Evaluating Infectious Disease Forecasts with Allocation Scoring Rules

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

CoE call
23 Jan 2024

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

Overview of this talk

- **Introduction and motivation**
- A sequence of 3 example scores
- Illustrative application
- Conclusions

The main point — version 1

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 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.

The main point — version 1

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 (or lack of 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 main point — version 2

The purpose of this ~~paper~~ slide deck is to:

- Illustrate a standard setup for deriving proper scores starting from a decision making context
- Illustrate that resulting scores don't necessarily align with WIS
- Make an argument that:
  - using WIS is a choice rather than an automatic default
  - when possible, it would be good to choose forecast scores that align with decision making context.

Overview of this talk

- Introduction and motivation
- **A sequence of 3 example scores**
- Illustrative application
- Conclusions

Proper scores from decision making context

Setup:

- Suppose a decision maker must select an action $\mathbf{x}$
- The quality of a decision can be quantified in relation to an outcome $\mathbf{y}$
- $\mathbf{y}$ is unknown at the time of decision making, but a forecast F of $\mathbf{Y}$ is available.

The following recipe gives us a proper score:

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?
2. Get the Bayes act $\mathbf{x}^F$ for a probabilistic forecast F
   - If $\mathbf{Y} \sim F$, what action minimizes expected loss?
3. Score the forecast with the scoring rule $S(F, \mathbf{y}) = s(\mathbf{x}^F, \mathbf{y})$
   - The observed loss if you took the action suggested by F

Deriving scores from decision making foundations
Example 1

The decision making problem setup:

- How much of a protective measure should we purchase?
  - Example: we can purchase some number of doses of antiviral medications to give to people who get sick
  - $\mathbf{x}$ is the amount that we purchase (our task is to choose this)
  - Each unit that we buy has some cost C

Deriving scores from decision making foundations
Example 1

The decision making problem setup:

- How much of a protective measure should we purchase?
  - Example: we can purchase some number of doses of antiviral medications to give to people who get sick
  - $\mathbf{x}$ is the amount that we purchase (our task is to choose this)
  - Each unit that we buy has some cost C
- The outcome $\mathbf{y}$ is the realized need for the protective measure
  - Example: how many doses did we end up needing?

Deriving scores from decision making foundations
Example 1

The decision making problem setup:

- How much of a protective measure should we purchase?
  - Example: we can purchase some number of doses of antiviral medications to give to people who get sick
  - $\mathbf{x}$ is the amount that we purchase (our task is to choose this)
  - Each unit that we buy has some cost C
- The outcome $\mathbf{y}$ is the realized need for the protective measure
  - Example: how many doses did we end up needing?
- If we don't buy enough, we will incur a loss L for each unit of unmet need, for a total loss of $L \cdot (\mathbf{y} - \mathbf{x})$
  - Assume that this loss L is greater than the cost C for each dose

Deriving scores from decision making foundations
Example 1

The decision making problem setup:

- How much of a protective measure should we purchase?
  - Example: we can purchase some number of doses of antiviral medications to give to people who get sick
  - $\mathbf{x}$ is the amount that we purchase (our task is to choose this)
  - Each unit that we buy has some cost C
- The outcome $\mathbf{y}$ is the realized need for the protective measure
  - Example: how many doses did we end up needing?
- If we don't buy enough, we will incur a loss L for each unit of unmet need, for a total loss of $L \cdot (\mathbf{y} - \mathbf{x})$
  - Assume that this loss L is greater than the cost C for each dose
- $\mathbf{y}$ is unknown at the time of decision making, but a forecast F of $\mathbf{Y}$ is available.

Deriving scores from decision making foundations

Example 1

Deriving the score:

1. Define a loss function $s(\mathbf{x}, \mathbf{y})$
2. Get the Bayes act $\mathbf{x}^F$ for a probabilistic forecast F
3. Score the forecast with the scoring rule $S(F, \mathbf{y}) = s(\mathbf{x}^F, \mathbf{y})$

Deriving scores from decision making foundations

Example 1

Deriving the score:

1. Define a loss function $s(\mathbf{x}, \mathbf{y})$
2. Get the Bayes act $\mathbf{x}^F$ for a probabilistic forecast F
3. Score the forecast with the scoring rule $S(F, \mathbf{y}) = s(\mathbf{x}^F, \mathbf{y})$

1. Our loss function is $s_1(\mathbf{x}, \mathbf{y}) = C \cdot \mathbf{x} + L \cdot \max(\mathbf{y} - \mathbf{x}, 0)$
   - We purchase $\mathbf{x}$ units at cost C per unit
   - We incur losses of $L \cdot (\mathbf{y} - \mathbf{x})$ if $\mathbf{y} > \mathbf{x}$, and 0 otherwise

Deriving scores from decision making foundations

Example 1

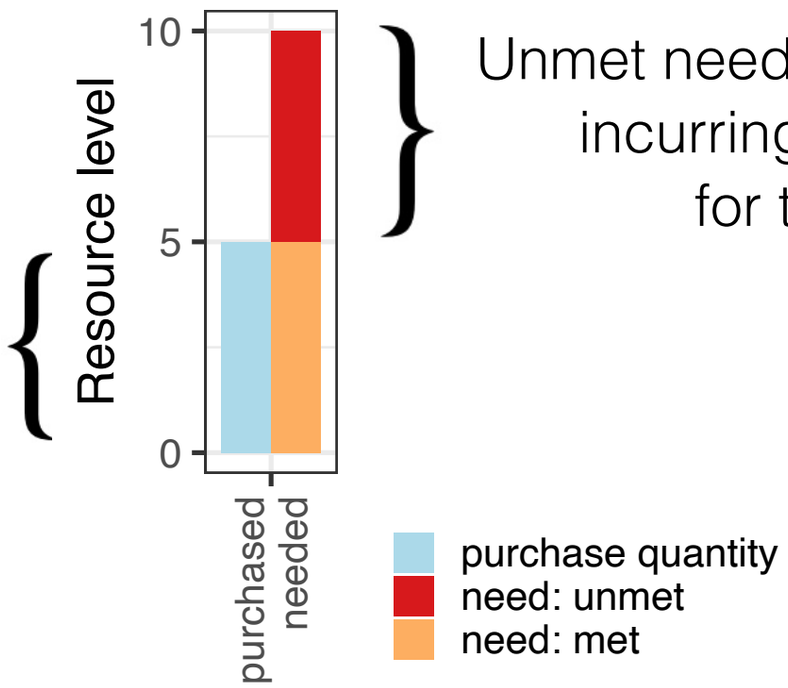
Deriving the score:

1. Define a loss function $s(\mathbf{x}, \mathbf{y})$
2. Get the Bayes act $\mathbf{x}^F$ for a probabilistic forecast F
3. Score the forecast with the scoring rule $S(F, \mathbf{y}) = s(\mathbf{x}^F, \mathbf{y})$

1. Our loss function is $s_1(\mathbf{x}, \mathbf{y}) = C \cdot \mathbf{x} + L \cdot \max(\mathbf{y} - \mathbf{x}, 0)$
   - We purchase $\mathbf{x}$ units at cost C per unit
   - We incur losses of $L \cdot (\mathbf{y} - \mathbf{x})$ if $\mathbf{y} > \mathbf{x}$, and 0 otherwise
   - Example: Suppose $C = 1$, $L = 4$, $\mathbf{x} = 5$, and $\mathbf{y} = 10$. The total loss is

$$s_1(5,10) = 1 \cdot 5 + 4 \cdot \max(10 - 5, 0) = 25$$

Purchased $\mathbf{x} = 5$ units at cost $C = 1$ each, for total costs of 5



Unmet need of $(10 - 5) = 5$ units incurring loss $L = 4$ each, for total loss of 20

Deriving scores from decision making foundations

Example 1

Deriving the score:

1. Define a loss function $s(\mathbf{x}, \mathbf{y})$
2. Get the Bayes act $\mathbf{x}^F$ for a probabilistic forecast F
3. Score the forecast with the scoring rule $S(F, \mathbf{y}) = s(\mathbf{x}^F, \mathbf{y})$

1. Our loss function is $s_1(\mathbf{x}, \mathbf{y}) = C \cdot \mathbf{x} + L \cdot \mathbf{max}(\mathbf{y} - \mathbf{x}, 0)$
2. The Bayes act is the amount of the protective measure that minimizes the expected loss according to the forecast distribution: $\mathbf{min}_x \mathbb{E}_F[s_1(x, Y)]$
   - Can show that $\mathbf{x}^F = F^{-1}(1 - C/L)$, a quantile of the forecast distribution at the probability level $\tau = 1 - C/L$

Deriving scores from decision making foundations

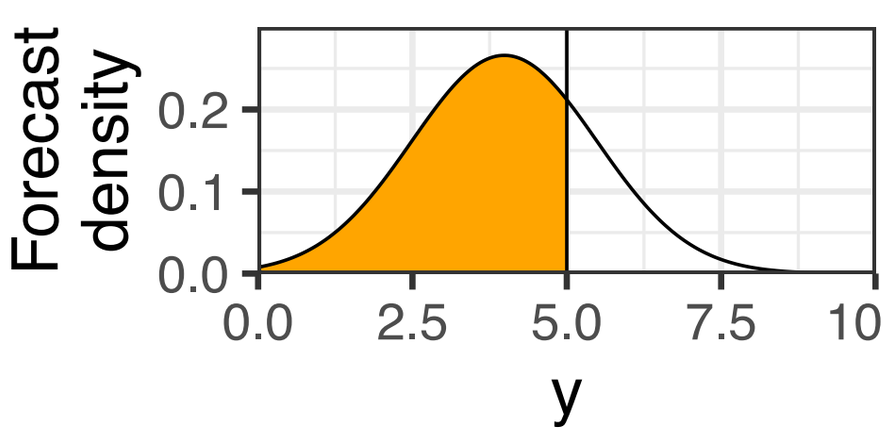
Example 1

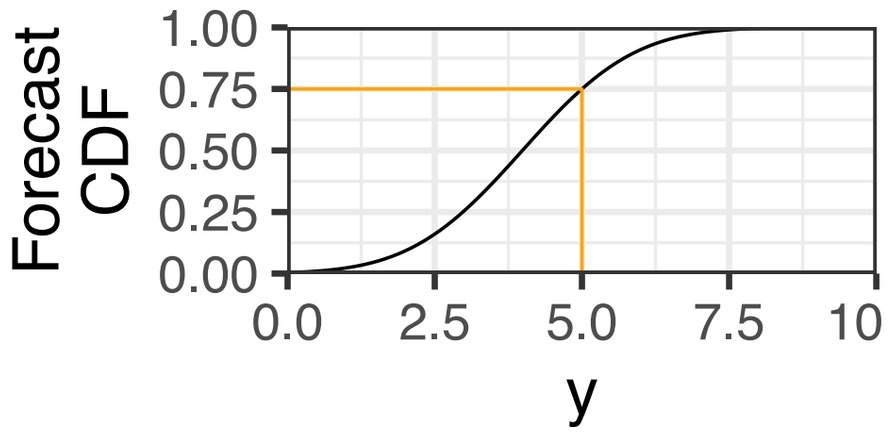
Deriving the score:

1. Define a loss function $s(\mathbf{x}, \mathbf{y})$
2. Get the Bayes act $\mathbf{x}^F$ for a probabilistic forecast F
3. Score the forecast with the scoring rule $S(F, \mathbf{y}) = s(\mathbf{x}^F, \mathbf{y})$

1. Our loss function is $s_1(\mathbf{x}, \mathbf{y}) = C \cdot \mathbf{x} + L \cdot \mathbf{max}(\mathbf{y} - \mathbf{x}, 0)$
2. The Bayes act is the amount of the protective measure that minimizes the expected loss according to the forecast distribution: $\mathbf{min}_x \mathbb{E}_F[s_1(x, Y)]$

- Can show that $\mathbf{x}^F = F^{-1}(1 - C/L)$, a quantile of the forecast distribution at the probability level $\tau = 1 - C/L$
- Example: with $C = 1$ and $L = 4$, we get $\tau = 1 - 1/4 = 0.75$
  - $\mathbf{x}^F$ is the 75th percentile of the forecast distribution
  - Suppose $F = Normal(3.988, 1.5^2)$. Then $x^F = 5$.





Deriving scores from decision making foundations

Example 1

Deriving the score:

1. Define a loss function $s(\mathbf{x}, \mathbf{y})$
2. Get the Bayes act $\mathbf{x}^F$ for a probabilistic forecast F
3. Score the forecast with the scoring rule $S(F, \mathbf{y}) = s(\mathbf{x}^F, \mathbf{y})$

1. Our loss function is $s_1(\mathbf{x}, \mathbf{y}) = C \cdot \mathbf{x} + L \cdot \max(\mathbf{y} - \mathbf{x}, 0)$
2. The Bayes act is $\mathbf{x}^F = F^{-1}(1 - C/L)$
3. Once $\mathbf{y}$ is observed, we score the forecast with

$$\begin{aligned} S_1(F, \mathbf{y}) &= s_1(\mathbf{x}^F, \mathbf{y}) \\ &= C \cdot F^{-1}(1 - C/L) + L \cdot \max(\mathbf{y} - F^{-1}(1 - C/L), 0) \end{aligned}$$

- Based on the purchase amount suggested by the forecast, what was the total purchase cost and losses due to unmet need?

Deriving scores from decision making foundations

Example 1

Deriving the score:

1. Define a loss function $s(\mathbf{x}, \mathbf{y})$
2. Get the Bayes act $\mathbf{x}^F$ for a probabilistic forecast F
3. Score the forecast with the scoring rule $S(F, \mathbf{y}) = s(\mathbf{x}^F, \mathbf{y})$

1. Our loss function is $s_1(\mathbf{x}, \mathbf{y}) = C \cdot \mathbf{x} + L \cdot \max(\mathbf{y} - \mathbf{x}, 0)$
2. The Bayes act is $\mathbf{x}^F = F^{-1}(1 - C/L)$
3. Once $\mathbf{y}$ is observed, we score the forecast with

$$S_1(F, \mathbf{y}) = s_1(\mathbf{x}^F, \mathbf{y})$$
$$= C \cdot F^{-1}(1 - C/L) + L \cdot \max(\mathbf{y} - F^{-1}(1 - C/L), 0)$$

- Based on the purchase amount suggested by the forecast, what was the total purchase cost and losses due to unmet need?
- Example: $C = 1, L = 4$
  - $F = Normal(3.988, 1.5^2)$. Then $x^F = 5$ doses would be purchased.
  - We observe a need for $\mathbf{y} = 10$ doses
  - The forecast's score is $s_1(\mathbf{x}^F, \mathbf{y}) = 1 \cdot 5 + 4 \cdot \max(10 - 5, 0) = 25$

Deriving scores from decision making foundations

Example 1

Recapping:

1. Our loss function is $s_1(\mathbf{x}, \mathbf{y}) = C \cdot \mathbf{x} + L \cdot \mathbf{max}(\mathbf{y} - \mathbf{x}, 0)$
   - Total purchase costs and losses due to unmet need
2. The Bayes act is $\mathbf{x}^F = F^{-1}(1 - C/L)$
   - A quantile of the forecast distribution
3. Once $\mathbf{y}$ is observed, we score the forecast with the loss that would have occurred if we had followed the forecast's recommendation:

$$
\begin{aligned}
S_1(F, \mathbf{y}) &= s_1(\mathbf{x}^F, \mathbf{y}) \\
&= C \cdot F^{-1}(1 - C/L) + L \cdot \max(\mathbf{y} - F^{-1}(1 - C/L), 0)
\end{aligned}
$$

Question:

Are you comfortable using a forecaster's performance on this score to inform:

- Whether to use their predictions in decisions about resource purchases?
- Whether to use their predictions to support other decisions?
- A general understanding of whether that forecaster is "good"?

Deriving scores from decision making foundations

Example 2

The decision making problem setup:

- We must decide how much of a resource to purchase
- We know that a loss of $L = 1000$ will be incurred for each unit of unmet need
- But we don't exactly know the resource cost per unit — it will be one of the following 7 values, with equal probability:

index, j	C_j	L	$\tau_j = 1 - C_j/L$
1	975	1000	0.025
2	900	1000	0.100
3	750	1000	0.250
4	500	1000	0.500
5	250	1000	0.750
6	100	1000	0.900
7	25	1000	0.975

Deriving scores from decision making foundations

Example 2

The decision making problem setup:

- We must decide how much of a resource to purchase
- We know that a loss of $L = 1000$ will be incurred for each unit of unmet need
- But we don't exactly know the resource cost per unit — it will be one of the following 7 values, with equal probability:

index, j	C_j	L	$\tau_j = 1 - C_j/L$
1	975	1000	0.025
2	900	1000	0.100
3	750	1000	0.250
4	500	1000	0.500
5	250	1000	0.750
6	100	1000	0.900
7	25	1000	0.975

Score for evaluation: average the scores (from Ex. 1) for each cost level:

$$S_2(F, \mathbf{y}) = \frac{1}{7} \sum_j S_1(F, \mathbf{y}; C_j, L)$$

Deriving scores from decision making foundations
Example 2

Recapping:

- A decision must be made about how much of a resource to purchase
- 7 possible costs per unit resource purchased, equally likely
  - (Note, I think you can also motivate this by averaging across 7 “decision making scenarios” with different costs)
- Score for evaluation: average the scores (from Ex. 1) for each cost level:

$$S_2(F, \mathbf{y}) = \frac{1}{7} \sum_k S_1(F, \mathbf{y}; C_k, L)$$

Question:

Are you comfortable using a forecaster’s performance on this score to inform:

- Whether to use their predictions in decisions about resource purchases?
- Whether to use their predictions to support other decisions?
- A general understanding of whether that forecaster is “good”?

Deriving scores from decision making foundations

Example 3

The decision making problem setup:

- How should we allocate K units of resources across n locations?
  - e.g., there are approximately $K = 15{,}000$ ventilators in the US
  - How could they be allocated to n healthcare facilities or states?
- $\mathbf{x} = (x_1, \ldots, x_n)$: Amount of resources allocated to each location
  - Require that $x_i \geq 0, \sum x_i = K$
- $\mathbf{y} = (y_1, \ldots, y_n)$: Resource need in each location (observed)
  - e.g., the observed number of patients in need of a ventilator in each facility or state
- $\mathbf{Y} = (Y_1, \ldots, Y_n)$: Resource need in each location (future, random)
- $\mathbf{F} = (F_1, \ldots, F_n)$: Forecasts of resource need in each location

Deriving scores from decision making foundations

Example 3

Deriving the score:

1. Define a loss function $s(\mathbf{x}, \mathbf{y})$
2. Get the Bayes act $\mathbf{x}^F$ for a probabilistic forecast F
3. Score the forecast with the scoring rule $S(F, \mathbf{y}) = s(\mathbf{x}^F, \mathbf{y})$

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

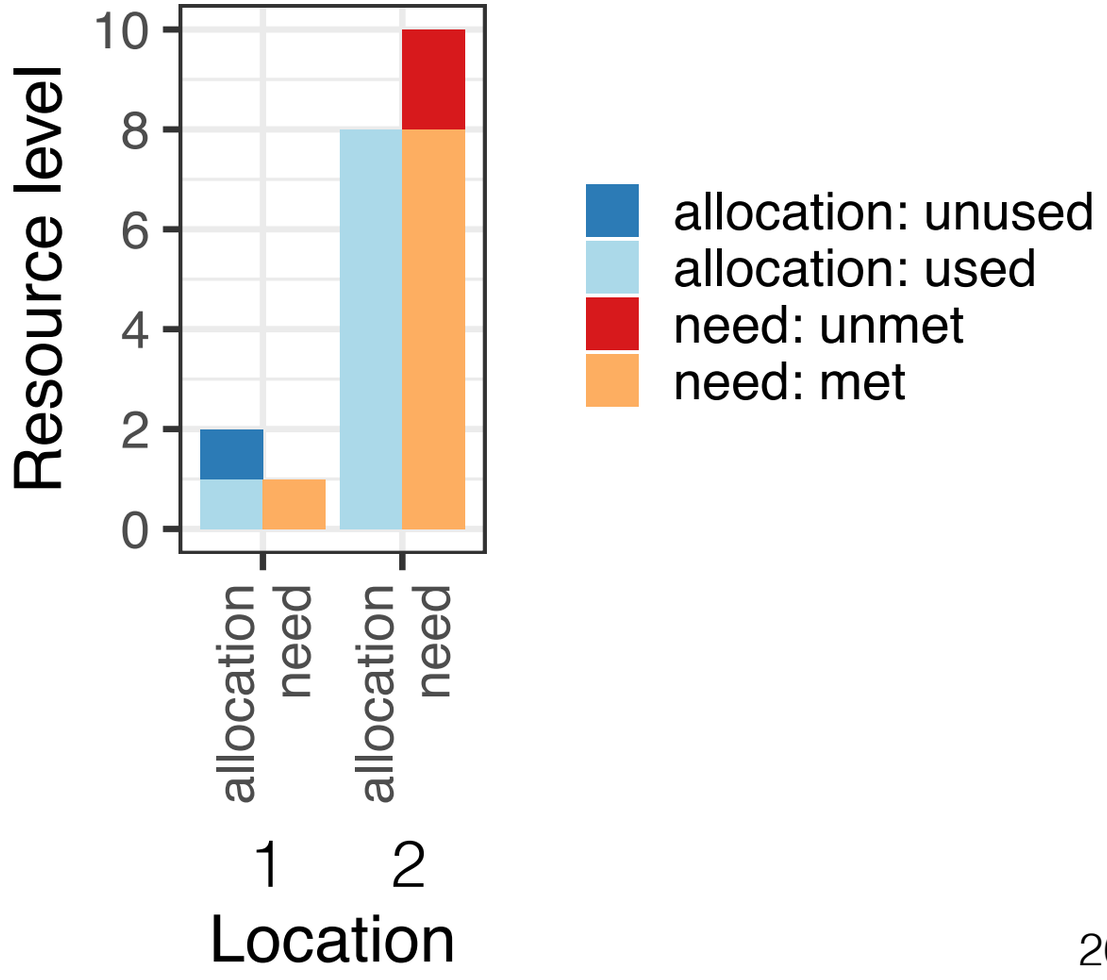
The allocation score — recipe step 1

- 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 $s(\mathbf{x}, \mathbf{y}) = 0 + (10 - 8) = 2$



The allocation score — recipe step 2

- Step 2: Get the Bayes act $\mathbf{x}^F$ for a probabilistic forecast F
  - If $\mathbf{Y} \sim F$, what action minimizes expected loss?

- With our loss, Bayes act is the allocation $\mathbf{x}$ that minimizes

$$\mathbb{E}_F[s(\mathbf{x}, \mathbf{Y})] = \sum \mathbb{E}_{F_i}[\max(Y_i - x_i, 0)]$$

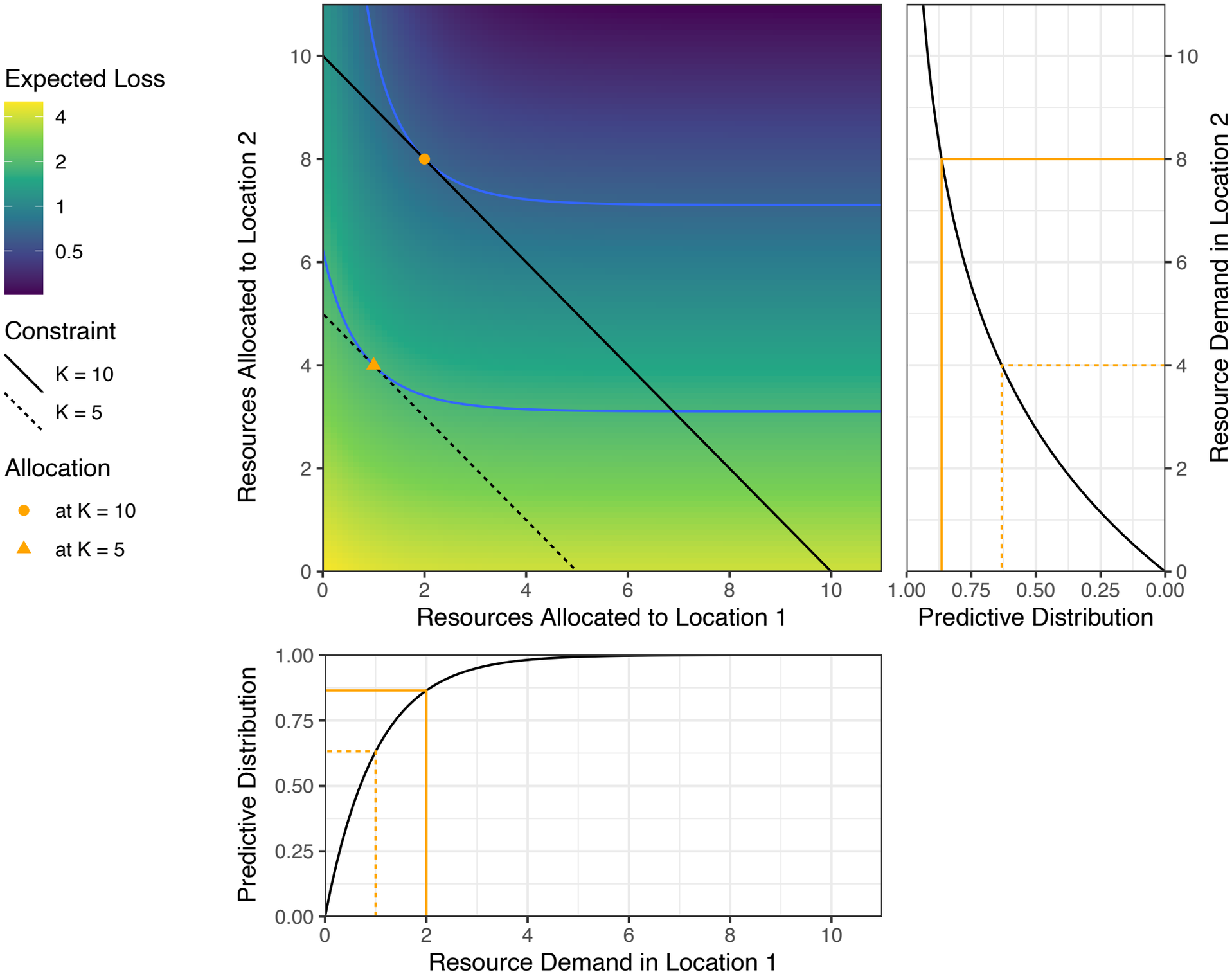
  - According to the forecast, what allocation results in the smallest expected total resource shortage?

- The Bayes act corresponds to a vector of quantiles from the marginal distributions at a quantile level τ that's shared for all locations:

$$x_i^F = F_i^{-1}(\tau) \text{ for some } \tau \in [0,1] \text{ such that } \sum F_i^{-1}(\tau) = K$$

The allocation score — recipe step 2

- Illustration: 2 locations with forecasts $F_1 = Exp(1/1)$ and $F_2 = Exp(1/4)$
- Allocations are proportional to the scale parameters 1 and 4



The allocation score — recipe step 3

- Step 3: Score the forecast with the scoring rule $S(F, \mathbf{y}) = s(\mathbf{x}^F, \mathbf{y})$
  - The observed loss if you took the action suggested by F

- Our "raw" allocation score is

$$S^{raw}(F, \mathbf{y}) = \sum \max(y_i - x_i^F, 0)$$

  - If you used the allocation suggested by F, what total unmet need would have resulted?

Regret and an Oracle adjustment

- Note: If the resource constraint is much less than resource needs,
  - any allocation will result in a lot of unmet need
  - any forecast will be assigned a high raw allocation score
- We'd like to know: could we have done better than the allocation suggested by F?
- We define the allocation score to be the "avoidable unmet need" that results from using the allocation suggested by F
  - How much better could we have done with another allocation than the one suggested by F?

$$S_3(F, \mathbf{y}) = S^{raw}(F, \mathbf{y}) - S^{raw}(F^{Oracle}, \mathbf{y})$$

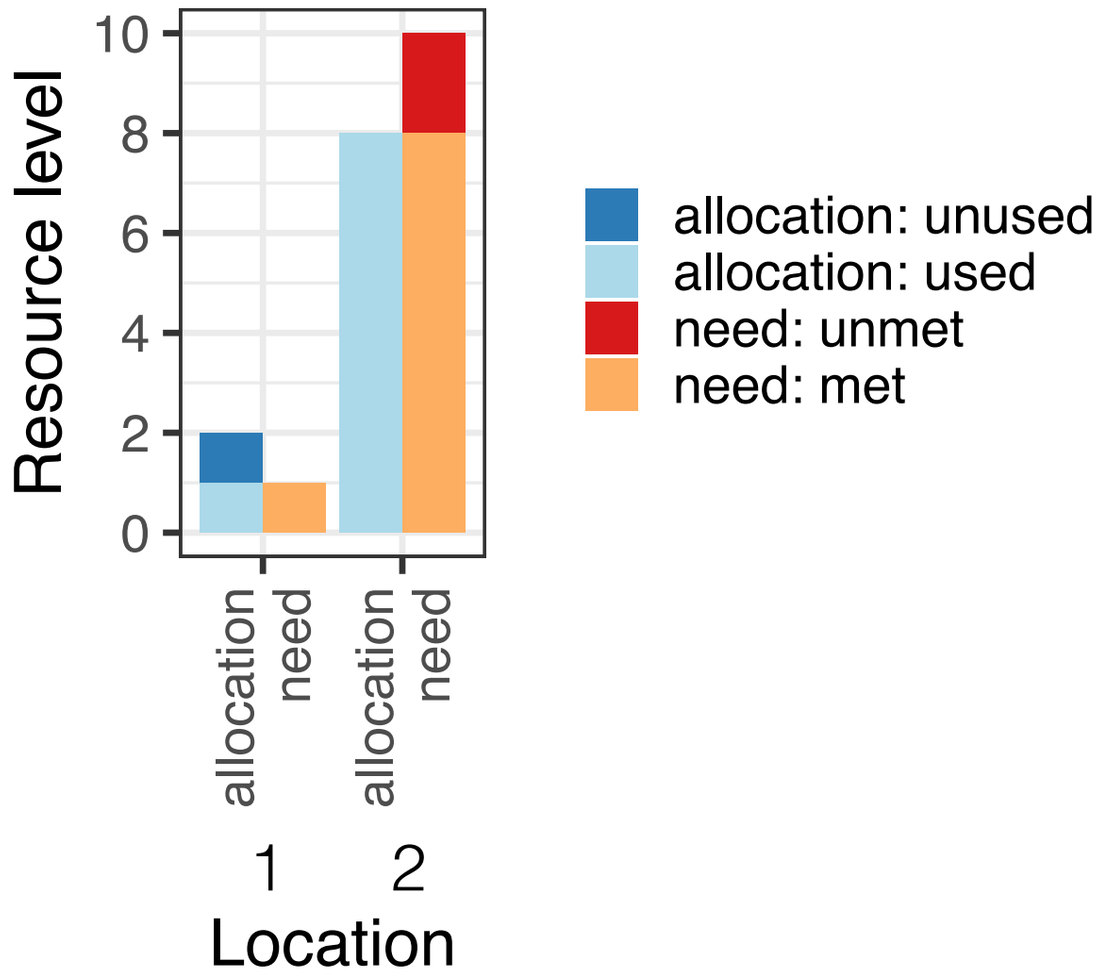
- F^{Oracle} represents an Oracle forecast that knows $\mathbf{y}$ in advance
- $S^{raw}(F^{Oracle}, \mathbf{y})$ is the "unavoidable unmet need"
- $S_3(F, \mathbf{y})$ is the "avoidable unmet need" that results from using allocations suggested by F

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



Deriving scores from decision making foundations

Example 3

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?

Question:

Are you comfortable using a forecaster's performance on this score to inform:

- Whether to use their predictions in decisions about resource allocations?
- Whether to use their predictions to support other decisions?
- A general understanding of whether that forecaster is "good"?

Deriving scores from decision making foundations
Wrapping Up

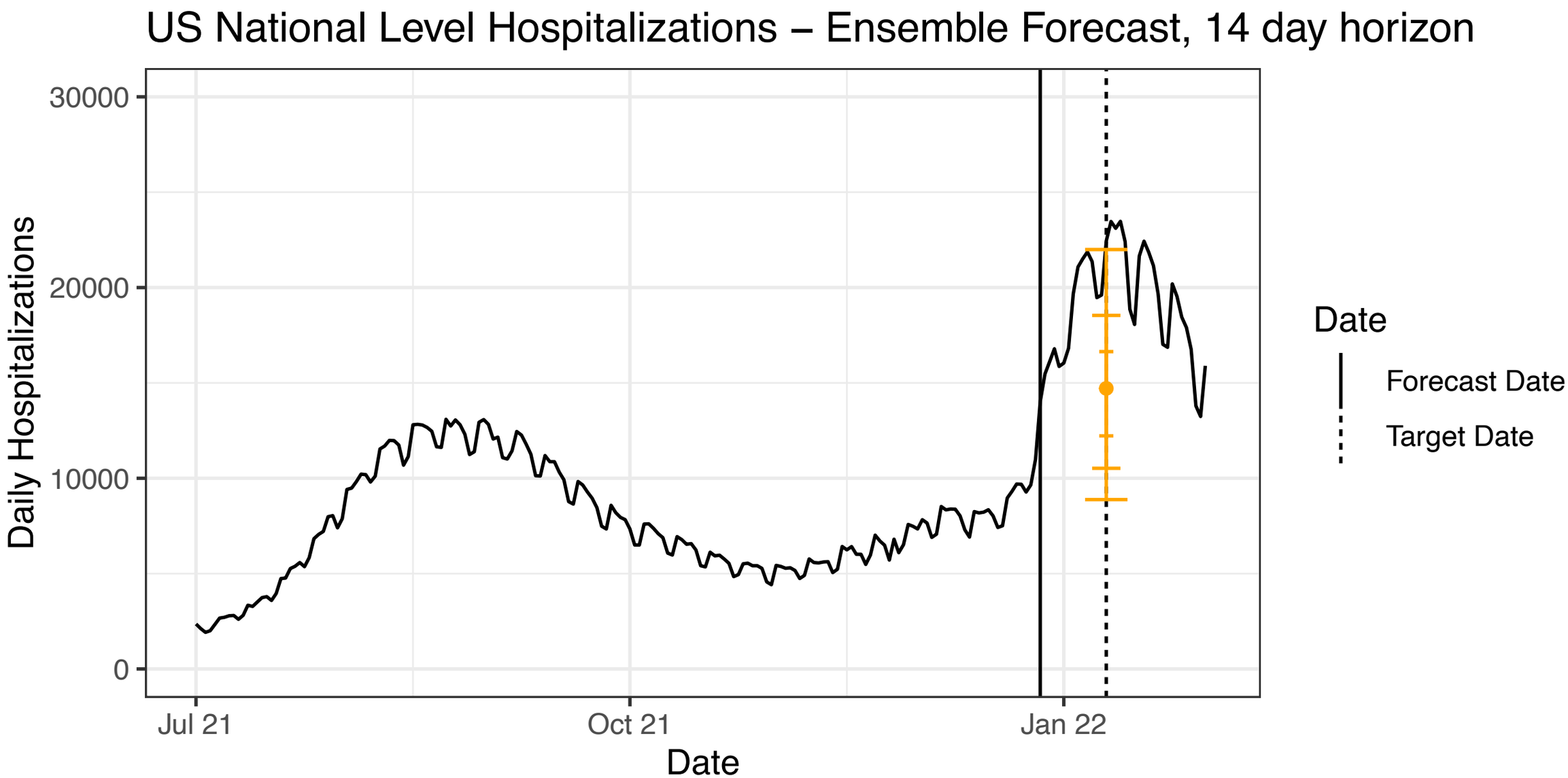
Decision Context	Score interpretation	Score name
• **x**: amount of a resource to buy • cost **C** per unit purchased • loss **L** per unit unmet need	Total costs and losses due to unmet need if you purchased the amount suggested by forecast	**Pinball loss** (Up to additive/ multiplicative constants)
• **x**: amount of a resource to buy • unknown cost **C** per unit purchased, 7 possible values • loss **L** per unit unmet need	Costs/losses if you purchased the amount suggested by the forecast, averaged across 7 possible costs	**WIS** (Up to additive/ multiplicative constants)
• **x**: resource allocation • Total of **K** units of resources available	Avoidable unmet need if you used the allocation suggested by the forecast	**Allocation score**

Overview of this talk

- Introduction and motivation
- A sequence of 3 example scores
- **Illustrative application**
- Conclusions

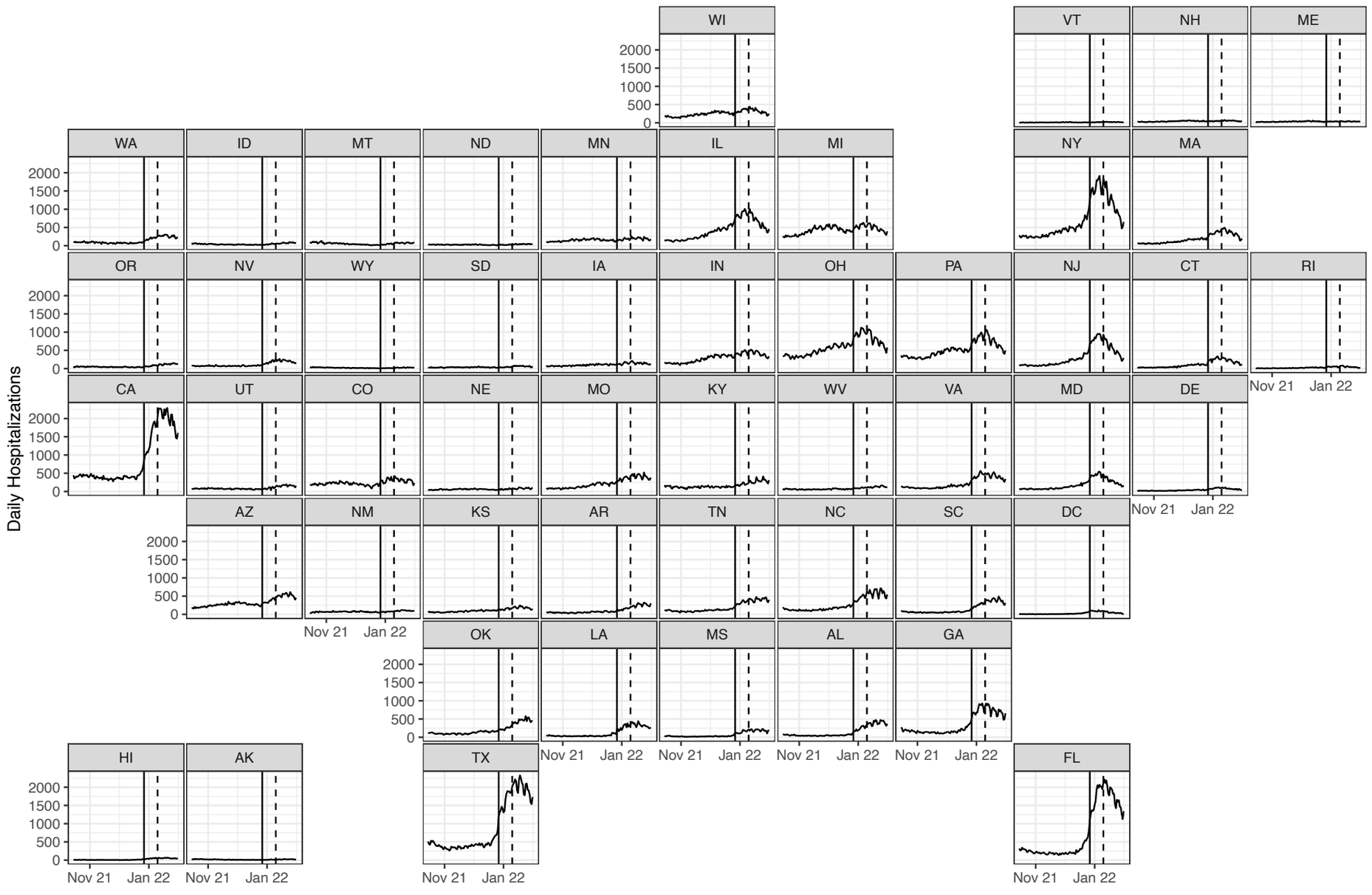
A “simple” example

- Allocation of federal resources to the US states heading into the Omicron hospitalizations wave



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, roughly the number of ventilators in the US (noting that we are working with forecasts of hospital admissions, not patients in need of a ventilator)



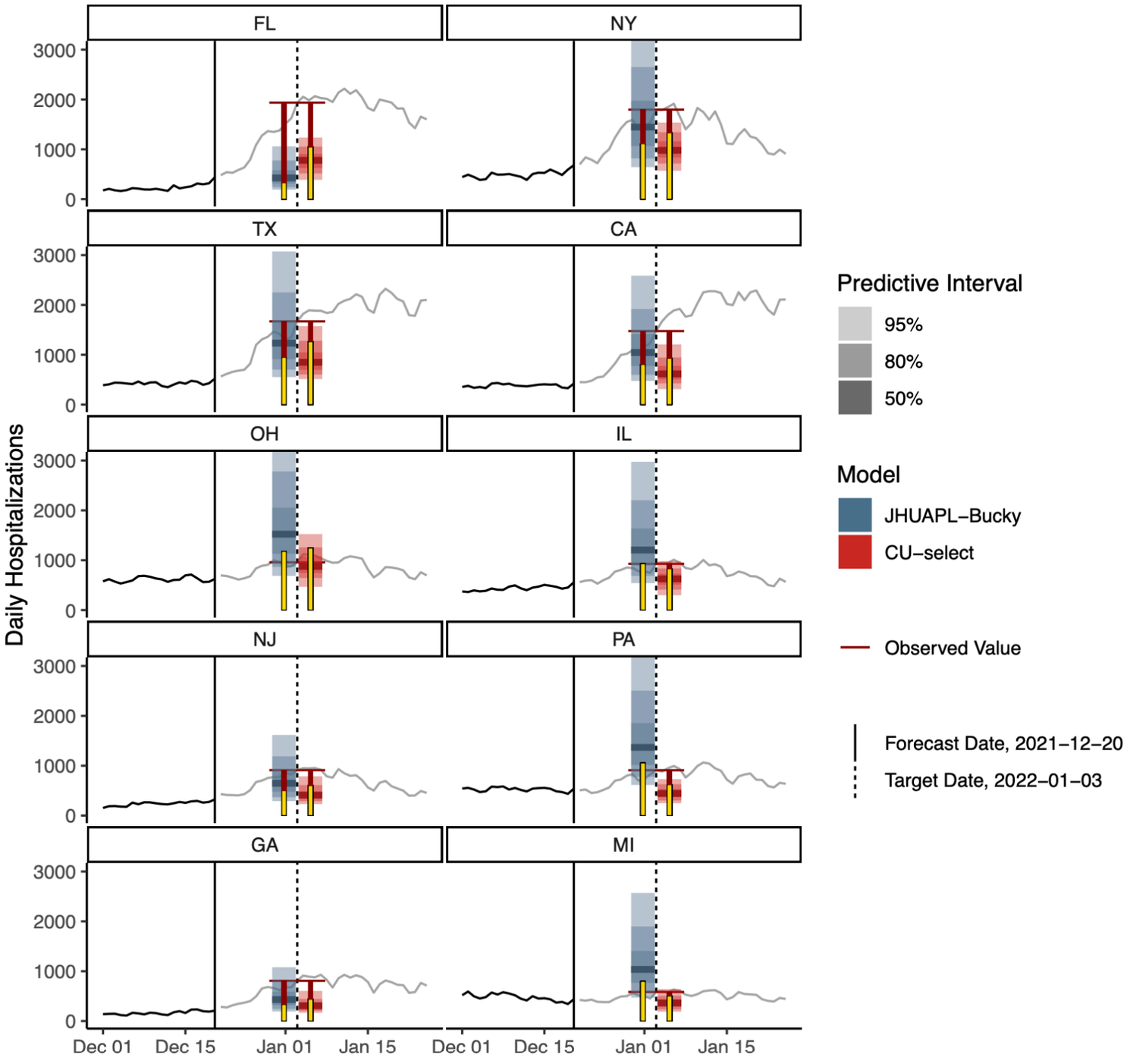
Models ranks by WIS and allocation score differ

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

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
- CU-select: Best allocation score, middling WIS
- Overall, JHUAPL-Bucky better captures level in individual locations
- But JHUAPL-Bucky under predicts FL by more and over predicts in several states (OH, IL, PA, MI) resulting in misused resources.
- CU-select under predicts in a consistent way across locations, getting relative resource need about right



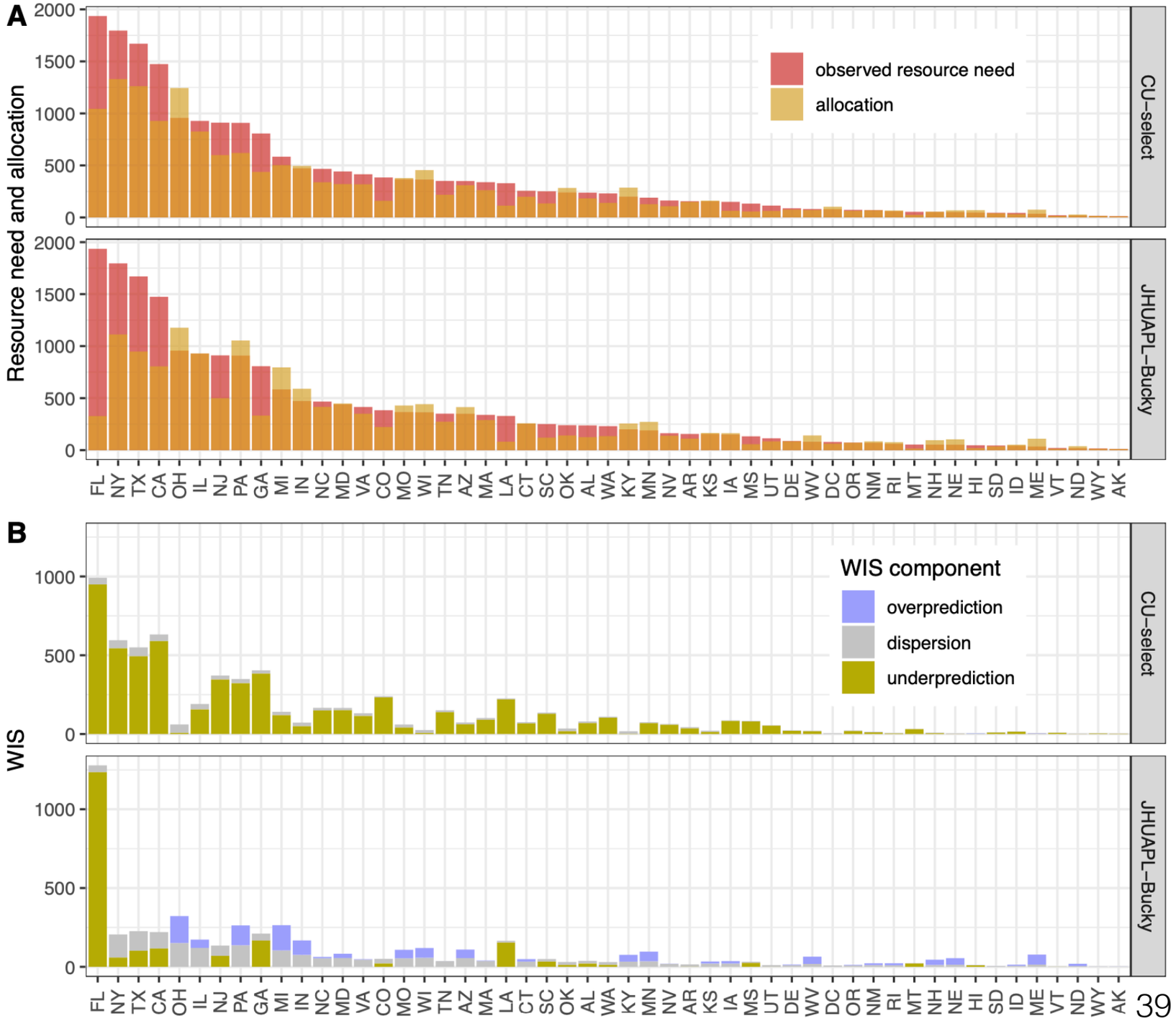
A closer look at JHUAPL-Bucky and CU-select

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

- Overall, JHUAPL-Bucky better captures level in individual locations

- But JHUAPL-Bucky under predicts FL by more and over predicts in several states, resulting in misused resources.

- CU-select under predicts in a consistent way across locations, getting relative resource need about right



Overview of this talk

- Introduction and motivation
- A sequence of 3 example scores
- Illustrative application
- **Conclusions**

Limitations, future work

- In practice, decision makers use many inputs alongside model-based predictions to inform decisions
- We don't measure value-added of forecasts to an existing decision-making process
- In many (most?) instances, it's challenging to quantify the loss associated with a decision
- We do not account for important considerations such as equity/ fairness of allocations
- We do not account for other broader elements of the decision-making context, such as the balance of multiple mitigation measures, increasing the resource constraint K, etc.
- It would be valuable to consider other decision-making contexts

How can we justify a choice of scoring rule?

How can we justify a choice of scoring rule?

1. **Choose something convenient**
   - "We're collecting forecasts in quantile format, so we'll use WIS"
   - "We're collecting forecast trajectories (as samples from a multivariate distribution) so we'll use energy score"

How can we justify a choice of scoring rule?

1. **Choose something convenient**
   - "We're collecting forecasts in quantile format, so we'll use WIS"
   - "We're collecting forecast trajectories (as samples from a multivariate distribution) so we'll use energy score"

2. **Make a (semi-formal?) argument that you like the properties of a score**
   - "CRPS/WIS rewards forecasts that are 'close to' the observed data, which is what we think epidemiologists care about"
   - "CRPS/WIS on a log scale measures skill at predicting local growth rates and is scale invariant, properties that we think epidemiologists care about"
   - "Log score penalizes forecasts where the observed value is far in the tail, and we think this is important"

How can we justify a choice of scoring rule?

1. **Choose something convenient**
   - "We're collecting forecasts in quantile format, so we'll use WIS"
   - "We're collecting forecast trajectories (as samples from a multivariate distribution) so we'll use energy score"

2. **Make a (semi-formal?) argument that you like the properties of a score**
   - "CRPS/WIS rewards forecasts that are 'close to' the observed data, which is what we think epidemiologists care about"
   - "CRPS/WIS on a log scale measures skill at predicting local growth rates and is scale invariant, properties that we think epidemiologists care about"
   - "Log score penalizes forecasts where the observed value is far in the tail, and we think this is important"

3. **Start from a decision making context, derive scores specific to that setting:**
   - Different settings lead to different scores
   - Not all decision making settings are amenable to this approach (e.g., how to quantify the loss due to an inaccurate public communication??)
   - Food for thought: if you didn't like some aspect of scores we discussed today — could this be addressed by being more careful about quantifying the loss associated with a particular action?

Thanks!

Acknowledgments to co-authors:
Aaron Gerding, Nicholas G. Reich, Ben Rogers

Acknowledgments to others who have offered input:
Matt Biggerstaff, Rebecca Borchering, Rosa Ergas, Melissa Kerr, Jeff Shaman

Allocation scores vary most when the constraint K is similar to total need

- Note: raw allocation scores for different forecasts will tend to be similar if the resource constraint is either very large or very small:

Resource constraint size	Locations receive…	Raw allocation scores are…
K is very large	Many resources	Small (locations get enough resources)
K is very small	Few resources	Large (locations get insufficient resources)

- In either of these extremes,
  - the Oracle can't improve much on allocations suggested by another forecast
  - So (Oracle-adjusted) allocation scores tend to be small
- On the other hand, if we have just enough resources, we'd better send them to the right places!

Working with forecasts in quantile format

- In the application to Hub forecasts, we infer distributions from quantiles using
  - a monotonic spline to interpolate the quantiles
  - parametric assumptions about tail behavior
- Our current thinking (to be thought through carefully): The score is still proper, but we believe that the “quantiles” elicited by this process are not quantiles.
- More thought would be needed to either adjust the score or the forecast representation for use as an official scoring metric for a Hub.