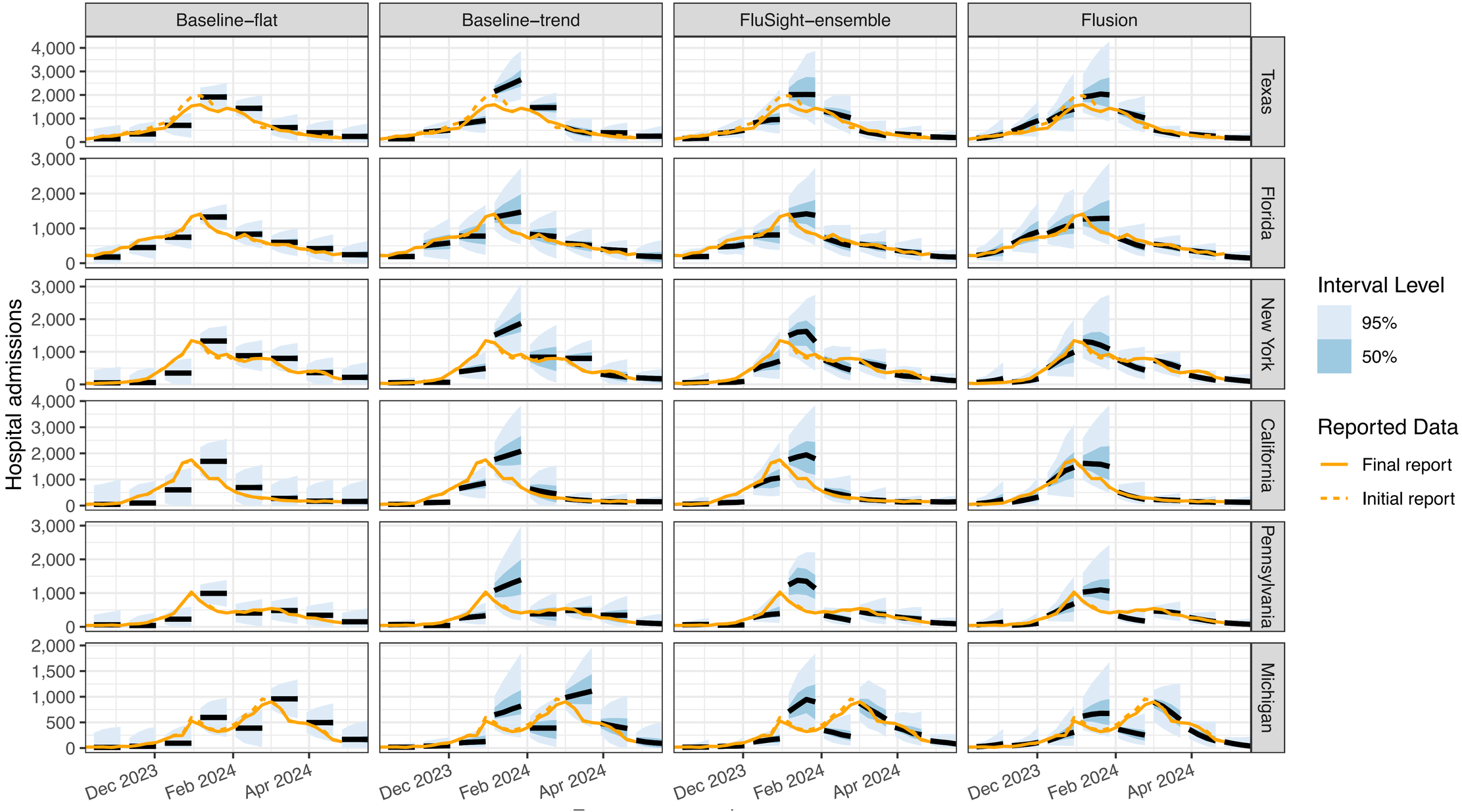
Rapid fire research talks

Evan Ray, **Nick Reich**
University of Massachusetts, Amherst

SISMID
July 18, 2025

Flusion

FluSight forecasts, 2023/24 season



Overall Results: FluSight 2023/24 season

Higher rank
Better Performance

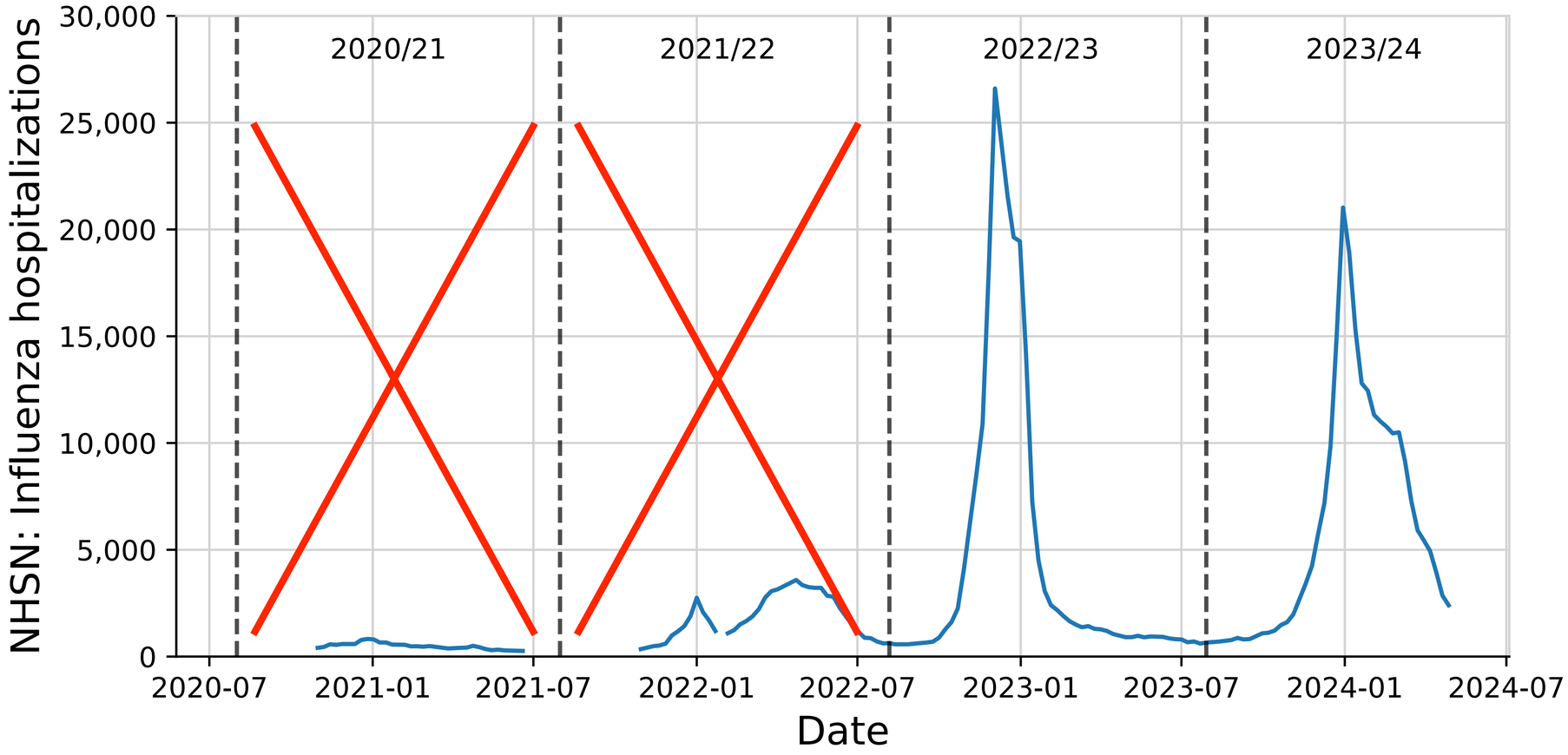
Lower rank
Worse Performance

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

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

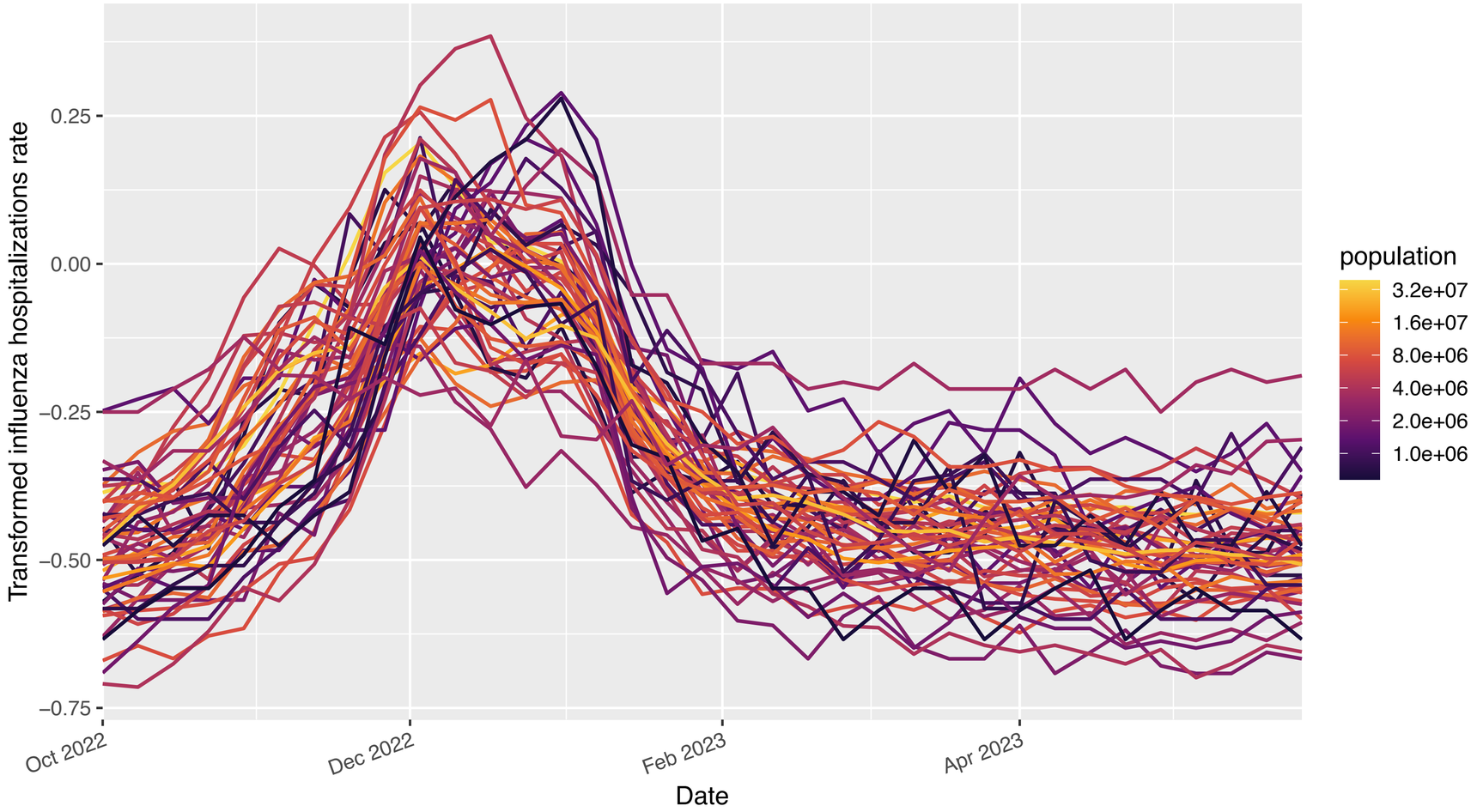
A central data challenge

- 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



Data transformations for AR model

- What we did (could be refined/simplified):
  - Convert to a hospitalization rate per 100k population
  - Take the fourth root, with an offset of 0.325
  - Center and scale by the per-location mean and 95th percentile



Considerations for parameter setup

$$\tilde{Z}_{l,t} \mid \tilde{z}_{l,t-1}, \ldots, \tilde{z}_{l,t-J}, \varepsilon_{l,t} = \sum_{j=1}^{J} \alpha_{l,j} \tilde{z}_{l,t-j} + \varepsilon_{l,t}$$

$$\varepsilon_{l,t} \sim \text{Normal}(0, \sigma^2_{\varepsilon,l})$$

- Note that the "regression coefficients", the $\alpha_{l,j}$, are specific to each location l and lag j.
- Because there is such little data for any given location, we decided to **estimate one set of coefficients across all locations**. That is, we set $\alpha_{l,j} = \alpha_j$.

- But... we also **estimated separate variance parameters** $\sigma^2_{\varepsilon,l}$ for each location, since noise levels depend strongly on population size.

- We set $J = 8$ based on intuition/guess. 🤷 🤞 😤

AR performance in overall results ('23/'24)

AR models with pooling;
AR models without pooling, p=2 or p=4

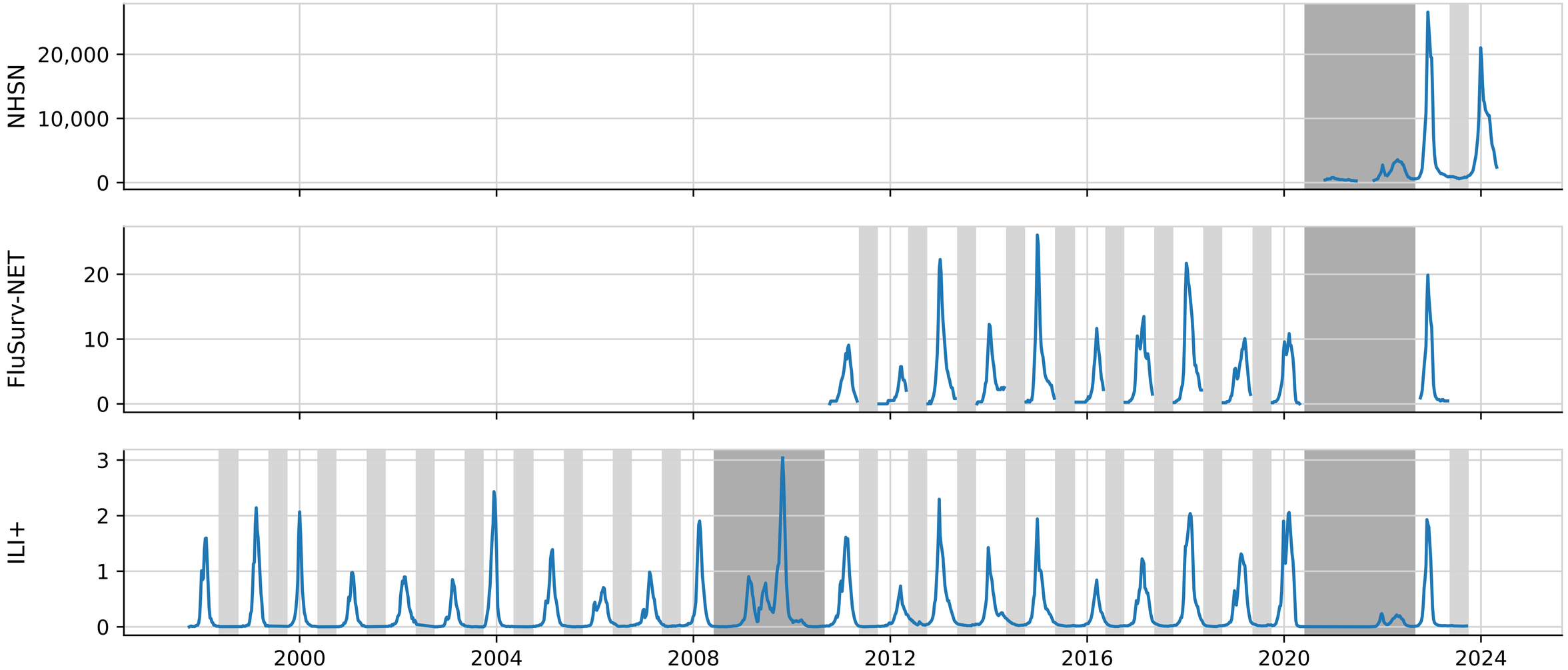
Model	% Submitted	MWIS	rMWIS	MAE	rMAE	50% Cov.	95% Cov.
Flusion	99.9	**29.6**	**0.610**	**45.6**	**0.670**	0.583	0.967
FluSight-ensemble	100.0	35.5	0.731	55.4	0.814	0.516	0.926
Other Model #1	100.0	35.6	0.731	54.0	0.792	0.558	**0.940**
Other Model #2	89.1	40.4	0.773	61.5	0.840	0.479	0.908
Other Model #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

Lower rank
Worse Performance

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

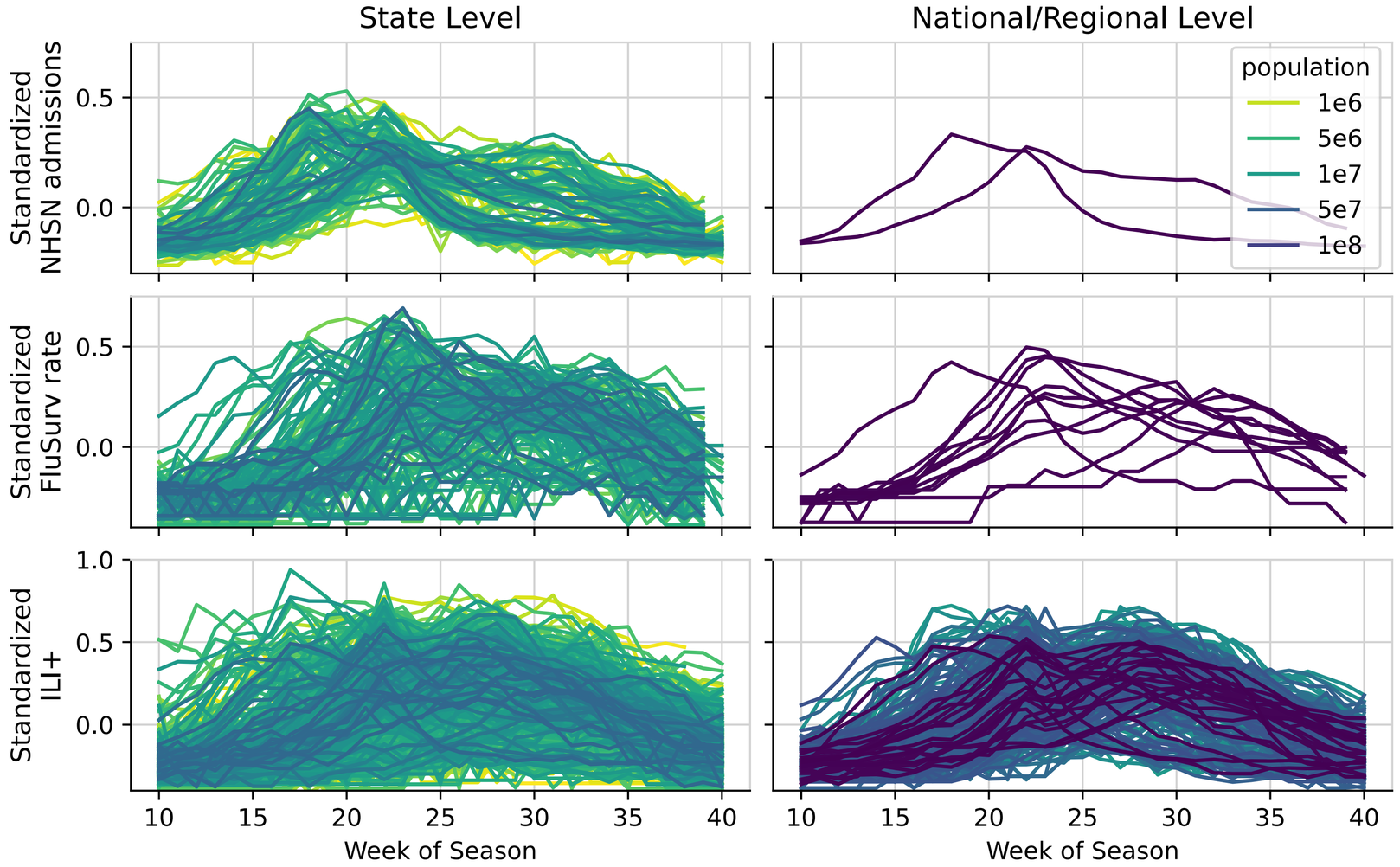
Actually, we have more data!

- 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

- We apply the same transformations we discussed for the AR model
  - Fourth root: stabilize variance across different times
  - Center and scale: put the data on a similar scale for different locations, signals

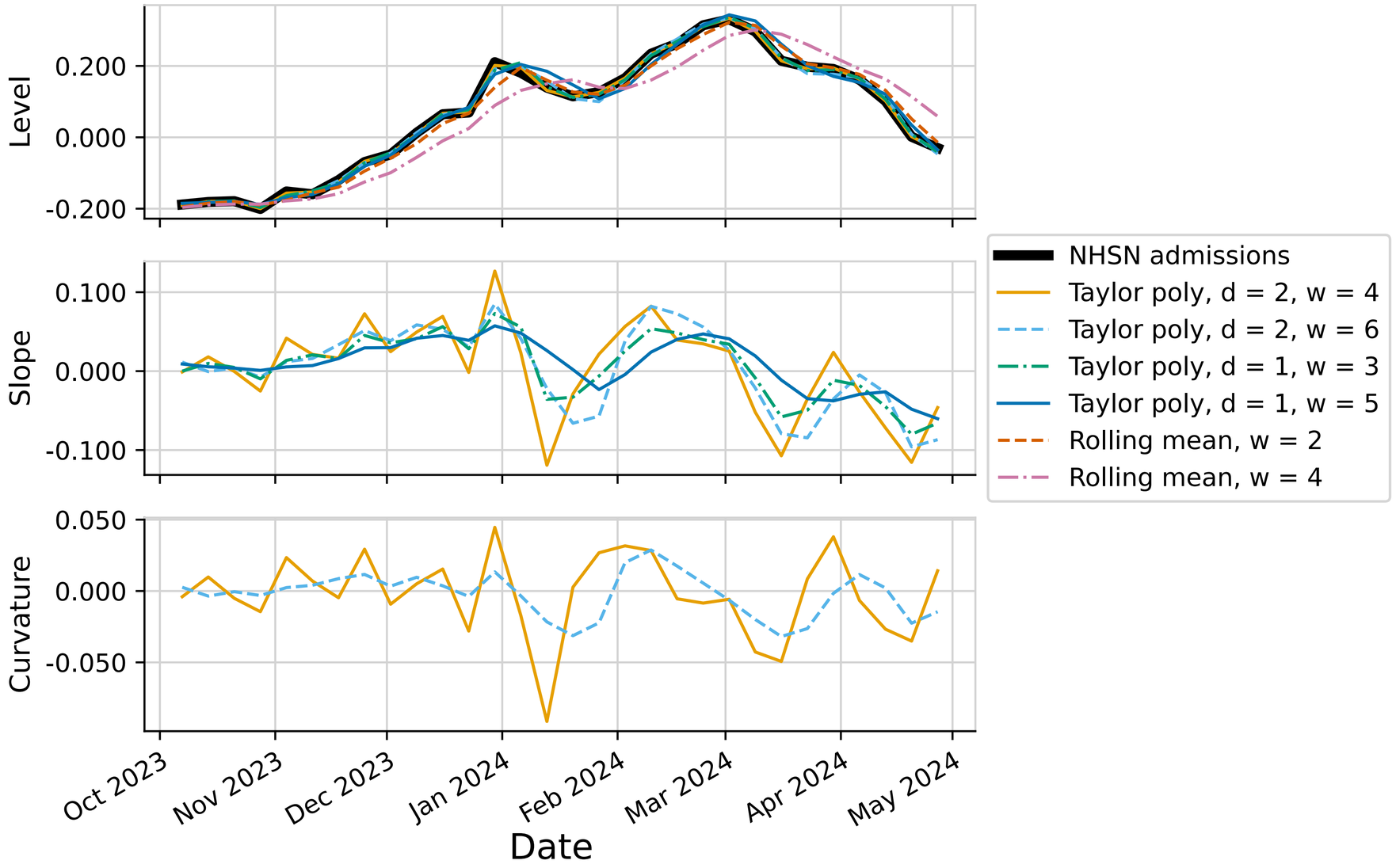


Flusion: an ensemble of 3 models

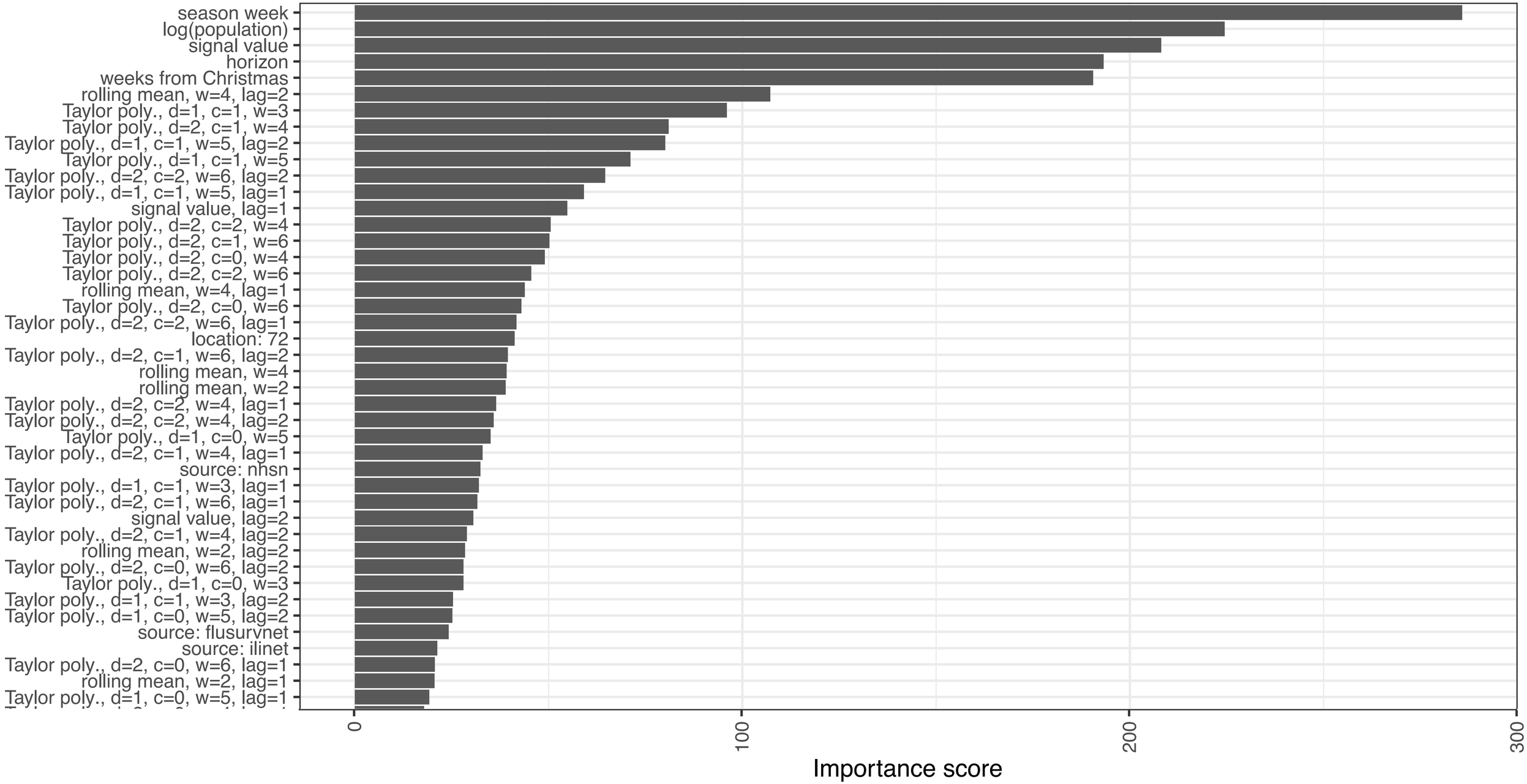
- Three component models:
  1. GBQR: A gradient boosting quantile regression model
     - Learns a mapping $f_\alpha(x)$ from features x to a predictive quantile at each quantile level α
     - Used 114 features:
       - Measures of local level, trend, curvature in the signal
       - One-hot encoding 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

GBQR: local level, slope, curvature features

- Example for Michigan, 2023/24 season

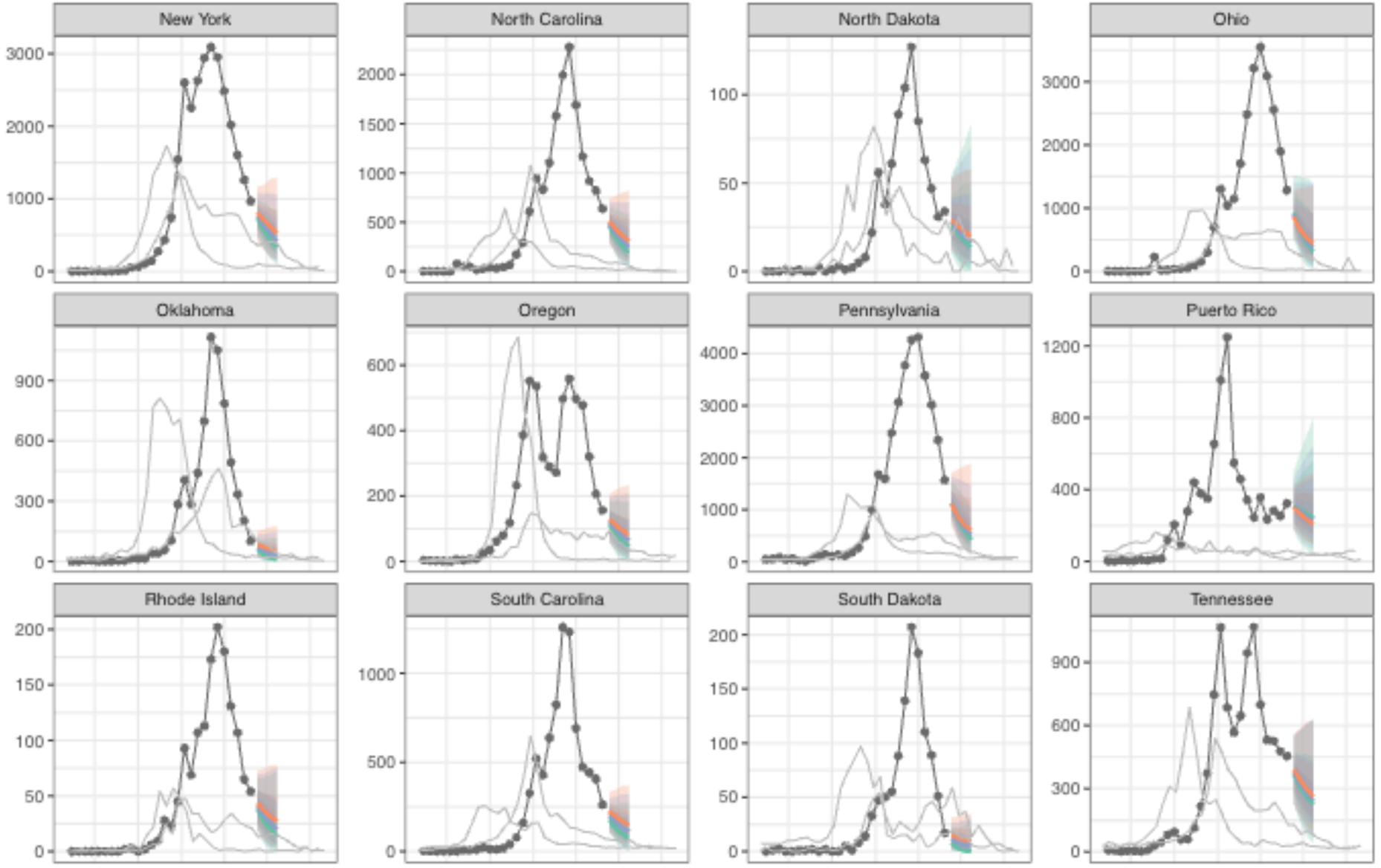


Feature Importance score: number of tree splits using feature



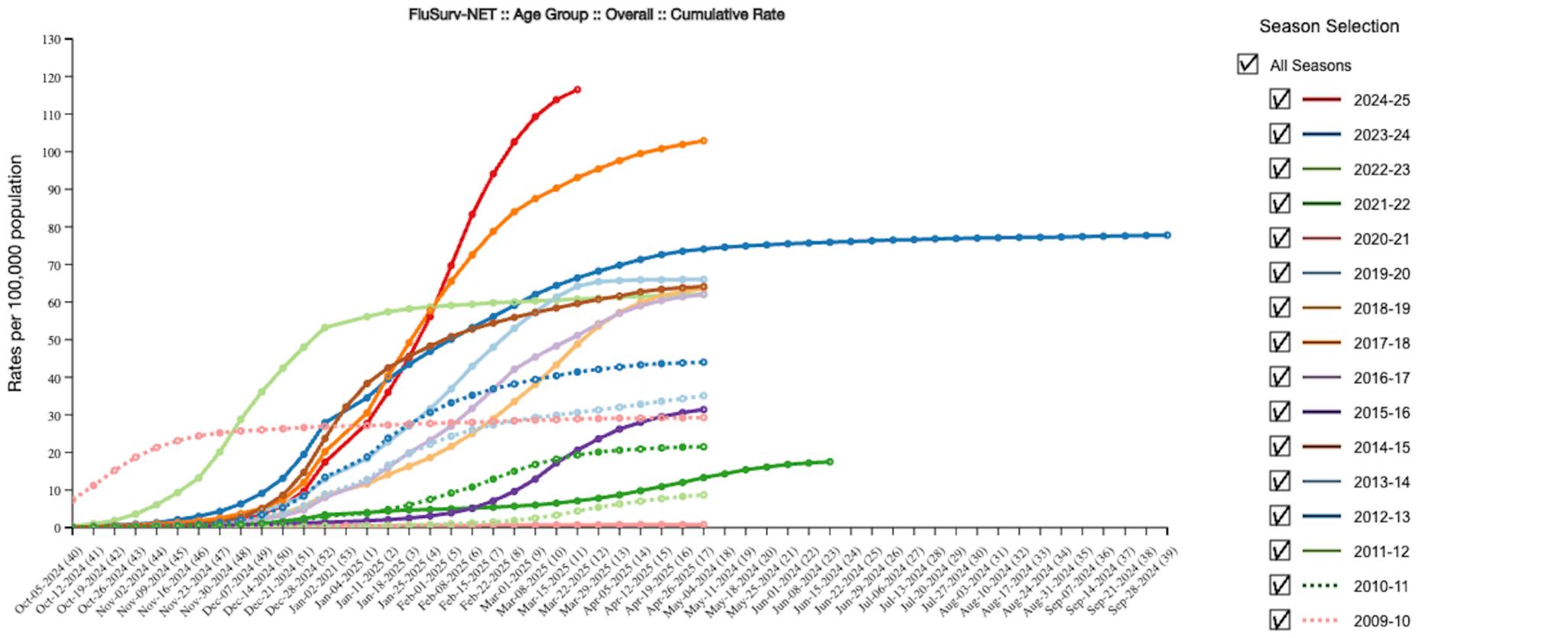
The 2024/2025 season was big (for hospitalizations)

- For NHSN data, this season's peak was generally 1.5-3 times larger than recent years.



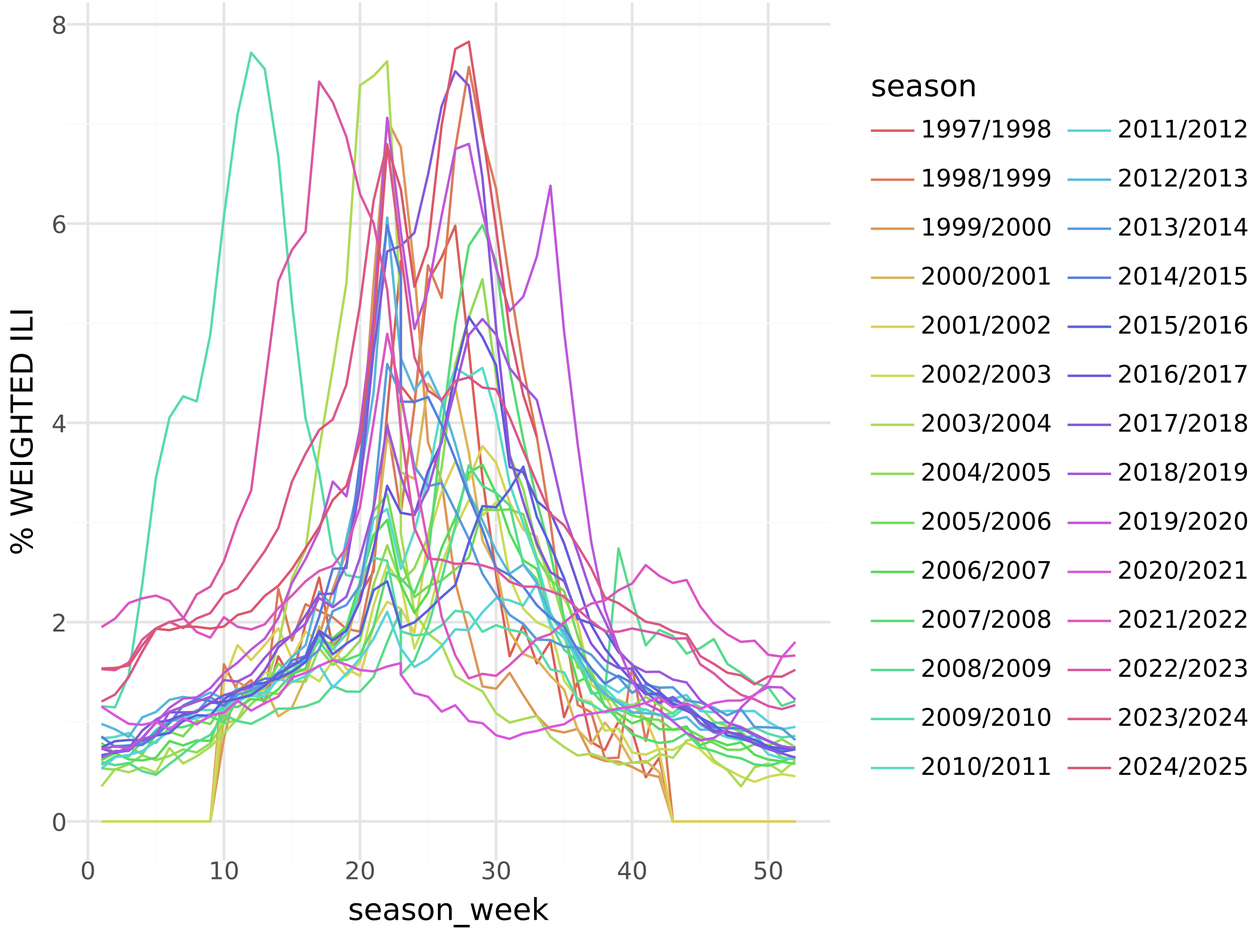
The 2024/2025 season was big (for hospitalizations)

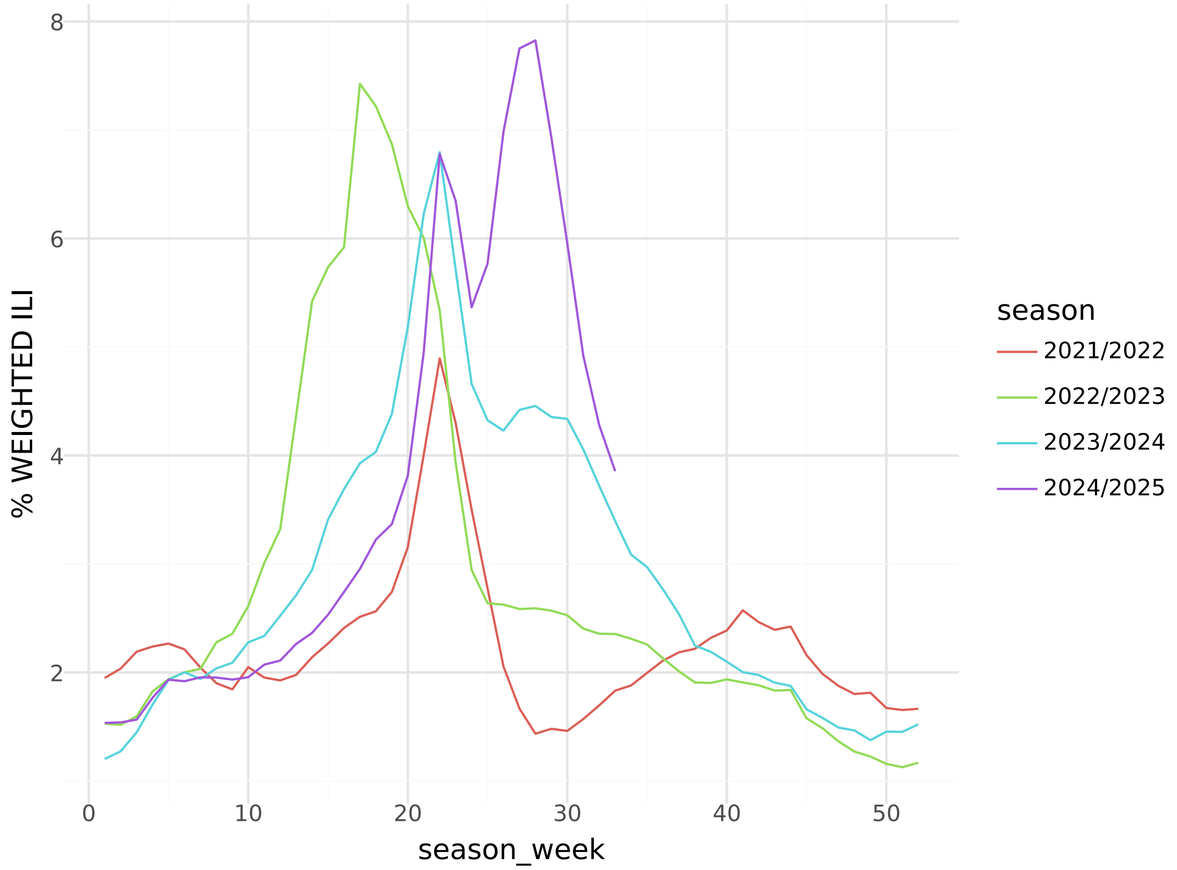
- Here's national level data for FluSurv-NET - red is current season



The 2024/2025 season was not an outlier for ILI

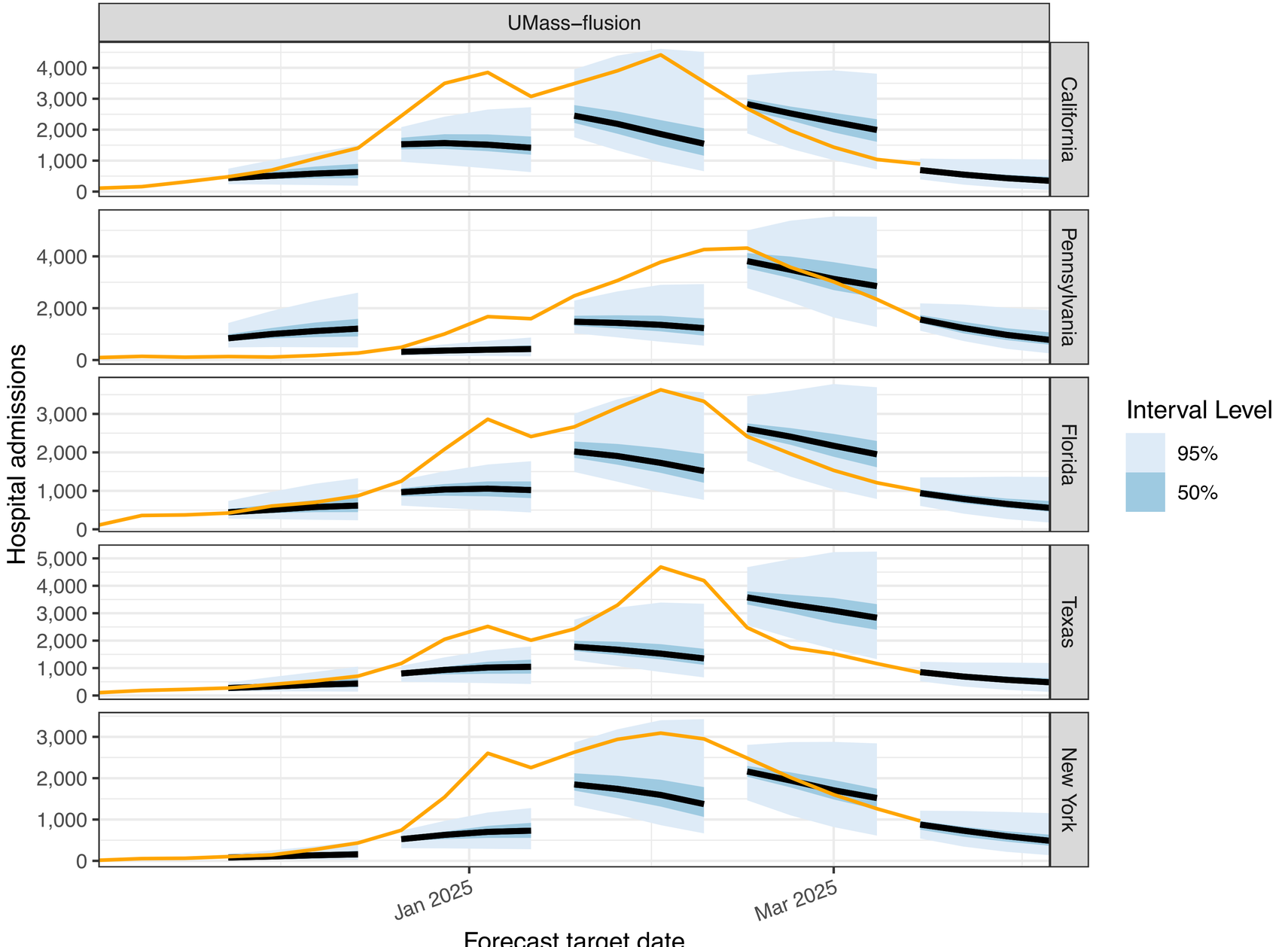
- ILI Net data. This season was a little larger than usual, but not huge.





We missed this season

- Here are some forecasts for this season. (5 states with highest total hospitalizations)
- We consistently under-predicted early in the season.



Flusion remained a top model in ’24/‘25

- Among unique models that submitted for all locations and for the entire season, Flusion was the second best individual model.
- Flusion was slightly worse than one *ensemble*, slightly better than three.

Model	Rel. WIS	WIS	Rel. MAE	MAE	50% Cov.	95% Cov.	N
PSI-PROF	0.62	128.6	0.75	196.8	47.4	87.0	4160
FluSight-lop_norm *	0.62	130.2	0.73	194.6	47.2	91.1	4160
UMass-flusion	0.65	137.5	0.72	191.3	30.3	76.1	4140
FluSight-trained_med *	0.66	127.8	0.74	188.3	52.2	86.9	3120
NEU_ISI-AdaptiveEnsemble	0.66	139.6	0.72	190.8	24.6	62.9	3916
FluSight-ensemble *	0.67	140.8	0.74	196.6	38.2	74.0	4160
CMU-TimeSeries	0.67	139.7	0.74	197.8	54.0	88.1	4160
FluSight-trained_mean *	0.67	131.5	0.77	196.6	53.4	90.7	3120
CEPH-Rtrend_fluH	0.67	139.8	0.71	187.4	37.3	72.3	4140
fjordhest-ensemble	0.67	143.0	0.74	198.3	30.3	67.5	3960
NIH-Flu_ARIMA	0.68	138.6	0.75	194.2	20.9	63.8	4128
UGA_flucast-Copycat	0.68	141.0	0.80	210.6	31.3	77.7	3732
CU-ensemble	0.70	145.8	0.77	203.1	38.0	78.1	4160

https://reichlab.io/flusight-dashboard/

* = ensemble model that combines models submitted to FluSight

The 2024/2025 season was big (for hospitalizations)

Summing up

- We missed for a big season
  - This season was unusually severe for hospitalizations
  - We generally missed throughout the first 2/3 of the season
  - To be useful to public health, forecasts really need to capture severe seasons —> there is more work to do!

Limitations of our setup

- Need to reconsider transformation scheme, since unusually severe seasons still look unrealistic/out-of-range after transformation
- The ILI data (which makes up the majority of our training set) had pretty different dynamics from the NHSN data (our target).

Thank you.

Acknowledgments to **all the folks who helped build these things.**

Overview of this talk

- Preview of results
- Our model
  - Data
  - Model Setup
- Ablation Experiments
- Early take on 2024/2025 results
- **Conclusions**

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 two 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. New data streams can be informed by historical surveillance systems.

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 $\{q_k : k = 1,\ldots,K\}$ denote a set of predictive quantiles at levels $\alpha_1, \ldots, \alpha_K$.

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

  - $$QS_{\alpha_k}(q_k, z_i) = \alpha_k \max(y - q_k, 0) + (1 - \alpha_k)\max(q_k - y, 0)$$

- 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 $MAE^{m}_{\mathcal{J}_{m,m'}}$ and $MAE^{m'}_{\mathcal{J}_{m,m'}}$
2. For model m, compute the geometric mean of ratios of MAEs for m compared to all other models

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

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

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

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

Considerations for parameter setup

$$\tilde{Z}_{l,t} \mid \tilde{z}_{l,t-1}, \ldots, \tilde{z}_{l,t-J}, \varepsilon_{l,t} = \sum_{j=1}^{J} \alpha_{l,j} \tilde{z}_{l,t-j} + \varepsilon_{l,t}$$

$$\varepsilon_{l,t} \sim \text{Normal}(0, \sigma^2_{\varepsilon,l})$$

- A classic Bayesian solution is to use a hierarchical prior for the $\alpha_{l,j'}$ perhaps:

$$\alpha_{l,j} \mid \alpha_j, \xi \sim \text{Normal}(\alpha_j, \xi^2)$$

$$\alpha_j \mid \psi \sim \text{Normal}(0, \psi^2)$$

- Taking this to an extreme, we set all $\alpha_{l,j} = \alpha_{j'}$ i.e. share parameters across locations ($\xi = 0$)
  - These are "global parameters" in the framework of Montero-Manso and Hyndman (2021).
- Kept separate variance parameters $\sigma^2_{\varepsilon,l}$ for each location, since noise levels depend strongly on population size.
  - Used half-Cauchy$(0,1)$ priors for all standard deviation parameters $\sigma_{\varepsilon,l}$ and ψ.
- We set $J = 8$ based on intuition/guess

ARX Model

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

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

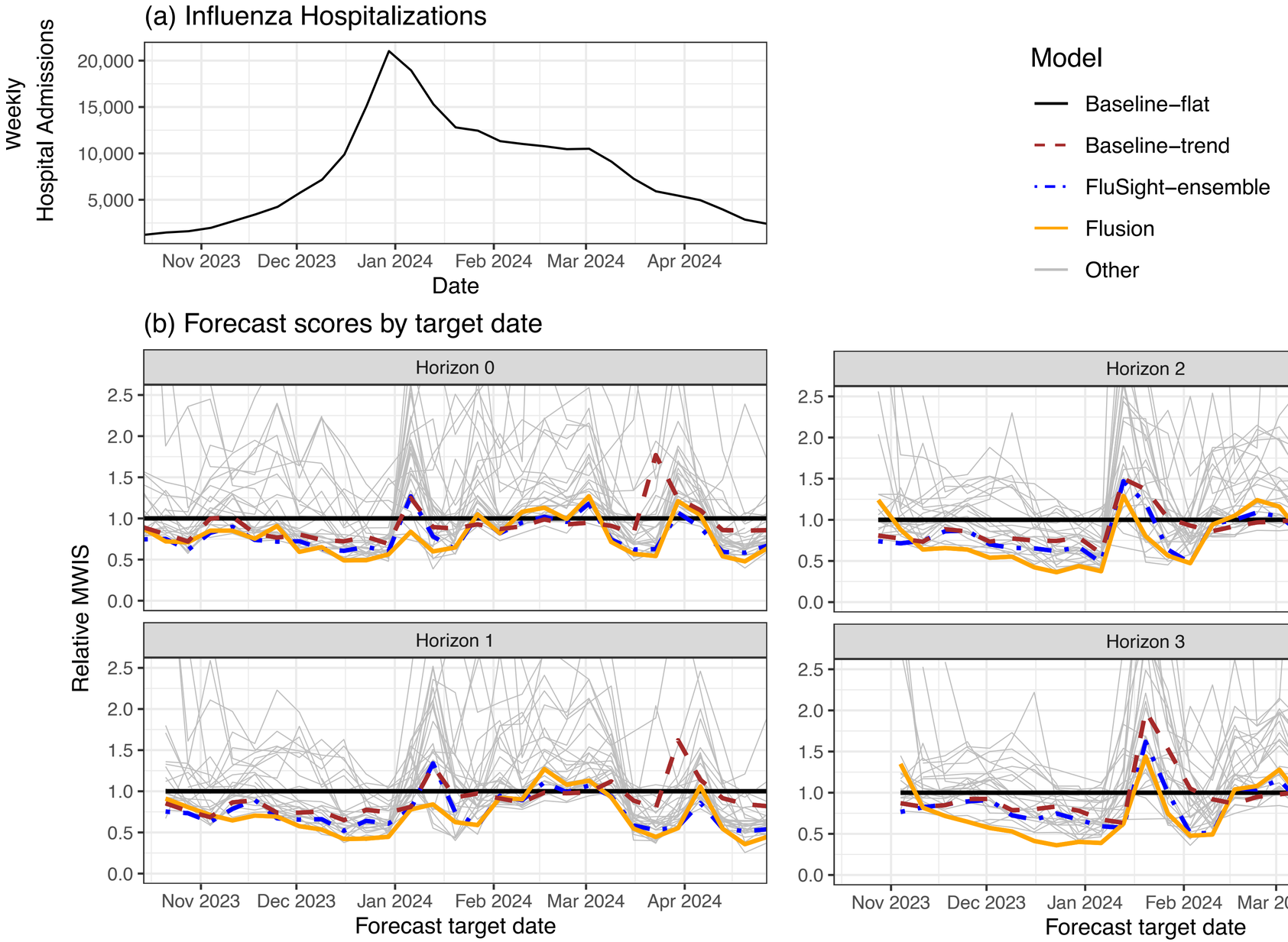
$$X_{l,t} \mid x_{l,t-1}, \ldots, x_{l,t-J}, \nu_{l,t} = \sum_{j=1}^{J} \gamma_j x_{l,t-j} + \nu_{l,t}$$

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

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

- 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



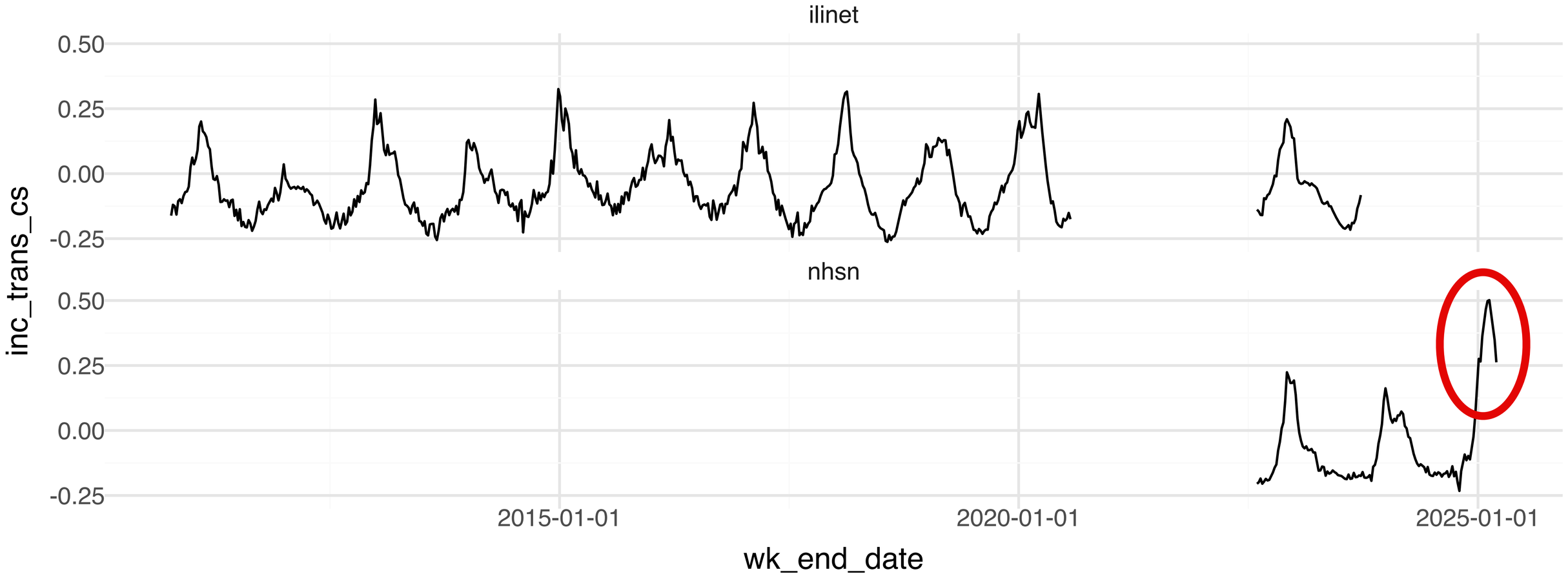
Attempting to capture holiday effects

$$\tilde{Z}_{l,t} \mid \tilde{z}_{l,t-1}, \ldots, \tilde{z}_{l,t-J}, \varepsilon_{l,t} = \sum_{j=1}^{J} \alpha_{l,j} \tilde{z}_{l,t-j} + \varepsilon_{l,t}$$

$$\varepsilon_{l,t} \sim \text{Normal}(0, \sigma^2_{\varepsilon,l})$$

The transformed data our models see

- Both models (AR, GBQR) see data that has been "globally scaled" for each location and data source.
  - A single center and scale parameter is used for all time points
- **Transformed data** from ILI Net and NHSN for Pennsylvania:



Overview of this talk

- Preview of results
- Our model
  - Data
  - Model Setup
- Ablation Experiments
- Early take on 2024/2025 results
- **Conclusions**

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 two 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. New data streams can be informed by historical surveillance systems.

Thank you.

Acknowledgments to **Evan Ray**
for conducting most of the work presented here
and drafting the slides.

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 $\{q_k : k = 1,\ldots,K\}$ denote a set of predictive quantiles at levels $\alpha_1, \ldots, \alpha_K$.

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

  - $$QS_{\alpha_k}(q_k, z_i) = \alpha_k \max(y - q_k, 0) + (1 - \alpha_k)\max(q_k - y, 0)$$

- 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 $MAE^{m}_{\mathcal{J}_{m,m'}}$ and $MAE^{m'}_{\mathcal{J}_{m,m'}}$
2. For model m, compute the geometric mean of ratios of MAEs for m compared to all other models

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

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

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

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

Considerations for parameter setup

$$\tilde{Z}_{l,t} \mid \tilde{z}_{l,t-1}, \ldots, \tilde{z}_{l,t-J}, \varepsilon_{l,t} = \sum_{j=1}^{J} \alpha_{l,j} \tilde{z}_{l,t-j} + \varepsilon_{l,t}$$

$$\varepsilon_{l,t} \sim \text{Normal}(0, \sigma^2_{\varepsilon,l})$$

- A classic Bayesian solution is to use a hierarchical prior for the $\alpha_{l,j'}$ perhaps:

$$\alpha_{l,j} \mid \alpha_j, \xi \sim \text{Normal}(\alpha_j, \xi^2)$$

$$\alpha_j \mid \psi \sim \text{Normal}(0, \psi^2)$$

- Taking this to an extreme, we set all $\alpha_{l,j} = \alpha_{j'}$ i.e. share parameters across locations ($\xi = 0$)
  - These are "global parameters" in the framework of Montero-Manso and Hyndman (2021).
- Kept separate variance parameters $\sigma^2_{\varepsilon,l}$ for each location, since noise levels depend strongly on population size.
  - Used half-Cauchy$(0,1)$ priors for all standard deviation parameters $\sigma_{\varepsilon,l}$ and ψ.
- We set $J = 8$ based on intuition/guess

ARX Model

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

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

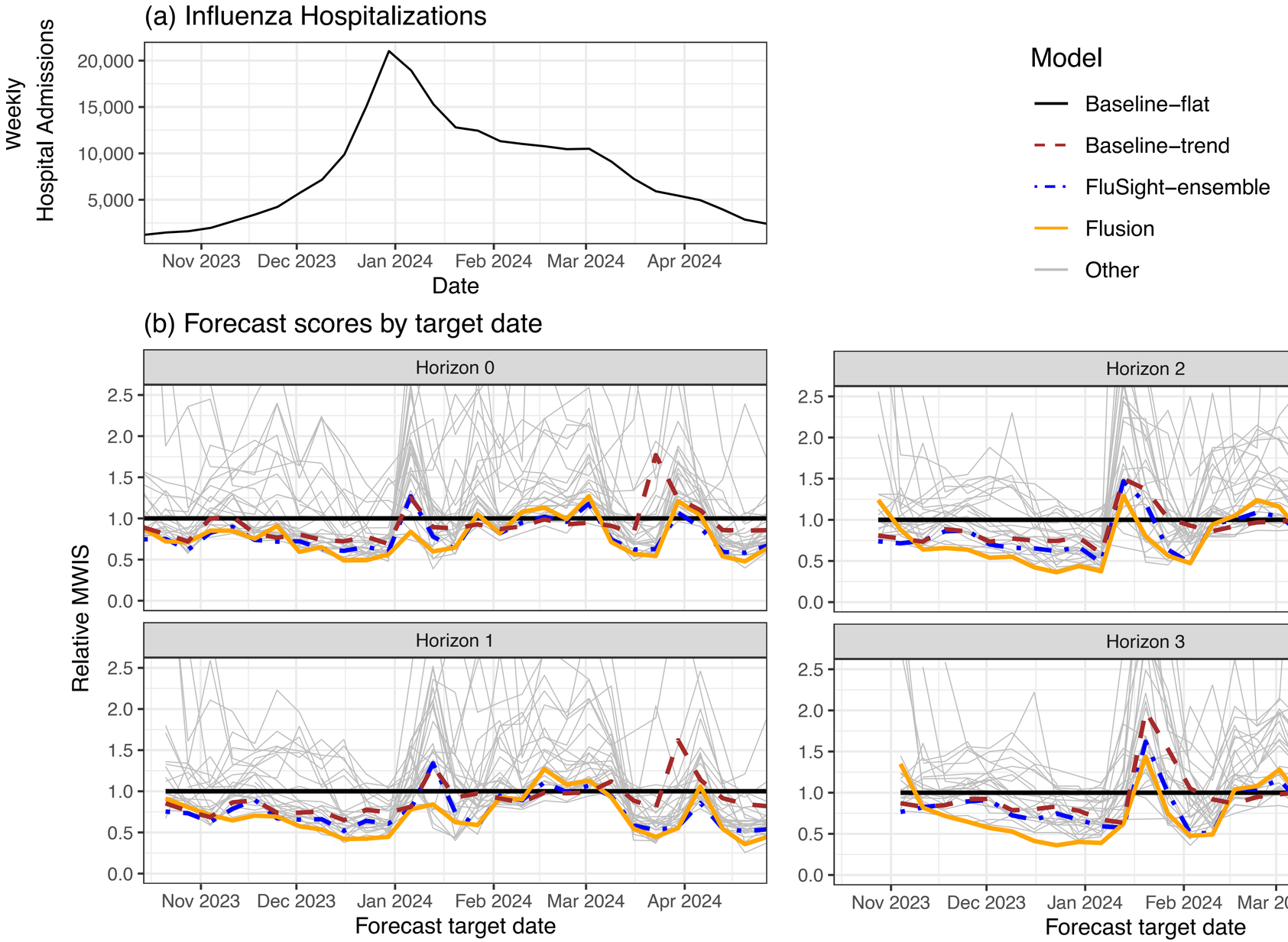
$$X_{l,t} \mid x_{l,t-1}, \ldots, x_{l,t-J}, \nu_{l,t} = \sum_{j=1}^{J} \gamma_j x_{l,t-j} + \nu_{l,t}$$

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

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

- 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



Attempting to capture holiday effects

$$\tilde{Z}_{l,t} \mid \tilde{z}_{l,t-1}, \ldots, \tilde{z}_{l,t-J}, \varepsilon_{l,t} = \sum_{j=1}^{J} \alpha_{l,j} \tilde{z}_{l,t-j} + \varepsilon_{l,t}$$

$$\varepsilon_{l,t} \sim \text{Normal}(0, \sigma^2_{\varepsilon,l})$$

The transformed data our models see

- Both models (AR, GBQR) see data that has been "globally scaled" for each location and data source.
  - A single center and scale parameter is used for all time points
- **Transformed data** from ILI Net and NHSN for Pennsylvania: