

UMassAmherst

School of Public Health
& Health Sciences

Biostatistics and Epidemiology

The computational science of infectious disease forecasting

Nicholas G. Reich

International Symposium on Forecasting

29 June 2021

 covid19forecasthub.org

 reichlab.io

 [reichlab](https://twitter.com/reichlab)



Reich
Lab | AT UMASS
AMHERST

Talk outline

1. Epidemic forecasting: some background
2. The "Hub" approach in epidemic forecasting
3. Adventures in ensemble building for COVID-19
4. Optimizing infrastructure for Hubs
5. Closing thoughts on the computational science of forecasting

Epidemic forecasting: some background



Reich Lab

Background

- Since 2011, in UMass-Amherst Dept of Biostats and Epidemiology.
- We focus on application of statistical models and software development in **real-time, operational collaborations** with public health agencies.
- Collaborators include US CDC, Thailand Ministry of Public Health, NYC Dept of Health, Massachusetts Dept of Public Health, Veterans Health Administration.
- In 2019, named a **CDC Influenza Forecasting Center of Excellence**.
- Since early 2020, our team has led the **COVID-19 Forecast Hub**, a large-scale collaborative effort to build ensemble forecasts that are used and disseminated by the US CDC.
- Since mid-2020, Dr. Evan Ray has co-led the lab with me.

What are models used to predict?

An epidemiological modeler's view on predicting the past, the near future, and the far future.

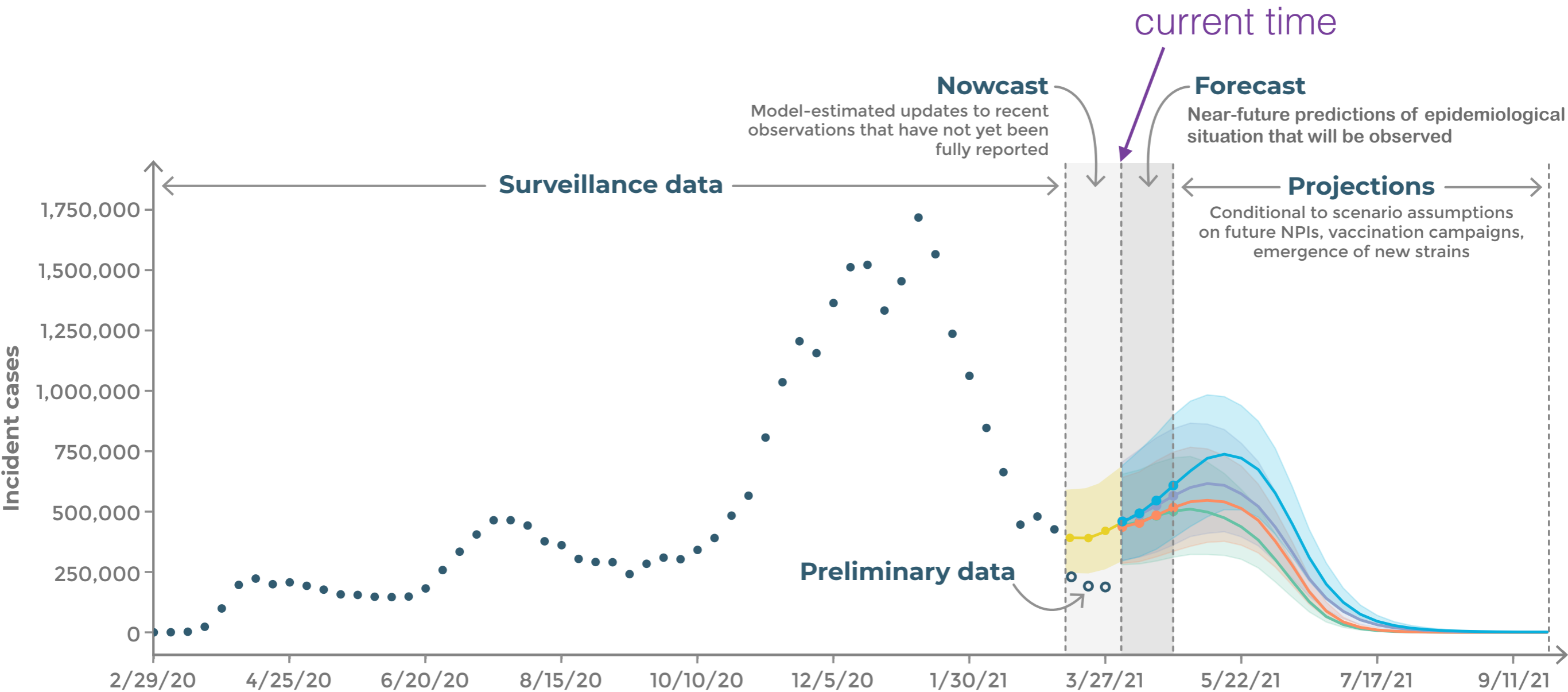


Image credit: Nicole Samay, Alex Vespignani,
via the Scenario Modeling Hub, <https://covid19scenariomodelinghub.org/>

Epidemic forecast feedback loop

- Weather forecasts don't impact the weather.
- An outbreak forecast could impact an outbreak.



2018: vector-control activities to prevent dengue in Thailand
courtesy of Sapon Iamsirithaworn, Thailand Department of Disease Control



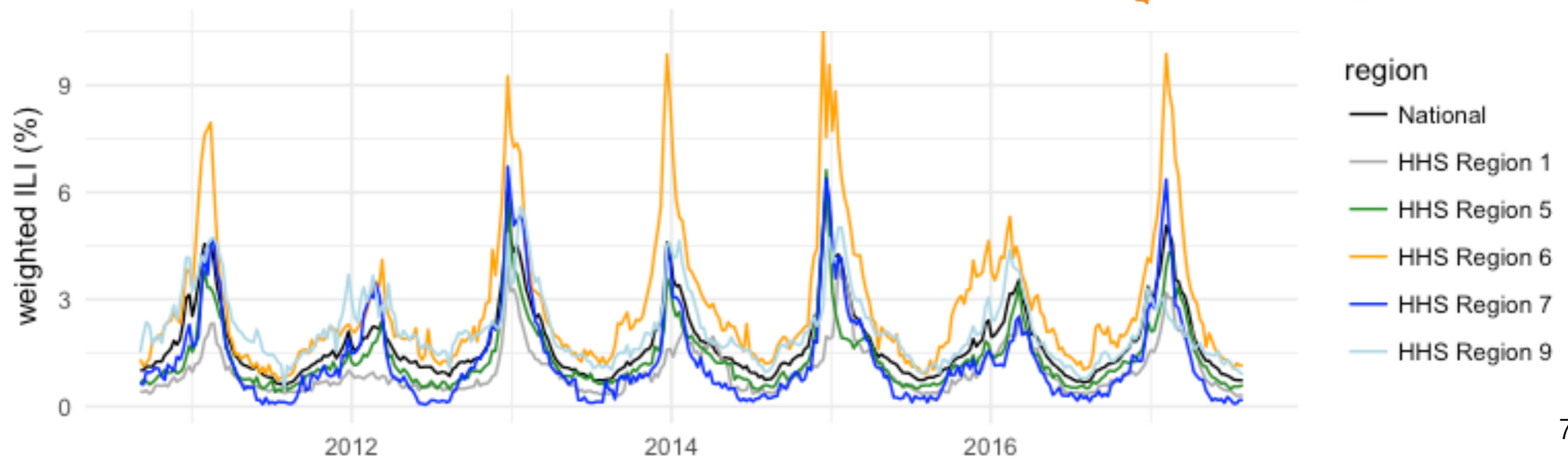
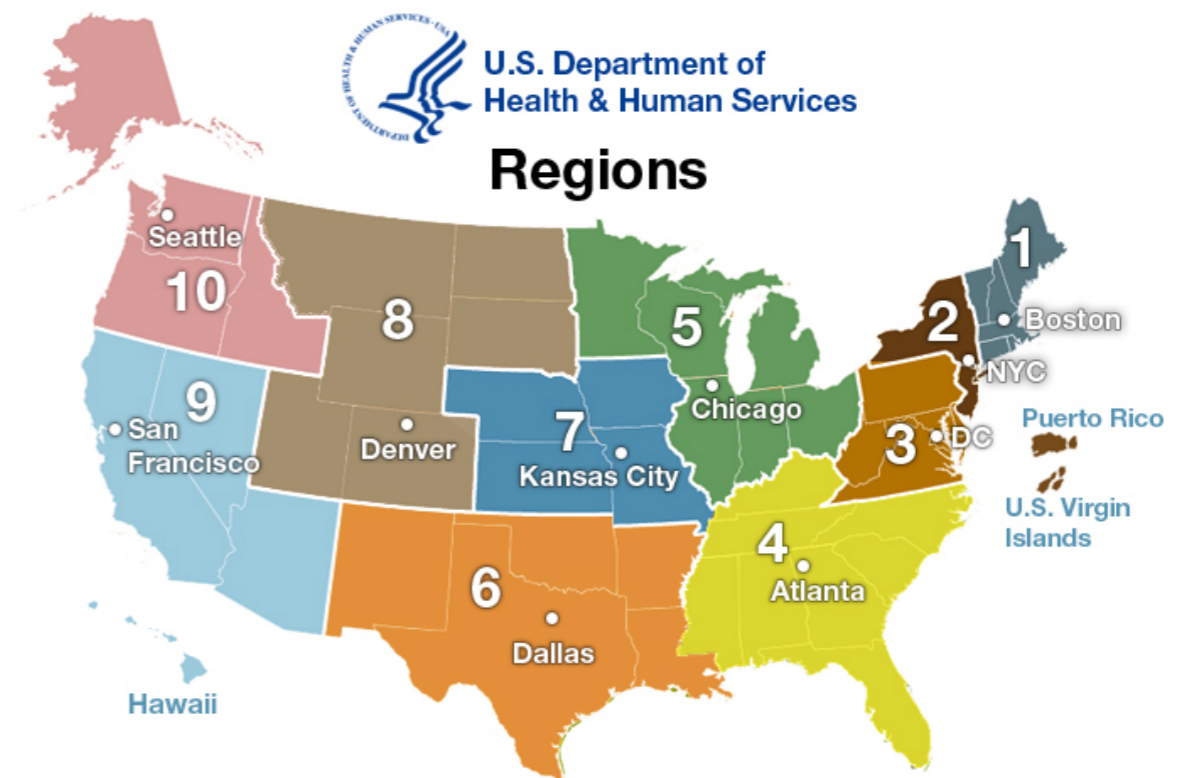
2014: US military troops heading to Liberia to assist with Ebola outbreak.
image: defense.gov

Typical epidemic forecasting setup

e.g. CDC FluSight challenges: U.S. national, regional, state level
Running annually since 2013.

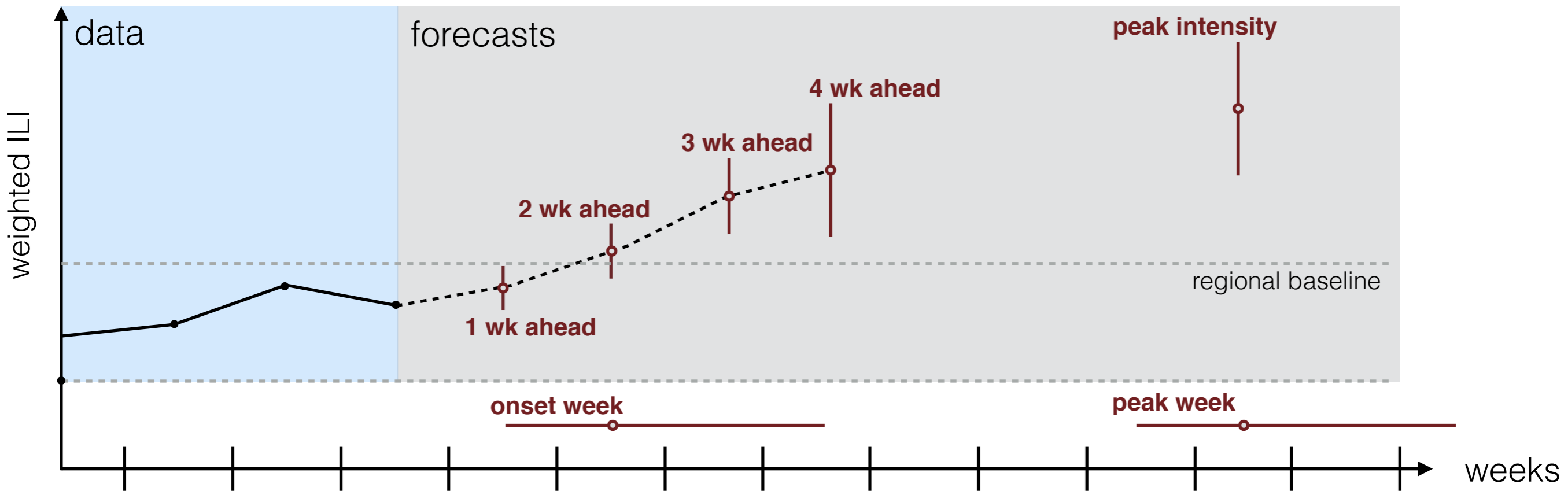
Target variable "weighted ILI":

The % of all outpatient visits with primary complaint of influenza-like illness (ILI), weighted by state population.



Targets with public health relevance

from annual CDC FluSight forecasting challenge



Biggerstaff et al. 2016, *BMC Inf Dis*. <https://doi.org/10.1186/s12879-016-1669-x>

McGowan et al. 2019, *Sci Rep*. <https://doi.org/10.1038/s41598-018-36361-9>

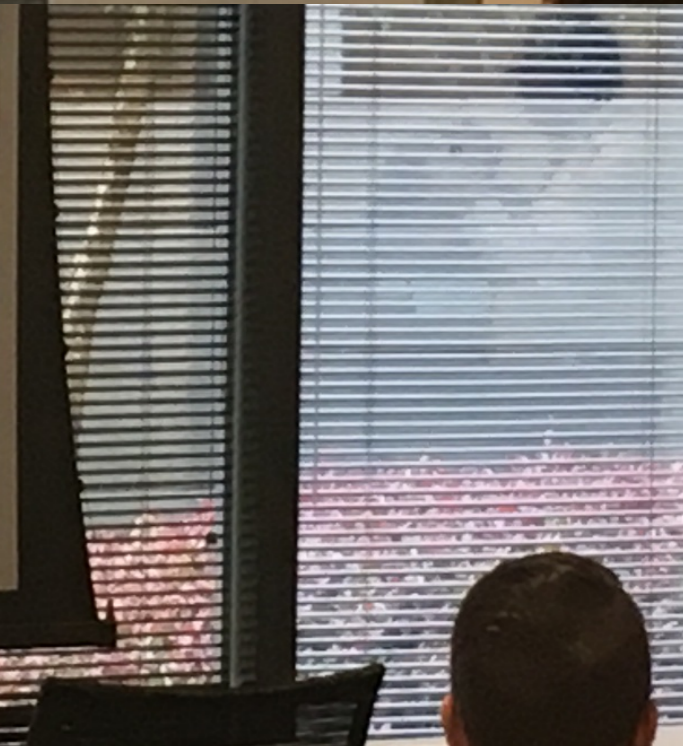
Lutz et al. 2019. *BMC Pub Hlth*. <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6902553/>

Dan Jernigan, Director of Influenza Division, CDC
September 2018

Forecasting Applications

- Informing healthcare providers
 - Outpatient clinic staffing
 - Emergency Department staffing and triage
 - Hospital general ward and ICU bed planning
- Informing pharmacies
 - Antiviral and symptom-reducing drug supplies
- Informing parents
 - Push messages on warning signs of severe influenza
 - Improved situational awareness for enhancing flu prevention actions
- Informing Schools
 - Prepare for increased absenteeism and potential for reactive school closures
- Informing Businesses
 - Alert for higher potential for absenteeism or caring for ill children
- Pandemic response
- Improving situational awareness through media

Influenza Division CDC



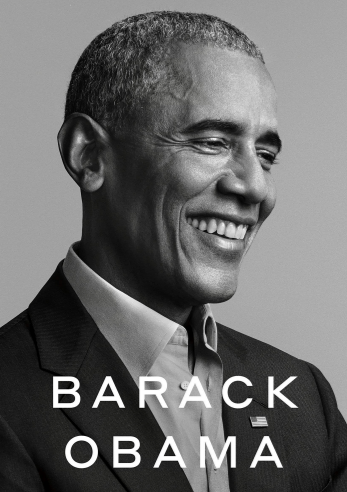
Why probabilistic forecasting?

“My emphasis on process was born of necessity. What I was quickly discovering about the presidency was that **no problem that landed on my desk, foreign or domestic, had a clean, 100 percent solution.** If it had, someone else down the chain of command would have solved it already. Instead, **I was constantly dealing with probabilities:** a 70 percent chance, say, that a decision to do nothing would end in disaster; a 55 percent chance that this approach versus that one *might* solve the problem (with a 0 percent chance that it would work out exactly as intended); a 30 percent chance that whatever we chose wouldn't work at all, along with a 15 percent chance that it would make the problem worse.

In such circumstances, chasing after the perfect solution led to paralysis. On the other hand, going with your gut too often meant letting preconceived notions or the path of least political resistance guide a decision--with cherry picked facts used to justify it. But with a sound process--one in which I was able to empty out my ego and really listen, **following the facts and logic as best I could** and considering them alongside my goals and my principles--I realized **I could make tough decisions and still sleep easy at night, knowing at a minimum that no one in my position, given the same information, could have made the decision any better.”**

–Barack Obama, *A Promised Land*, p 294

A PROMISED LAND



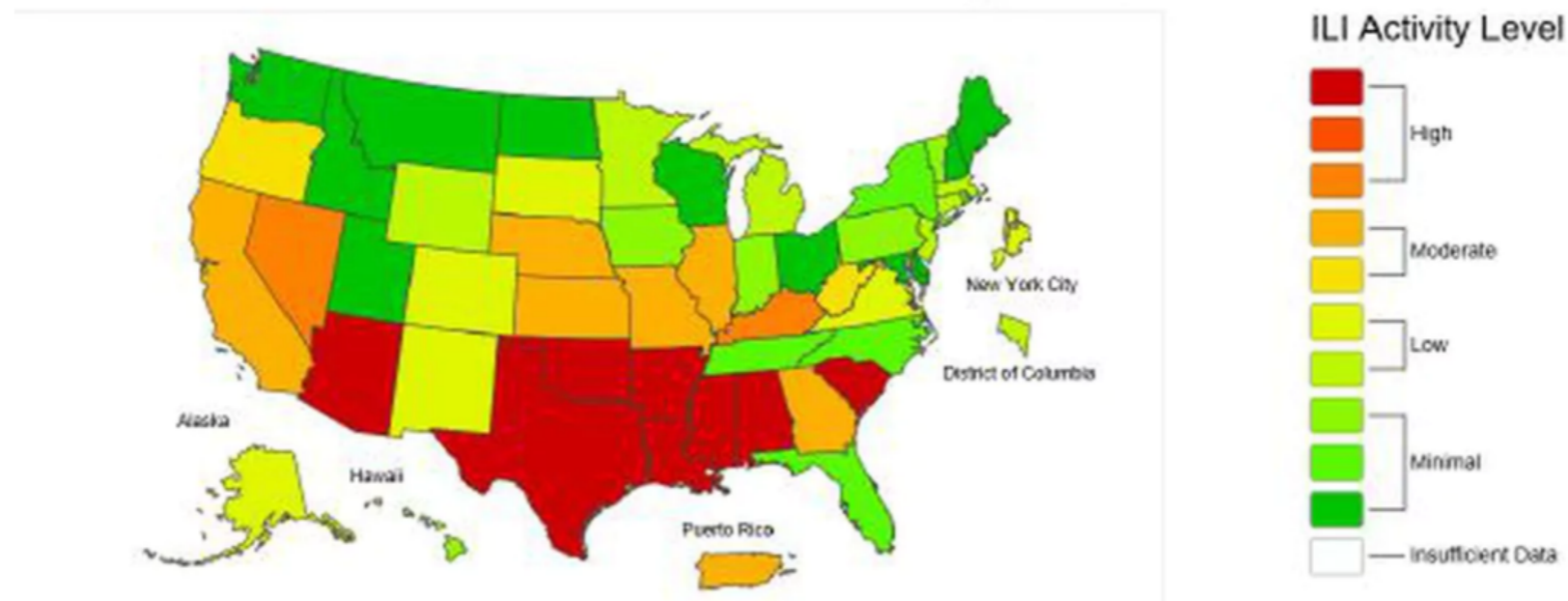
BARACK
OBAMA

Why this may be a bad flu season, especially around the holidays

By **Lena H. Sun** December 22, 2017 

CDC's flu forecasters say there's a 30 percent chance the season will peak around the end of December and a 60 percent chance that the greatest incidence will be by late January, Jernigan said. Generally, flu season peaks near the end of February.

Influenza-Like Illness (ILI) Activity Level Indicator Determined by Data Reported to ILINet
2017-18 Influenza Season Week 50 ending Dec 16, 2017



The "Hub" approach in epidemic forecasting

Model coordination is key

- There have been numerous government-coordinated outbreak forecasting efforts (flu, Ebola, chikungunya, Zika, dengue, etc...).
- One consistent finding across all efforts:

Combining models into an "ensemble" provides more consistent forecasts than any single model.

Flu: Reich et al. 2019, *PLOS Comp Bio*. <https://doi.org/10.1371/journal.pcbi.1007486>

Flu: McGowan et al. 2019, *Sci Rep*. <https://doi.org/10.1038/s41598-018-36361-9>

Dengue: Johansson et al. 2019, *PNAS*.

Ebola: Viboud et al. 2018, *Epidemics*.

COVID-19: Cramer et al. 2020, *medrxiv*.

The "Hub" idea is not new

- The idea: coordinated modeling between groups to inform policy and/or develop knowledge about a system.
- Different than a competition: involving coordination between groups, often in real-time.

Climate

ipcc

ipcc.ch

Ecology

FISHERIES & MARINE ECOSYSTEM
FISH-MIP
MODEL INTERCOMPARISON PROJECT

isimip.org/about/marine-ecosystems-fisheries/

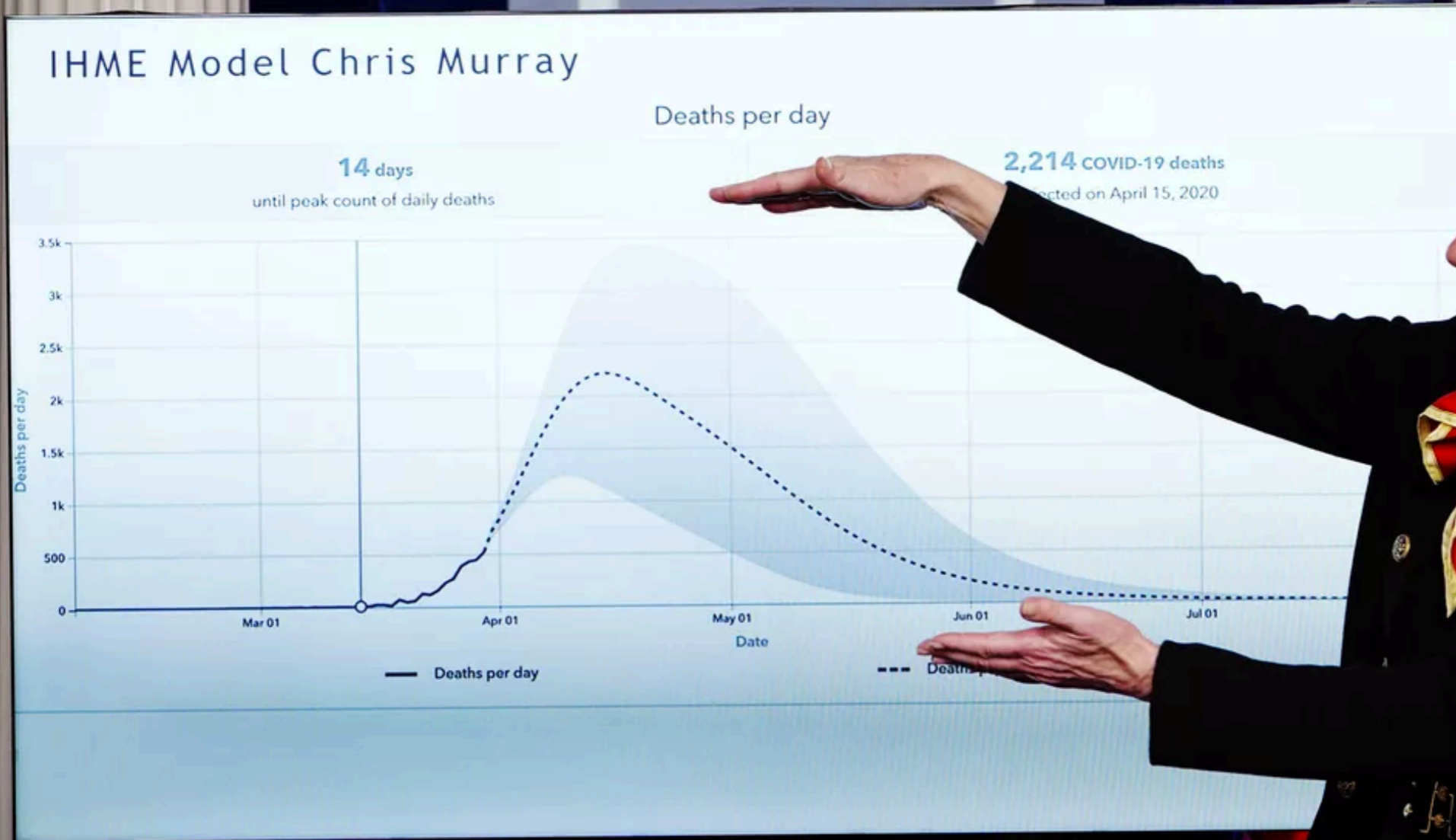
Space Science



COMMUNITY
COORDINATED
MODELING
CENTER

ccmc.gsfc.nasa.gov

Policy makers needed >1 model



early April 2020



covid19forecasthub.org

Launched April 6, 2020

1. Provide decision-makers and general public with reliable information about where the pandemic is headed in the next month.
2. Assess reliability of forecasts and gain insight into which modeling approaches do well.
3. Create a community of infectious disease modelers underpinned by an open-science ethos.



COVID-19

ForecastHub

By the numbers

- Each week the Forecast Hub receives forecasts of weekly incident **cases, hospitalizations and deaths** in the US due to COVID-19 from dozens of groups.
- The Hub builds an **ensemble that combines quantile-based predictive distributions** from these models for 1 through 4 week ahead forecasts.
- To date, we have curated data from **101 models: over 4,000 submissions and 57 million predictions.**

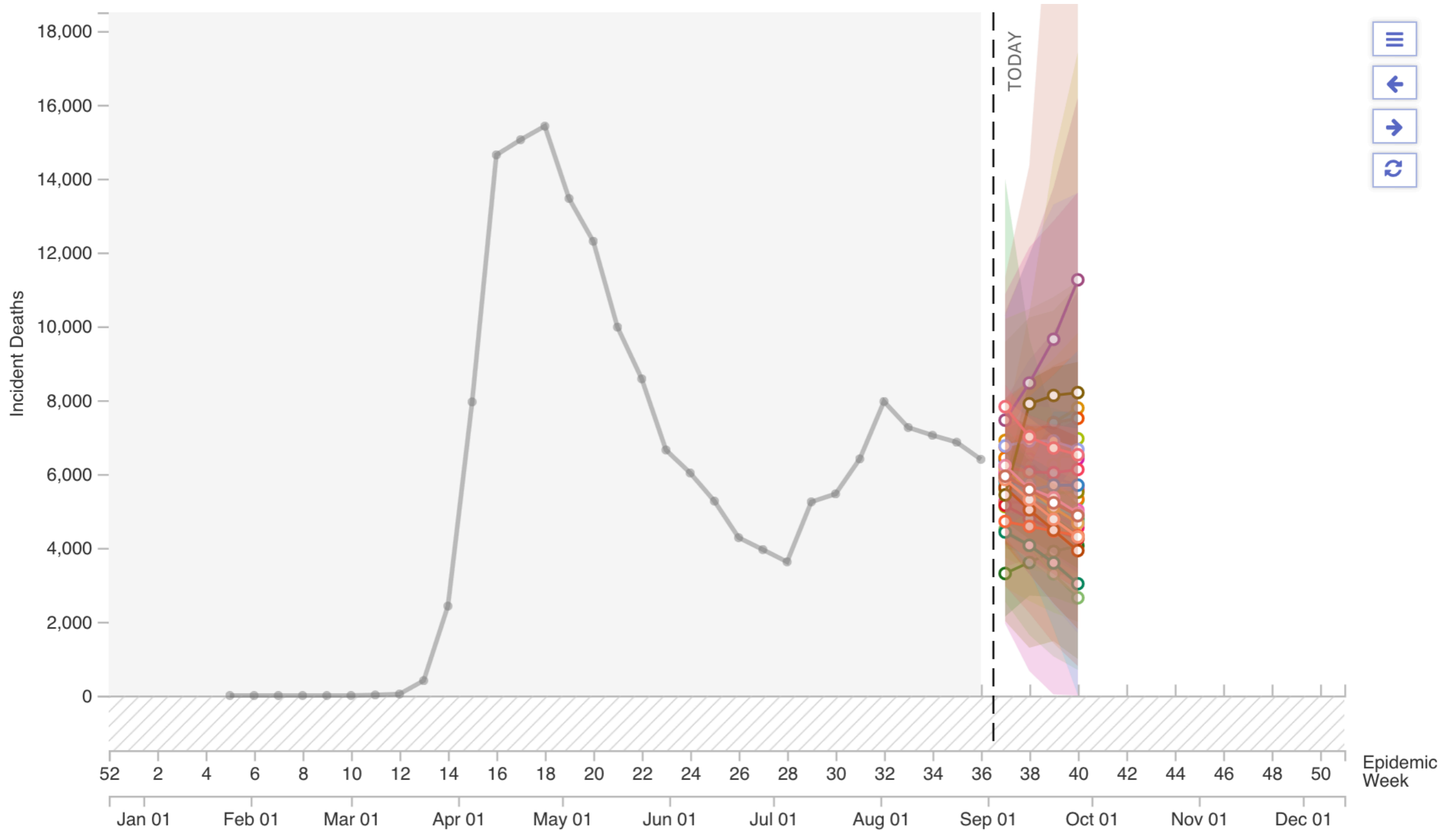
Data on GitHub and on Zoltar.

<https://github.com/reichlab/covid19-forecast-hub/>

<https://zoltardata.com/project/44>

Demo Visualization

<https://viz.covid19forecasthub.org/>





COVID-19 ForecastHub

Forecast data from the COVID-19 Forecast Hub is shared directly with the CDC, and published on the CDC website weekly.

COVID-19 Forecasts: Deaths

Updated Nov. 19, 2020 [Print](#)



Observed and forecasted new and total reported COVID-19 deaths as of November 16, 2020.

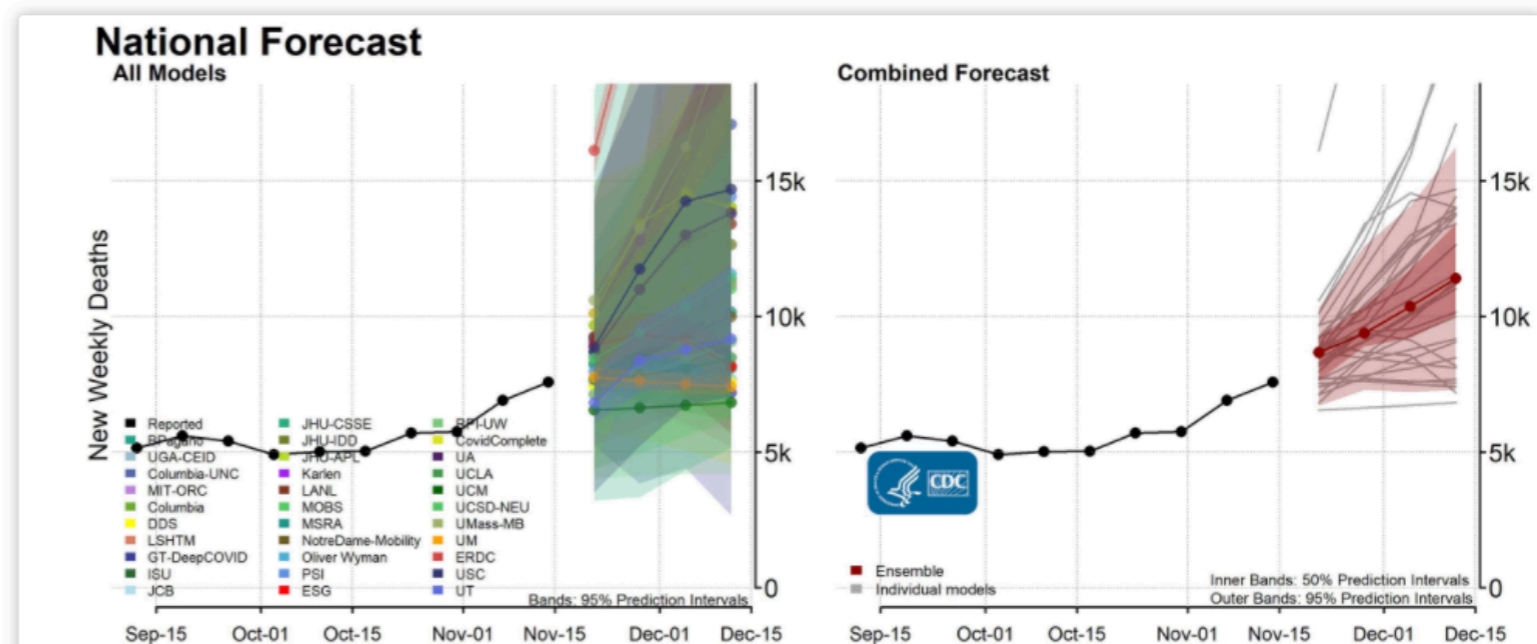
Interpretation of Forecasts of New and Total Deaths

- This week CDC received forecasts of COVID-19 deaths over the next 4 weeks from 36 modeling groups that were included in the ensemble forecast. Of the 36 groups, 33 provided forecasts for both new and total deaths, two groups forecasted total deaths only, and one forecasted new death only.
- This week's national [ensemble forecast](#) predicts that the number of newly reported COVID-19 deaths will likely increase over the next four weeks, with 7,300 to 16,000 new deaths likely to be reported in the week ending December 12, 2020. The national ensemble predicts that a total of 276,000 to 298,000 COVID-19 deaths will be reported by this date.
- The state- and territory-level ensemble forecasts predict that over the next 4 weeks, the number of newly reported deaths per week will likely increase in 36 jurisdictions, which are indicated in the forecast plots below. Trends in numbers of future reported deaths are uncertain or predicted to remain stable in the other states and territories.

On This Page

- [National Forecast](#)
- [State Forecasts](#)
- [Ensemble Forecast](#)
- [Forecast Assumptions](#)

National Forecast



Individual COVID-19 models vary

roughly ordered by date of first submission

- IHME-CurveFit: "**hybrid modeling approach** to generate our forecasts, which incorporates elements of statistical and disease transmission models."
- MOBS-GLEAM_COVID: "The GLEAM framework is based on **a metapopulation approach** in which the world is divided into geographical subpopulations. Human **mobility between subpopulations is represented on a network.**"
- UMass-MechBayes: "**classical compartmental models from epidemiology**, prior distributions on parameters, models for time-varying dynamics, models for partial/noisy observations of confirmed cases and deaths."
- UT-Mobility: "For each US state, **we use local data from mobile-phone GPS traces** made available by [SafeGraph] to quantify the changing impact of social-distancing measures on 'flattening the curve.' "
- GT-DeepCOVID: "This **data-driven deep learning model** learns the dependence of hospitalization and mortality rate on various detailed syndromic, demographic, mobility and clinical data."
- Google Cloud AI: "a novel approach that integrates **machine learning** into **compartmental disease modeling** to predict the progression of COVID-19"
- Facebook AI: "**recurrent neural networks** with a vector autoregressive model and train the joint model with a specific regularization scheme that increases the **coupling between regions**"
- CMU-TimeSeries: "A **basic AR-type time series model** fit using lagged values of case counts and deaths as features. No assumptions are made regarding reopening or governmental interventions."

Forecast Skill: Weighted Interval Score

- Consider a single $(1 - \alpha) \times 100\%$ predictive interval $[l, u]$ for the observed response y . The interval score is:

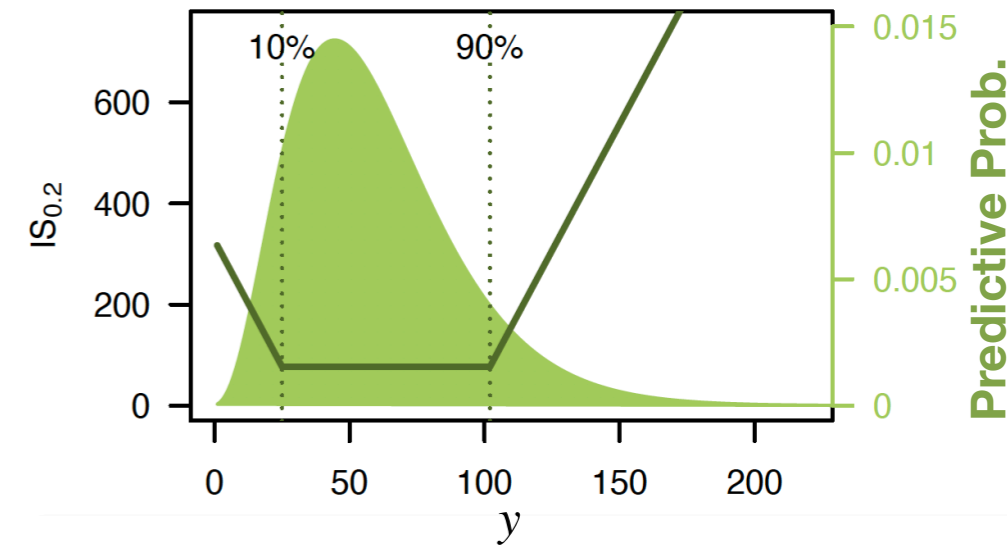
$$\mathbf{IS}_\alpha(F, y) = \underbrace{(u - l)}_{\text{Width of interval}} + \frac{2}{\alpha} \cdot \underbrace{(l - y) \cdot \mathbf{1}(y < l)}_{\text{Penalty if interval is too high}} + \frac{2}{\alpha} \cdot \underbrace{(y - u) \cdot \mathbf{1}(y > u)}_{\text{Penalty if interval is too low}},$$

Width of interval

Penalty if interval is too high

Penalty if interval is too low

- Smaller \mathbf{IS}_α is better



Forecast Skill: Weighted Interval Score

- Consider a single $(1 - \alpha) \times 100\%$ predictive interval $[l, u]$ for the observed response y . The interval score is:

$$\mathbf{IS}_\alpha(F, y) = \underbrace{(u - l)}_{\text{Width of interval}} + \frac{2}{\alpha} \cdot \underbrace{(l - y) \cdot \mathbf{1}(y < l)}_{\text{Penalty if interval is too high}} + \frac{2}{\alpha} \cdot \underbrace{(y - u) \cdot \mathbf{1}(y > u)}_{\text{Penalty if interval is too low}},$$

Width of interval

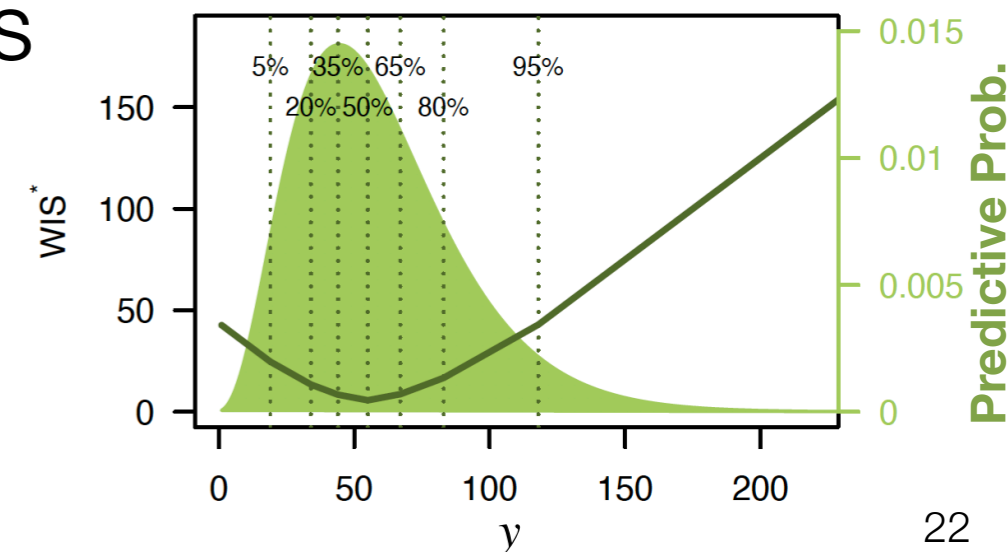
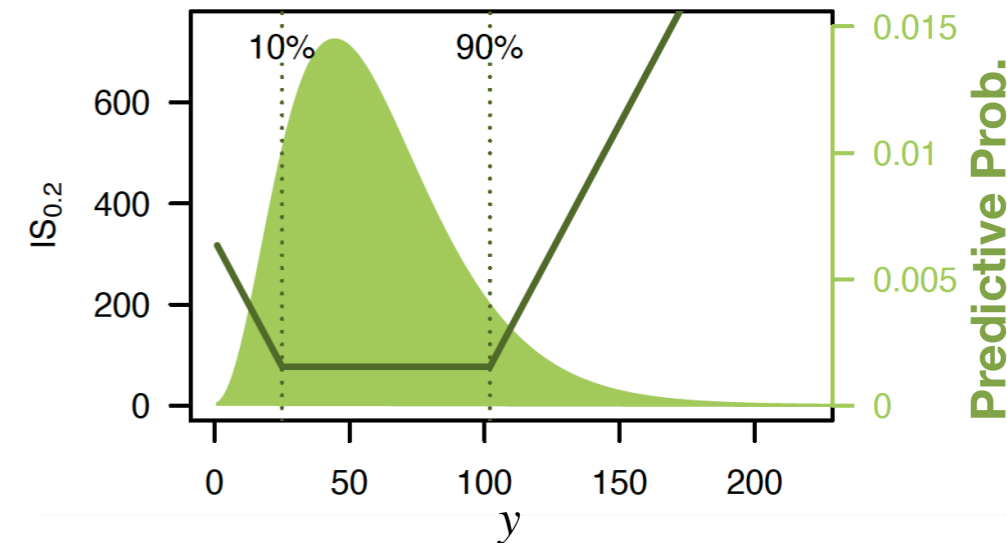
Penalty if interval is too high

Penalty if interval is too low

- Smaller \mathbf{IS}_α is better
- For multiple predictive intervals, we compute a weighted average of \mathbf{IS}_α

$$\mathbf{WIS}_{\alpha_{0:K}}(F, y) = \frac{1}{K + 1} \times \left(w_0 \times 2 \times |y - m| + \sum_{k=1}^K (w_k \times \mathbf{IS}_{\alpha_k}(F, y)) \right).$$

- We use weights $w_i = \frac{\alpha_i}{2}$, in which case $\mathbf{WIS} \approx \text{CRPS}$ (continuous ranked probability score)
- The resulting score is **proper**: in expectation, it is minimized by the true predictive distribution.
- Equivalent to pinball loss.



Defining relative WIS

- For each pair of models m and m' , we compute the pairwise relative WIS

$$\theta_{m,m'} = \frac{\text{mean WIS of model } m}{\text{mean WIS of model } m'}$$

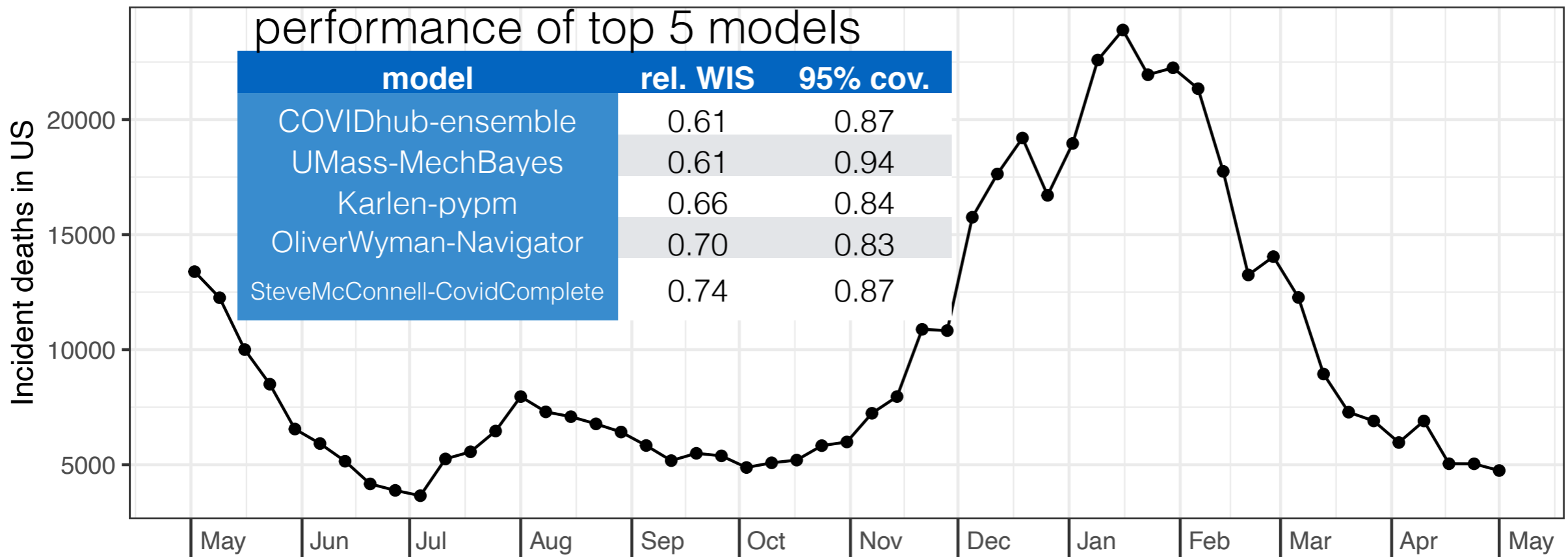
based on the *available overlap of forecast targets*.

- We take a geometric mean of all pairwise relative WIS values:

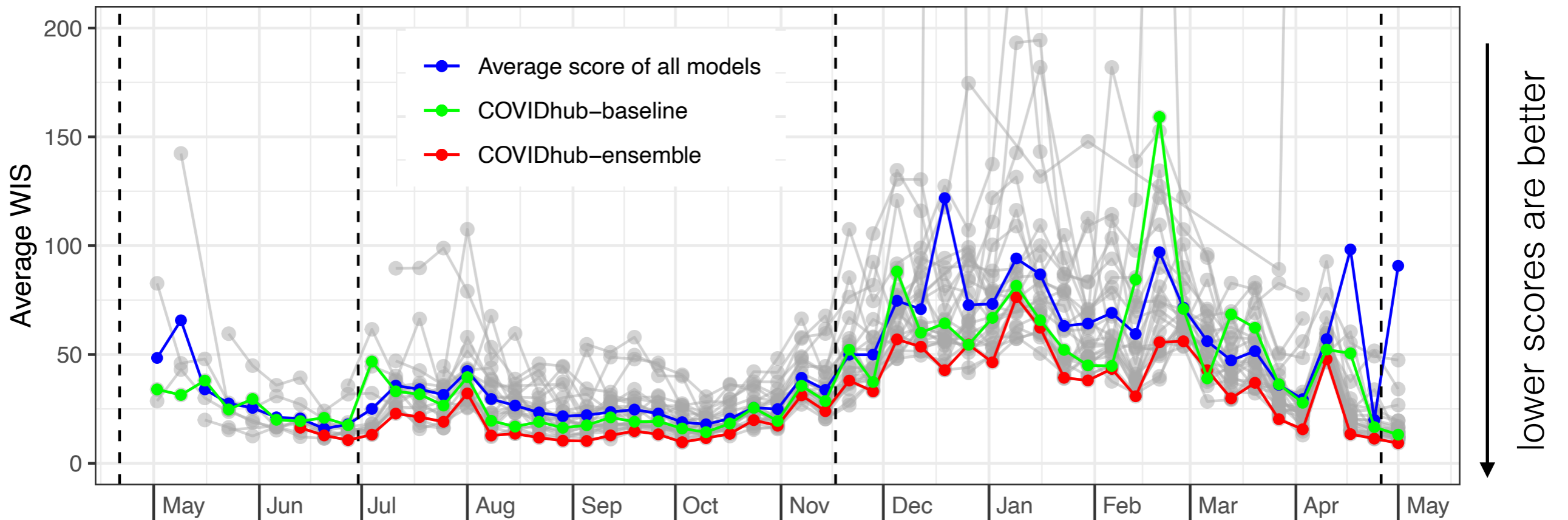
$$\theta_m = \left(\prod_{m'=1}^M \theta_{m,m'} \right)^{1/M}$$

- Then, θ_m is a measure of relative WIS that describes the relative performance of model m , adjusted for the difficulty of the forecasts model m made. It assumes that no model can gain an advantage by focusing on just some targets.
- We define $\theta_m^* = \theta_m / \theta_B$ where θ_B is the relative WIS for the baseline model.

A: Observed weekly COVID-19 deaths in the US

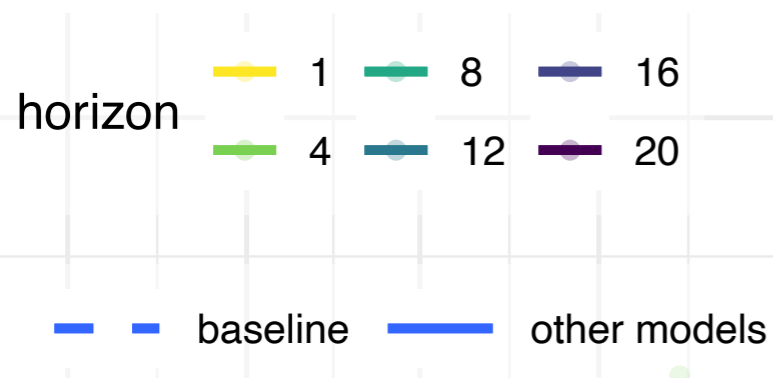
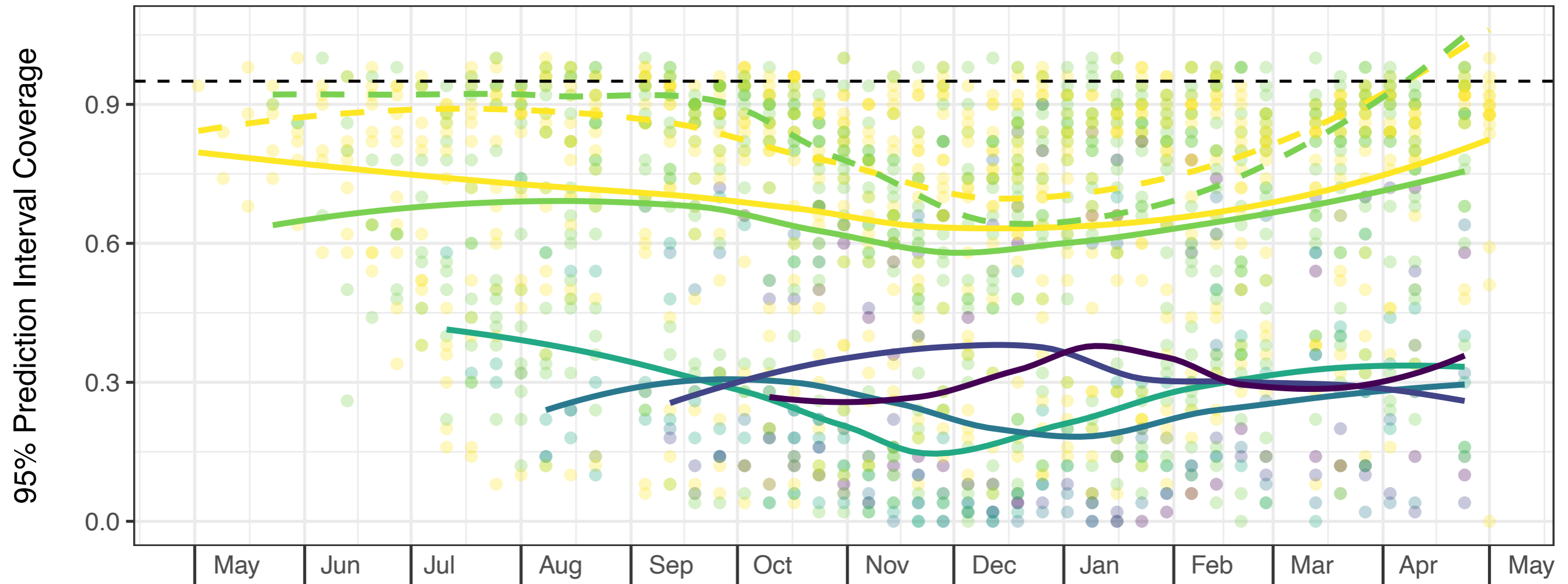


B: Average 1-week ahead weighted interval scores by model



Errors increase with horizon

C: 95% prediction interval coverage across time, stratified by forecast horizon



a point is the empirical coverage rate for a single model across all locations in a given week, colored by horizon.

lines are smooth trends through all points for a given horizon.



COVID-19 ForecastHub

<https://covid19forecasthub.org/>

Team: Martha Zorn, Nutcha Wattanachit, Serena Wang, Ariane Stark, Apurv Shah, Nicholas Reich, Evan Ray, Jarad Niemi, Khoa Le, Abdul Kanji, Dasuni Jayawardena, Yuxin Huang, Katie House, Aaron Gerding, Estee Cramer, Matt Cornell, Alvaro J. Castro Rivadeneira, Andrea Brennen, Johannes Bracher

* underline denotes ensemble contributor

US CDC Collaborators: Matthew Biggerstaff, Michael Johansson, Velma Lopez, Rachel Slayton, Jo Walker

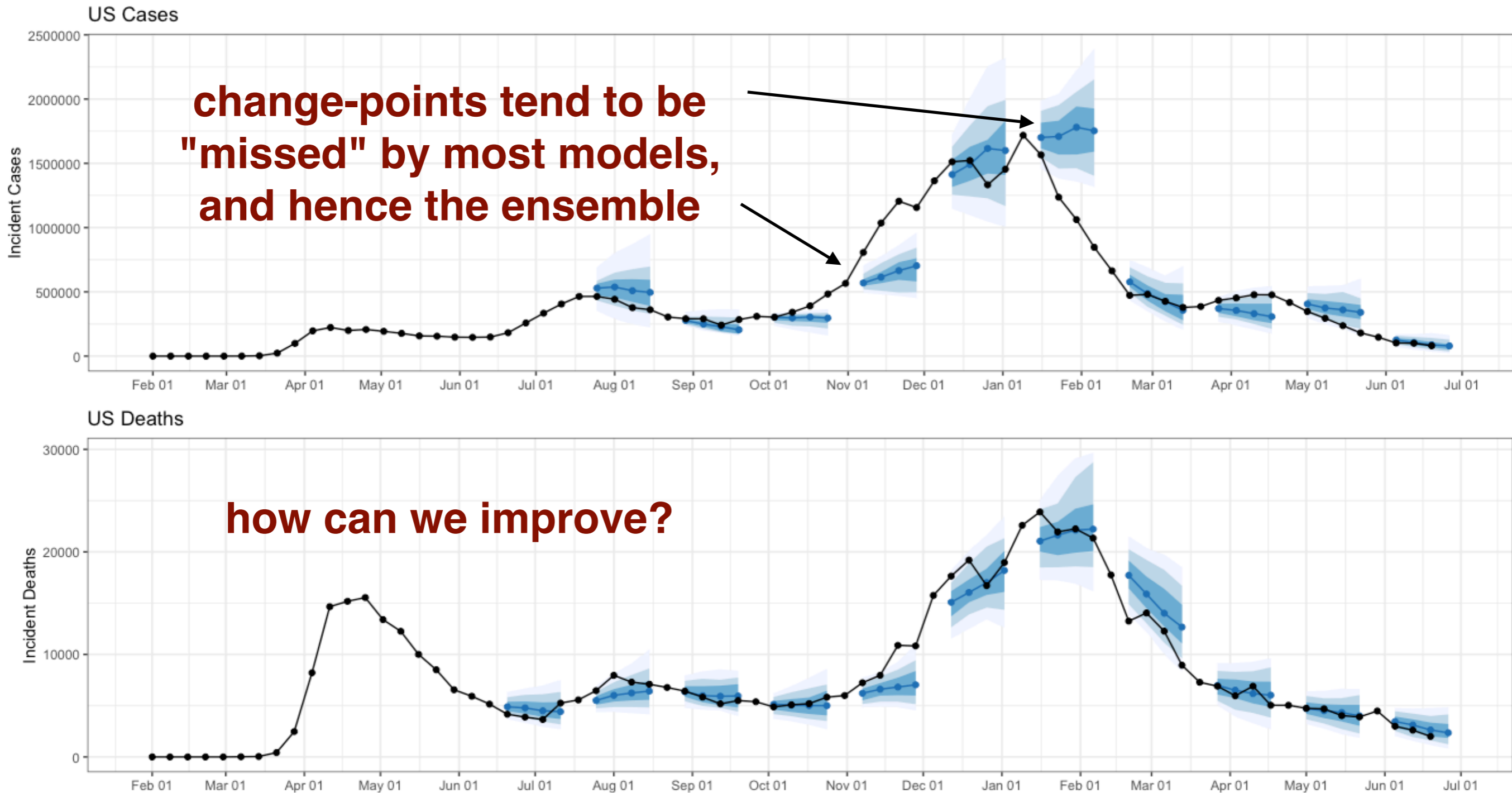
Ensemble “advisors”: Jacob Bien, Logan Brooks, Sebastian Funk, Tilmann Gneiting, Anja Muhlemann, Aaron Rumack, Ryan Tibshirani

Modeling groups: Over 80 groups at various institutions have contributed forecasts to the hub

Adventures in ensemble building for COVID-19

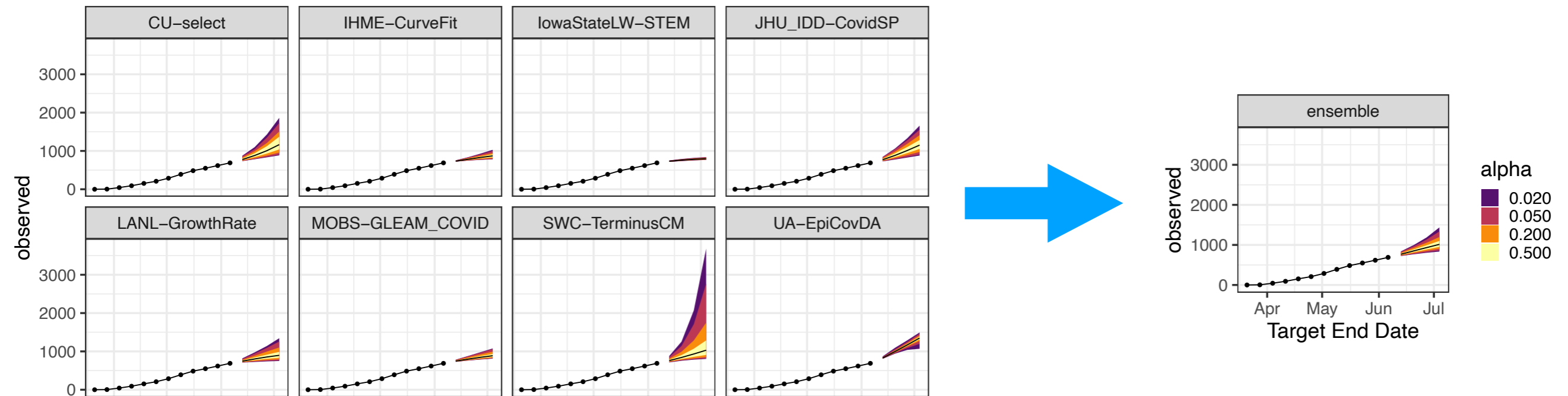
Forecasts have missed change-points

(especially for case forecasts)



Building the Ensemble: View 1

Alabama



- For each combination of spatial unit s , time point t , and forecast horizon h , teams are required to submit $K=23$ (or 7) quantiles of a predictive distribution:
 $\hat{P}(Y \leq q_{s,t,h,1}^m) = 0.01, \hat{P}(Y \leq q_{s,t,h,2}^m) = 0.025, \dots, \hat{P}(Y \leq q_{s,t,h,12}^m) = 0.5, \dots, \hat{P}(Y \leq q_{s,t,h,23}^m) = 0.99$

The predictive median

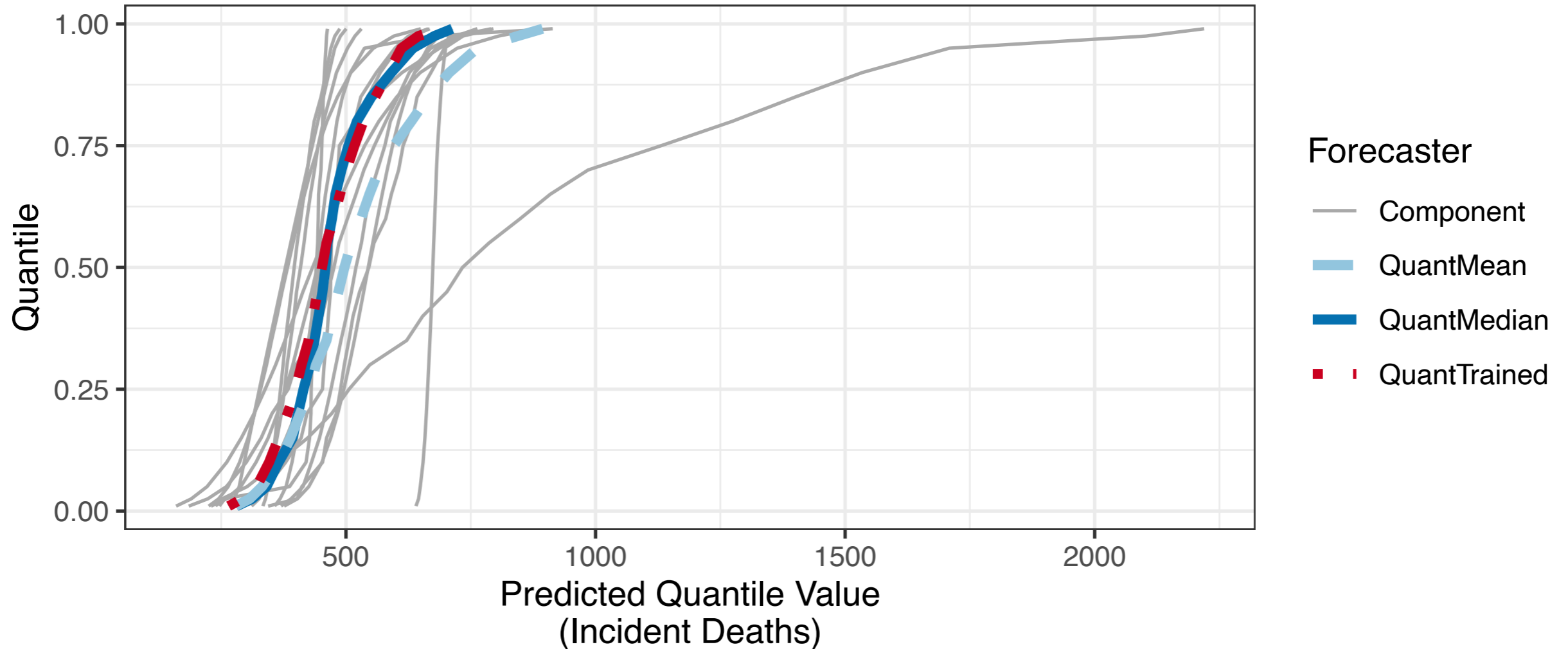
Limits of a 98% prediction interval

- The predictive quantiles for the ensemble are a combination of component predictions at each quantile level:

$$q_{s,t,h,k} = f(q_{s,t,h,k}^1, \dots, q_{s,t,h,k}^M) \text{ for each } k = 1, \dots, 23$$

Building an Ensemble: View 2

- The pairs $(q_{s,t,h,k}^m, \widehat{P}(Y_{s,t,h} \leq q_{s,t,h,k}^m))$ fall along the predictive CDF for model m



- Three options for the combination function f :

- QuantMean: $q_{s,t,h,k} = \frac{1}{M} \sum_{m=1}^M q_{s,t,h,k}^m$

Used through July 21, 2020

- QuantMedian: $q_{s,t,h,k} = \mathbf{median}(q_{s,t,h,k}^1, \dots, q_{s,t,h,k}^M)$

Used starting July 28, 2020

- QuantTrained: $q_{s,t,h,k} = \mathbf{weighted\ median}(q_{s,t,h,k}^1, \dots, q_{s,t,h,k}^M)$

Experimental version
submitted since

What is the optimal ensemble?

What is the optimal ensemble?

April 2020: We started simple.

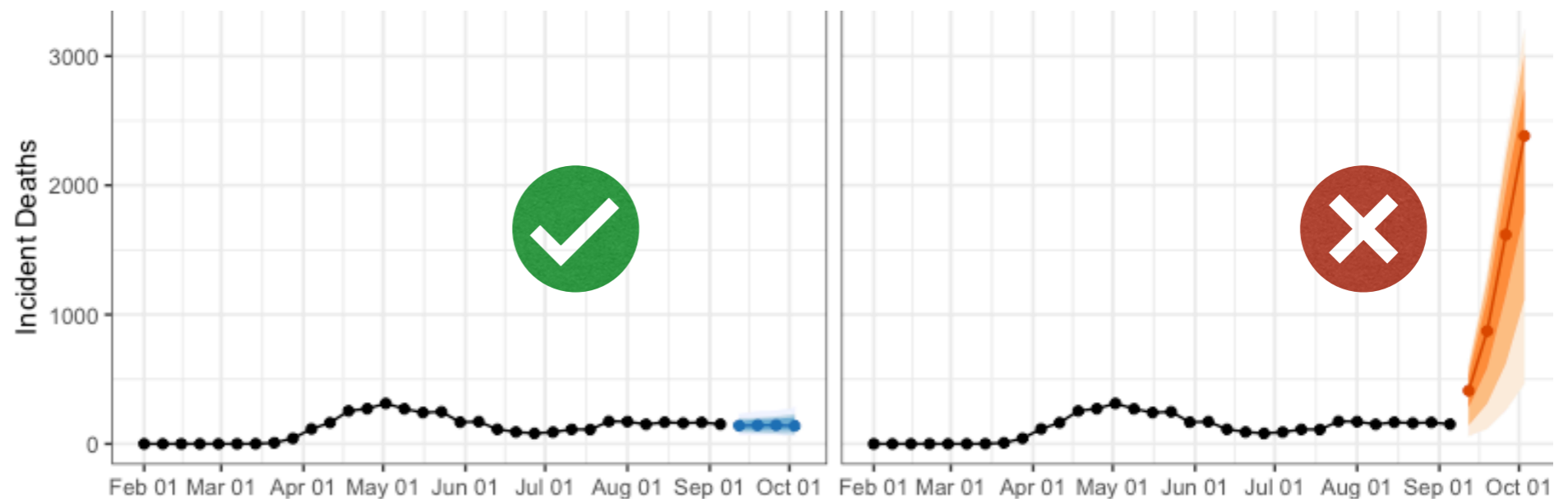
Equal-weighted mean

What is the optimal ensemble?

We cannot ask our public health collaborators to defend ensemble forecasts that "explode" because 1-2 models go off the rails.

"Robust"

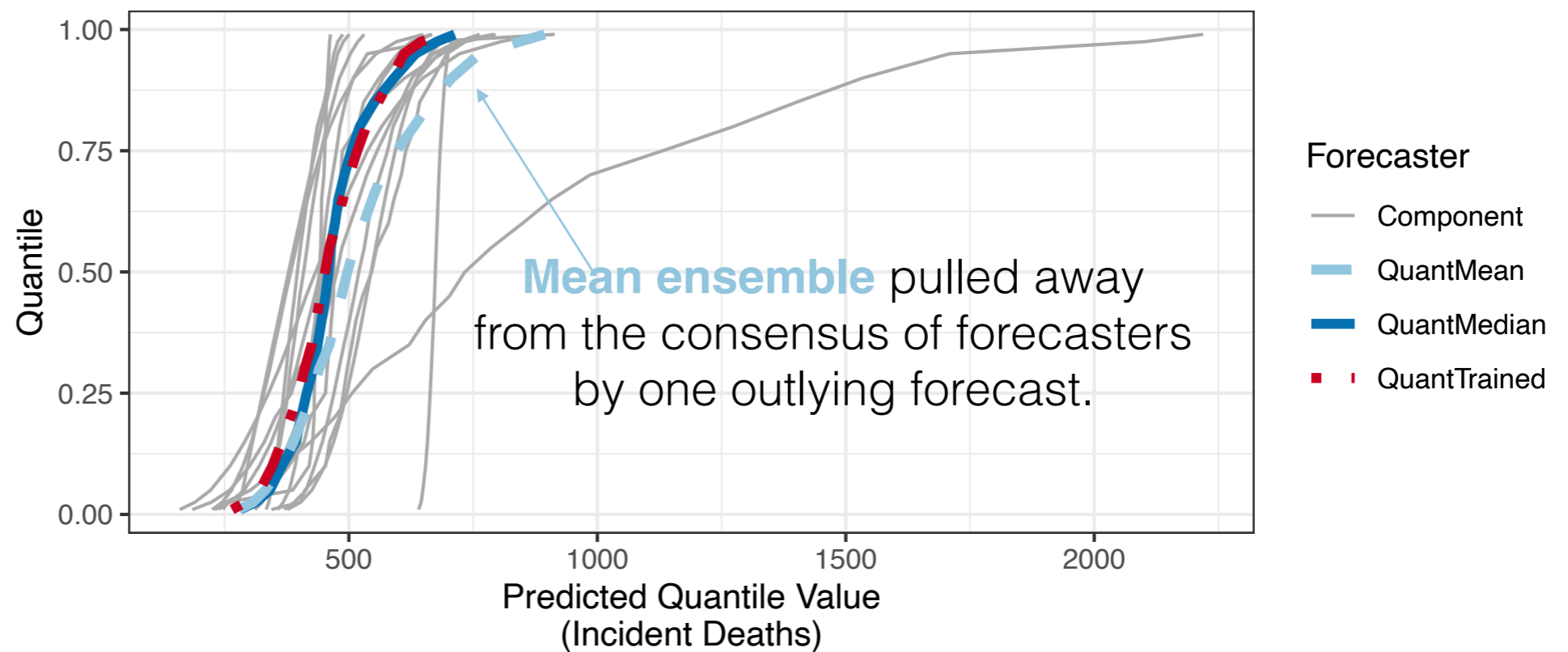
(i.e. ensemble does not "blow up")





What is the optimal ensemble?

We cannot ask our public health collaborators to defend ensemble forecasts that "explode" because 1-2 models go off the rails.

"Robust"
(i.e. ensemble does not "blow up")



What is the optimal ensemble?

"Robust" (i.e. ensemble does not "blow up")	No	 Equal-weighted mean
	Yes	 Median

In operation from April 6 - July 21, 2020.

Used starting July 28, 2020.

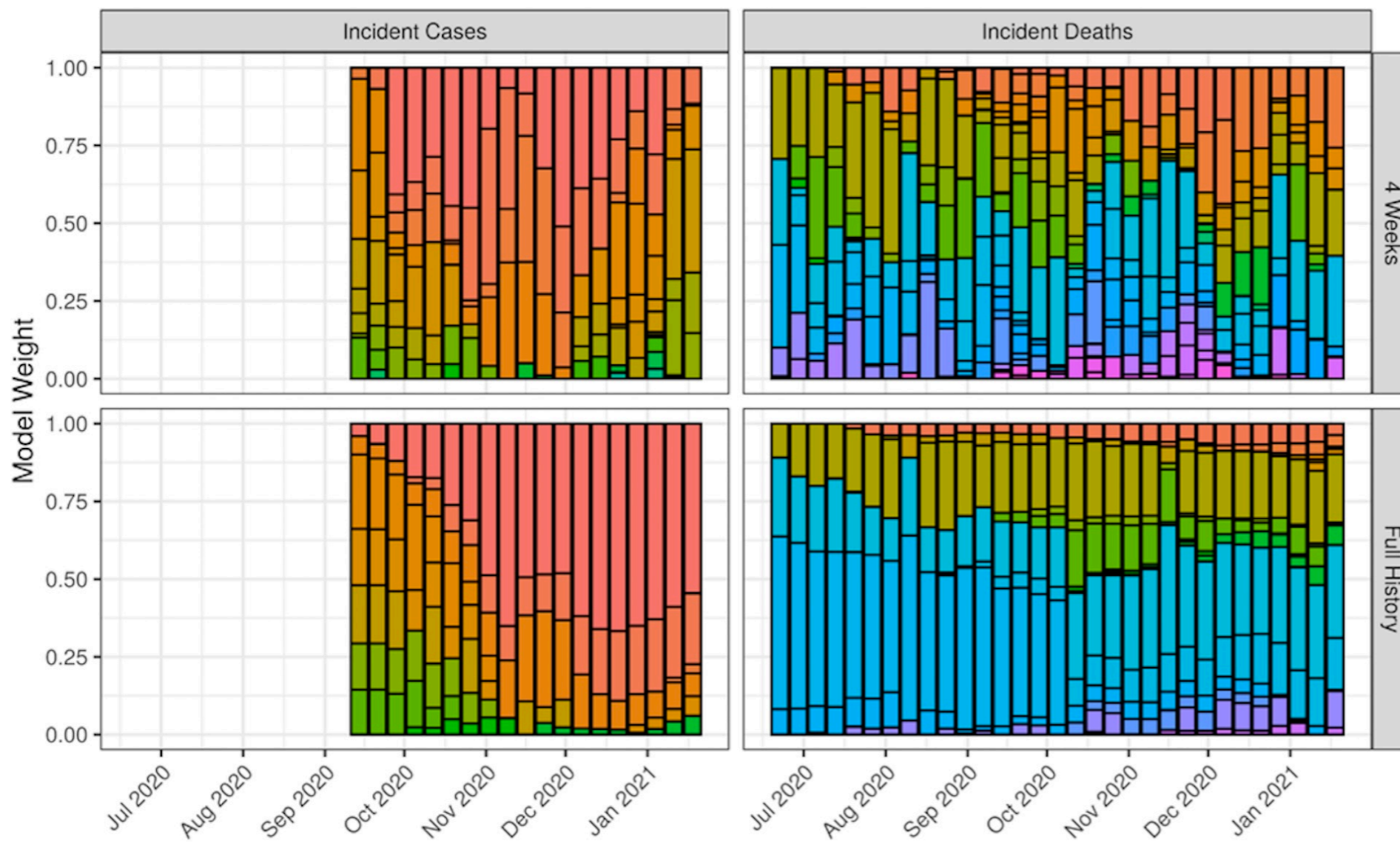
This change to a median fixed an occasional, high cost error. It also made it so we didn't have to manually curate models. But shouldn't weighting be able to fix this costly error?

What is the optimal ensemble?

... shouldn't weighting be able to fix this costly error?

"Trained"
(i.e. component forecasts are weighted)

Model weights estimated for each week's ensemble, optimized for low WIS scores



Lots of subtleties!

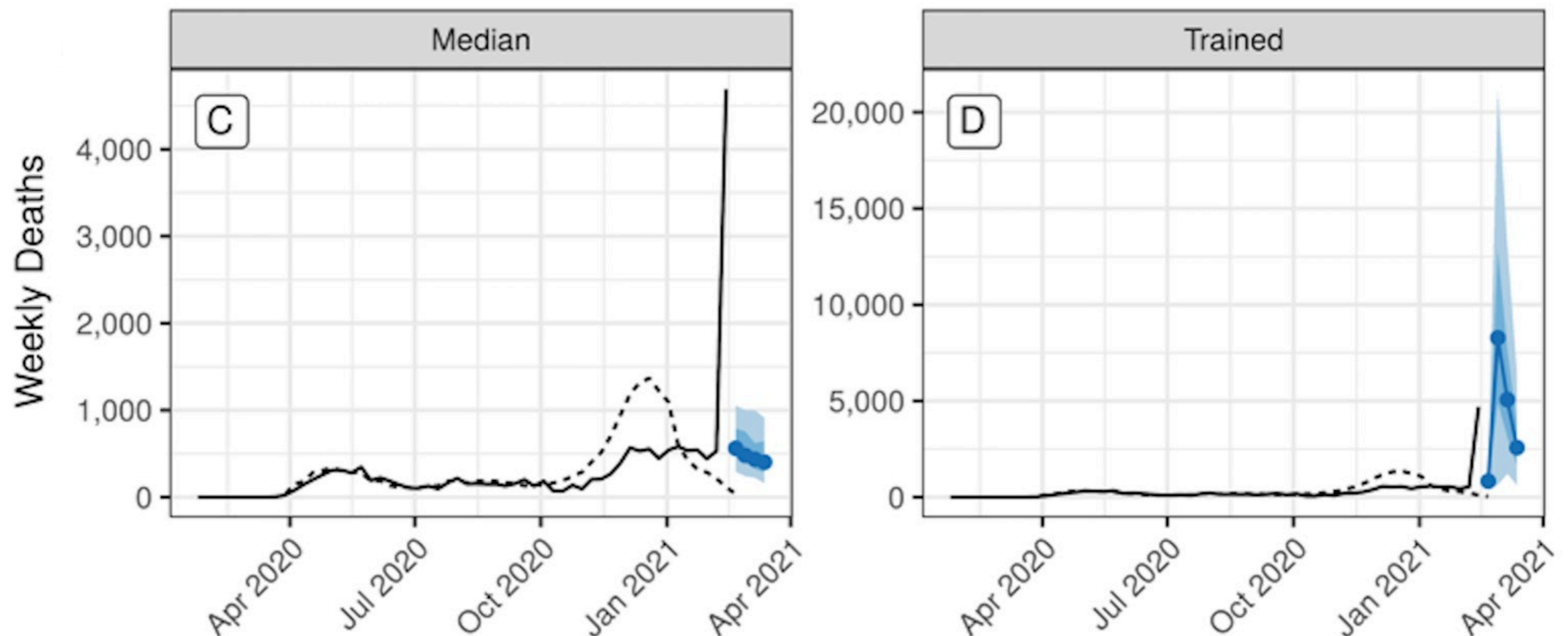
- Need estimation for different targets.
- Include all models, or just "top" models?
- Same weights for all quantiles?
- How much model history to include?
- How to deal with models that don't submit for one week?

What is the optimal ensemble?





... shouldn't weighting be able to fix this costly error?

"Trained"
(i.e. component forecasts are weighted)




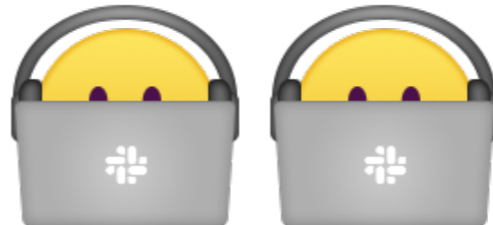
Ultimately, weighted means were still occasionally not robust!



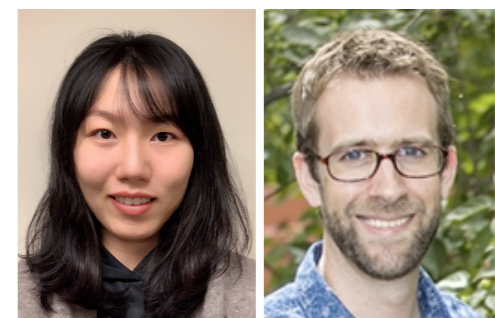
What is the optimal ensemble?

		"Trained" (i.e. component forecasts are weighted)	
		No	Yes
"Robust" (i.e. ensemble does not "blow up")	No	 Equal-weighted mean	 Variations on a weighted mean
	Yes	 Median	

What is the optimal ensemble?

		"Trained" (i.e. component forecasts are weighted)	
		No	Yes
"Robust" (i.e. ensemble does not "blow up")	No	 Equal-weighted mean	 Variations on a weighted mean
	Yes	 Median	

What about a weighted median?



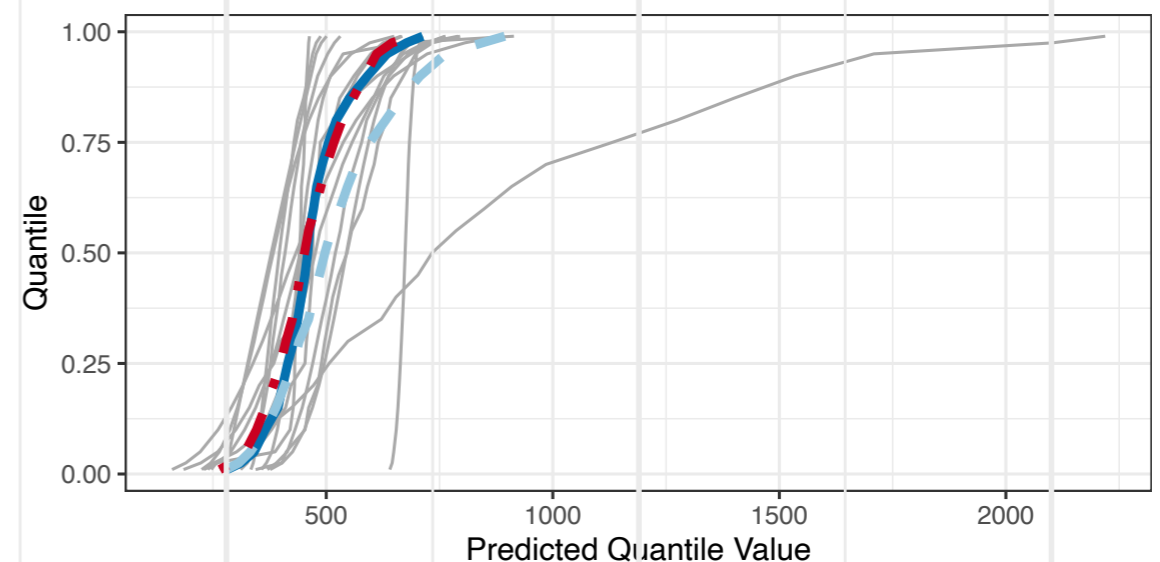
Serena & Evan

What is the optimal ensemble?

Brief detour: computing a weighted median from quantile forecasts.

Recall, the situation here is that we want to be computing a weighted median for each of the desired k quantiles.

So, we take the predicted k^{th} quantiles from models: q_k^1, \dots, q_k^M .
(We assume that the quantiles are in increasing order.)



950

1000

1050

1100

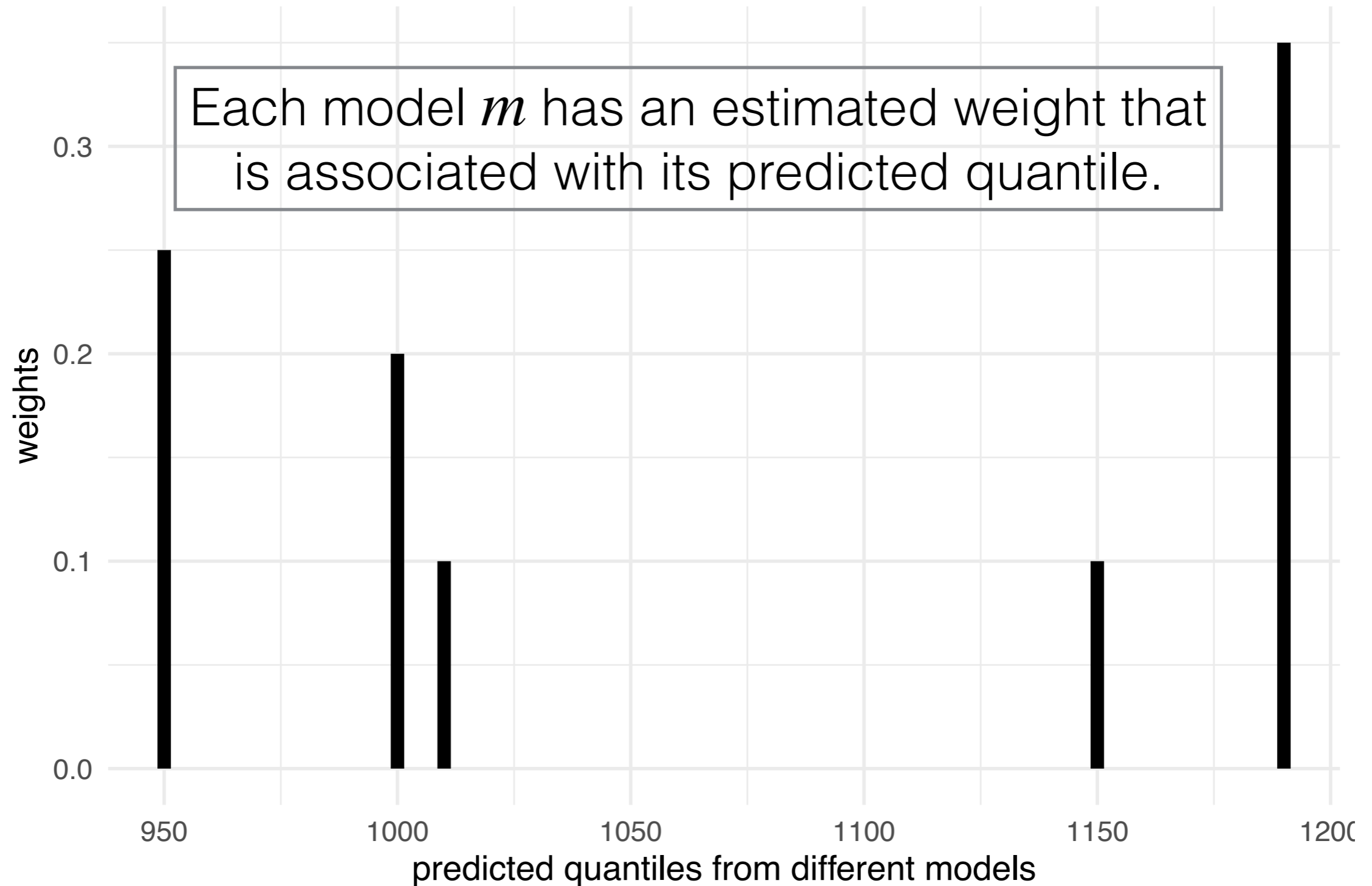
1150

1200

predicted quantiles from different models

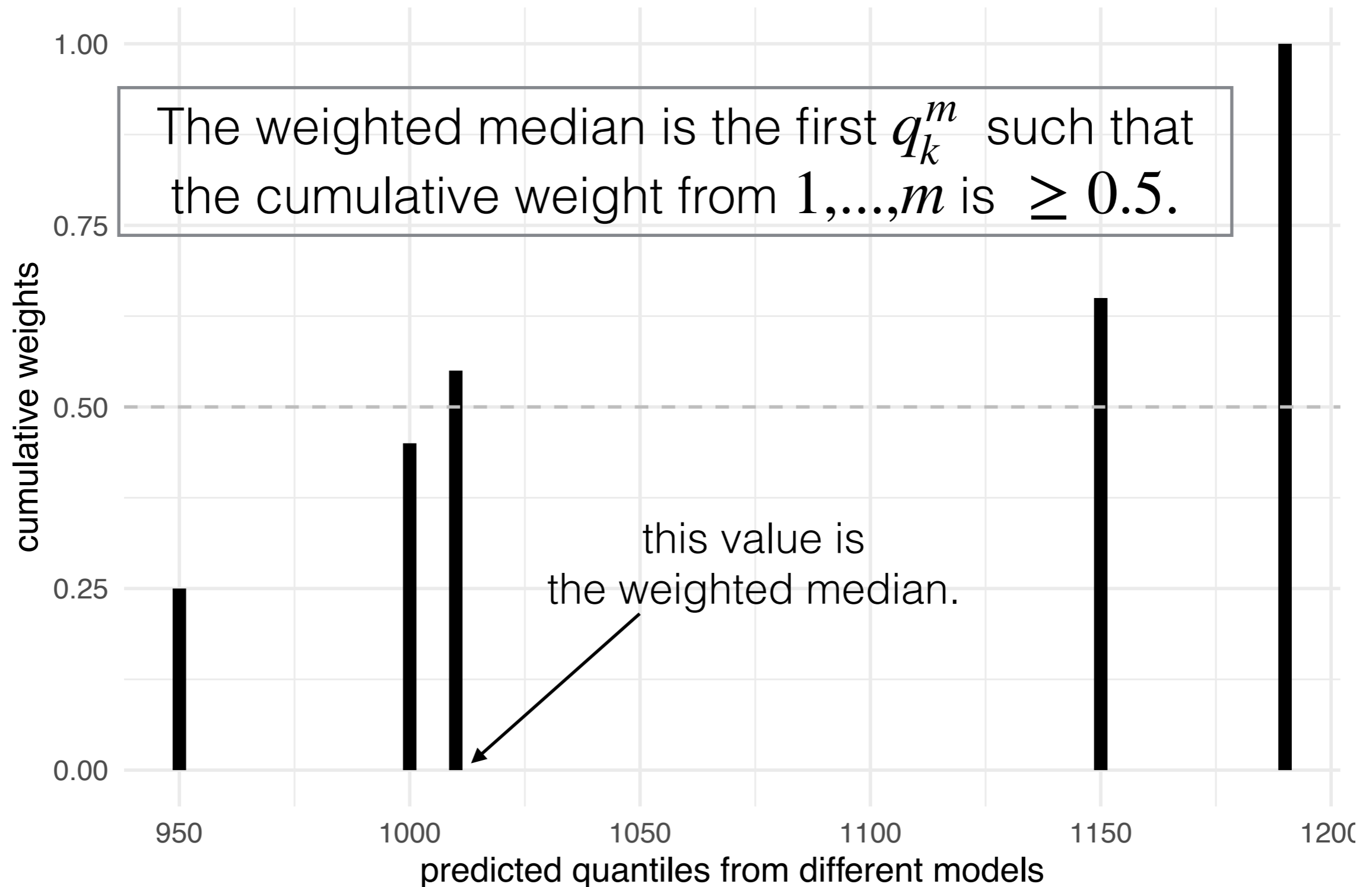
What is the optimal ensemble?

Brief detour: computing a weighted median from quantile forecasts.







What is the optimal ensemble?

Brief detour: computing a weighted median from quantile forecasts.



What is the optimal ensemble?

		"Trained" (i.e. component forecasts are weighted)	
		No	Yes
"Robust" (i.e. ensemble does not "blow up")	No	 Equal-weighted mean	 Variations on a weighted mean
	Yes	 Median	 Variations on a weighted median

- Median of best 5 or 10 individual models
- Weighted median, weights from a weighted mean ensemble
- Weighted median, weights based on relative WIS:

$$w_m = \frac{\exp(-\theta \cdot \mathbf{Relative\ WIS\ Model\ m})}{\sum_{j=1}^M \exp(-\theta \cdot \mathbf{Relative\ WIS\ Model\ j})}$$

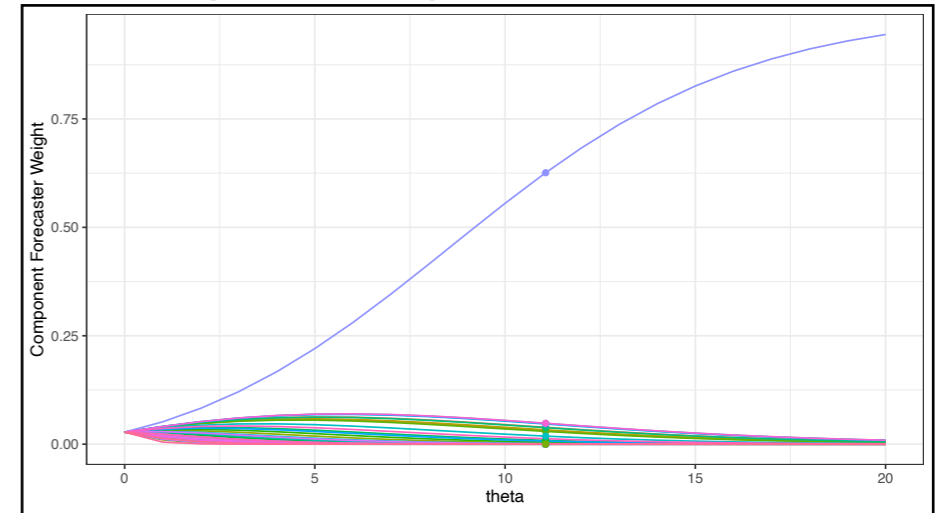
Relative WIS Weighted Median

- Idea: Introduce a single non-negative parameter θ determining model weights as a function of training set relative WIS

$$w_m = \frac{\exp(-\theta \cdot \mathbf{Relative\ WIS\ Model\ m})}{\sum_{j=1}^M \exp(-\theta \cdot \mathbf{Relative\ WIS\ Model\ j})}$$

- If model m is bad (high relative WIS), model m gets low weight; if model m is good (low relative WIS), model m gets high weight.

- θ controls how dispersed the weights are:
 - $\theta = 0 \implies$ all weights are equal
 - $\theta \rightarrow \infty \implies$ top model gets all the weight



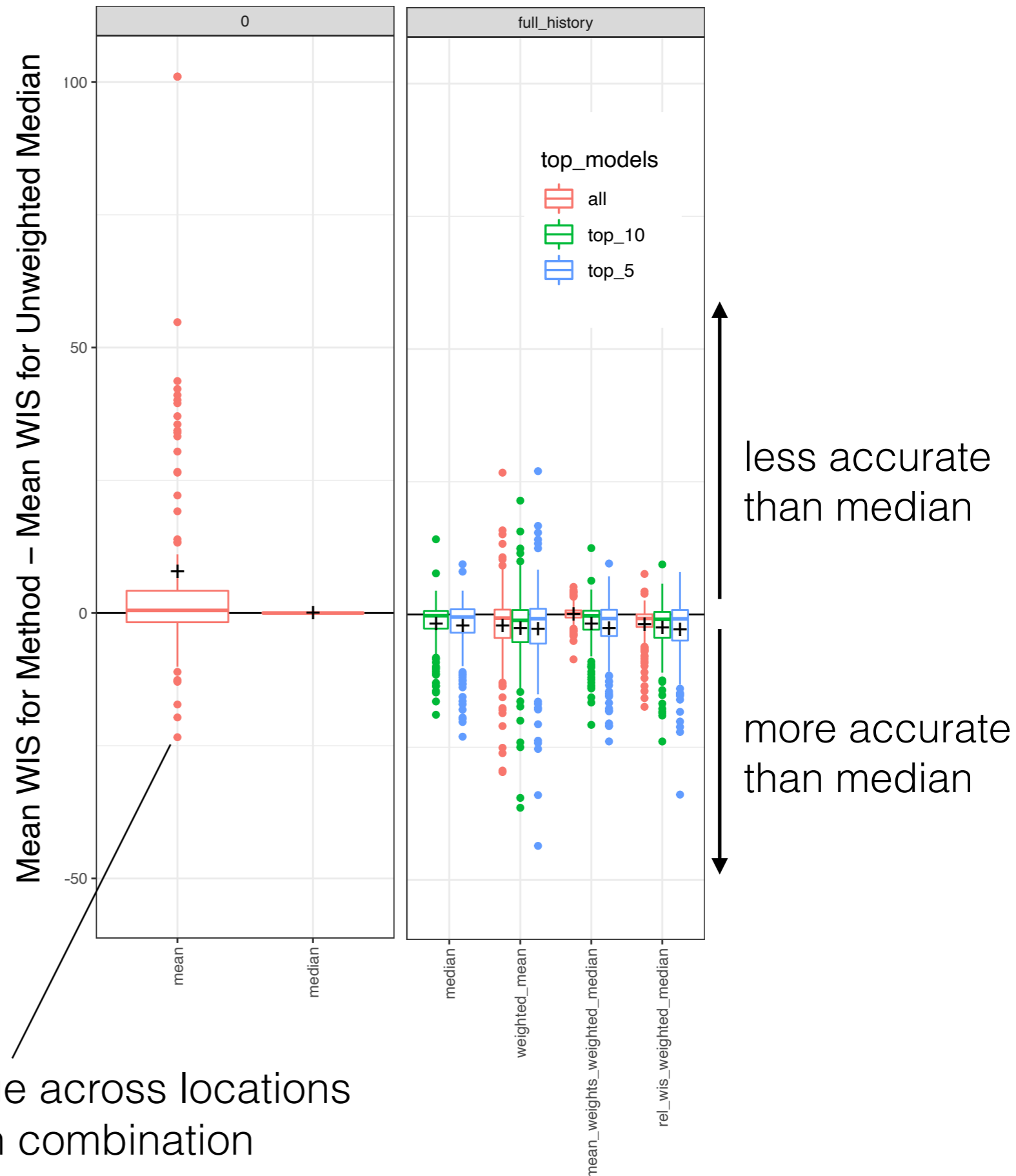
- Ensemble forecast is weighted median at each quantile level
- For estimation of θ , we use a grid-search approach.

Relative WIS Weighted Median

Using top 5 or 10 models, there appear to be consistent improvements with many methods over median of all models.





Weighted mean approach shows some outliers.

No formal tests performed.
(Open to suggestions of tests to do. Lots of correlation between observations!)



every point represents an average across locations for one forecast date and horizon combination





What is the optimal ensemble?

		"Trained" (i.e. component forecasts are weighted)	
		No	Yes
"Robust" (i.e. ensemble does not "blow up")	No	 Equal-weighted mean	 Variations on a weighted mean
	Yes	 Median	 Variations on a weighted median

- Median of best 5 or 10 individual models
- Weighted median, weights from a weighted mean ensemble
- Weighted median, weights based on relative WIS

Not a clear winner between these three, but we are using the last one of these at the Hub currently, estimating fresh weights and θ every week.

What is the optimal ensemble?

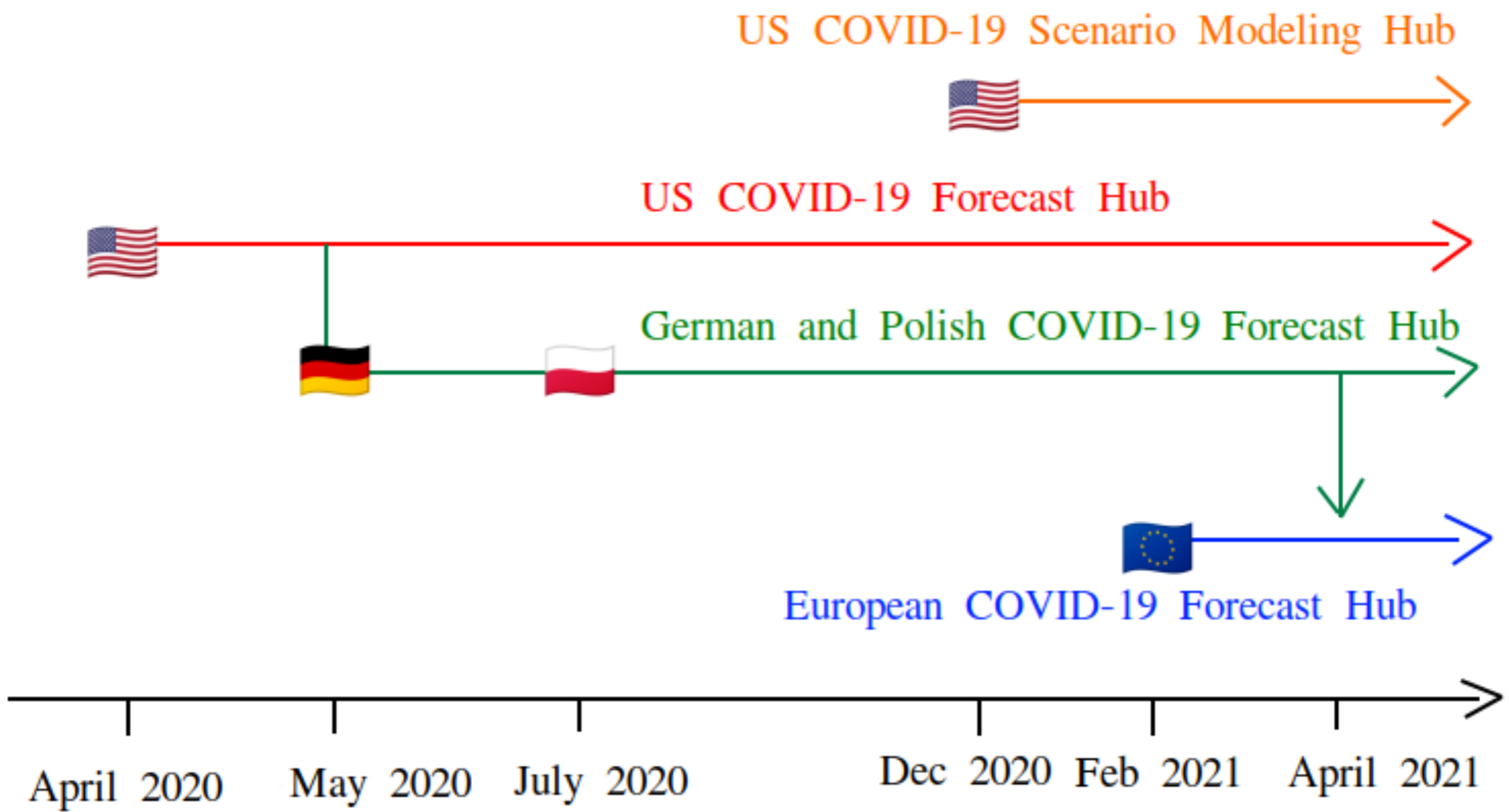
		"Trained" (i.e. component forecasts are weighted)	
		No	Yes
"Robust" (i.e. ensemble does not "blow up")	No	 Equal-weighted mean	 Variations on a weighted mean
	Yes	 Median	 Variations on a weighted median (weights based on relative WIS)

Key take-aways

1. Non-robust methods occasionally blow up; robust methods have better worst-case performance
2. Generally, trained methods have better mean performance:
 - a. Lower average MAE and WIS.
 - b. Closer to nominal interval coverage rates.

Optimizing infrastructure for Hubs

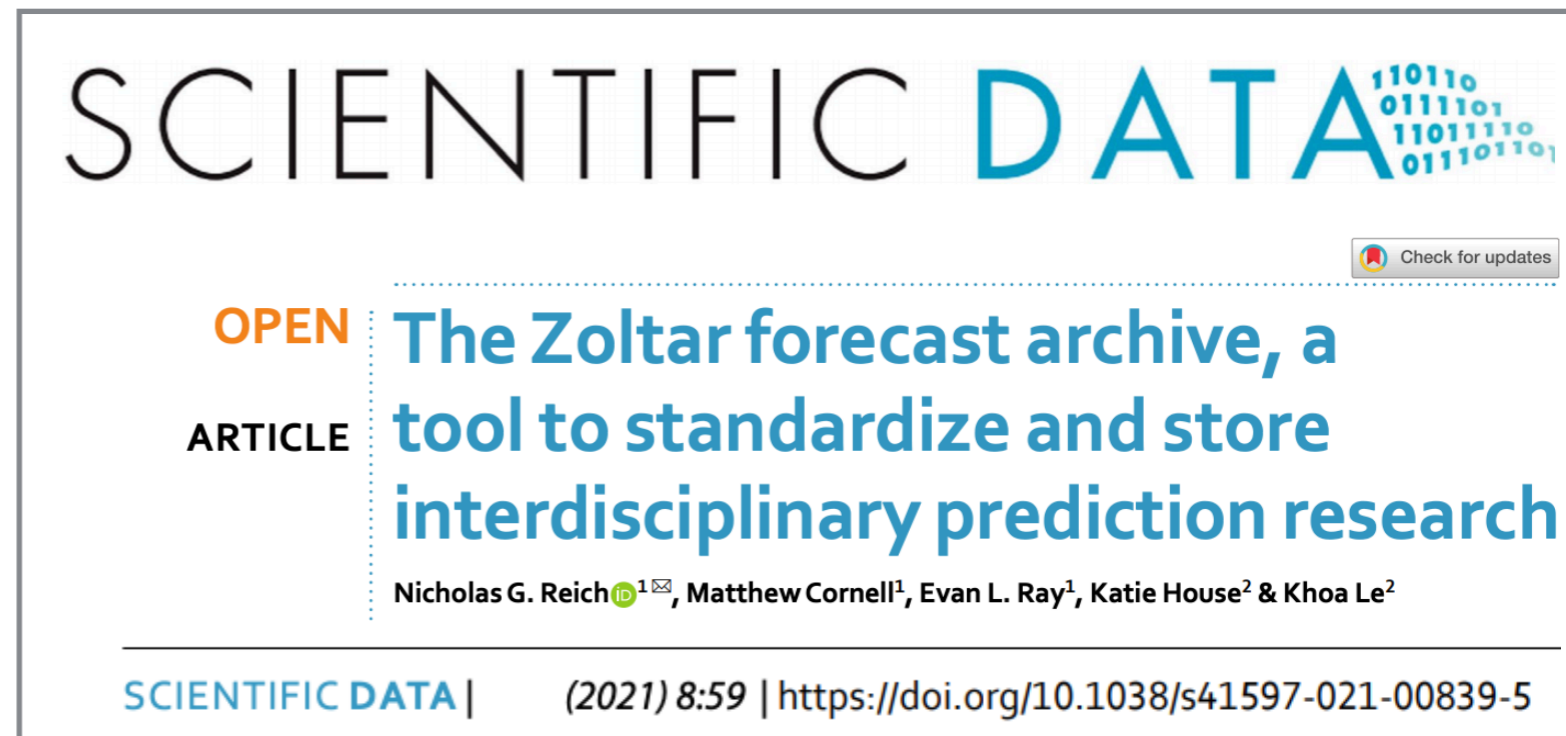
The Hub model has taken root





Scalable structure for Hubs

For coordinated probabilistic forecasting efforts to be sustainable and scalable, we need to develop a standard set of definitions, protocols, tools.

- What constitutes a probabilistic forecast?
- How can a forecast be represented?
- What tools do we need to evaluate, combine?




SCIENTIFIC DATA 

 Check for updates

OPEN **The Zoltar forecast archive, a tool to standardize and store interdisciplinary prediction research**

ARTICLE

Nicholas G. Reich ¹✉, Matthew Cornell¹, Evan L. Ray¹, Katie House² & Khoa Le²

SCIENTIFIC DATA | (2021) 8:59 | <https://doi.org/10.1038/s41597-021-00839-5>

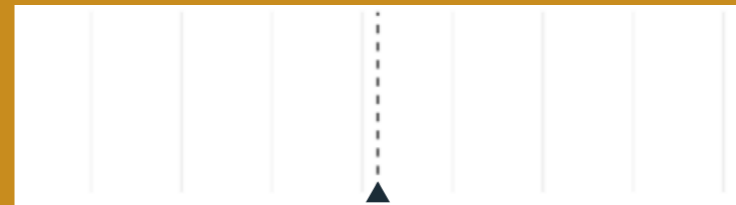
We define a "prediction" as a quantitative statement about unobserved data

Prediction

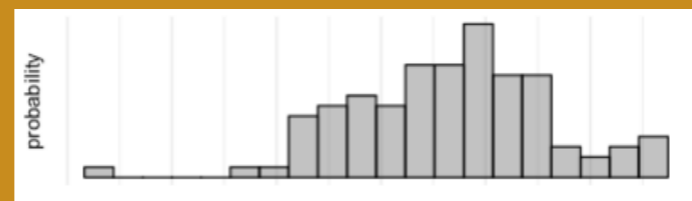
the collection of 1 or more prediction elements specific to one target and unit.

Prediction Elements

point



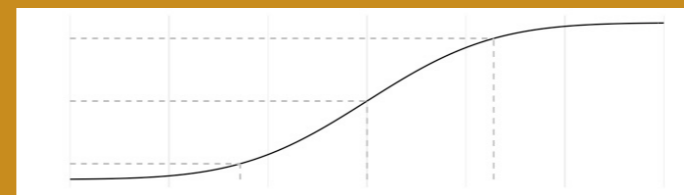
bin



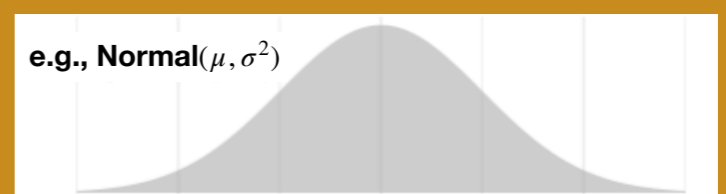
sample



quantile



named



example probabilistic prediction:







samples of the 1 week ahead confirmed COVID-19 cases in Florida
(the prediction element) (the target) (the unit)

We define a "forecast" as a collection of predictions

Forecast

metadata

- > the model that made the forecast
- > the forecast date for this forecast
- > the date the forecast was submitted

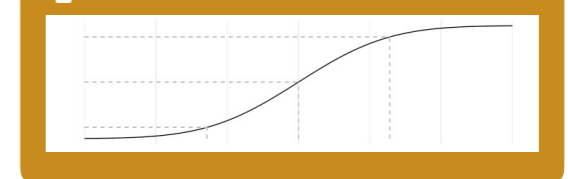
data			
	target 1	...	target T
unit 1		...	
unit 2		...	
...
unit K		...	

a prediction for a single unit-target pair may contain different elements, e.g. points and quantiles

point



quantile



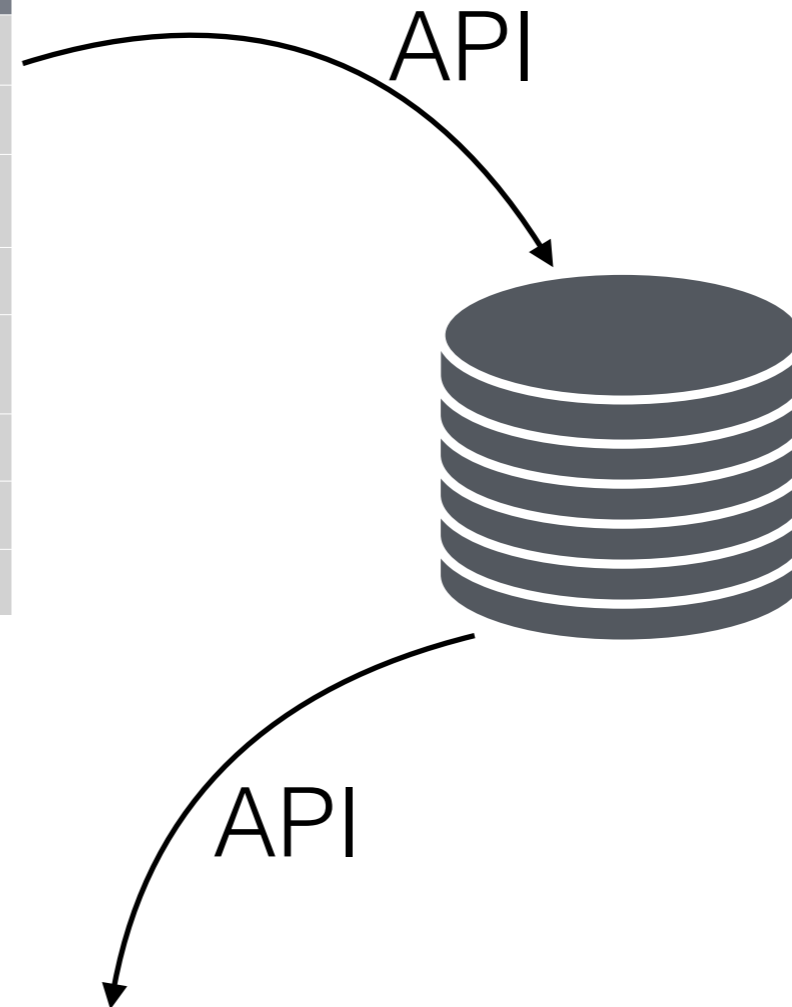
Forecast storage structure

forecast_id	target	unit	type	data
1	1 wk ahead case	US	quantile	{q: (0.1, 0.5, 0.9), value: (290, 320, 330)}
1	1 wk ahead case	US	point	{value: 320}
1	2 wk ahead case	US	sample	{value: (315, 310, 322, 333, ...)}
1	1 wk ahead case	US	quantile	{q: (0.025, 0.5, 0.975), value: (310, 320, 325)}
1	1 wk ahead case	DE	named	{family: "norm", param1: 300, param2: 22}
2	1 wk ahead case	US	quantile	{q: (0.1, 0.5, 0.9), value: (290, 320, 330)}
2	1 wk ahead case	US	point	{value: 320}
...

E.g. we think of a forecast in a "tibble"-like structure, with predictions stored in cells as lists of data.

Infrastructure simplifies Hub tasks

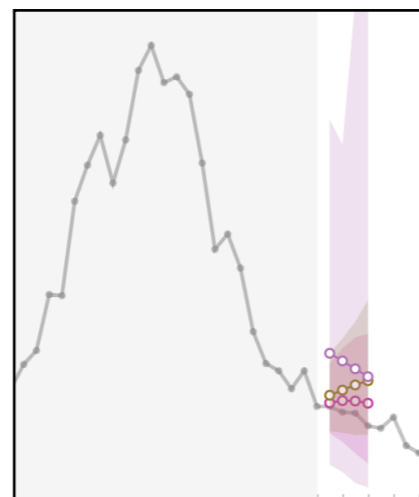
forecast_id	target	unit	type	data
1	1 wk ahead case	US	quantile	{q: (0.1, 0.5, 0.9), value: (290, 320, 330)}
1	1 wk ahead case	US	point	{value: 320}
1	2 wk ahead case	US	sample	{value: (315, 310, 322, 333, ...)}
1	1 wk ahead case	US	quantile	{q: (0.025, 0.5, 0.975), value: (310, 320, 325)}
1	1 wk ahead case	DE	named	{family: "norm", param1: 300, param2: 22}
2	1 wk ahead case	US	quantile	{q: (0.1, 0.5, 0.9), value: (290, 320, 330)}
2	1 wk ahead case	US	point	{value: 320}
...



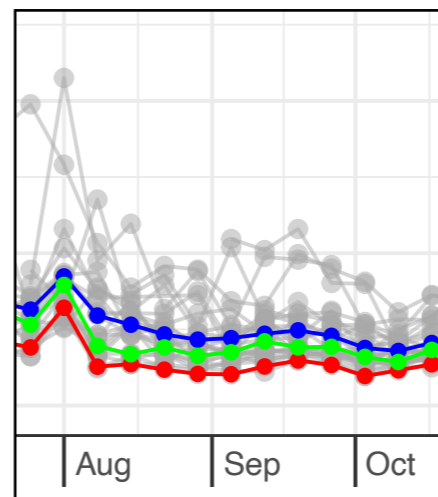
extract

ahead case	US
ahead case	US
ahead case	US
ahead case	a
2	c
3	a
3	b
3	c

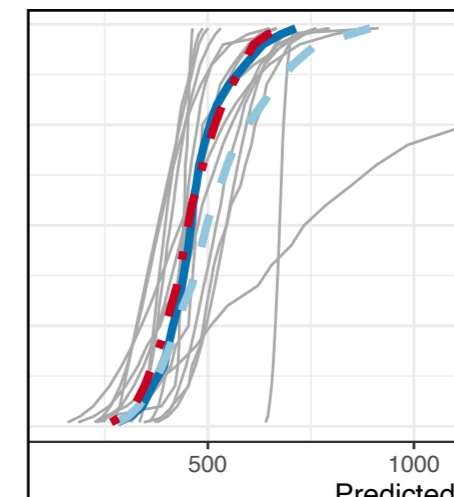
viz



eval



combine



Infrastructure simplifies Hub tasks

forecast_id	target	unit	type	data
1	1 wk ahead case	US	quantile	{q: (0.1, 0.5, 0.9), value: (290, 320, 330)}
1	1 wk ahead case	US	point	{value: 320}
1	{value: (315, 310, ...)}
1
1
2	1 wk ahead case	US	quantile	{q: (0.1, 0.5, 0.9), value: (290, 320, 330)}
2	1 wk ahead case	US	point	{value: 320}
...



COVID-19
ForecastHub

API



zoltardata.com

zoltr (R), zoltpy (python) libraries

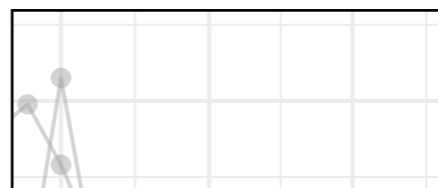
extract

ahead case	US
ahead case	US
ahead case	US

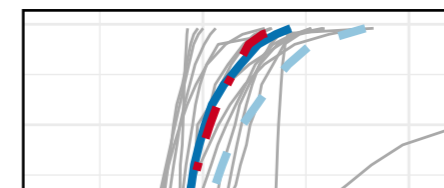
viz



eval

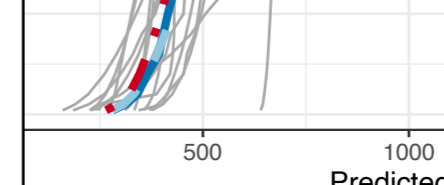
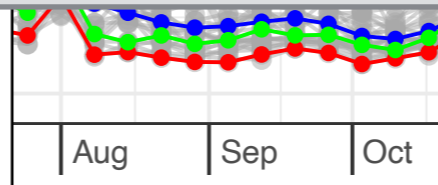


combine



covidHubUtils R package

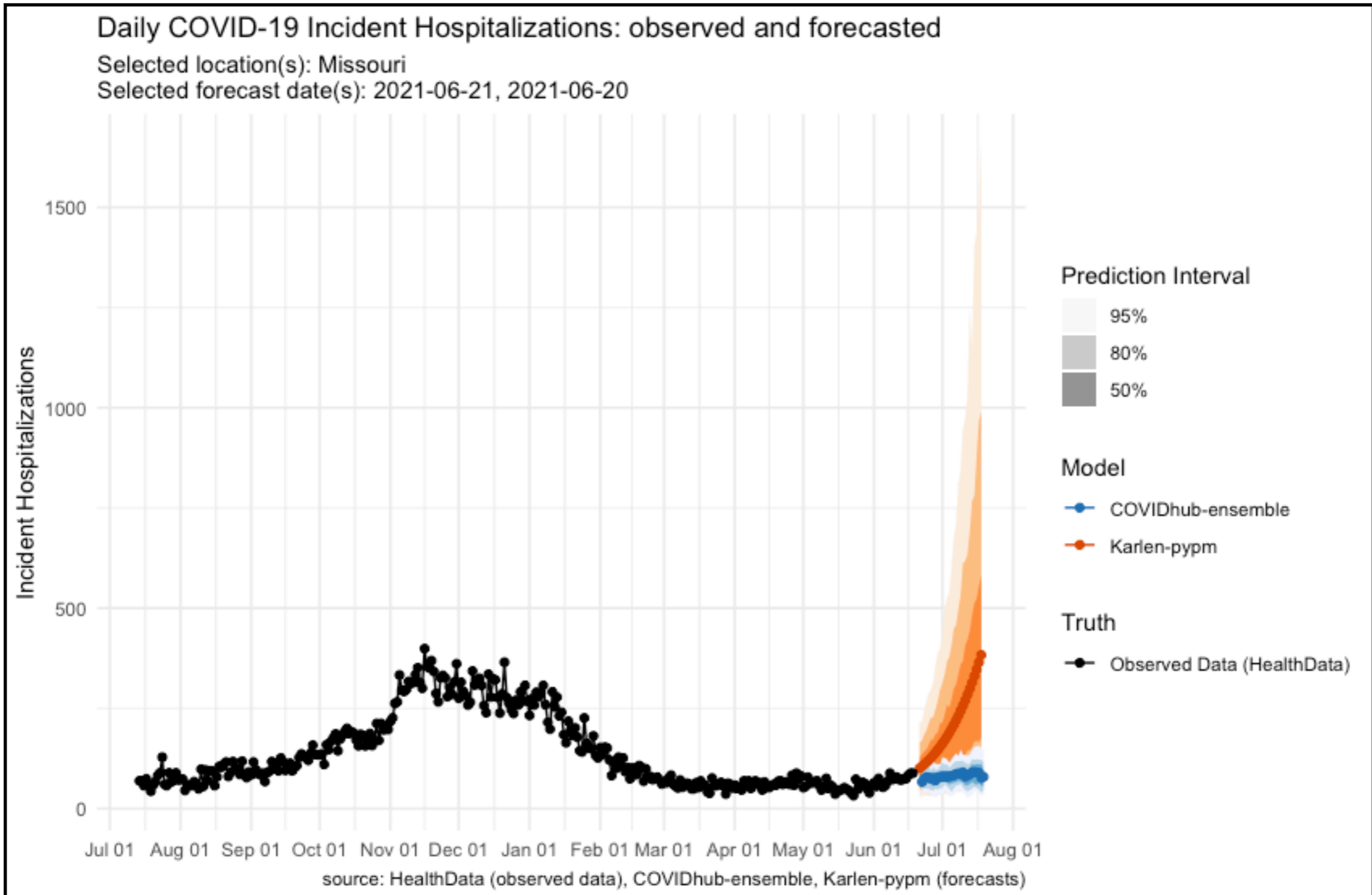
3	a
3	b
3	c



```
library(covidHubUtils)
```

```
forecast_data <- load_forecasts(  
  models = c("COVIDhub-ensemble", "Karlen-pypm"),  
  locations = "29",  
  targets = paste(1:28, "day ahead inc hosp"),  
  forecast_dates = c("2021-06-20", "2021-06-21"))
```

```
plot_forecasts(forecast_data, truth_source="HealthData", fill_by_model = TRUE)
```



Closing thoughts

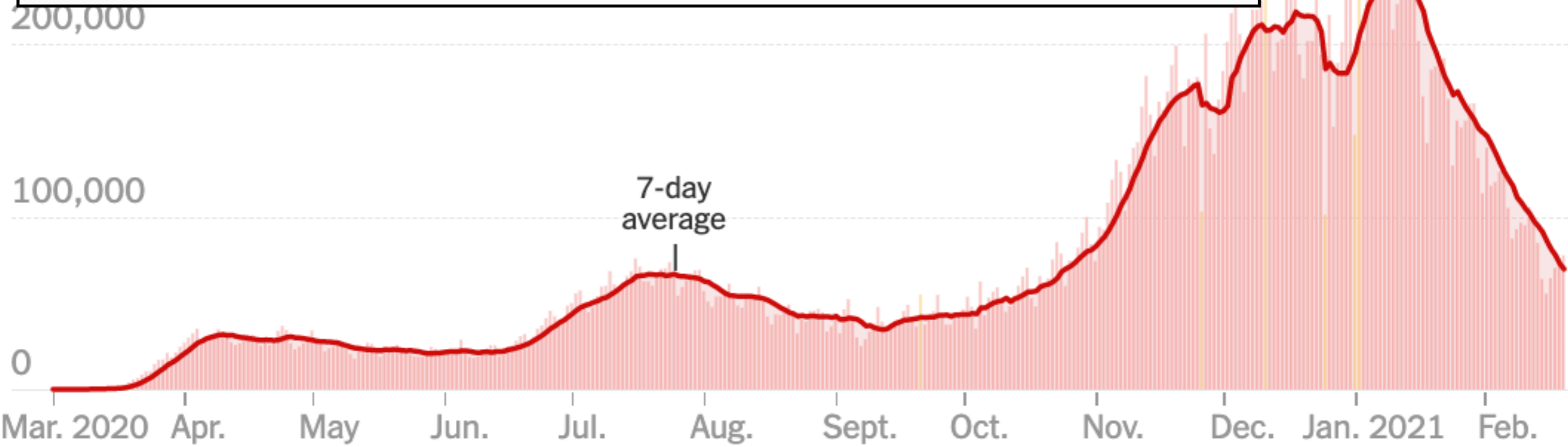
Ensemble in use by gov't officials

CNN health Food Fitness Wellness Parenting Vital Signs LIVE TV Edition Q

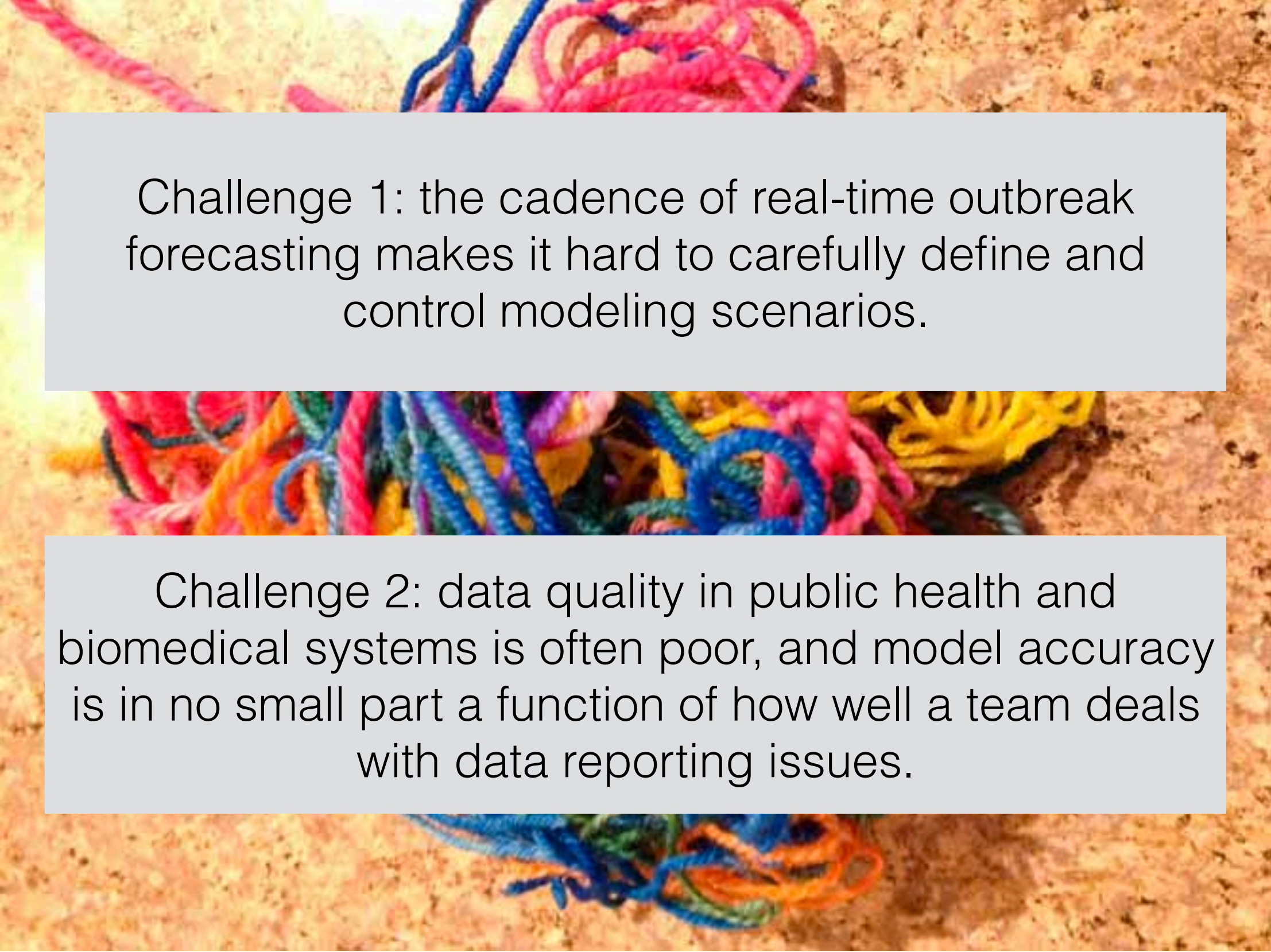
Biden says Covid-19 death toll will likely top 500,000 next month, asks Americans to wear mask for his first 100 days

By **Dakin Andone** and **Madeline Holcombe**, CNN
Updated 4:14 PM ET, Thu January 21, 2021

Biden's warning of tens of thousands more deaths in just a few weeks echoes the latest ensemble forecast by the US Centers for Disease Control and Prevention, which projects the death toll could reach up to 508,000 by February 13.



Disentangling knowledge from the hub



Challenge 1: the cadence of real-time outbreak forecasting makes it hard to carefully define and control modeling scenarios.

Challenge 2: data quality in public health and biomedical systems is often poor, and model accuracy is in no small part a function of how well a team deals with data reporting issues.

Learning from and with other hubs

What are the similarities with related model coordination efforts in other fields?

(these and many others...)

Climate

The logo for the Intergovernmental Panel on Climate Change (IPCC), consisting of the lowercase letters "ipcc" in a blue, sans-serif font.

ipcc.ch

Ecology

The logo for the Fisheries & Marine Ecosystem Model Intercomparison Project (FISH-MIP). It features the text "FISHERIES & MARINE ECOSYSTEM" in red above "FISH-MIP" in large blue letters, with a small fish icon between "FISH" and "MIP". Below this is "MODEL INTERCOMPARISON PROJECT" in red.

isimip.org/about/marine-ecosystems-fisheries/

Space Science



COMMUNITY
COORDINATED
MODELING
CENTER

ccmc.gsfc.nasa.gov

Key open questions

- What data signals can improve COVID-19 forecasting? (new preprint from CMU Delphi group)
- Does it say anything about pandemic predictability and/or model quality and diversity that it is hard to beat a simple median model?
- What, if any, general conclusions can we take away from Hub data about accuracy of different model structures, at different horizons, etc...?

Take-aways

1. It is always important to look at multiple models.
2. Pandemic forecasting is hard, especially at change-points.
3. No model is reliably well-calibrated at horizons longer than 4 weeks ahead.
4. Simple ensemble methods work quite well.
5. Efficient model coordination requires good infrastructure.

Thank you!

With acknowledgments to Evan Ray, all members of Reich Lab and COVID-19 Forecast Hub, CDC collaborators, modeling contributors.

The Reich Lab and the COVID-19 Forecast Hub have been supported by the National Institutes of General Medical Sciences (R35GM119582) and the US Centers for Disease Control and Prevention (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 US CDC.



The Reich Lab and the COVID-19 Forecast Hub have been supported by the National Institutes of General Medical Sciences (R35GM119582) and the US Centers for Disease Control and Prevention (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 US CDC.

Promote open infrastructure

Support and value development of portable, reusable, open-source infrastructure for data and modeling.

Some COVID-19 examples, supported by CDC, European CDC, among others:

Forecast and Modeling Hubs

Standardized model outputs to support public health decision-makers.



<https://covid19forecasthub.eu/>



COVID-19
ForecastHub

<https://covid19forecasthub.org/>



COVID-19
ScenarioModelingHub

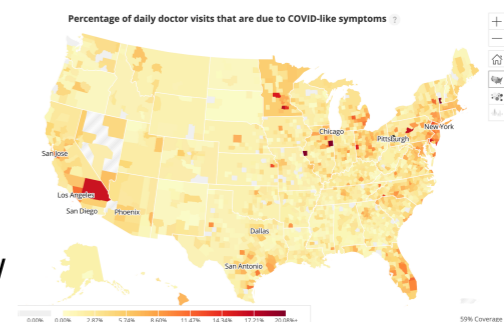
<https://covid19scenariomodelinghub.org/>

Data infrastructure

APIs for public health data.

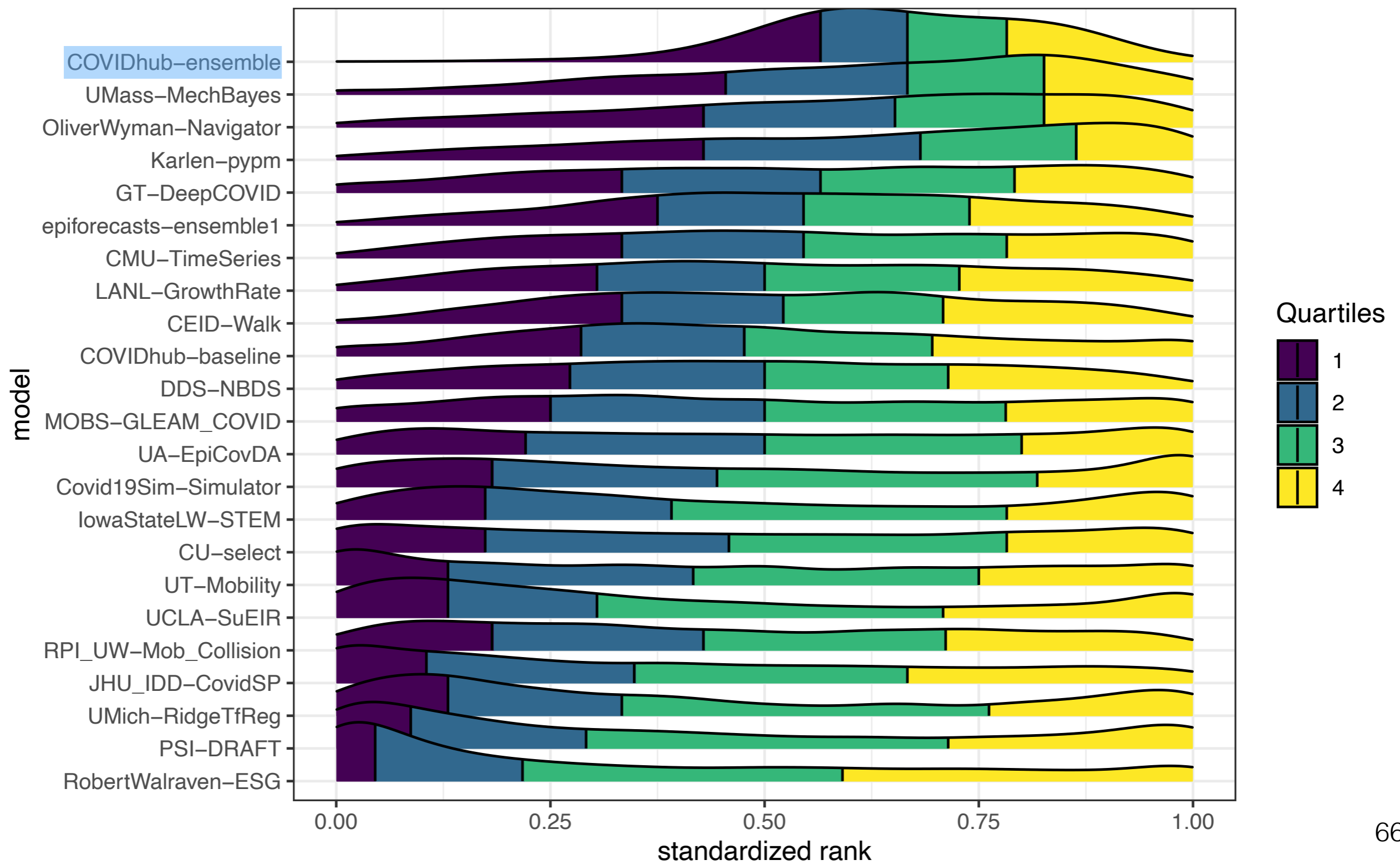
COVIDcast

<https://covidcast.cmu.edu/>



Hub ensemble is most consistent

Across over 10,000 predictions, the ensemble is ranked in the top half of all forecasts for incident deaths over 75% of the time. No other model achieves this level of consistency.



Relative WIS Weighted Median

- Idea: Introduce a single parameter determining model weights as a function of training set relative WIS

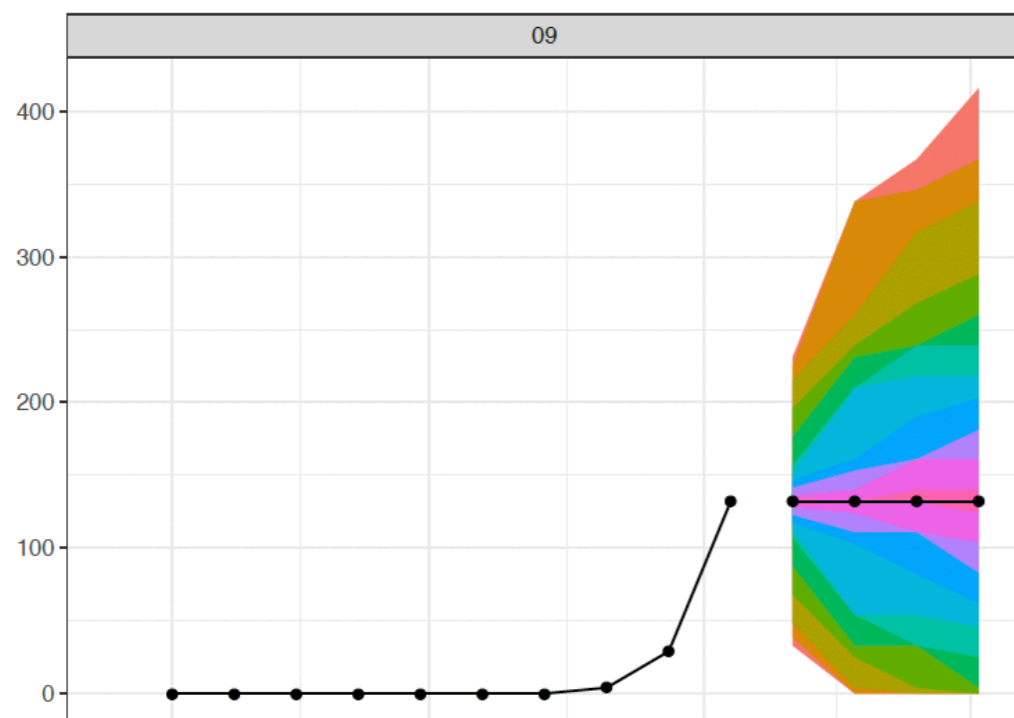
$$w_m = \frac{\exp(-\theta \cdot \mathbf{Relative\ WIS\ Model\ m})}{\sum_{j=1}^M \exp(-\theta \cdot \mathbf{Relative\ WIS\ Model\ j})}$$

- If theta is large and model m is bad (high relative WIS), model m gets low weight; if model m is good (low relative WIS), model m gets high weight
- Ensemble forecast is weighted median at each quantile level
- For estimation, objective is not differentiable; currently optimizing with simulated annealing (slow — roughly a day to run; could just use grid search?)
- Can also combine with the idea of using only models with lowest relative WIS in the past few weeks

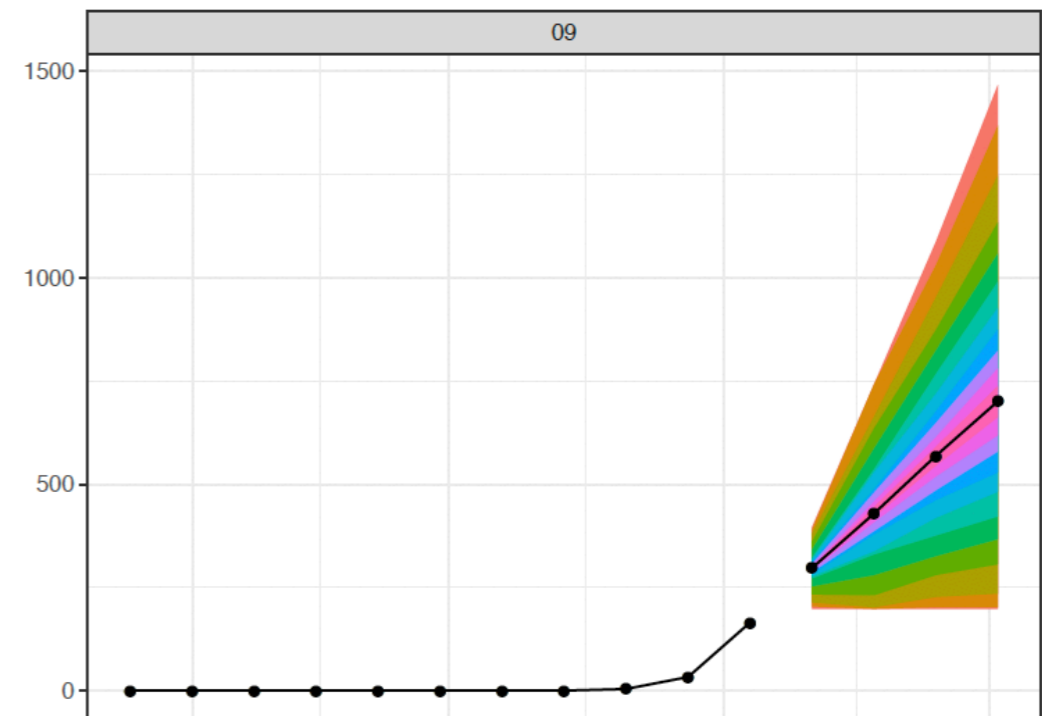
Baseline Model

- Different from flu forecasting baseline model! Not "seasonally" driven.
- Acknowledgment: idea adapted from a suggestion by Ryan Tibshirani (CMU).
- Goal: Median predicted incidence is most recent observed incidence.
- Predictions of cumulative deaths derived from predictions of incident deaths.

Incident Deaths

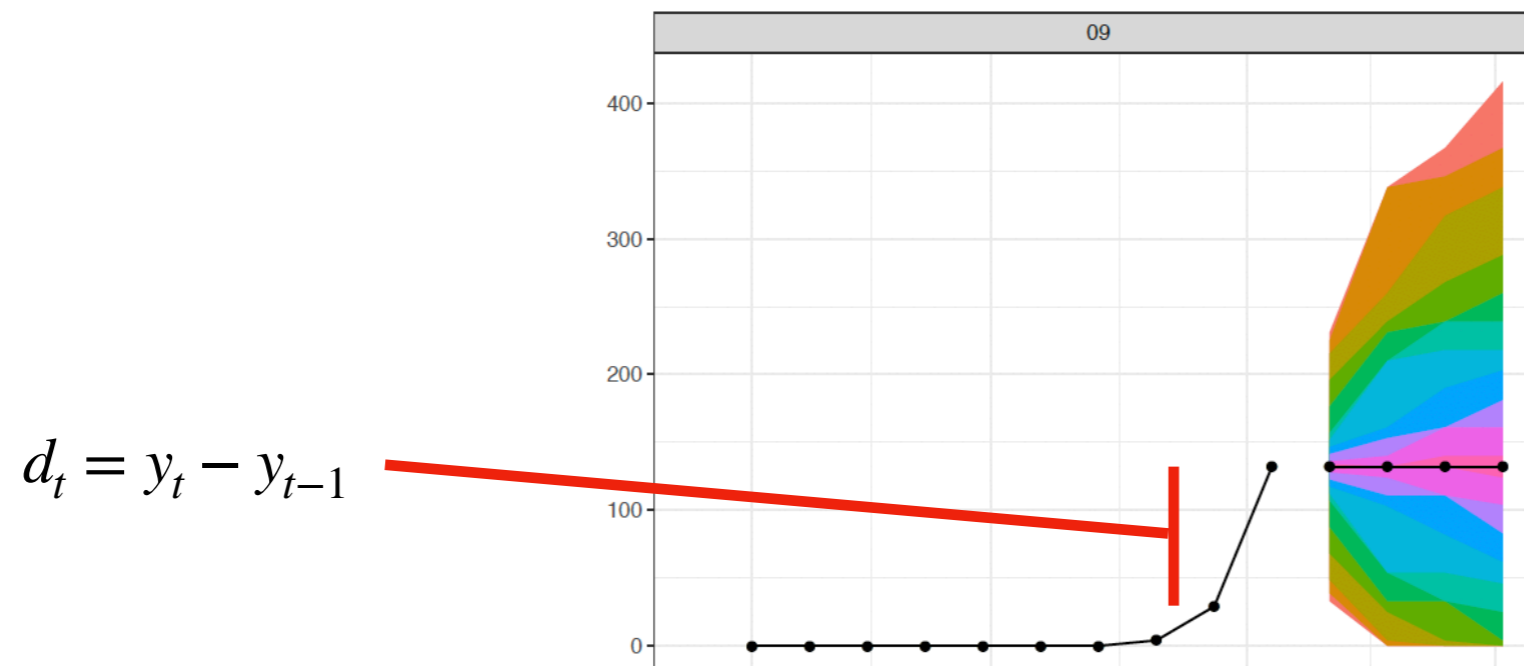


Cumulative Deaths



Baseline Model

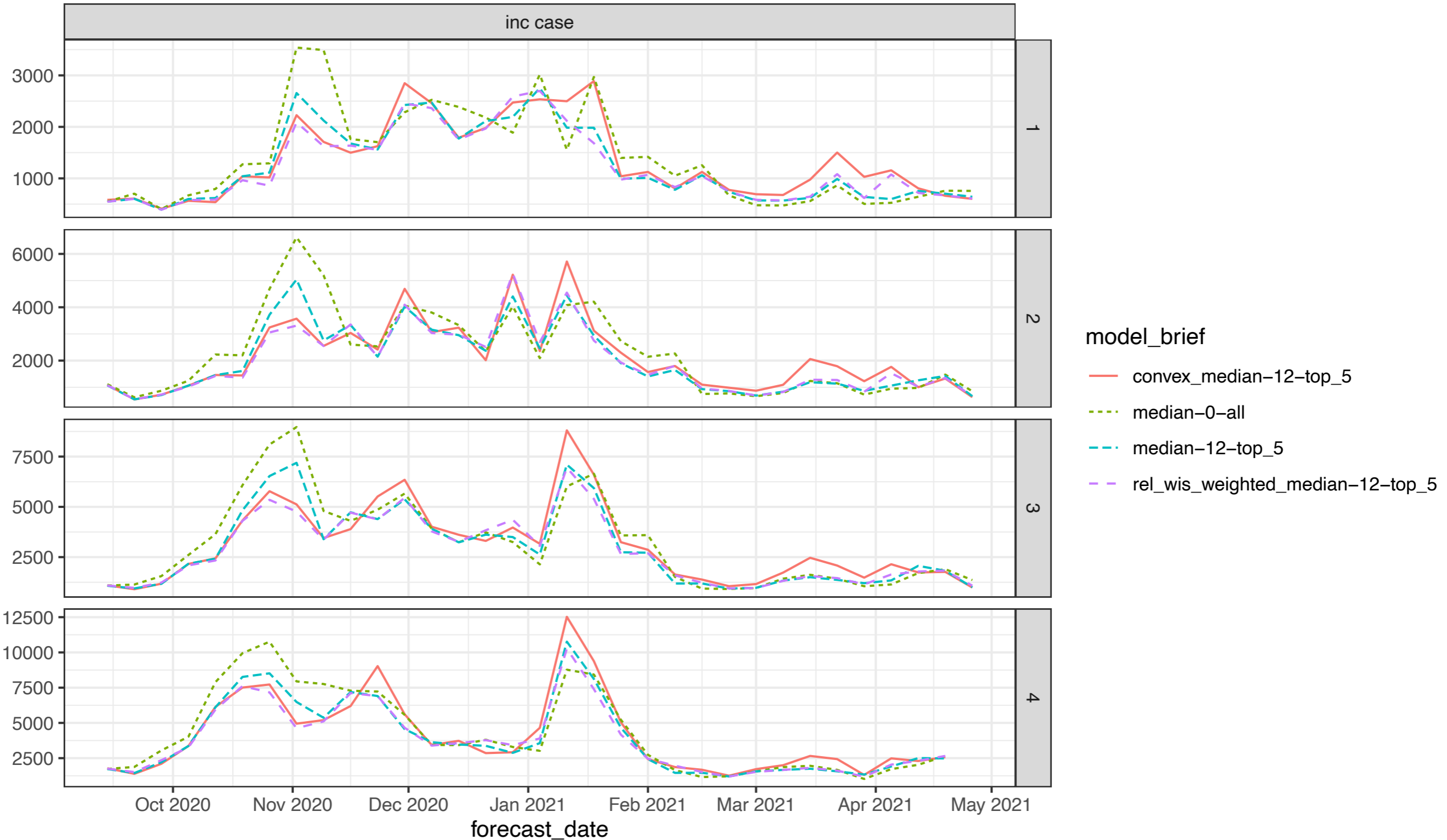
- Procedure:
 - Compute first differences of historical incidence:



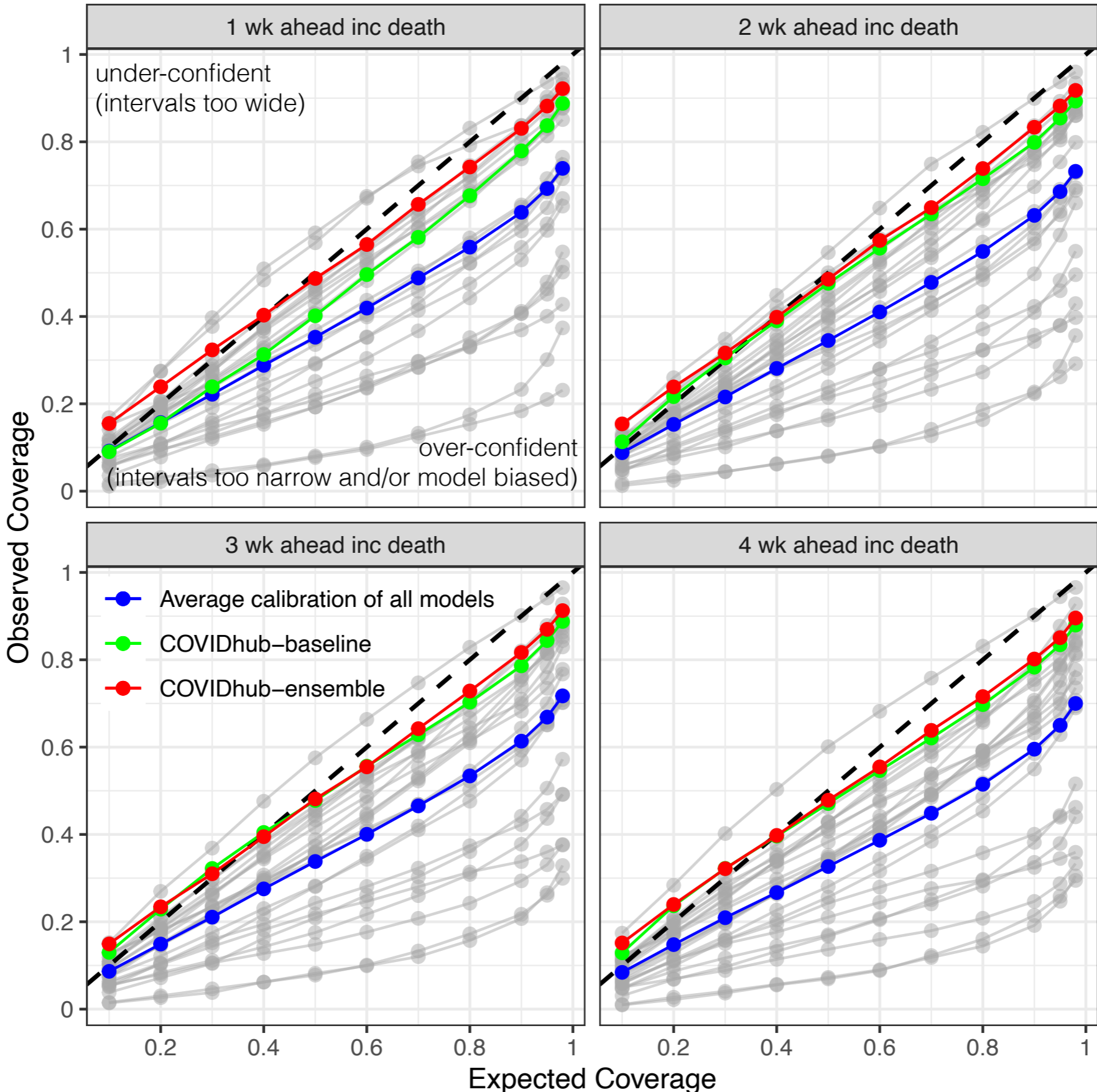
- Collect first differences and their negatives
- Sample first differences and add to last observed incidence; take quantiles of the resulting distribution
- Iterate for horizons > 1
- Adjustments for “niceness”:
 - Force median = last observed incidence
 - Truncate at 0

Evaluation Over Time – Cases

Inc Case

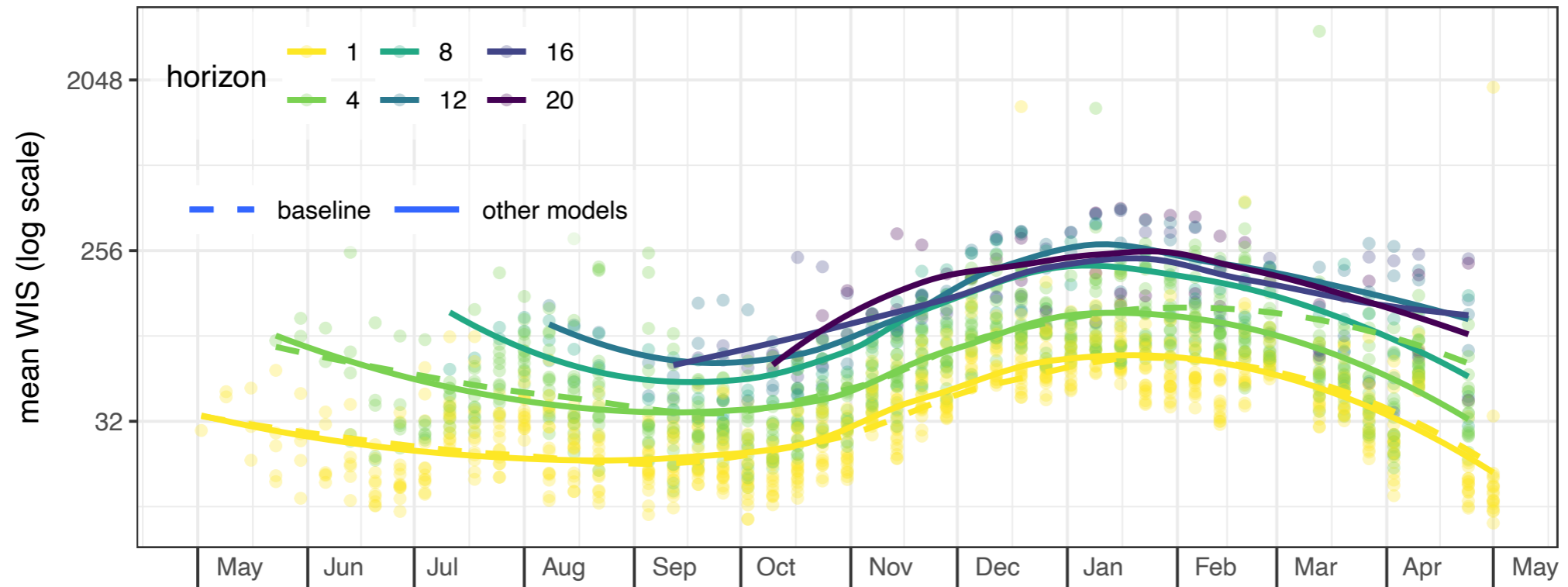


Model calibration – Deaths

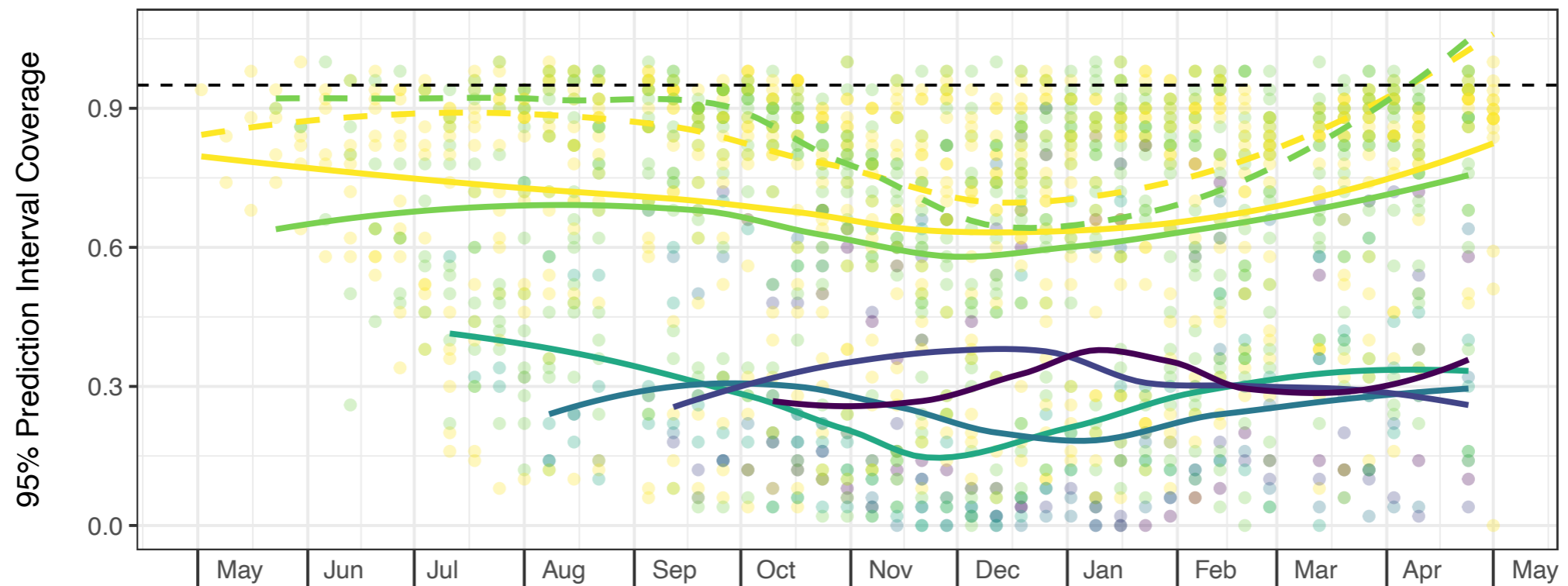


Errors increase with horizon

B: mean WIS across time, stratified by forecast horizon

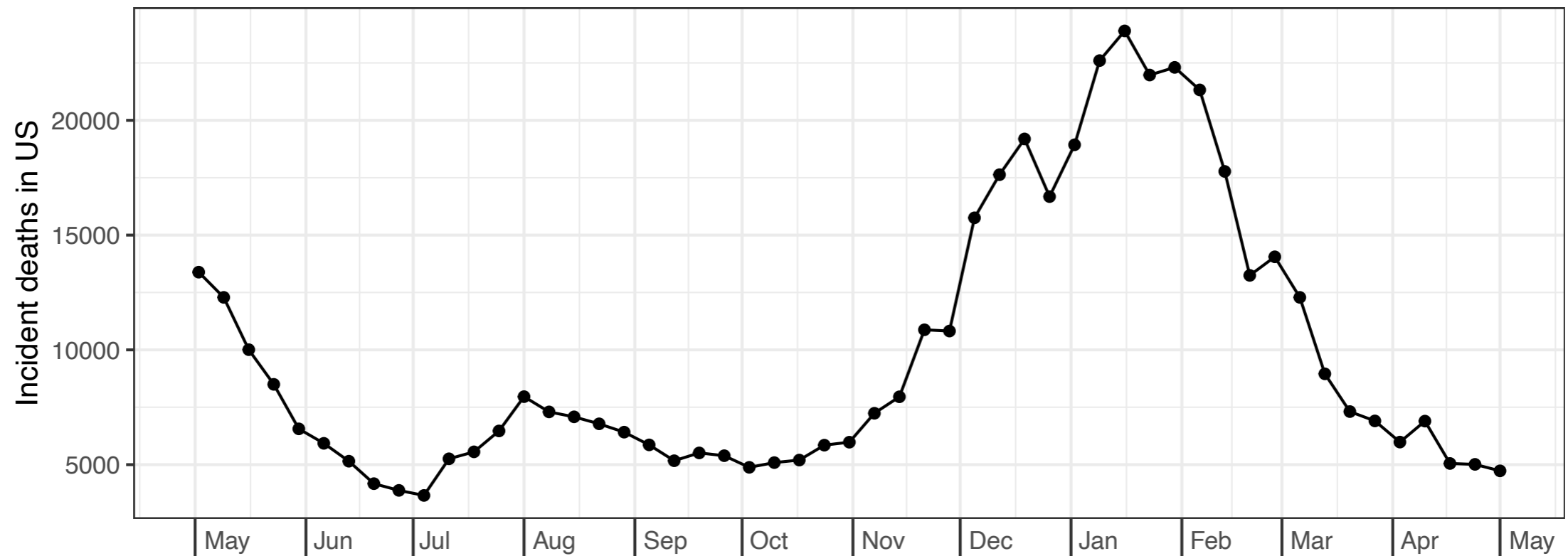


C: 95% prediction interval coverage across time, stratified by forecast horizon

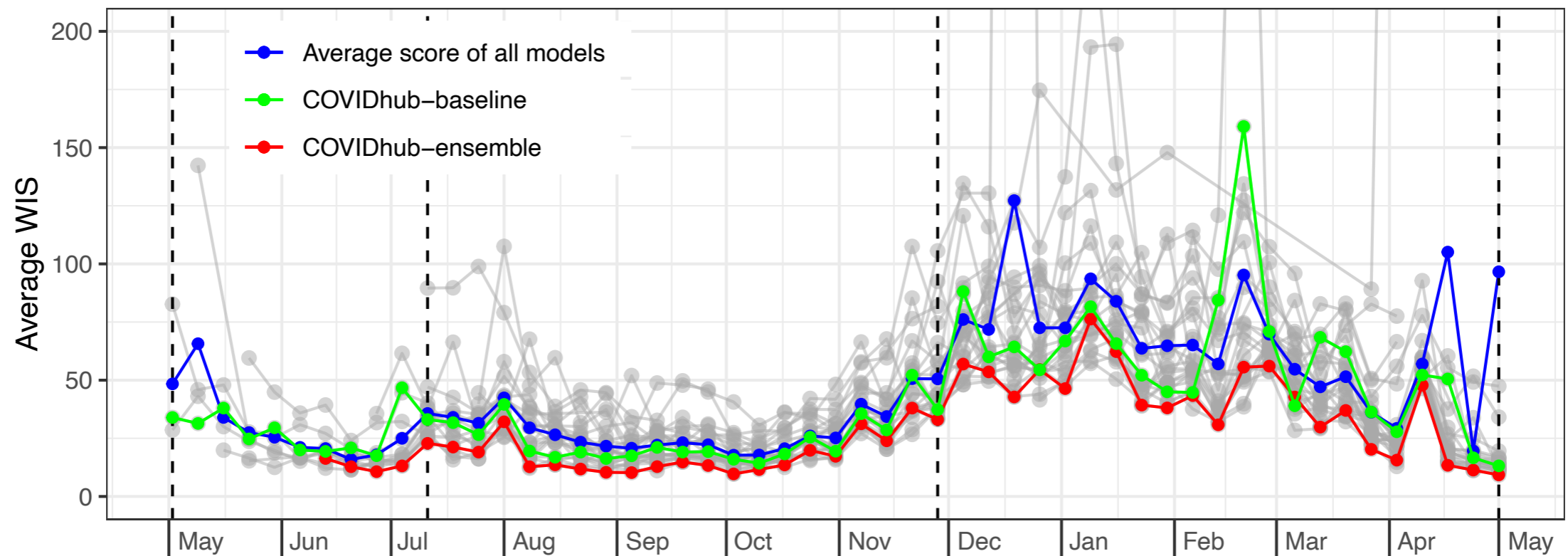


Errors over week – Deaths

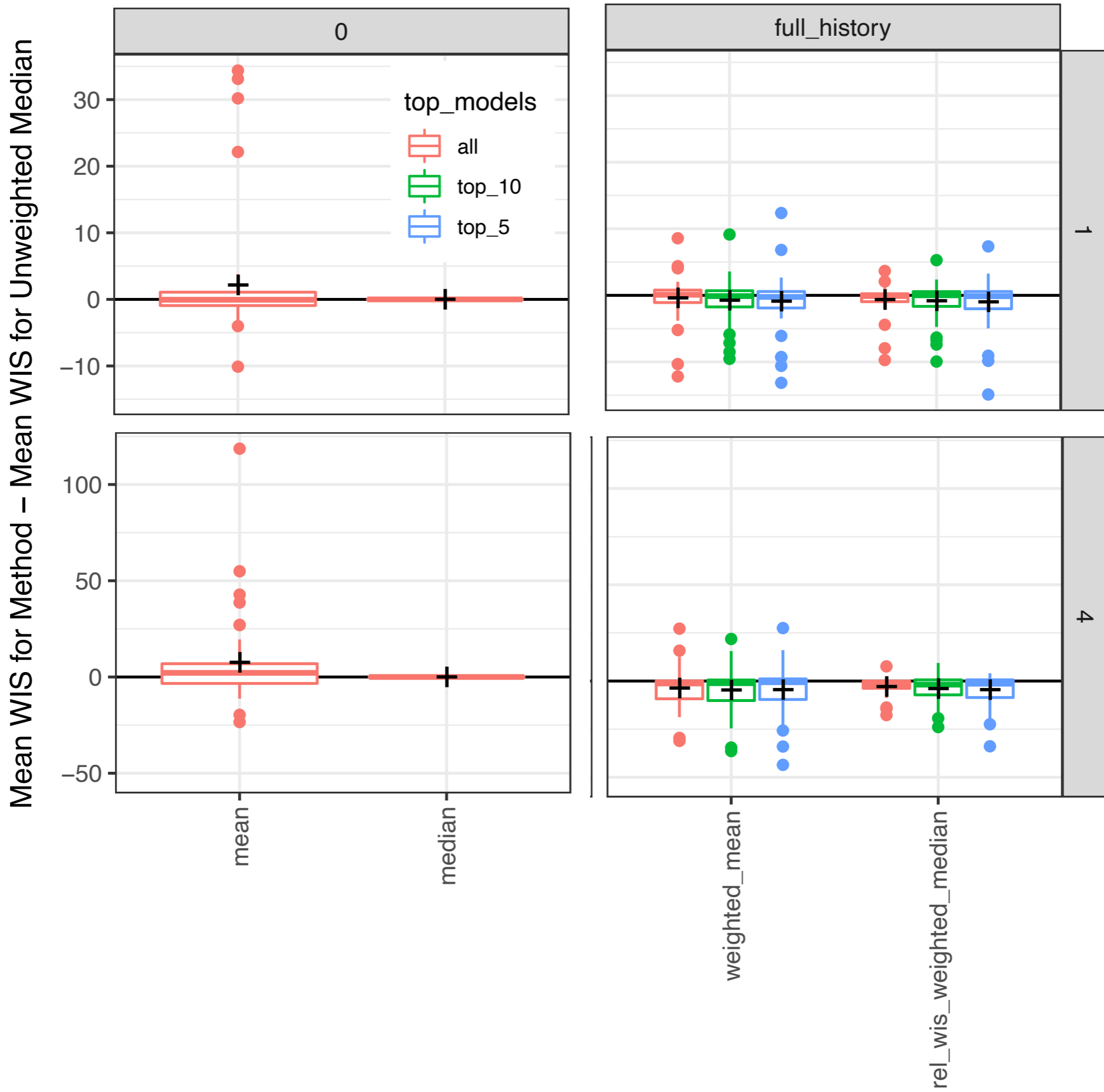
A: Observed weekly COVID-19 deaths in the US



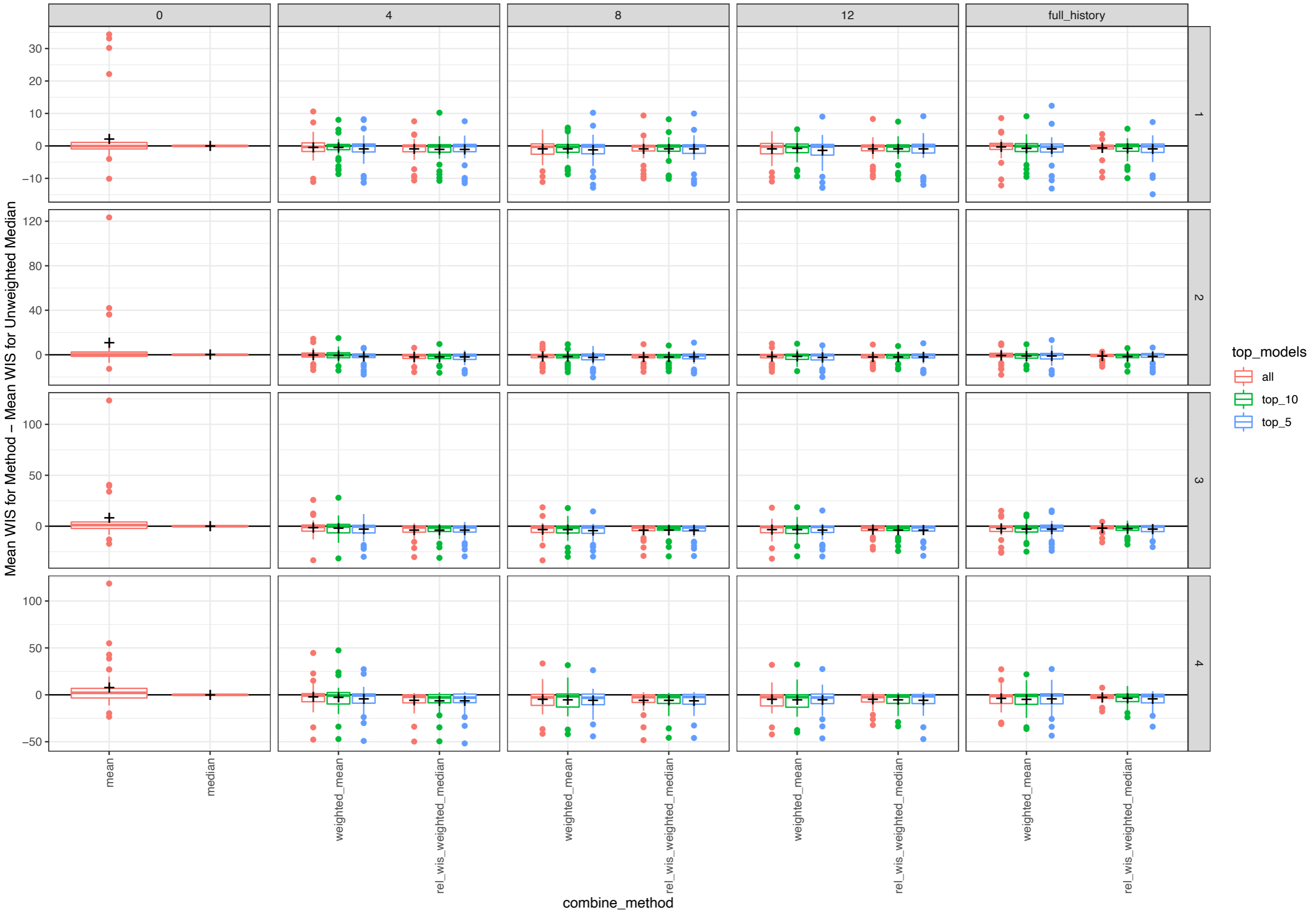
B: Average 1-week ahead weighted interval scores by model



inc_death, state



inc_death, state



inc_case, state

