Evaluating Infectious Disease Forecasts with Allocation Scoring Rules

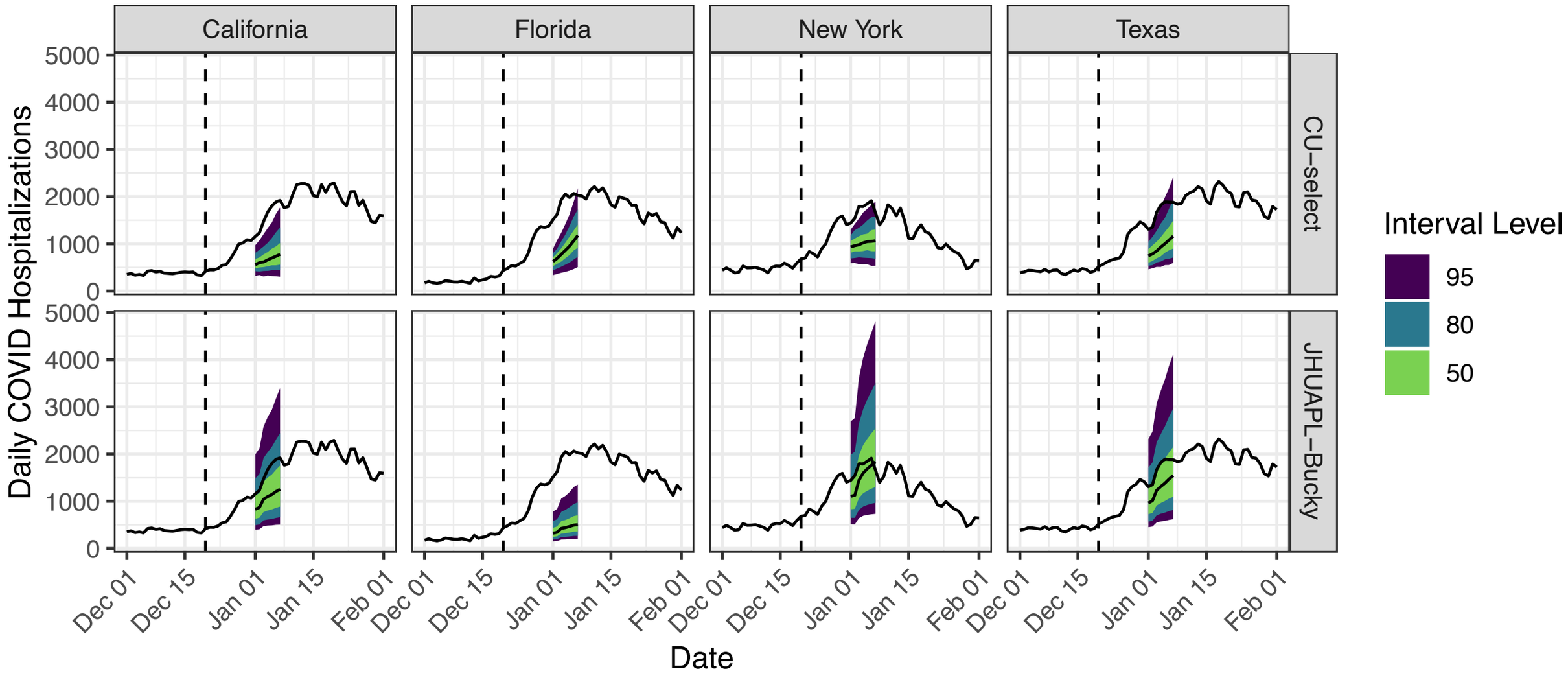
Aaron Gerding, **Nicholas G. Reich**, Ben Rogers, Evan L. Ray

UMass Statistics & Data Science Seminar
27 Feb 2025

UMassAmherst | School of Public Health & Health Sciences
Biostatistics and Epidemiology

Which of these forecasts is better?

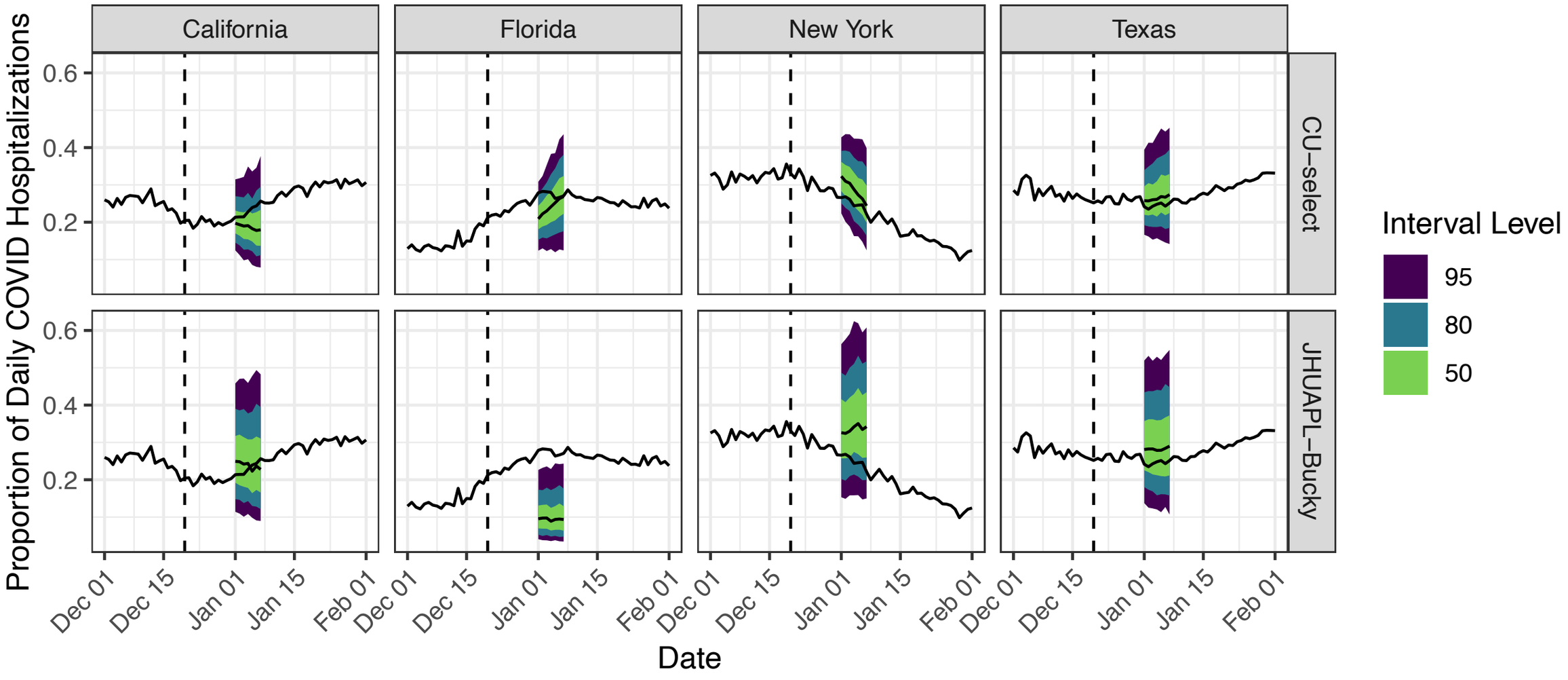
- Predictions are of COVID-19 hospitalizations in 4 states, leading into the Omicron wave in the US in winter 2021/2022



- Which model was more successful?
  - CU-select WIS: **417** ❌
  - JHUAPL-Bucky WIS: **380** 🏆

Which of these forecasts is better?

- Data: on each day, what proportion of hospitalizations were from each of these four states?
- Predictions were obtained from the predictions of daily hospitalization counts on previous slide, assuming independence across locations.



- Which model was more successful?

The main point

Common methods for forecast evaluation		Common (?) uses of forecasts
Log score		Situational awareness, public communications
WIS, CRPS	???	Planning expansions to hospital bed or ICU capacity
Empirical coverage rates		Site selection for vaccine trials
Allocation score		Allocation of limited medical supplies (e.g. ventilators, oxygen)

- It is not clear how well standard forecast scores measure the value of forecasts for public health decision making.
- We'll develop a proper score that is specifically tuned to measuring the value of forecasts for decisions about resource allocations.

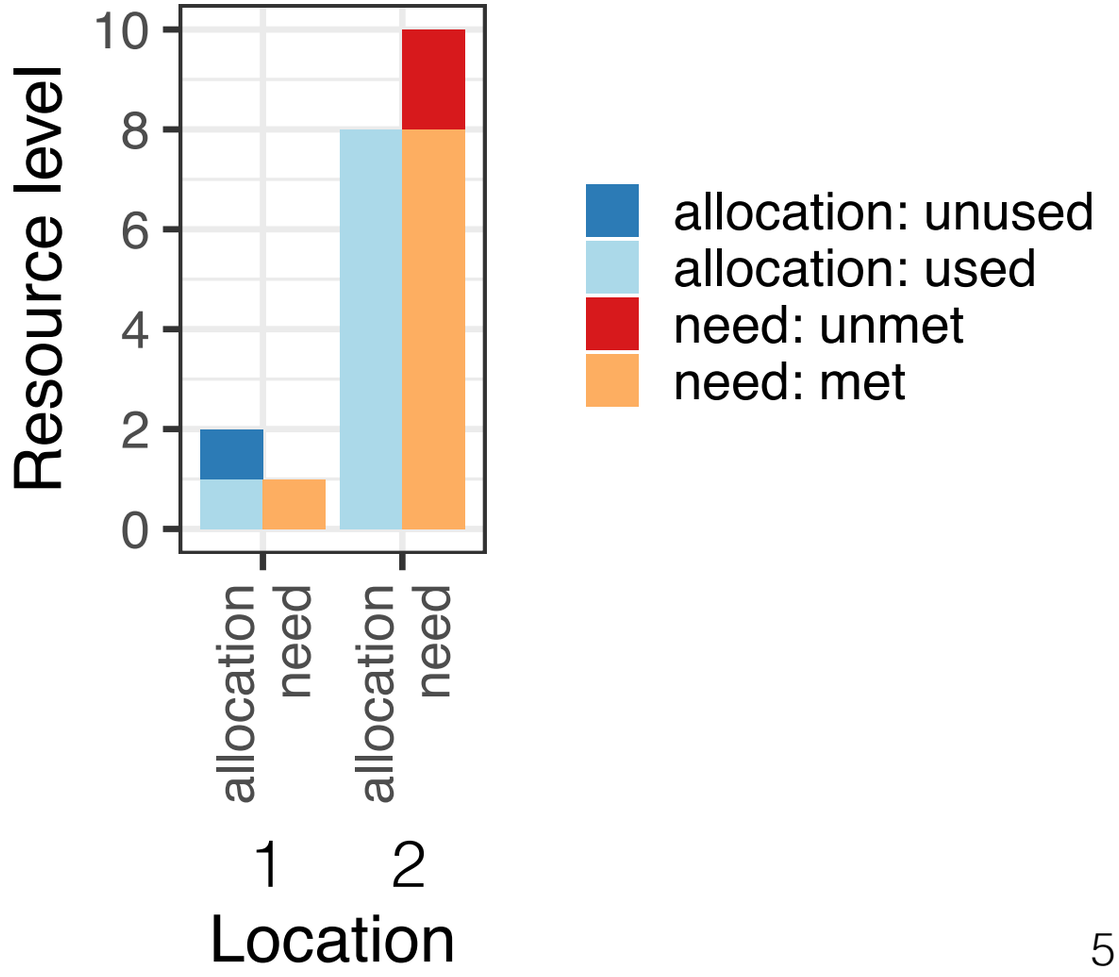
The allocation score — recipe step 1

- Step 1: Define a loss function $s(\mathbf{x}, \mathbf{y})$
  - If you take the action $\mathbf{x}$ and observe the outcome $\mathbf{y}$, what are your losses?
- Our loss measures the total amount of unmet need across all locations:

$$s(\mathbf{x}, \mathbf{y}) = \sum \max(y_i - x_i, 0)$$

  - If realized need in location i is greater than allocated resources, add $y_i - x_i$
  - Otherwise, enough resources allocated to location i; no contribution to loss

- Example:
  - two locations, $K = 10$ units of resources
  - allocate $\mathbf{x} = (2, 8)$ units of resources to locations 1 and 2
  - eventually the value $\mathbf{y} = (1, 10)$ is observed
  - unmet need is $s(\mathbf{x}, \mathbf{y}) = 0 + (10 - 8) = 2$

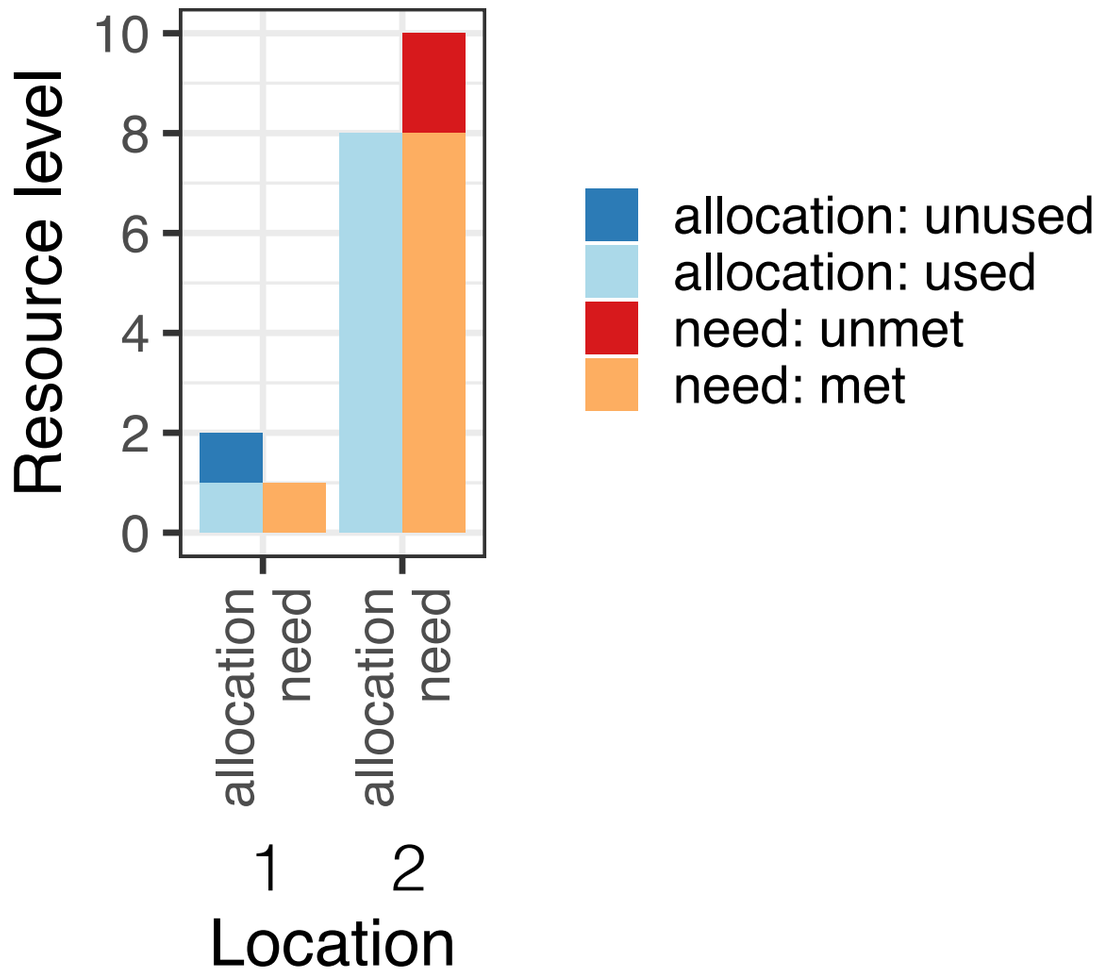


Oracle adjustment — Example

- How much better could we have done with another allocation than the one suggested by F?

$$S(F, \mathbf{y}) = S^{raw}(F, \mathbf{y}) - S^{raw}(F^{Oracle}, \mathbf{y})$$

- Example with $K = 10$, two locations, $\mathbf{x} = (2,8)$, and $\mathbf{y} = (1,10)$
- This allocation's raw score is 2
- Since total need is 11 and K=10, even the best possible allocation has a raw score of 1
- The "allocation regret" score is $(2 - 1) = 1$ units of avoidable loss



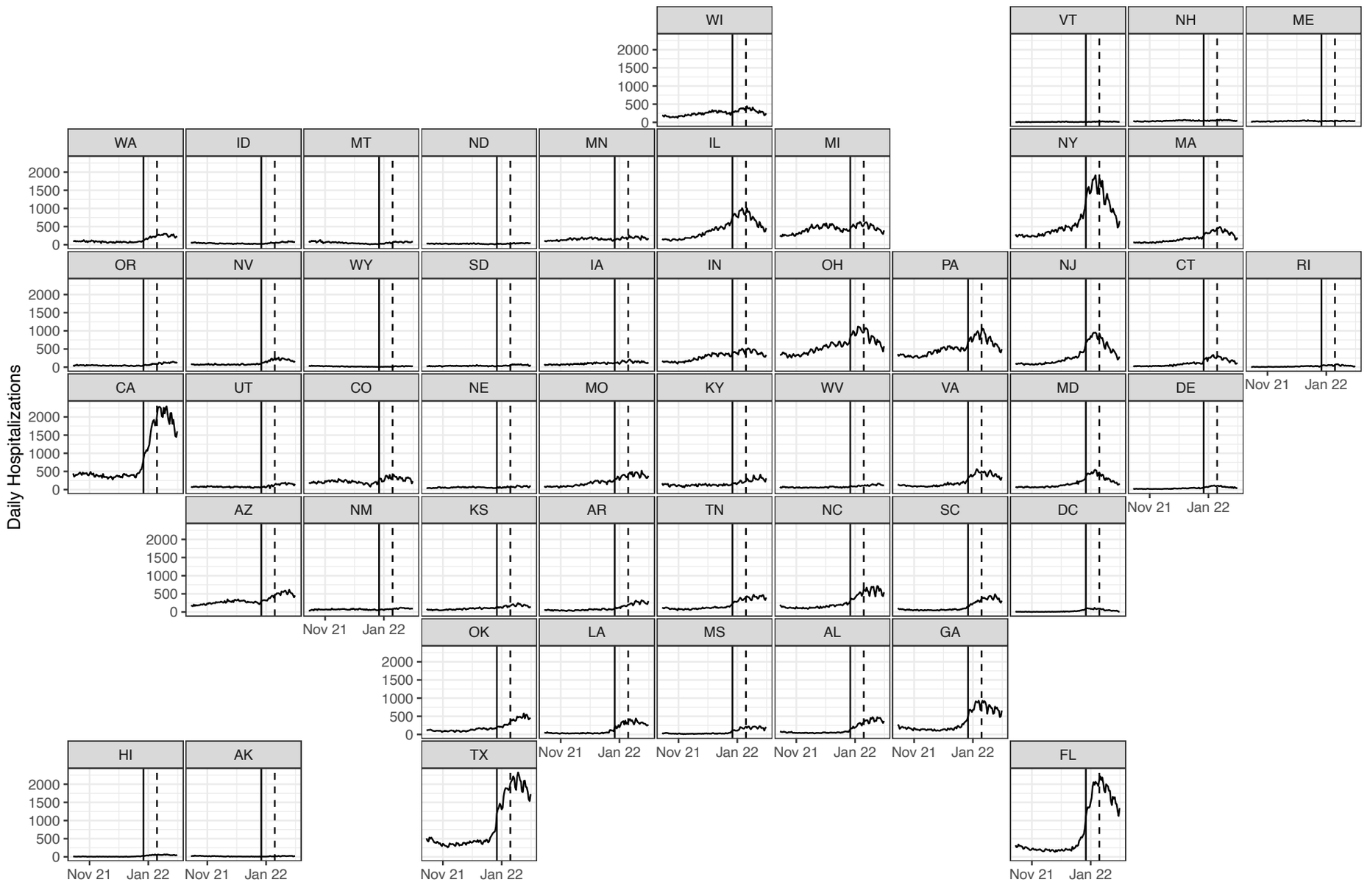
The allocation score — recapping

Recapping:

- How should we allocate K units of resources across n locations?
- $\mathbf{x} = (x_1, \ldots, x_n)$: Amount of resources allocated to each location
- $\mathbf{y} = (y_1, \ldots, y_n)$: Resource need in each location (observed)
- In each location, the Bayes act sets $x_i^F = F_i^{-1}(\tau)$ where $\sum x_i^F = K$
- Score for evaluation: "allocation regret":
  - how much avoidable loss resulted from using the allocation suggested by F?
  - How much better could we have done by using a different allocation?

A “simple” example

- Allocation of federal resources to the US states heading into the Omicron hospitalizations wave
- For illustrative purposes, we set K = 15,000



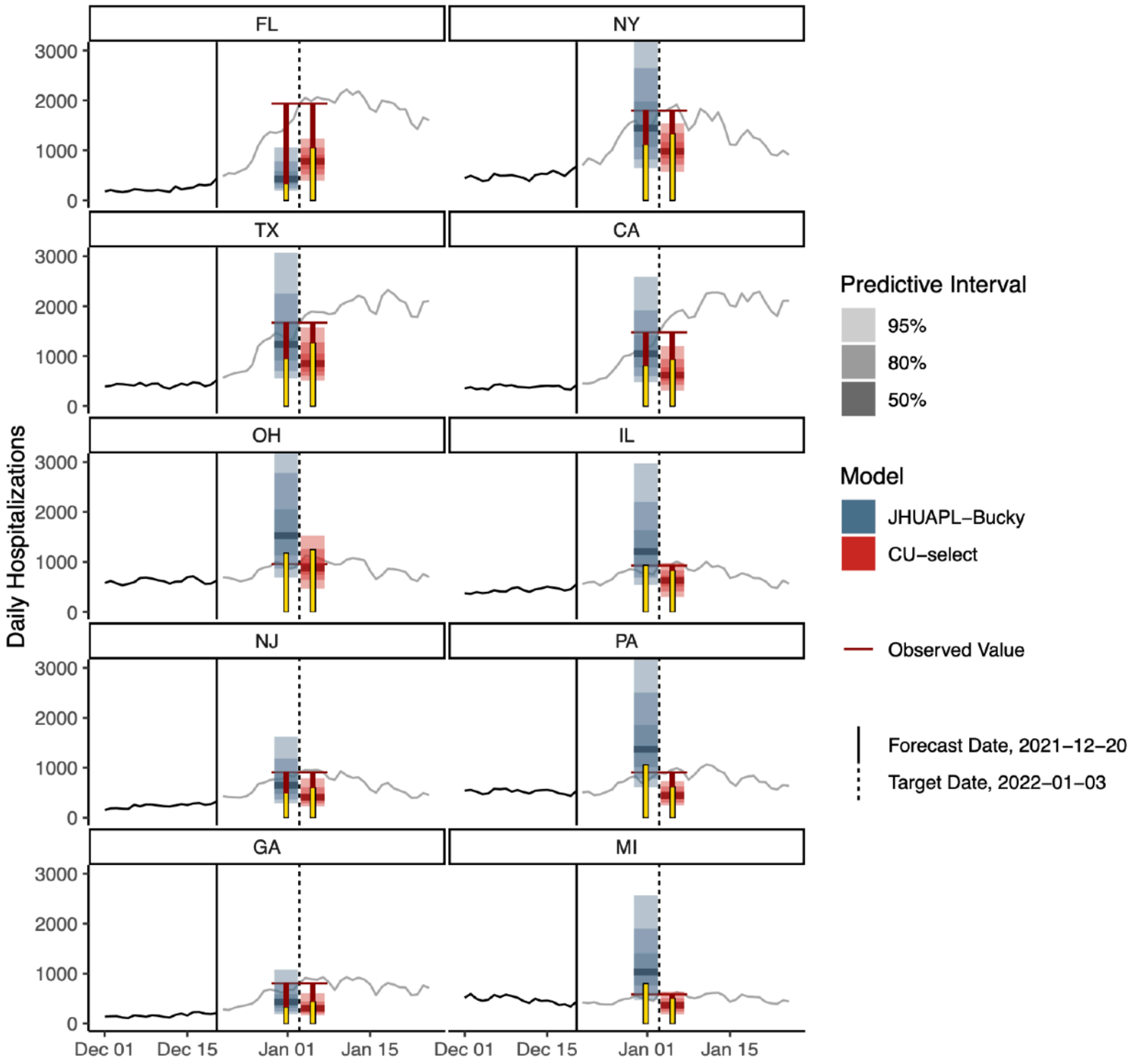
Models ranks by WIS and allocation score differ

- CU-select: best allocation score, middling WIS
- JHUAPL-Bucky: second-best WIS, poor allocation score
- USC-SI_kJalpha: pretty good according to both scores

Model	AS	MWIS
CU-select	669	133
COVIDhub-ensemble	873	159
USC-SI_kJalpha	995	91
JHUAPL-Gecko	1034	164
MUNI-ARIMA	1084	169
COVIDhub-trained_ensemble	1089	169
COVIDhub-baseline	1175	170
JHUAPL-Bucky	1358	102
JHUAPL-SLPHospEns	1540	129
UVA-Ensemble	2469	213

A closer look at JHUAPL-Bucky and CU-select

- JHUAPL-Bucky: 2nd best WIS, 3rd worst allocation score
  - In most states, captures magnitude of outcomes fairly well
  - Under predicted in FL, over predicted in states like MI and PA
- CU-select: Best allocation score, middling WIS
  - Consistently under predicted in all locations
  - Got relative resource need about right



A closer look at JHUAPL-Bucky and CU-select

- JHUAPL-Bucky: 2nd best WIS, 3rd worst allocation score
- CU-select: Best allocation score, middling WIS