Flusion:

Integrating multiple data sources for accurate influenza predictions

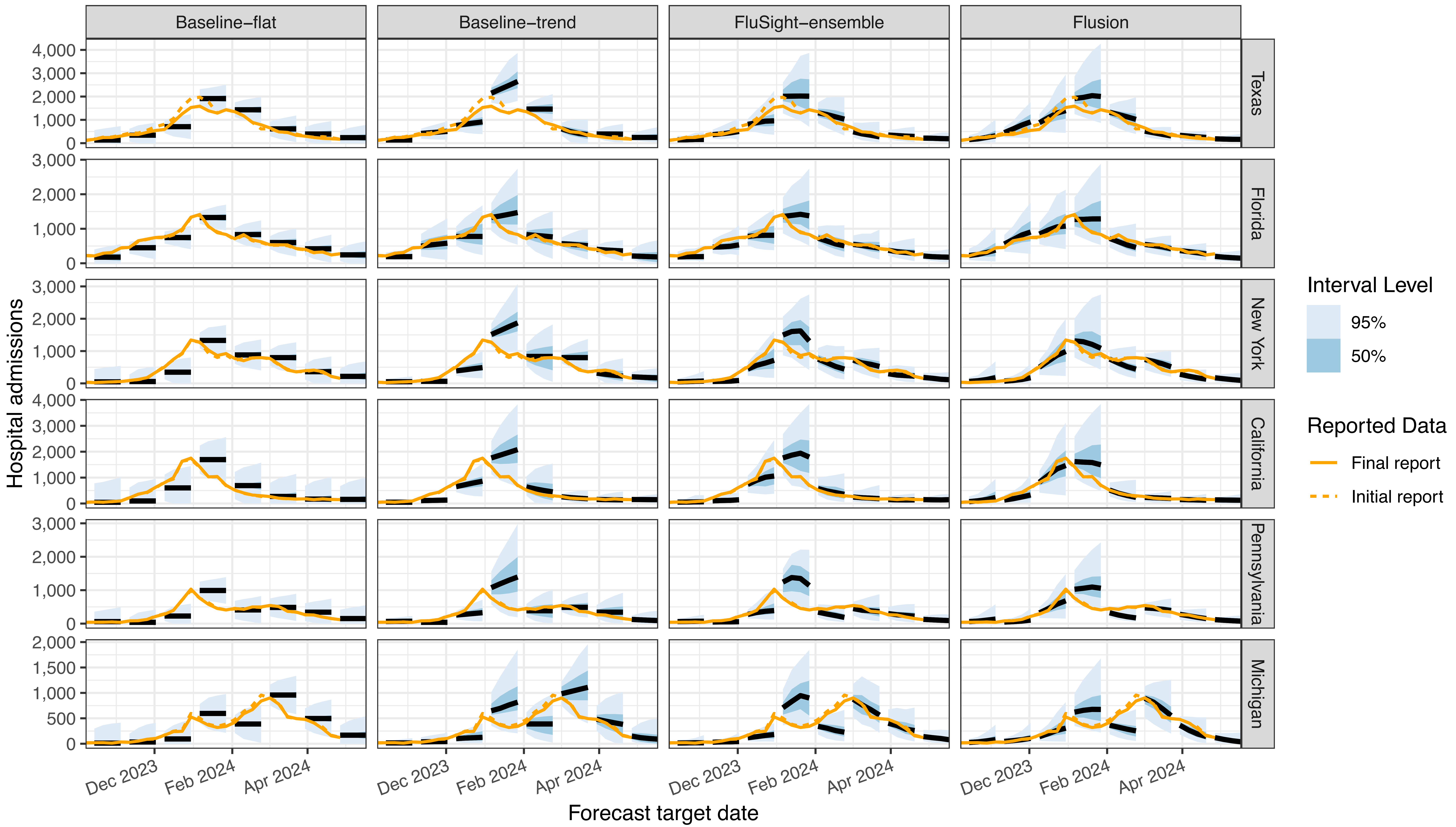
Evan L. Ray, Yijin Wang, Russel D. Wolfinger, Nicholas G. Reich
University of Massachusetts, Amherst

Yale Biostatistics Seminar
November 12, 2024

Overview of this talk

- **Motivation, preview of results**
- Modeling approaches
  - Model 1
  - Model 2
  - Flusion: an ensemble model
- Conclusions

First look: FluSight forecasts, 2023/24 season



Overall Results: FluSight 2023/24 season

Higher rank
Better 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

Lower rank
Worse Performance

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

Why Forecast Infectious Disease?

- Situational awareness for the public and stakeholders like health care providers.

https://www.cdc.gov/flu-forecasting/data-vis/04242024-flu-forecasts.html

CDC FluSight

EXPLORE TOPICS

SEARCH

APRIL 26, 2024

Flu Hospital Admission as of May 11, 2024

PURPOSE

This week's ensemble predicts that the number of new weekly laboratory confirmed influenza hospital admissions will likely decrease nationally, with 520 to 4,700 laboratory confirmed influenza hospital admissions likely reported in the week ending May 11, 2024.

Interpretation of forecasts

- Reported and forecasted new influenza hospital admissions as of April 24, 2024
- This week, 24 modeling groups contributed 29 forecasts that were eligible for inclusion in the ensemble forecasts for at least one jurisdiction. Contributing teams are listed below.

ON THIS PAGE

Why Forecast Infectious Disease?

- Situational awareness for the public and stakeholders like health care providers.

- Allocation of resources such as
  - antiviral treatments
  - hospital care staff
  - hospital beds

https://www.cdc.gov/flu-forecasting/about/index.html

Health Care Management Science (2021) 24:253–272
https://doi.org/10.1007/s10729-020-09542-0

From predictions to prescriptions: A data-driven response to COVID-19

Dimitris Bertsimas[1,2] · Leonard Boussioux[2] · Ryan Cory-Wright[2] · Arthur Delarue[2] · Vassilis Digalakis[2] · Alexandre Jacquillat[1,2] · Driss Lahlou Kitane[2] · Galit Lukin[2] · Michael Li[2] · Luca Mingardi[2] · Omid Nohadani[3] · Agni Orfanoudaki[2] · Theodore Papalexopoulos[2] · Ivan Paskov[2] · Jean Pauphilet[4] · Omar Skali Lami[2] · Bartolomeo Stellato[5] · Hamza Tazi Bouardi[2] · Kimberly Villalobos Carballo[2] · Holly Wiberg[2] · Cynthia Zeng[2]

"Our results have been used at the clinical level by several hospitals to triage patients, guide care management, plan ICU capacity, and re-distribute ventilators. At the policy level, they are currently supporting safe back-to-work policies at a major institution and vaccine trial location planning at Janssen Pharmaceuticals…"

Why Forecast Infectious Disease?

- Situational awareness for the public and stakeholders like health care providers.

- Allocation of resources such as
  - antiviral treatments
  - hospital care staff
  - hospital beds

https://www.cdc.gov/flu-forecasting/about/index.html

Health Care Management Science (2021) 24:253–272
https://doi.org/10.1007/s10729-020-09542-0

From predictions to prescriptions: A data-driven response to COVID-19

Dimitris Bertsimas[1,2] · Leonard Boussioux[2] · Ryan Cory-Wright[2] · Arthur Delarue[2] · Vassilis Digalakis[2] · Alexandre Jacquillat[1,2] · Driss Lahlou Kitane[2] · Galit Lukin[2] · Michael Li[2] · Luca Mingardi[2] · Omid Nohadani[3] · Agni Orfanoudaki[2] · Theodore Papalexopoulos[2] · Ivan Paskov[2] · Jean Pauphilet[4] · Omar Skali Lami[2] · Bartolomeo Stellato[5] · Hamza Tazi Bouardi[2] · Kimberly Villalobos Carballo[2] · Holly Wiber

“Our res
triage pa
distribut
back-to-
planning

nature COMMUNICATIONS

ARTICLE

Check for updates

https://doi.org/10.1038/s41467-021-23989-x OPEN

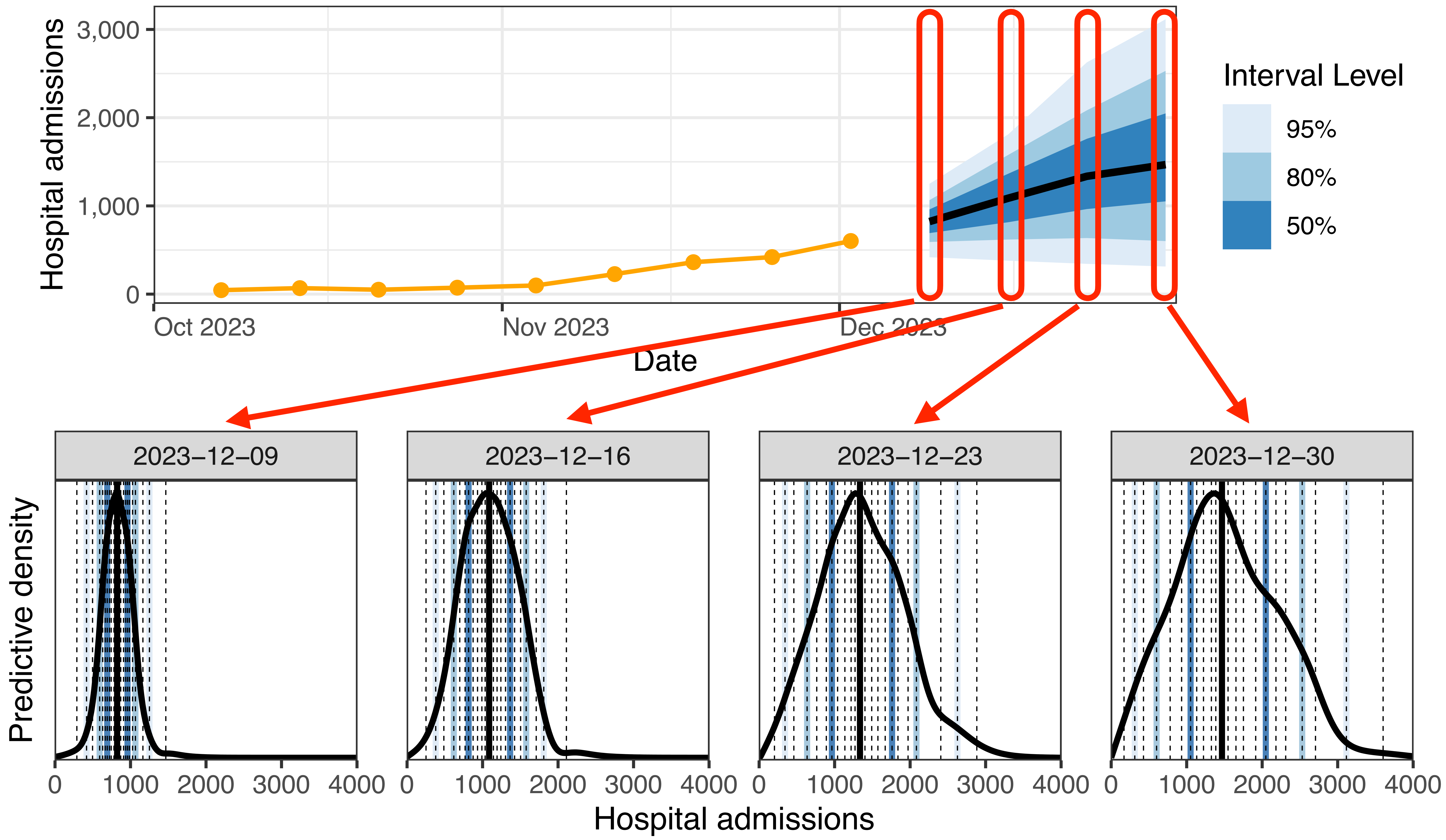
Design of COVID-19 staged alert systems to ensure healthcare capacity with minimal closures

Haoxiang Yang[1], Özge Sürer[2], Daniel Duque[2], David P. Morton[2], Bismark Singh[3], Spencer J. Fox[4], Remy Pasco[5], Kelly Pierce[6], Paul Rathouz[7], Victoria Valencia[7], Zhanwei Du[4], Michael Pignone[7], Mark E. Escott[8], Stephen I. Adler[8], S. Claiborne Johnston[7] & Lauren Ancel Meyers[4,9]✉

“…we describe the optimization and maintenance of the staged alert system that has guided COVID-19 policy in a large US city (Austin, Texas) since May 2020. As cities worldwide face future pandemic waves, our findings provide a robust strategy for tracking COVID-19 hospital admissions as an early indicator of hospital surges and enacting staged measures to ensure integrity of the health system, safety of the health workforce, and public confidence.”

Anatomy of a forecast

- On each week (time index t), available data report on weekly hospitalizations up to $t - 1$
- We submit predictive distributions for weeks t through $t + 3$
- Distributions are represented by their quantiles at 23 probability levels: 0.01, 0.025, 0.05, ..., 0.99

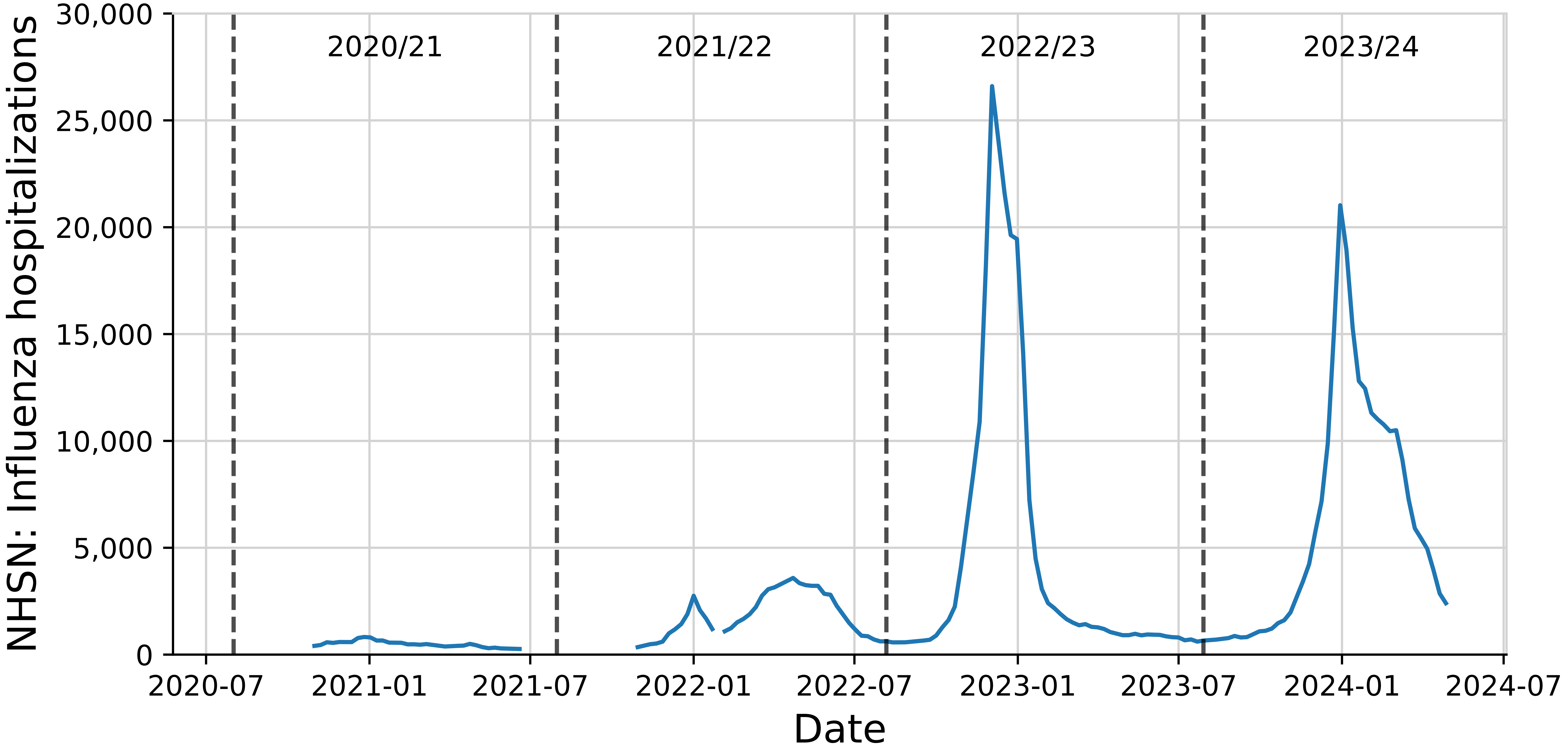


Overview of this talk

- Motivation, preview of results
- Modeling approaches
  - **Model 1**
  - Model 2
  - Flusion: an ensemble model
- Conclusions

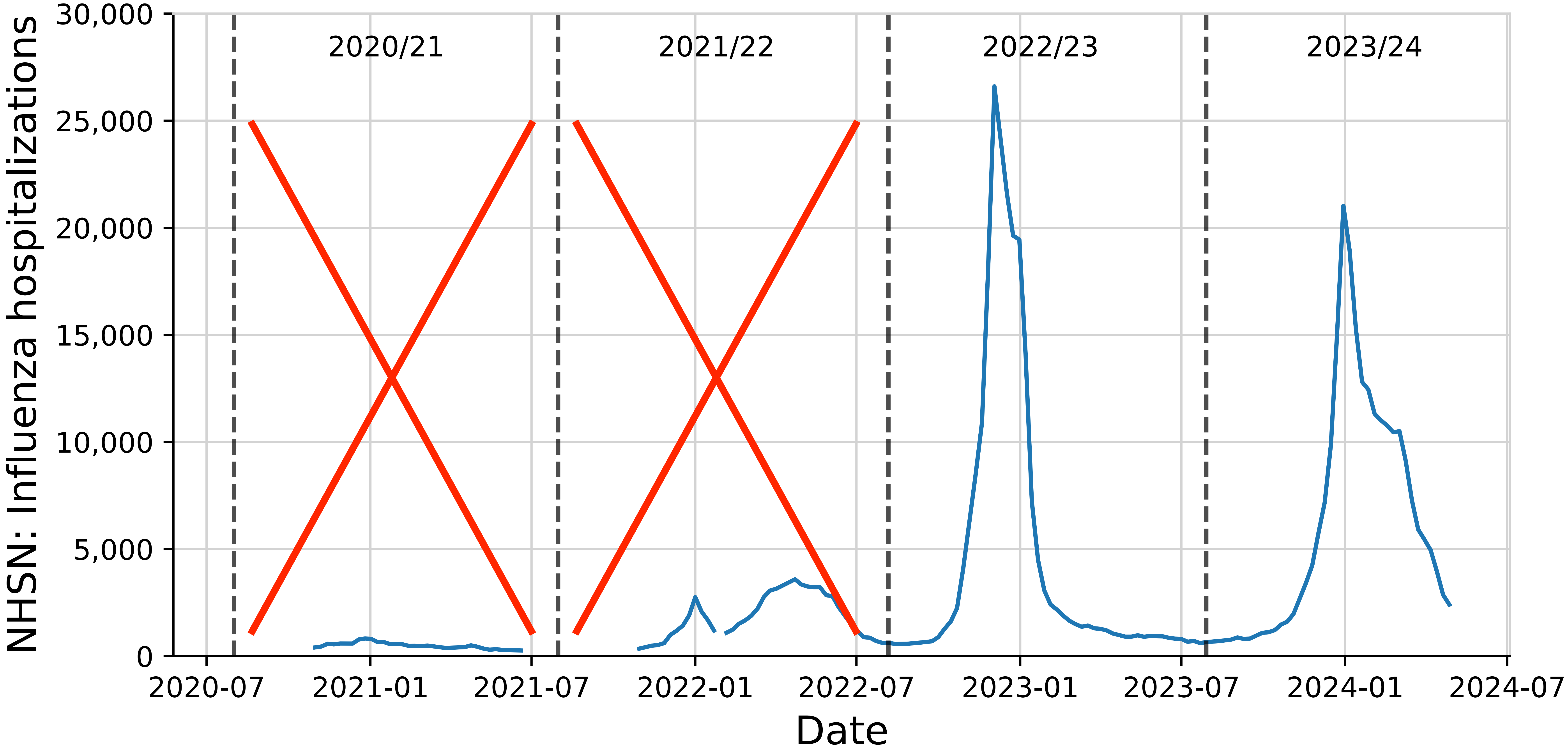
A central data challenge

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



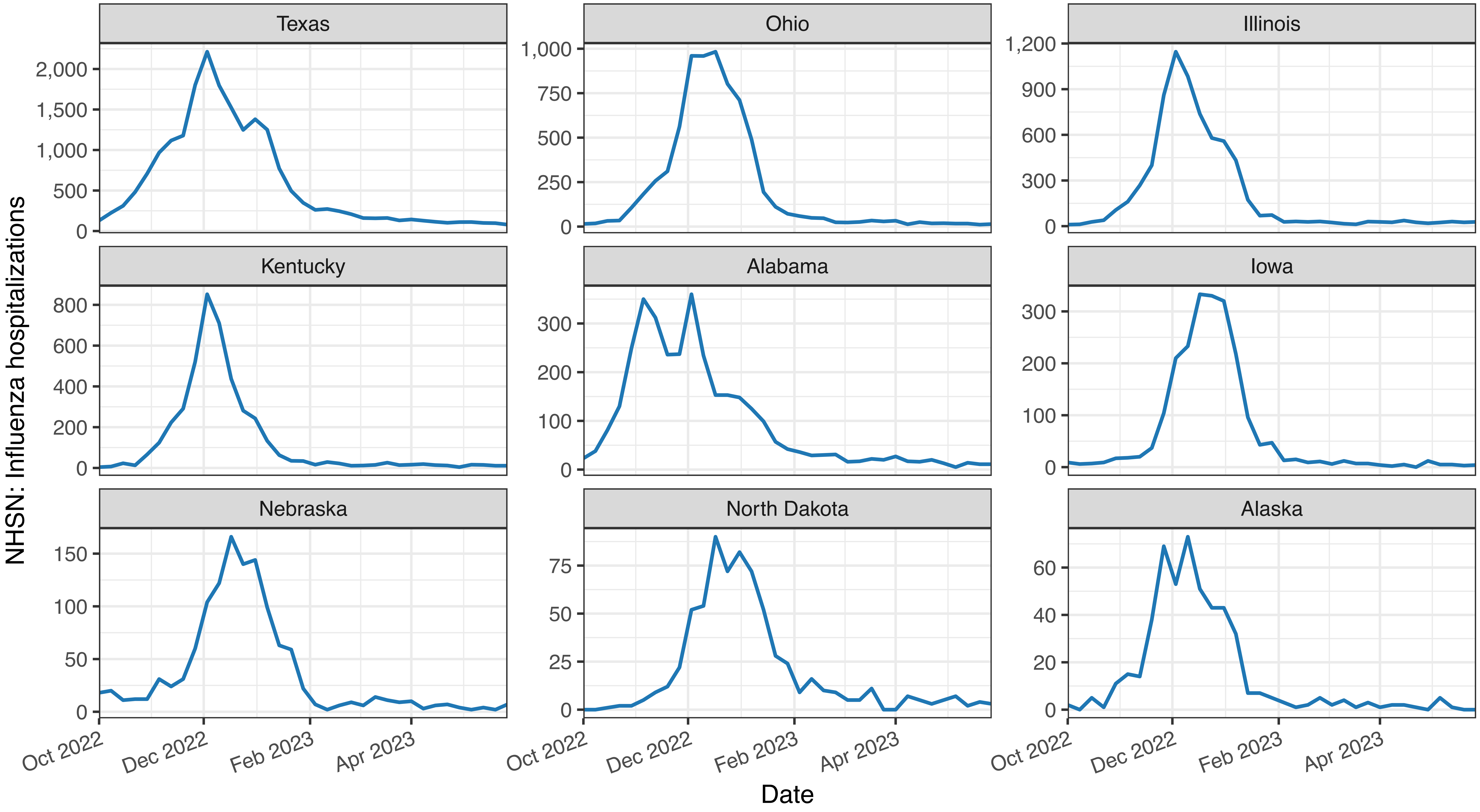
A central data challenge

- 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



What would you do?

- You have 1 season of available historical data for each of 52 states/jurisdictions (9 shown)
- What model would you try first?



My first model choice: autoregressive

- Notation: $Z_{l,t}$ is our modeled variable (hospitalizations*) for location l and time t
- With a Bayesian treatment:

$$Z_{l,t} \mid z_{l,t-1}, \ldots, z_{l,t-J} \sim \text{Normal}\left(\sum_{j=1}^{J} \alpha_j z_{l,t-j}, \sigma_{\varepsilon,l}^2\right)$$

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

$$\psi, \sigma_{\varepsilon,l} \sim \text{Half-Cauchy}(0,1)$$

My first model choice: autoregressive

- Notation: $Z_{l,t}$ is our modeled variable (hospitalizations*) for location l and time t
- With a Bayesian treatment:

$$Z_{l,t} \mid z_{l,t-1}, \ldots, z_{l,t-J} \sim \text{Normal}\left(\sum_{j=1}^{J} \alpha_j z_{l,t-j}, \sigma_{\varepsilon,l}^2\right)$$

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

$$\psi, \sigma_{\varepsilon,l} \sim \text{Half-Cauchy}(0,1)$$

- The α_j coefficients are shared across locations!
  - We have 35 observations per location from the 2022/23 season; that's not much!
  - Pooling gets us a better ratio of parameters to available data

My first model choice: autoregressive

- Notation: $Z_{l,t}$ is our modeled variable (hospitalizations*) for location l and time t
- With a Bayesian treatment:

$$Z_{l,t} \mid z_{l,t-1}, \ldots, z_{l,t-J} \sim \text{Normal}\left(\sum_{j=1}^{J} \alpha_j z_{l,t-j}, \sigma^2_{\varepsilon,l}\right)$$

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

$$\psi, \sigma_{\varepsilon,l} \sim \text{Half-Cauchy}(0,1)$$

- The α_j coefficients are shared across locations!
  - We have 35 observations per location from the 2022/23 season; that's not much!
  - Pooling gets us a better ratio of parameters to available data
- Kept separate variance parameters $\sigma^2_{\varepsilon,l}$ for each location
  - noise levels depend strongly on population size

My first model choice: autoregressive

- Notation: $Z_{l,t}$ is our modeled variable (hospitalizations*) for location l and time t
- With a Bayesian treatment:

$$Z_{l,t} \mid z_{l,t-1}, \ldots, z_{l,t-J} \sim \text{Normal}\left(\sum_{j=1}^{J} \alpha_j z_{l,t-j}, \sigma^2_{\varepsilon,l}\right)$$

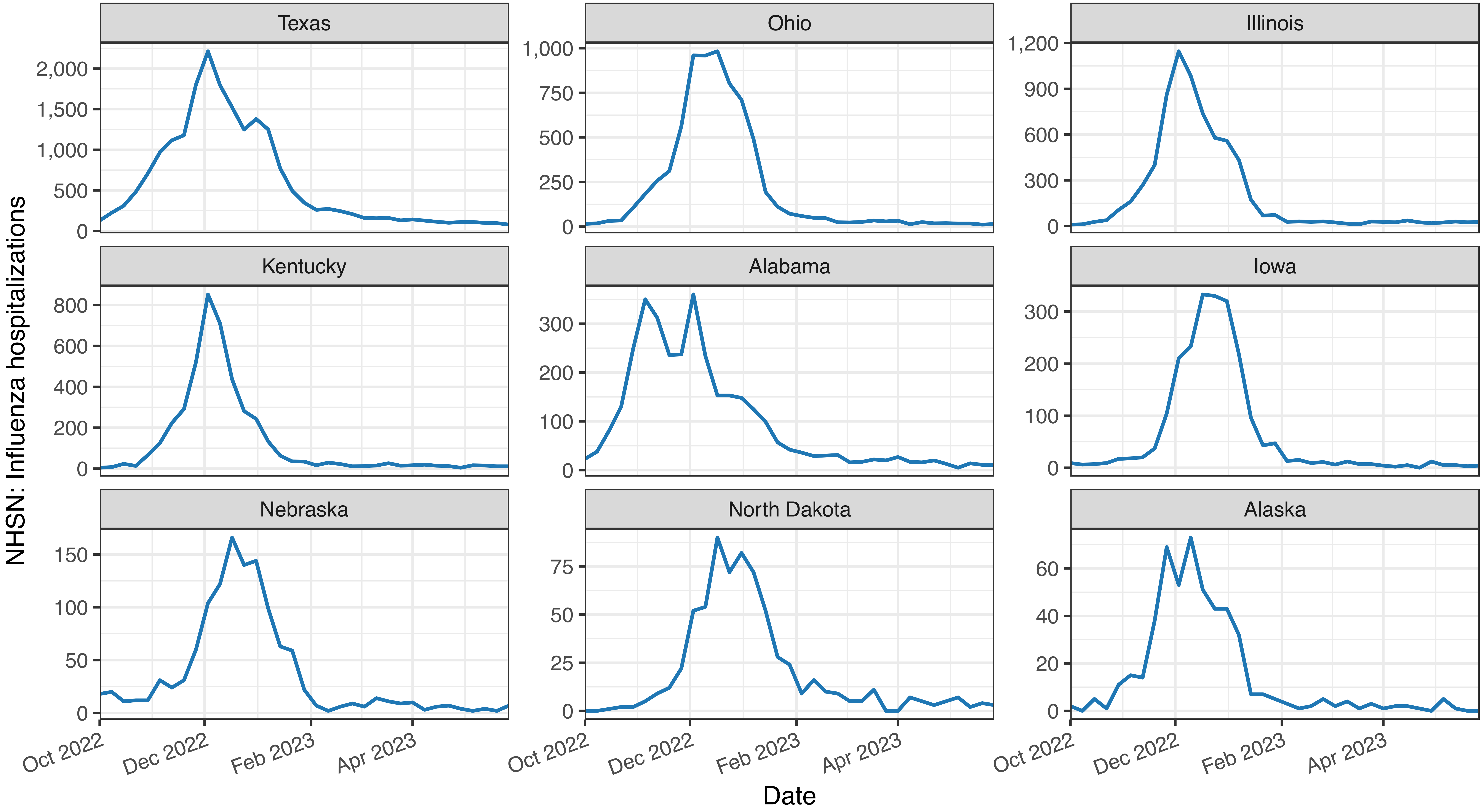
$$\alpha_j \mid \psi \sim \text{Normal}(0, \psi^2)$$

$$\psi, \sigma_{\varepsilon,l} \sim \text{Half-Cauchy}(0,1)$$

- The α_j coefficients are shared across locations!
  - We have 35 observations per location from the 2022/23 season; that's not much!
  - Pooling gets us a better ratio of parameters to available data
- Kept separate variance parameters $\sigma^2_{\varepsilon,l}$ for each location
  - noise levels depend strongly on population size
- Chose $J = 8$ based on intuition/a guess

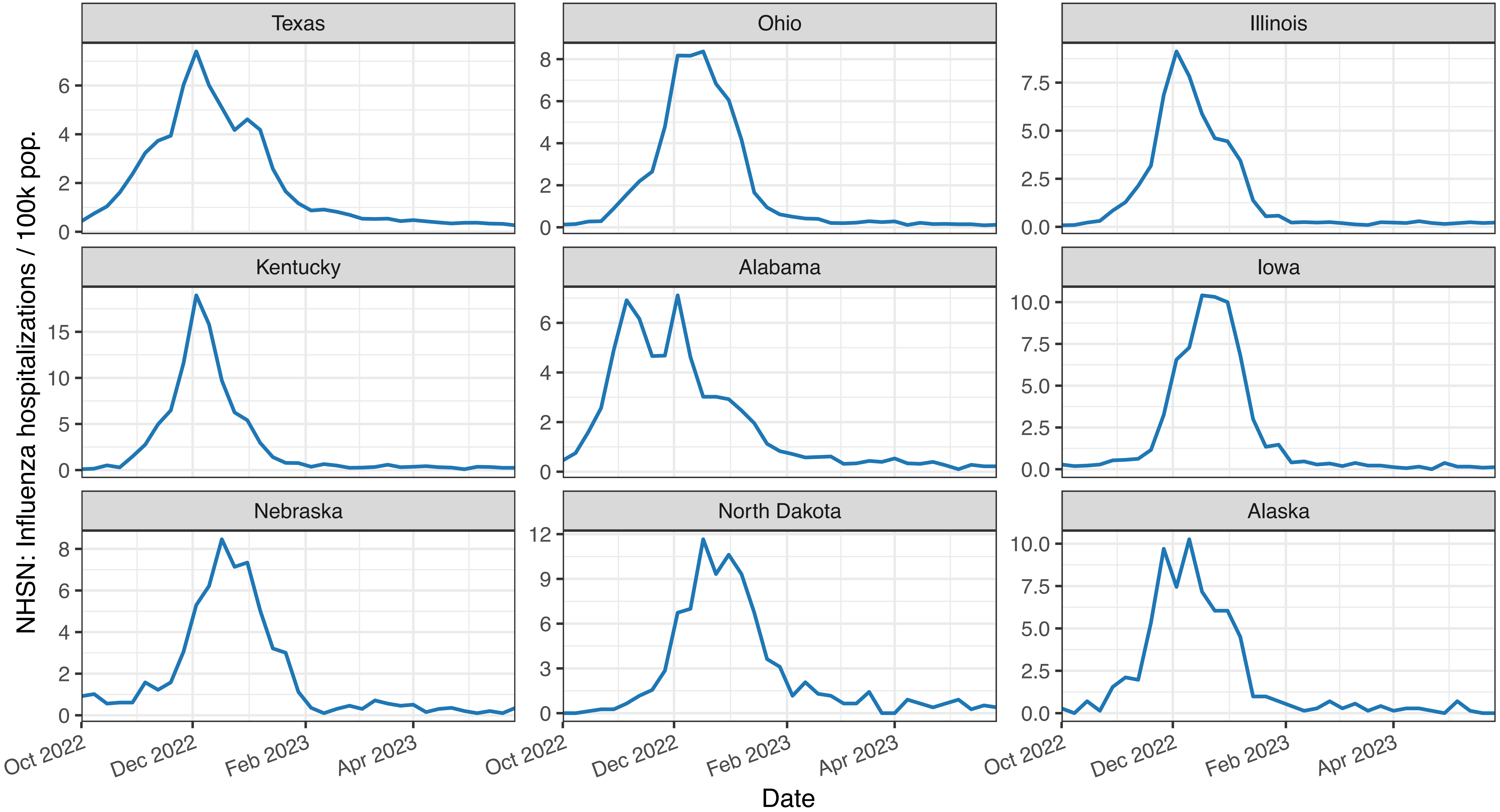
Data transformations for AR model

- Challenges:
  - Our target signal is hospital admissions, which varies in magnitude with state population
  - The signal has more variability around its trend near the peak than in the shoulders



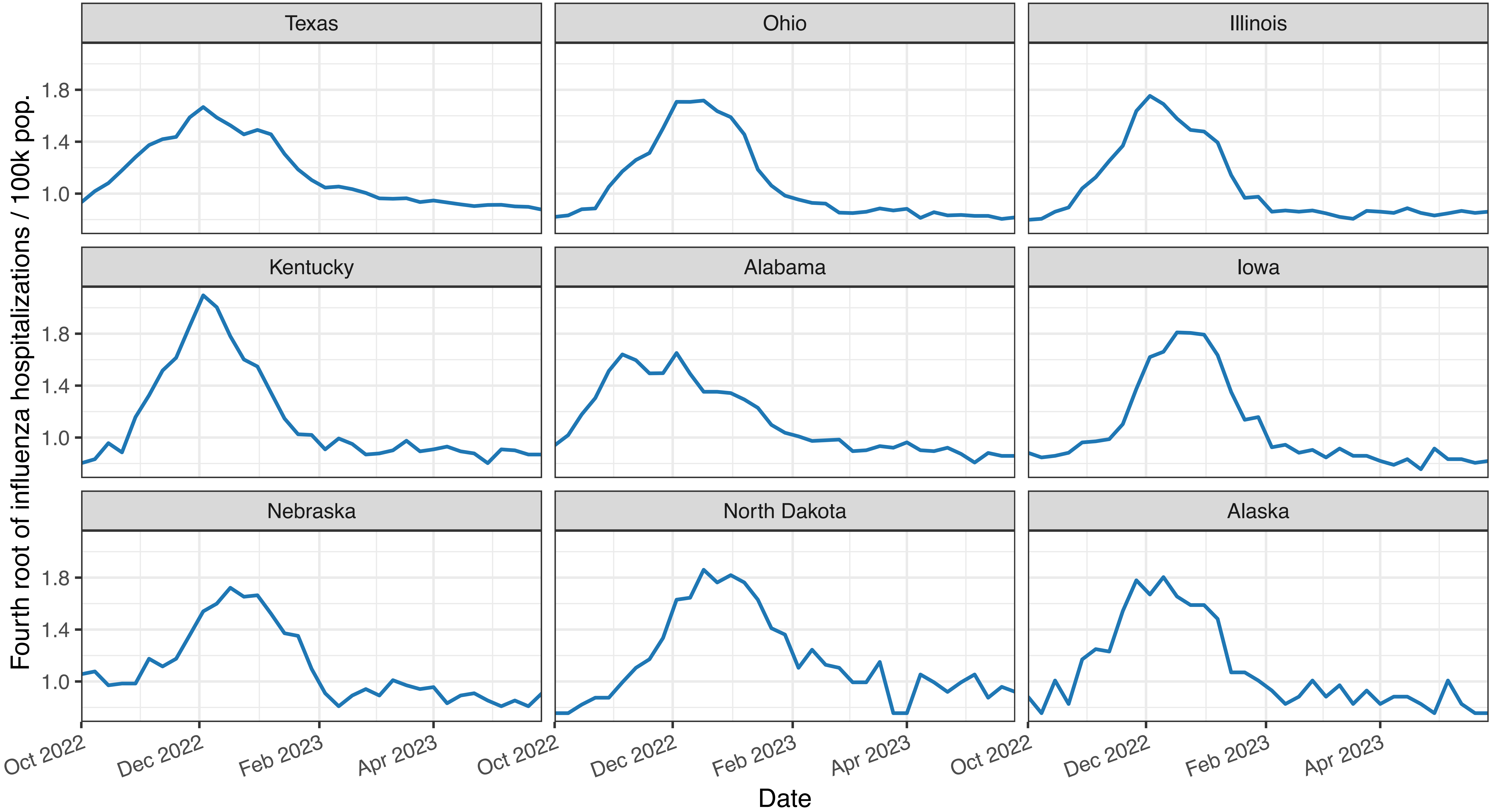
Data transformations for AR model

- What I did (I think this could be refined/simplified):
  - Convert to a hospitalization rate per 100k population



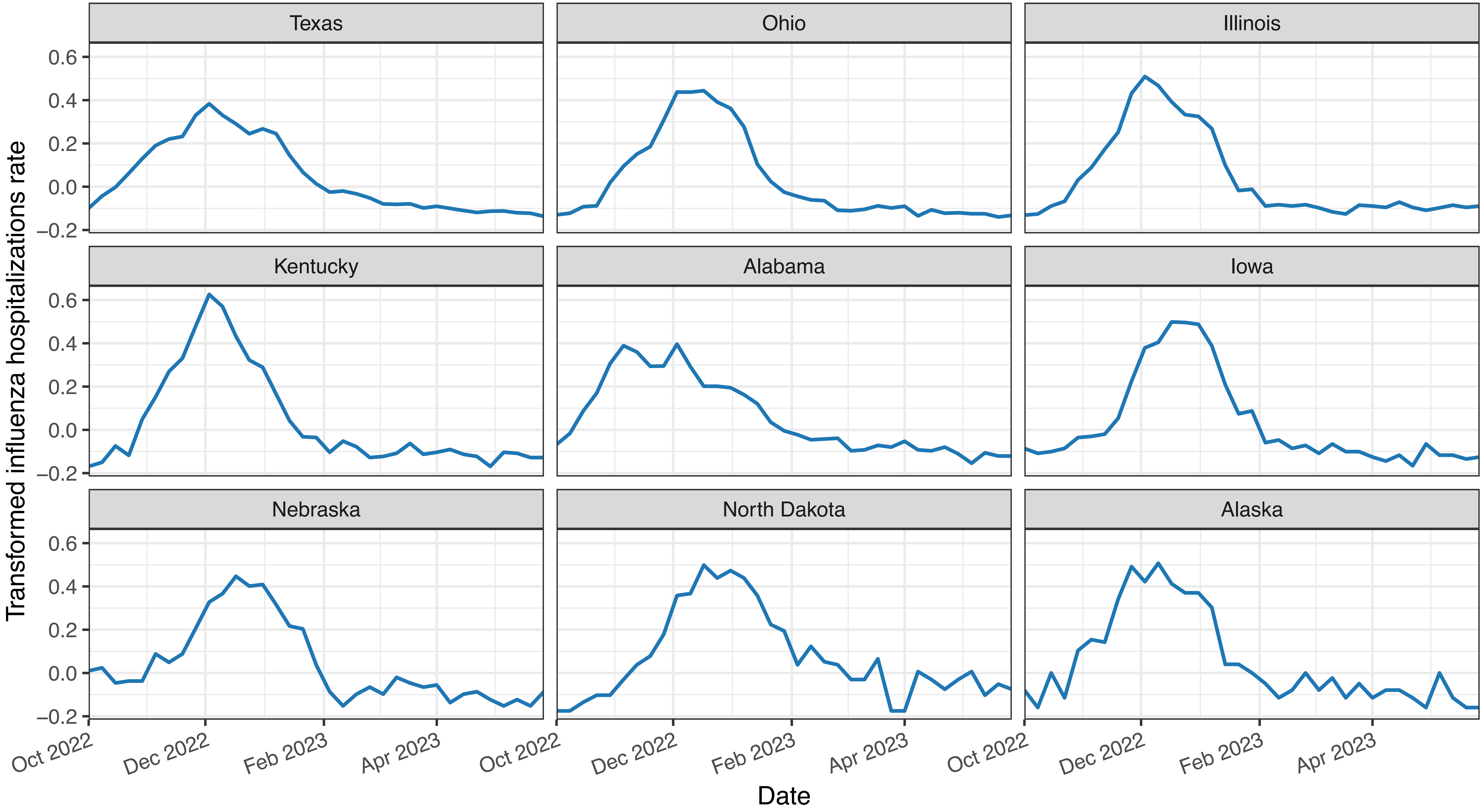
Data transformations for AR model

- What I did (I think this could be refined/simplified):
  - Convert to a hospitalization rate per 100k population
  - Take the fourth root, with an offset of 0.325

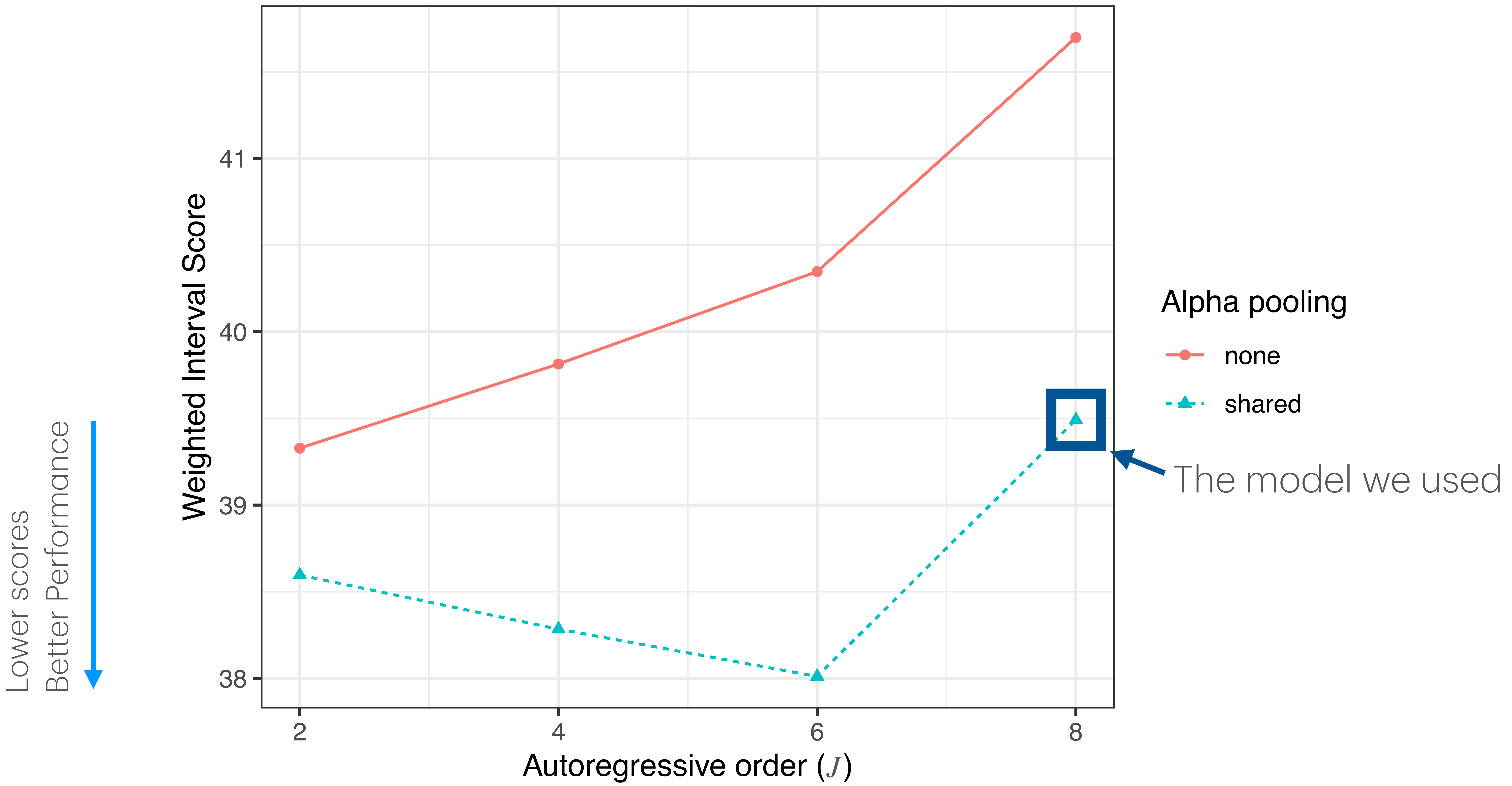


Data transformations for AR model

- What I did (I think this 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



Post-hoc evaluation results for AR model



AR performance in overall results

AR models with pooling; AR models without pooling, J=2 or J=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)

Model 1 Conclusion

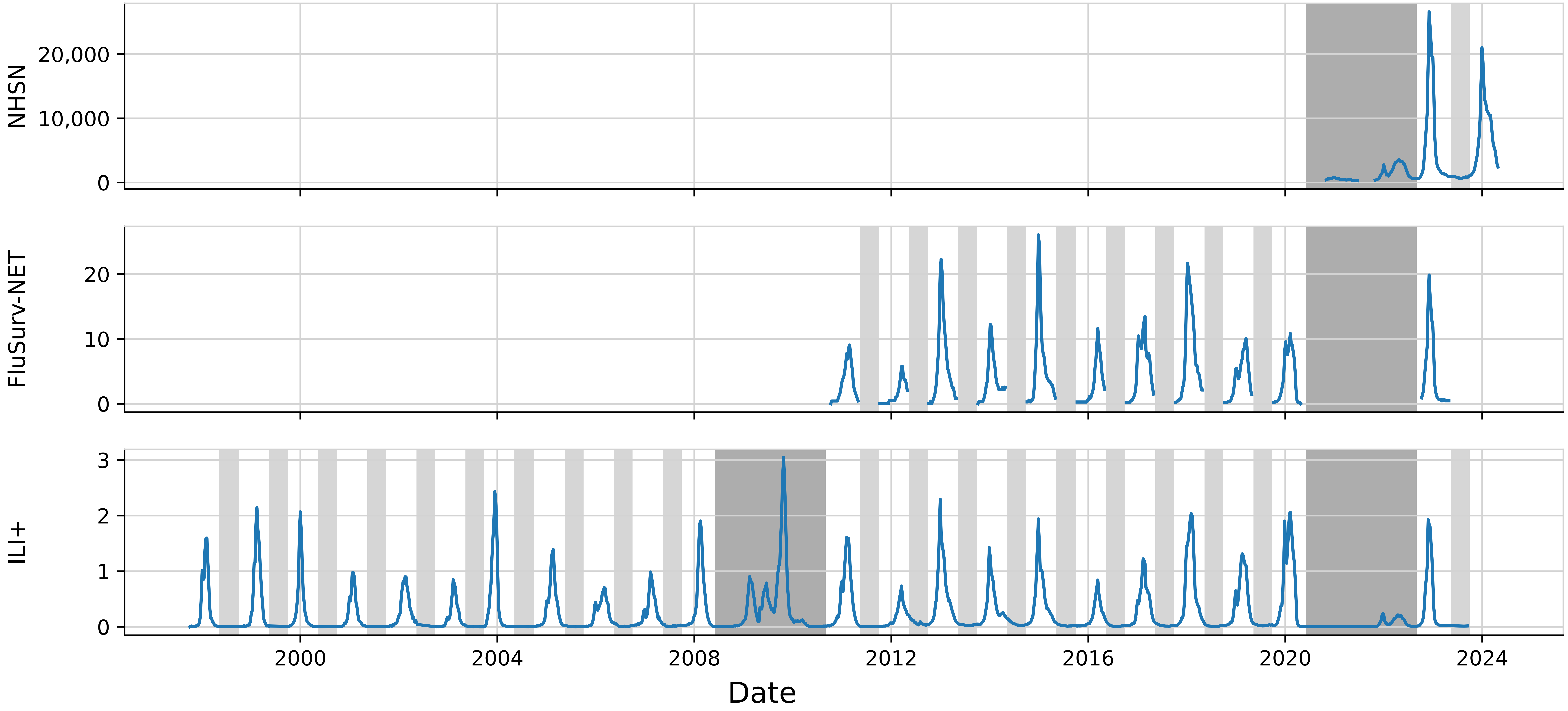
Sweating the details on a model from 1927
can generate some pretty decent forecasts

Overview of this talk

- Motivation, preview of results
- Modeling approaches
  - Model 1
  - **Model 2**
  - Flusion: an ensemble model
- Conclusions

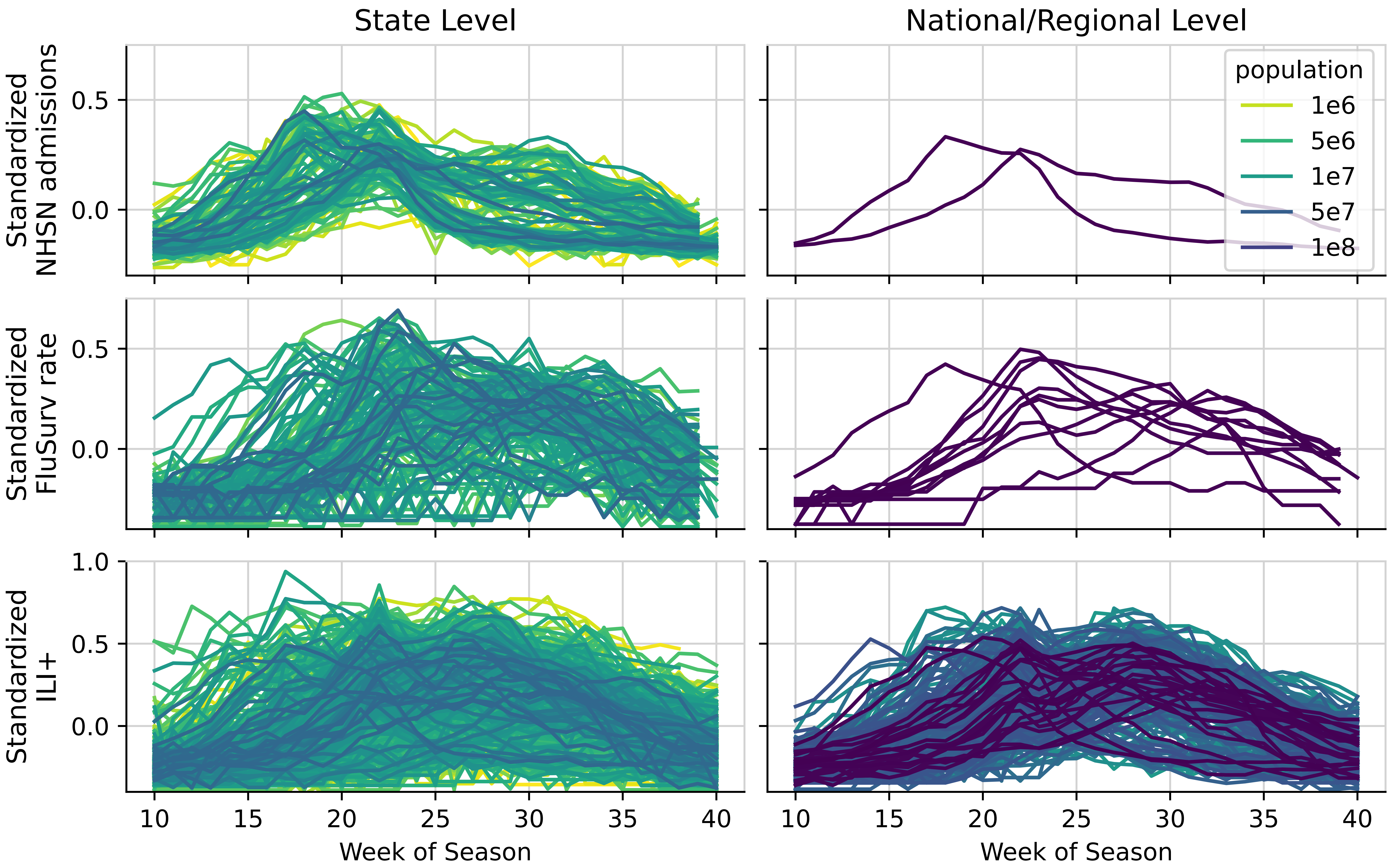
Actually, we have more data!

- 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

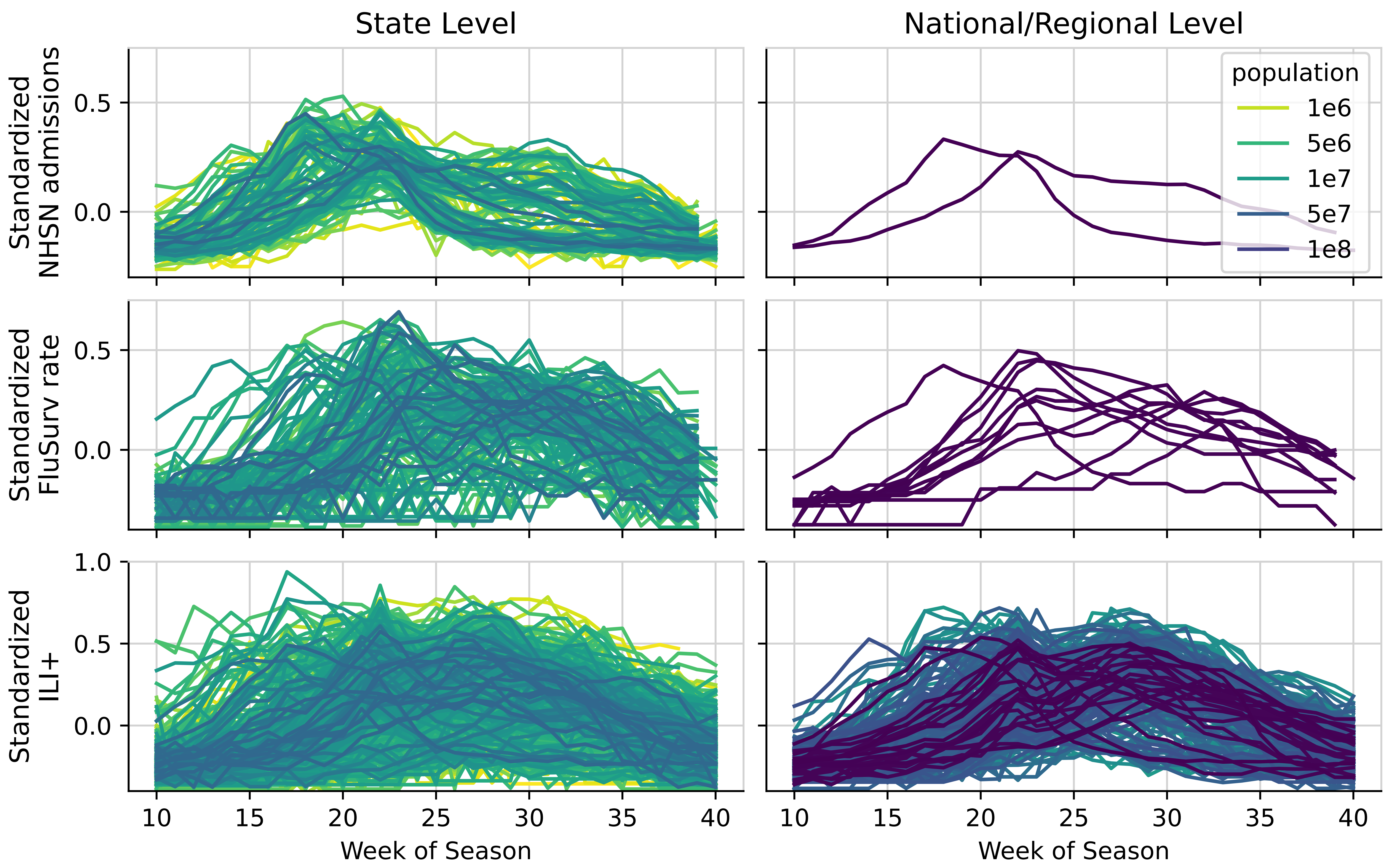


Data preprocessing

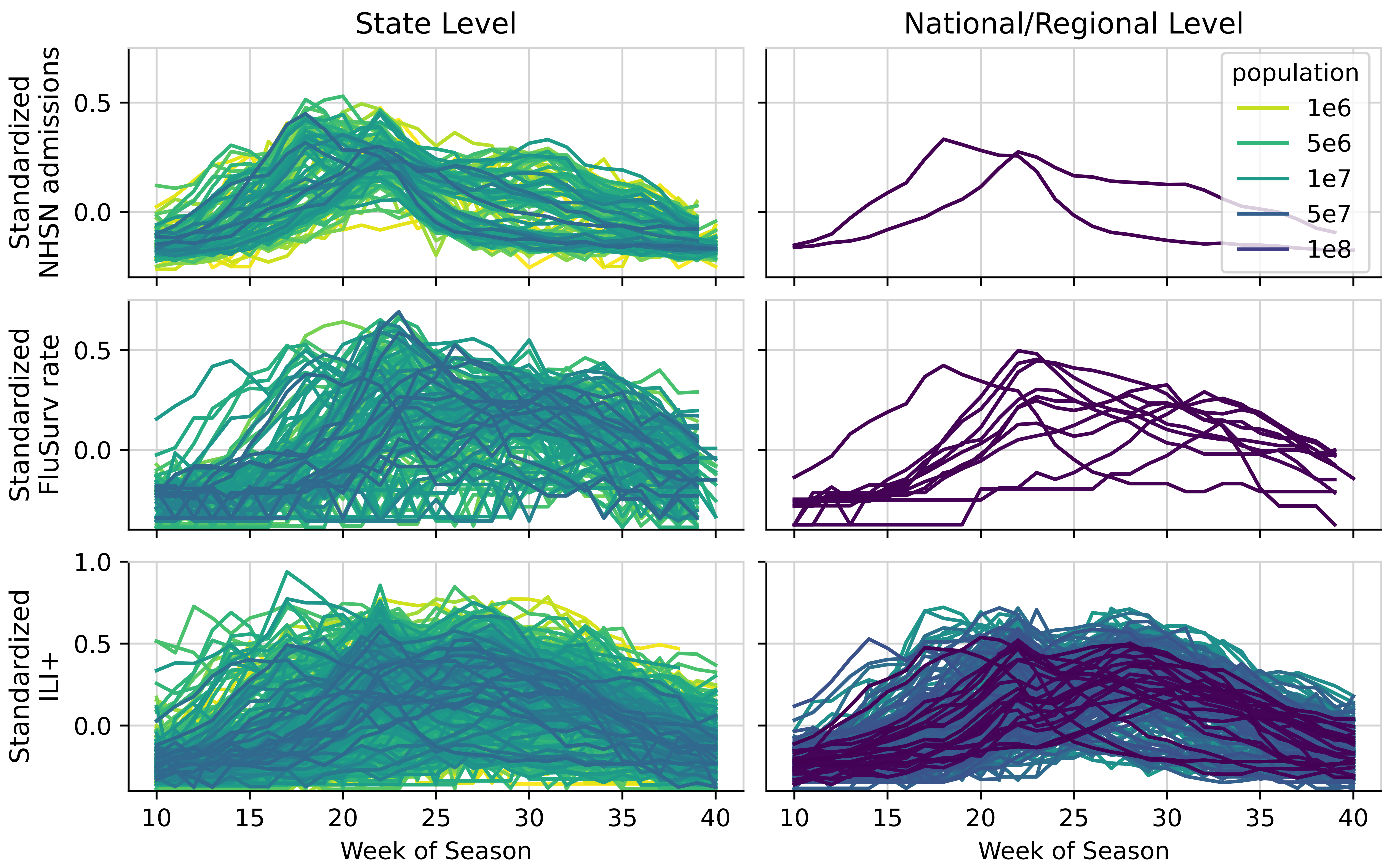
- We apply the same transformations we discussed for the AR model
  - Fourth root: stabilize variance across different times
  - Center and scale: put the data on a similar scale for different locations, signals



We have lots of data!



We have lots of data!



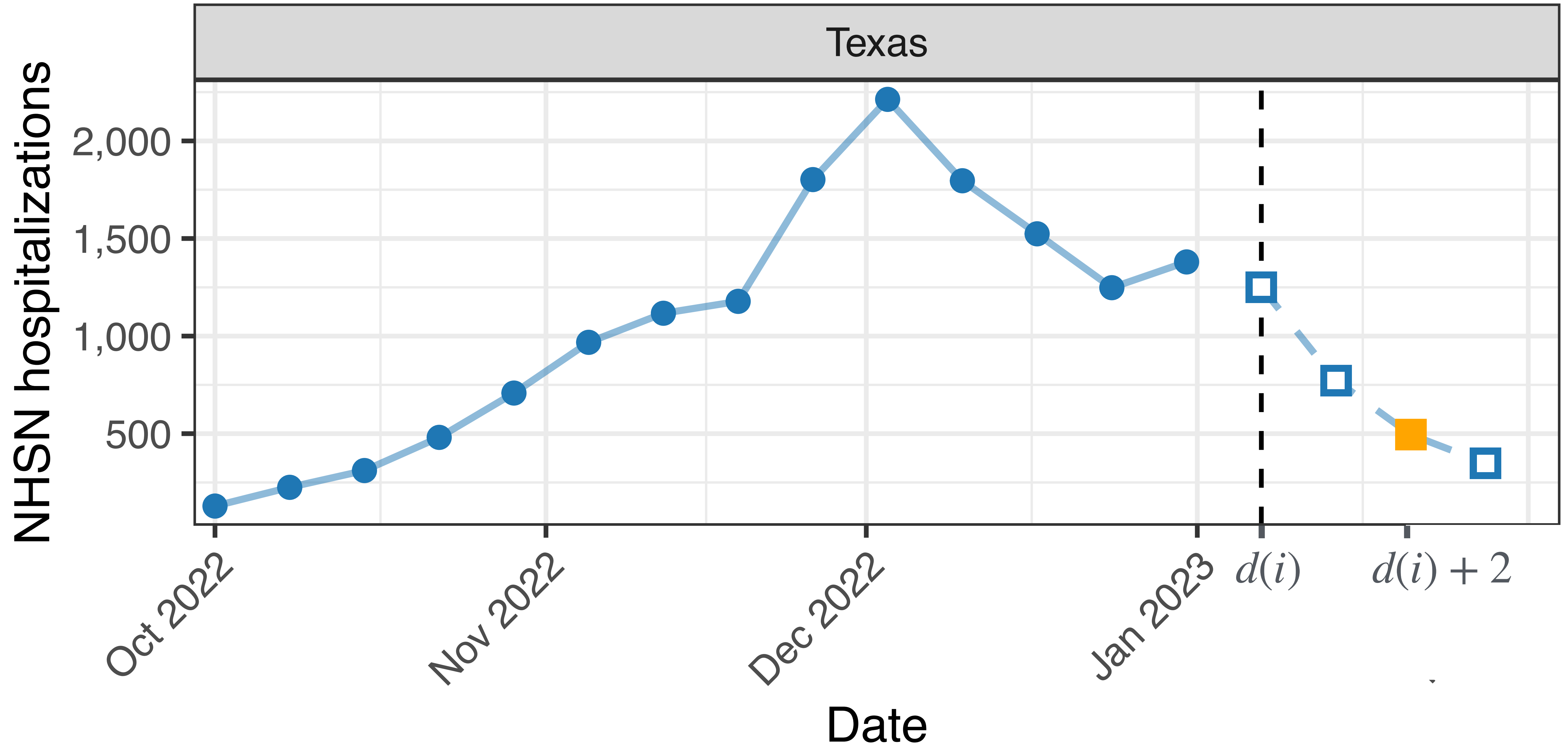
- **Let's do some machine learning**

Some notation

- Let i index ***forecast tasks*** defined by a combination of:
  - $s(i)$: data **s**ource (NHSN, FluSurv, ILI+)
  - $l(i)$: **l**ocation (which state, region, or national level unit)
  - $d(i)$: reference **d**ate, when we are making the forecast
  - $h(i)$: forecast **h**orizon: 0, 1, 2, or 3 weeks ahead
    - The target date for a prediction is $d(i) + h(i)$

Some notation

- Let i index ***forecast tasks*** defined by a combination of:
  - $s(i)$: data **s**ource (NHSN, FluSurv, ILI+)
  - $l(i)$: **l**ocation (which state, region, or national level unit)
  - $d(i)$: reference **d**ate, when we are making the forecast
  - $h(i)$: forecast **h**orizon: 0, 1, 2, or 3 weeks ahead
    - The target date for a prediction is $d(i) + h(i)$
- Example with $s(i)$ = NHSN, $l(i)$ = Texas, $d(i)$ = Jan. 7, 2023, $h(i)$ = 2



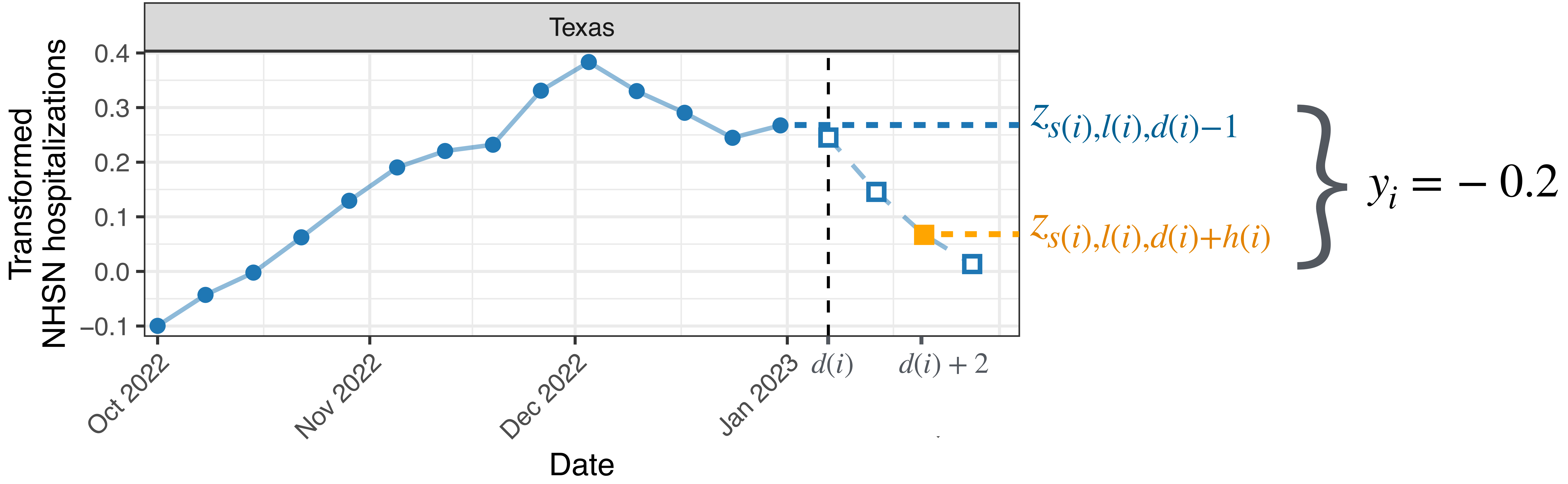
Set up as a regression problem

- For each task i, we form the pair (x_i, y_i):
  - x_i is a vector of features measuring how the signal $s(i)$ behaved in location $l(i)$ near date $d(i)$
    - See next slide
  - y_i is the prediction target

Set up as a regression problem

- For each task i, we form the pair (x_i, y_i):
  - x_i is a vector of features measuring how the signal $s(i)$ behaved in location $l(i)$ near date $d(i)$
    - See next slide
  - y_i is the prediction target
    - We set $y_i =$ difference in transformed data between the target date and the last observation

$$= z_{s(i),l(i),d(i)+h(i)} - z_{s(i),l(i),d(i)-1}$$



Features

- We used 114 features for each x_i

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

Features

- We used 114 features for each x_i
- Note: when predicting a target signal and location, features measure information only about that signal and location
- Example: $s(i)$ = NHSN, $l(i)$ = Texas
  - x_i does not contain any info. about FluSurv or ILI+
  - x_i does not contain any info. about New Mexico or Oklahoma

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

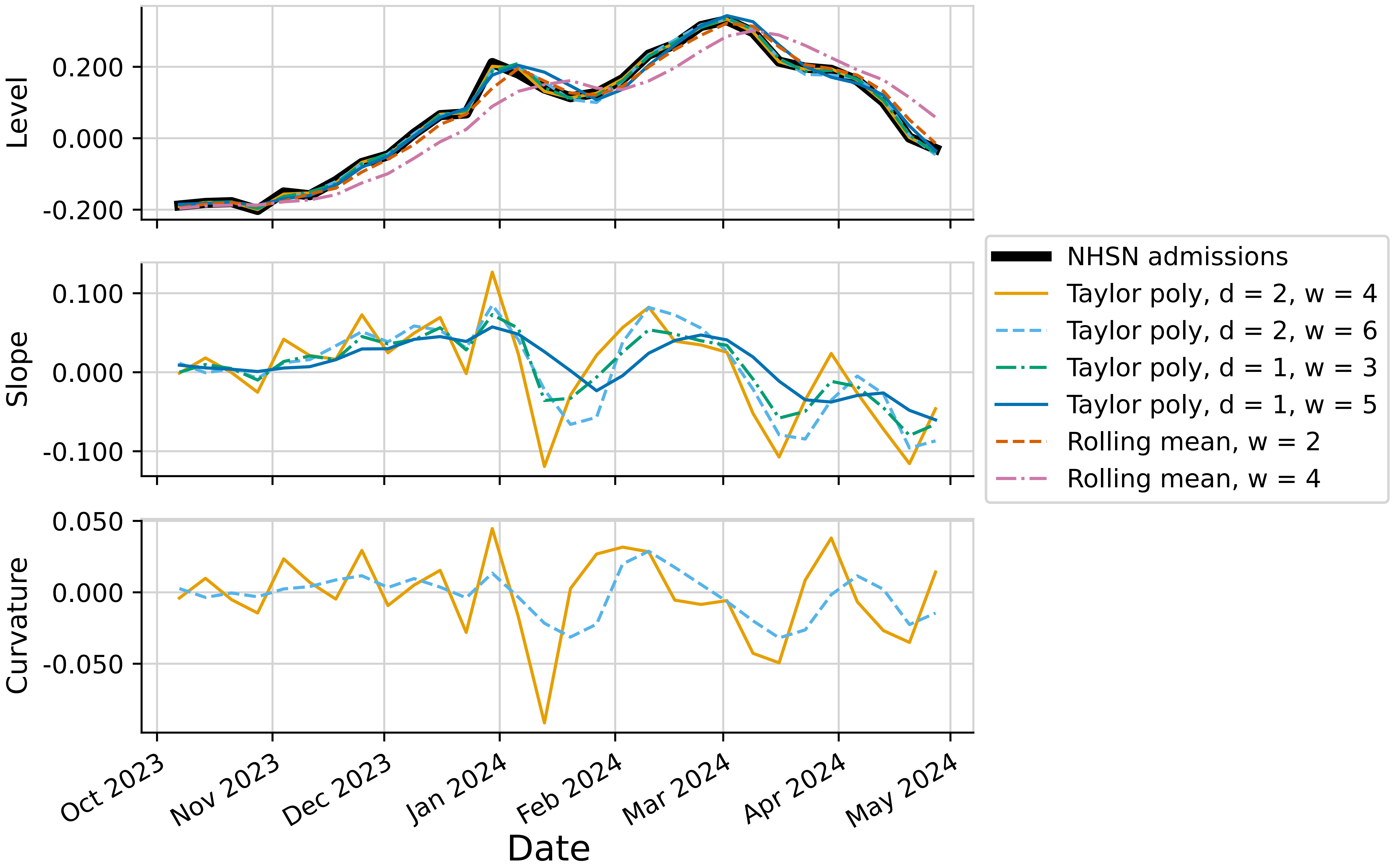
Features

- We used 114 features for each x_i
- Note: when predicting a target signal and location, features measure information only about that signal and location
- Example: $s(i)$ = NHSN, $l(i)$ = Texas
  - x_i does not contain any info. about FluSurv or ILI+
  - x_i does not contain any info. about New Mexico or Oklahoma
- However, the model is trained on examples (x_i, y_i) from all data sources and locations!

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

Local level, slope, and curvature features

- Example for Michigan, 2023/24 season



Gradient boosting quantile regression

- We now have a bunch of training examples (x_i, y_i)
- Goal: A predictive distribution for $Y_i \mid X_i = x_i$
  - In particular, we want quantiles of that distribution at the levels $\tau \in \{0.01, 0.025, \ldots, 0.99\}$

Gradient boosting quantile regression

- We now have a bunch of training examples (x_i, y_i)
- Goal: A predictive distribution for $Y_i \mid X_i = x_i$
  - In particular, we want quantiles of that distribution at the levels $\tau \in \{0.01, 0.025, \ldots, 0.99\}$
- We use gradient boosting quantile regression (GBQR)
  - For each τ, learn a flexible mapping $f_\tau(x)$ from features x to a predictive quantile at level τ
  - $f_\tau(x)$ takes the form of a sum of regression trees
  - Fit by targeting the quantile loss using the LightGBM package

Overview of this talk

- Motivation, preview of results
- Modeling approaches
  - Model 1
  - Model 2
  - **Flusion: an ensemble model**
- Conclusions

Flusion: an ensemble of 3 models

- Three component models:
  1. GBQR: A gradient boosting quantile regression model, all 114 features
  2. GBQR-no-level: Same as GBQR, but not allowed to see measures of local level of signal
     - Goal: introduce model diversity by including multiple variations on the model
     - This was unsuccessful
  3. ARX: Bayesian autoregressive model
     - This model also used 1 covariate: a spike function peaking at Christmas
     - This had essentially no impact on model performance

- Each model m produces a set of predictive quantiles for each $\tau \in \{0.01, 0.025, \dots, 0.99\}$

- At each quantile level τ, Flusion takes the average of the quantiles from these models:

$$\hat{q}_\tau^{(\text{Flusion})} = \frac{1}{3} \sum_m \hat{q}_\tau^{(m)}$$

Overall Results: FluSight 2023/24 season

Higher rank
Better 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

Lower rank
Worse Performance

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

Experiment A: Component models

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

Higher rank
Better 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

Top 4 include GBQR

- Primary driver was whether or not GBQR was included

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

Overview of this talk

- Motivation, preview of results
- Modeling approaches
  - Model 1
  - Model 2
- **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 2 other signals with a longer history

- Many model refinements are possible!
  - Use type or subtype-specific data to see multiple waves within 1 season
  - Use vaccination uptake and efficacy data to inform estimates of season severity
  - Improve handling of holiday effects
  - Allow the model to see contemporaneous data:
    - Other signals
    - Other locations
  - ...

Summary & conclusions

- Simple methods like AR models can carry you a long way
- But modern methods and careful use of multiple data sources really are valuable!
- 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!

- Questions?

- Acknowledgments to co-authors: Yijin Wang, Russel D. Wolfinger, Nicholas G. Reich

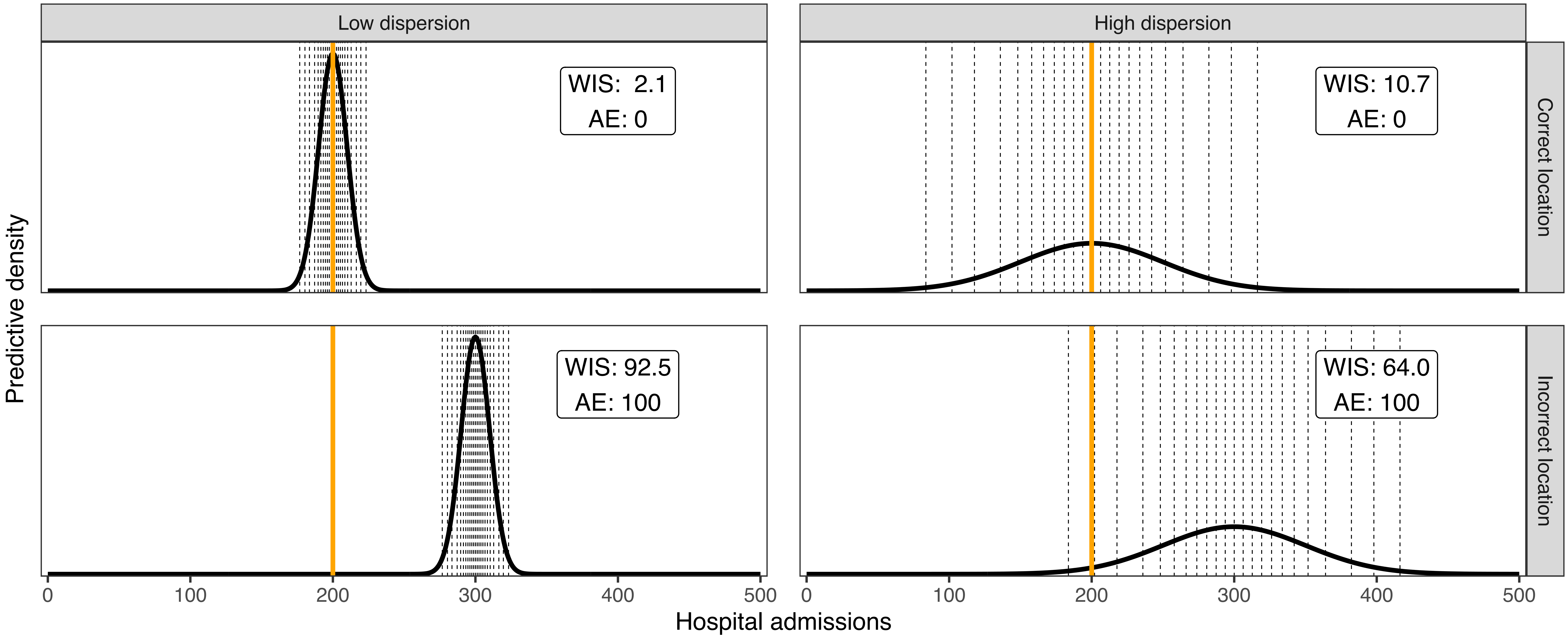
- Acknowledgments to funders:

  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.

- Preprint available on arXiv: https://arxiv.org/abs/2407.19054

Evaluation: Weighted Interval Score (WIS)

- WIS can be interpreted as a generalization of absolute error to a set of predictive quantiles
  - Measures the distance of the distribution from the observed outcome
  - Lower is better
  - Approximates CRPS as the number of quantiles increases; equivalent to average pinball loss



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 $\tau_1, \ldots, \tau_K$.
    $$WIS(\{q_k : k = 1,\ldots,K\}, y) = \frac{1}{K}\sum_k 2 \cdot QS_{\tau_k}(q_k, y)$$
  - $$QS_{\tau_k}(q_k, zy) = \tau_k \max(y - q_k, 0) + (1 - \tau_k) \max(q_k - y, 0)$$

- 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:

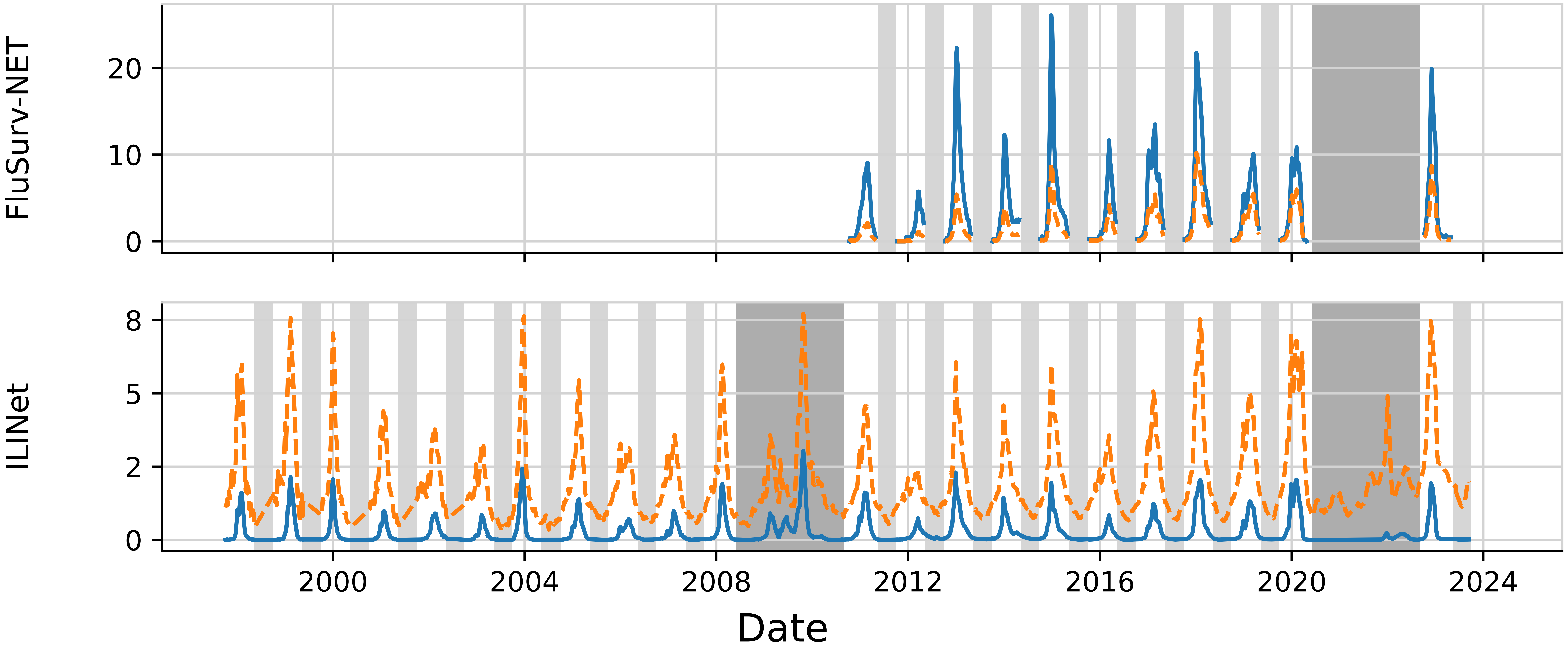
- different forecasters submit predictions for different locations and dates
- MAE and WIS are sensitive to the scale of the prediction target
- MAE and WIS values for forecasts in different locations and dates are not comparable

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 $MAE^{m}_{\mathcal{I}_{m,m'}}$ and $MAE^{m'}_{\mathcal{I}_{m,m'}}$
2. For model m, compute the geometric mean of ratios of MAEs for m compared to all other models
$$\theta^m = \left(\prod_{m' \neq m} \frac{MAE^{m}_{\mathcal{I}_{m,m'}}}{MAE^{m'}_{\mathcal{I}_{m,m'}}} \right)^{1/(M-1)}$$
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



Holiday effects in the ARX model

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

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

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

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

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

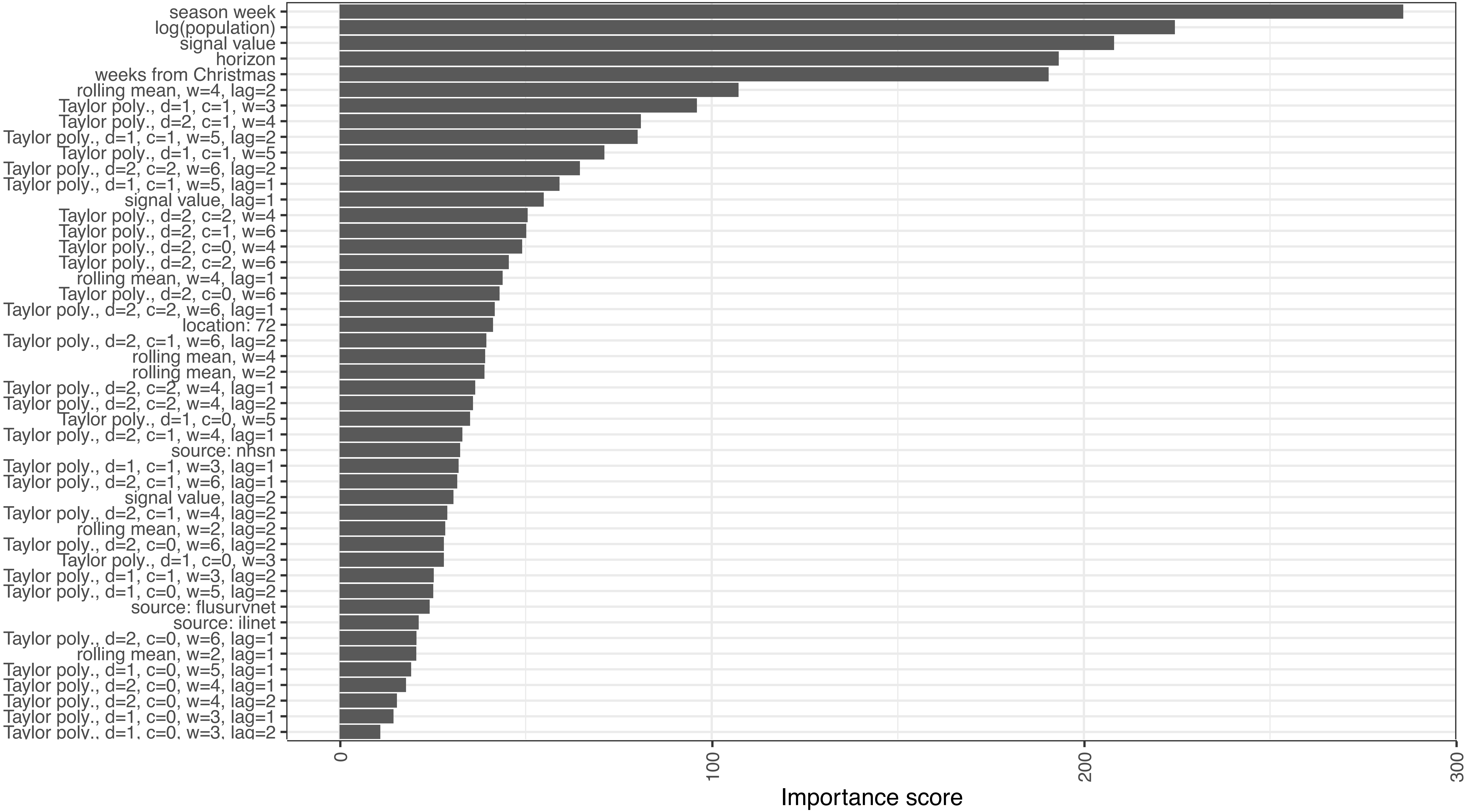
- Key idea for AR setup:
  - AR coefficients shared across locations (to avoid overfitting to limited data)
  - Separate variance innovation term per location (noise levels differ based on population)
- We used 1 covariate:
  - takes the value 3 on Christmas week
  - 2 one week before and one week after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

Intro. to gradient boosting (formulas)

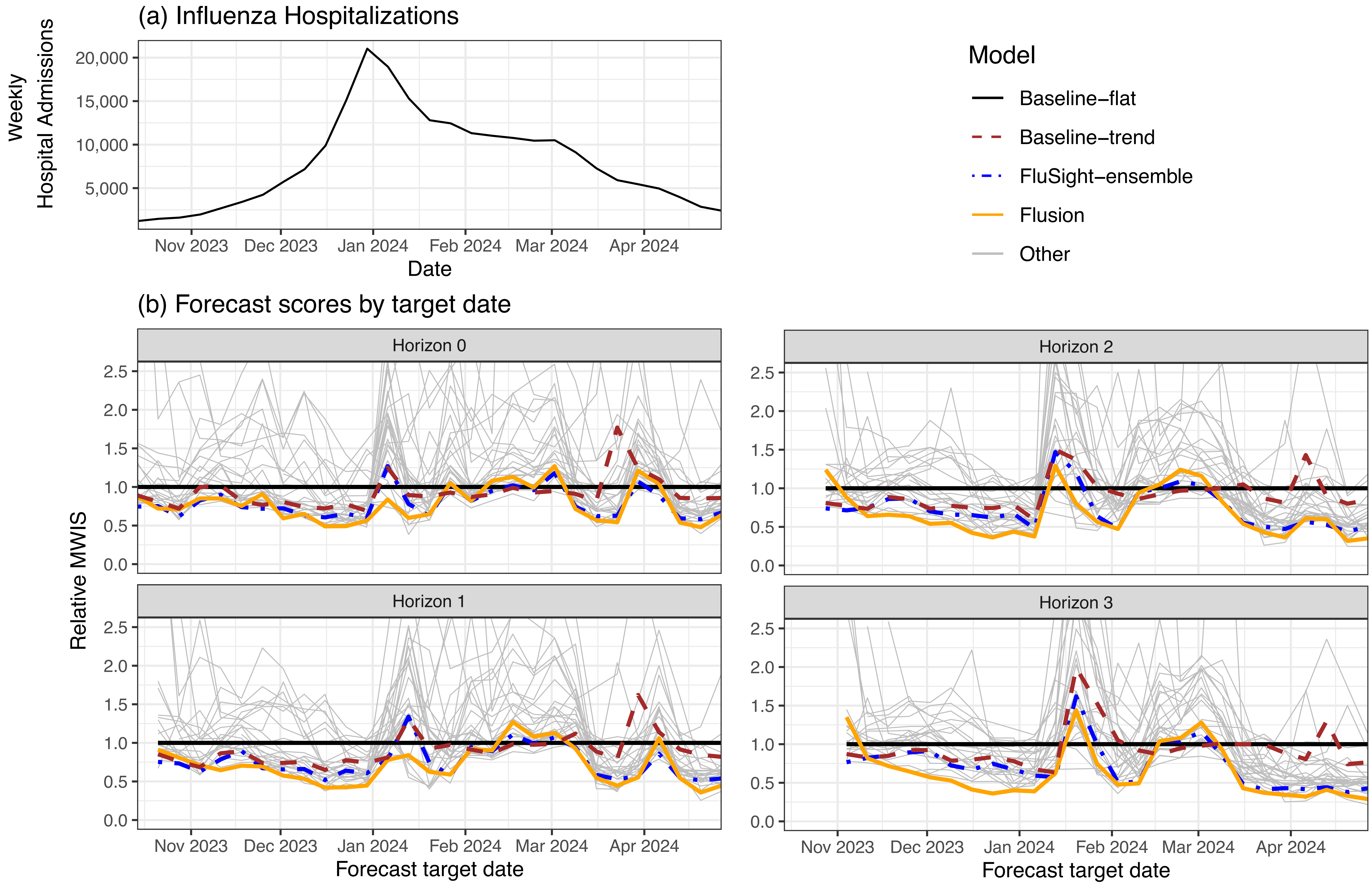
- Inputs:
    - Training set with pairs (x_i, y_i), $i = 1,\ldots,n$. Define $\mathbf{y} = (y_1, \ldots, y_n)^\mathsf{T}$
    - Loss function $L(\hat{y}_i, y_i)$: measures how well $\hat{y}_i$ estimates y_i (lower loss is better)
- Method:
    - We will construct the regression function $f(x) = \sum_b f^{(b)}(x)$ iteratively
        - each $f^{(b)}(x)$ will be a regression tree
    - After step $b-1$, we have the predictions $\hat{y}_i^{(b-1)} = \sum_{a=1}^{b-1} f^{(a)}(x_i)$, collected in the vector $\hat{\mathbf{y}}^{(b-1)}$
    - After step $b-1$, the total loss is $L_{tot}(\hat{\mathbf{y}}^{(b-1)}) = \sum_i L(\hat{y}_i^{(b-1)}, y)$
    - $\delta_i^{(b-1)} = -\dfrac{\partial}{\partial \hat{y}_i} L_{tot}(\hat{\mathbf{y}})\,|_{\hat{\mathbf{y}}=\hat{\mathbf{y}}^{(b-1)}}$ indicates how to change $\hat{y}_i^{(b-1)}$ so as to reduce the total loss.
    - We fit the next regression tree $f^{(b)}(x)$ to the pairs $(x_i, \delta_i^{(b-1)})$

Feature Importance

- Importance score: number of tree splits using feature



MWIS by date and forecast horizon



Quantile loss

- Suppose we have a set of observations $y_i \sim \mathcal{D}, i = 1, \ldots, n$
- We want to estimate the quantile q of the distribution $\mathcal{D}$ at probability level τ: $P_{\mathcal{D}}(Y \leq q) = \tau$
- $QS_\tau(q, y) = (1 - \tau)\max(q - y, 0) + \tau \max(y - q, 0)$

- For a fixed value of y, the derivatives with respect to q are:
- $\frac{\partial}{\partial q} QS_\tau(q, y) = -\tau$ if $y > q$ and $\frac{\partial}{\partial q} QS_\tau(q, y) = 1 - \tau$ if $y < q$

Optimizing quantile loss

- For a fixed y_i, the derivatives with respect to q are:
  - $\frac{\partial}{\partial q} QS_\tau(q, y_i) = -\tau$ if $q < y_i$ and $\frac{\partial}{\partial q} QS_\tau(q, y_i) = 1 - \tau$ if $q > y_i$
- Averaging across all i:

$$\frac{\partial}{\partial q} \frac{1}{n} \sum_i QS_\tau(q, y_i) = (\text{proportion with } q < y_i) \cdot (1 - \tau) - (\text{proportion with } q > y_i) \cdot \tau$$

- Imagine applying gradient descent to find q:
  - The derivative w.r.t. q is 0 if the proportions are "balanced", with (proportion with $y_i > q$) $= \tau$ and (proportion with $y_i < q$) $= 1 - \tau$
    - In other words, q is a τ-quantile of the sample y_i's
  - The derivative is negative if