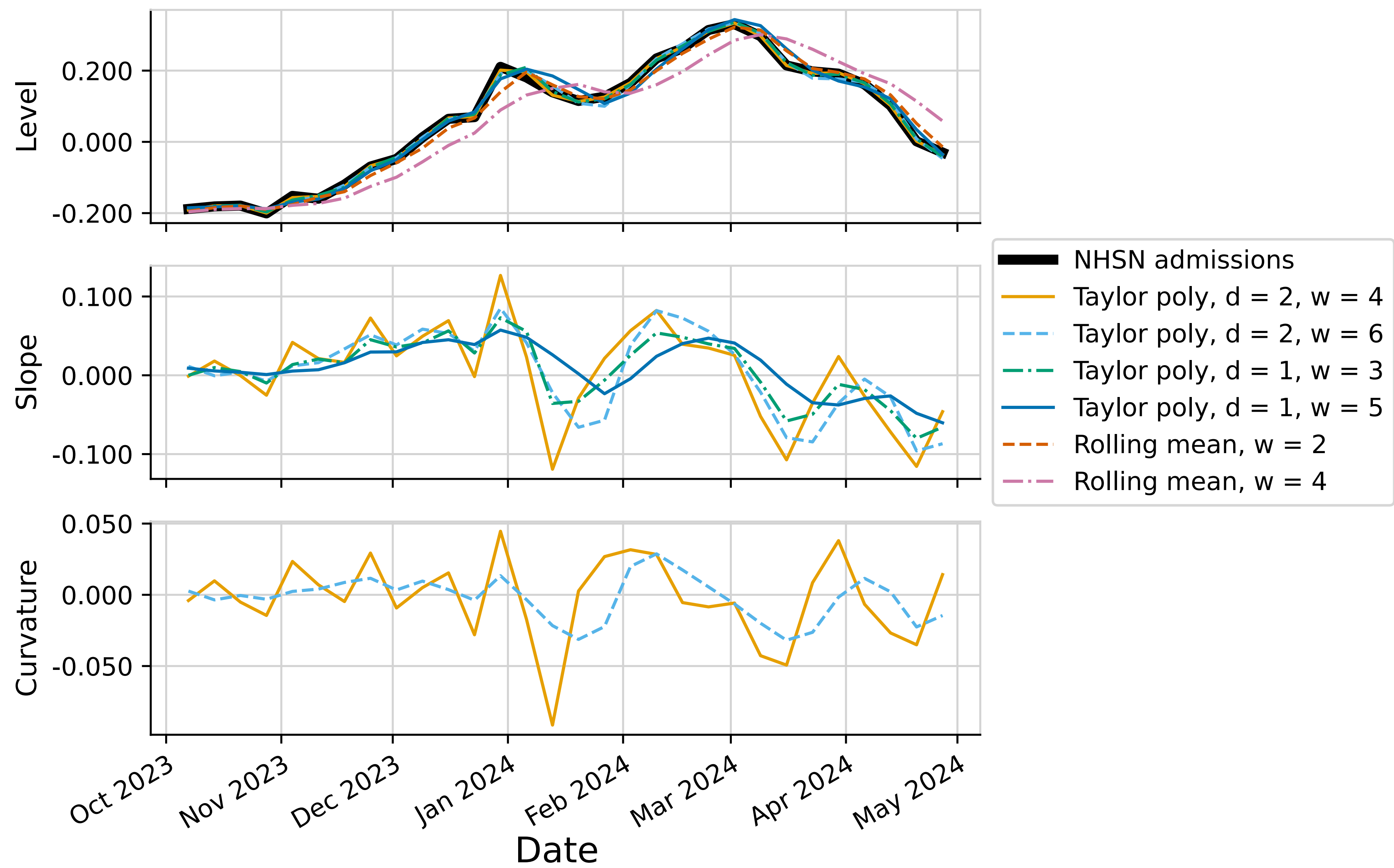
GBQR and GBQR-no-level models

- GBQR used 114 features
- GBQR-no-level omitted features from groups 8-12 that measure local level of the signal

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

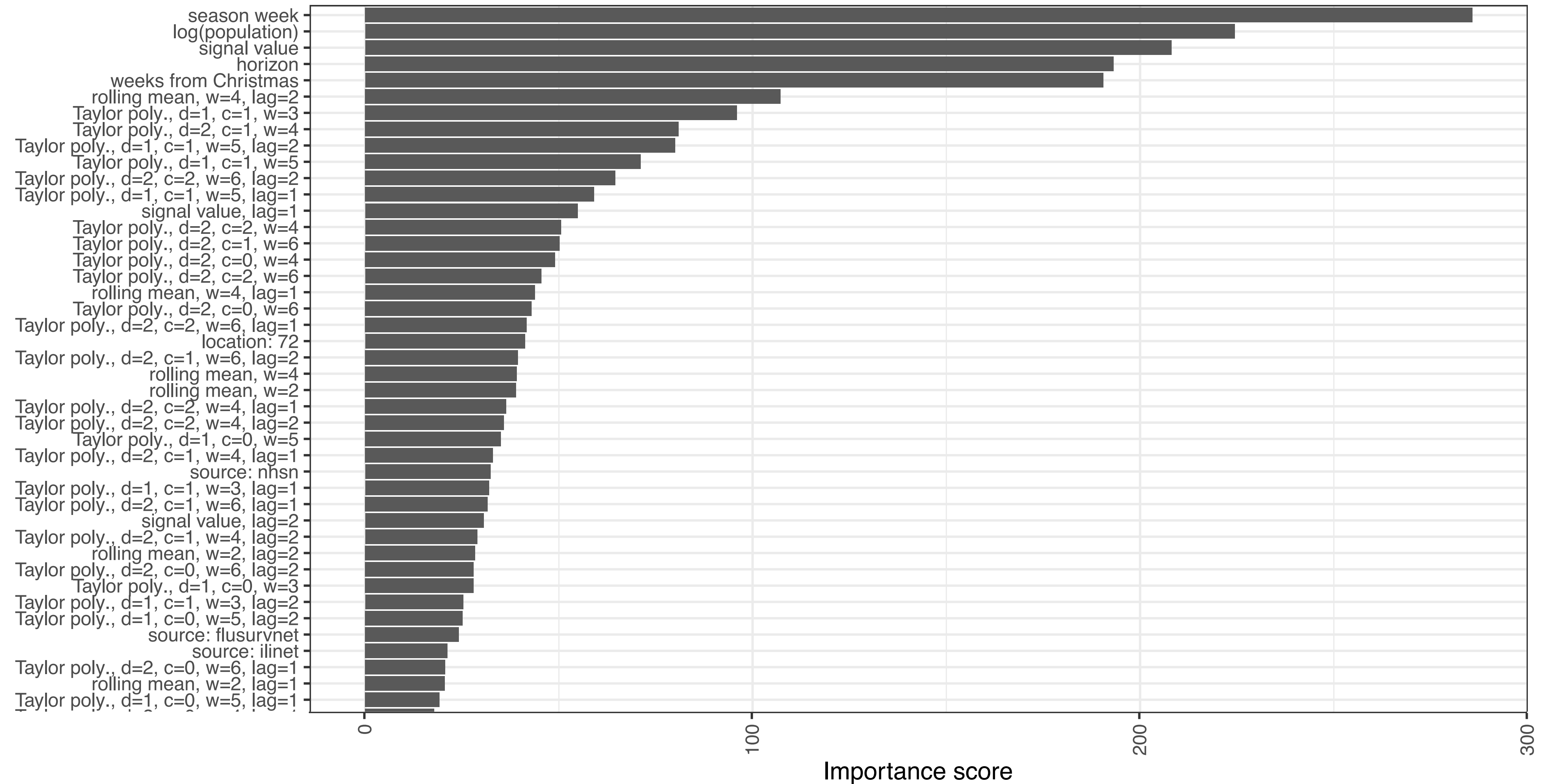
GBQR: local level, slope, curvature features

- Example for Michigan, 2023/24 season



Feature Importance

- Importance score: number of tree splits using feature



ARX Model

- We used a Bayesian specification of an autoregressive model (order $J = 8$) with covariates

$$Y_{l,t} \mid y_{l,t-1}, \dots, y_{l,t-J}, x_{l,t-1}, \dots, x_{l,t-J}, \varepsilon_{l,t} = \sum_{j=1}^J \alpha_j y_{l,t-j} + \sum_{j=1}^J \beta_j x_{l,t-j} + \varepsilon_{l,t}$$

$$X_{l,t} \mid x_{l,t-1}, \dots, x_{l,t-J}, \nu_{l,t} = \sum_{j=1}^J \gamma_j x_{l,t-j} + \nu_{l,t}$$

$$\varepsilon_{l,t} \sim \text{Normal}(0, \sigma_{\varepsilon,l})$$

$$\nu_{l,t} \sim \text{Normal}(0, \sigma_{\nu,l})$$

- Key idea for AR setup:
 - AR coefficients shared across locations (to avoid overfitting to limited data)
 - Separate variance innovation term per location (noise levels differ based on population)
- We used 1 covariate:
 - takes the value 3 on Christmas week
 - 2 one week before and one week after Christmas
 - 1 two weeks before and two weeks after Christmas
 - 0 otherwise

MWIS by date and forecast horizon

