COVID-19 Forecast Similarity Analysis

Johannes Bracher, Evan Ray, Nick Reich, Nutcha Wattanachit

08/02/2021

Cramér distance

Consider two predictive distributions F and G. Their *Cramér distance* or *integrated quadratic distance* is defined as

$$\begin{aligned} \text{CD}(F, G) &= \int_{-\infty}^{\infty} (F(x) - G(x))^2 dx && (1) \\ &= \mathbb{E}_{F,G}|X - Y| - 0.5\left[\mathbb{E}_F|X - X'| + \mathbb{E}_G|Y - Y'|\right], && (2) \end{aligned}$$

where $F(x)$ and $G(x)$ denote the cumulative distribution functions, X, X' are independent random variables following F, and Y, Y' are independent random variables following G. This formulation in (2) illustrates that the Cramér distance depends on the shift between F and G (first term) and the variability of both F and G (of which the two last expectations in above equation are a measure).

The Cramér distance is the divergence associated with the continuous ranked probability score (Thorarinsdottir 2013, Gneiting and Raftery 2007), which is defined by

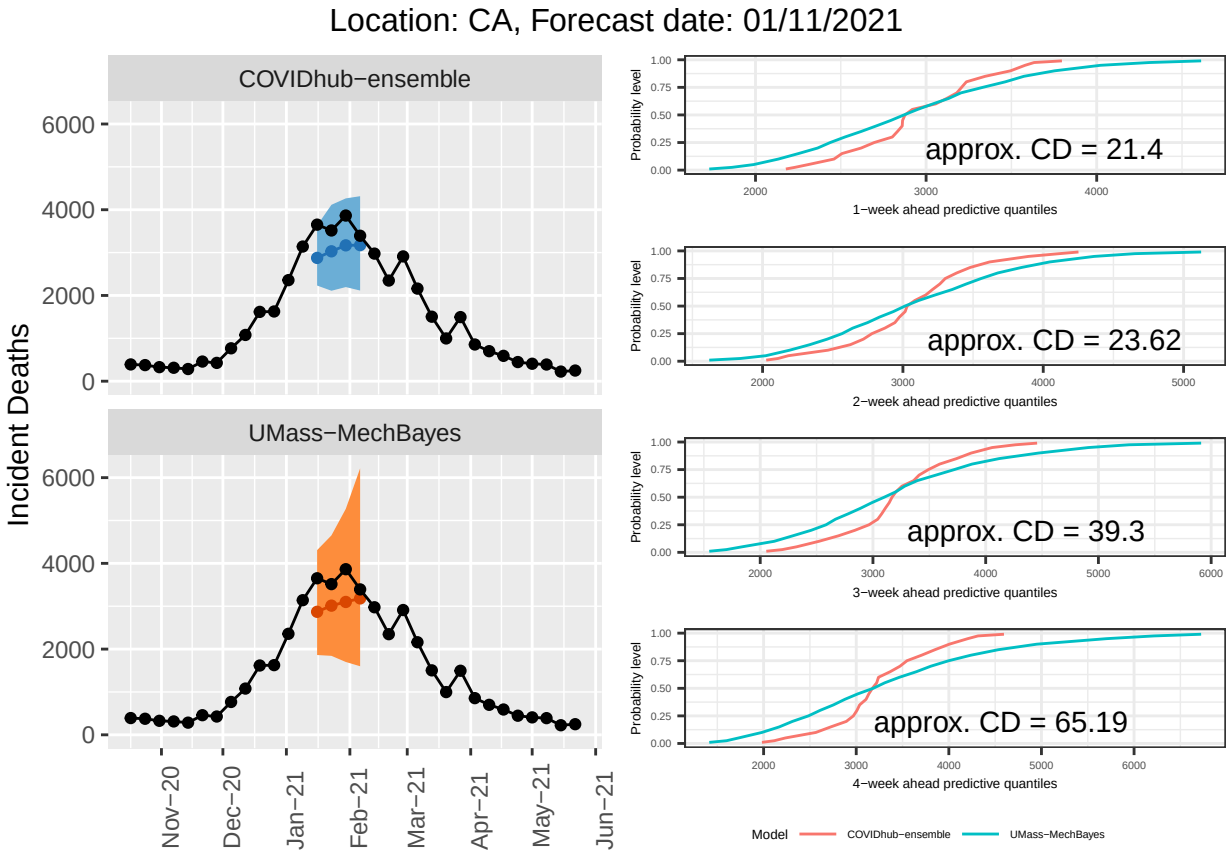
$$\begin{aligned} \text{CRPS}(F, y) &= \int_{-\infty}^{\infty} (F(x) - \mathbf{1}(x \geq y))^2 dx && (3) \\ &= 2 \int_0^1 \{(\mathbf{1}(y \leq q_k^F) - \tau_k\}(q_k^F - y) d\tau_k && (4) \end{aligned}$$

Forecast inclusion criteria

- Models: All models with complete submissions for the following criteria
- Targets: 1-4 wk ahead inc death and inc case
- Target end dates: Oct 19th, 2020 - May 24th,2021
- Probability levels: All
- Locations:
  - 5 states with highest cumulative deaths by February 27th, 2021: CA, FL, NY, PA, TX
  - 5 states with highest cumulative cases by February 27th, 2021: CA, FL, IL, NY, TX
  - 5 states with lowest cumulative deaths by February 27th, 2021: AK, HI, ME, VT, WY
  - 5 states with lowest cumulative cases by February 27th, 2021: DC, HI, ME, VT, WY

1-4 Week Ahead Incident Death Forecasts

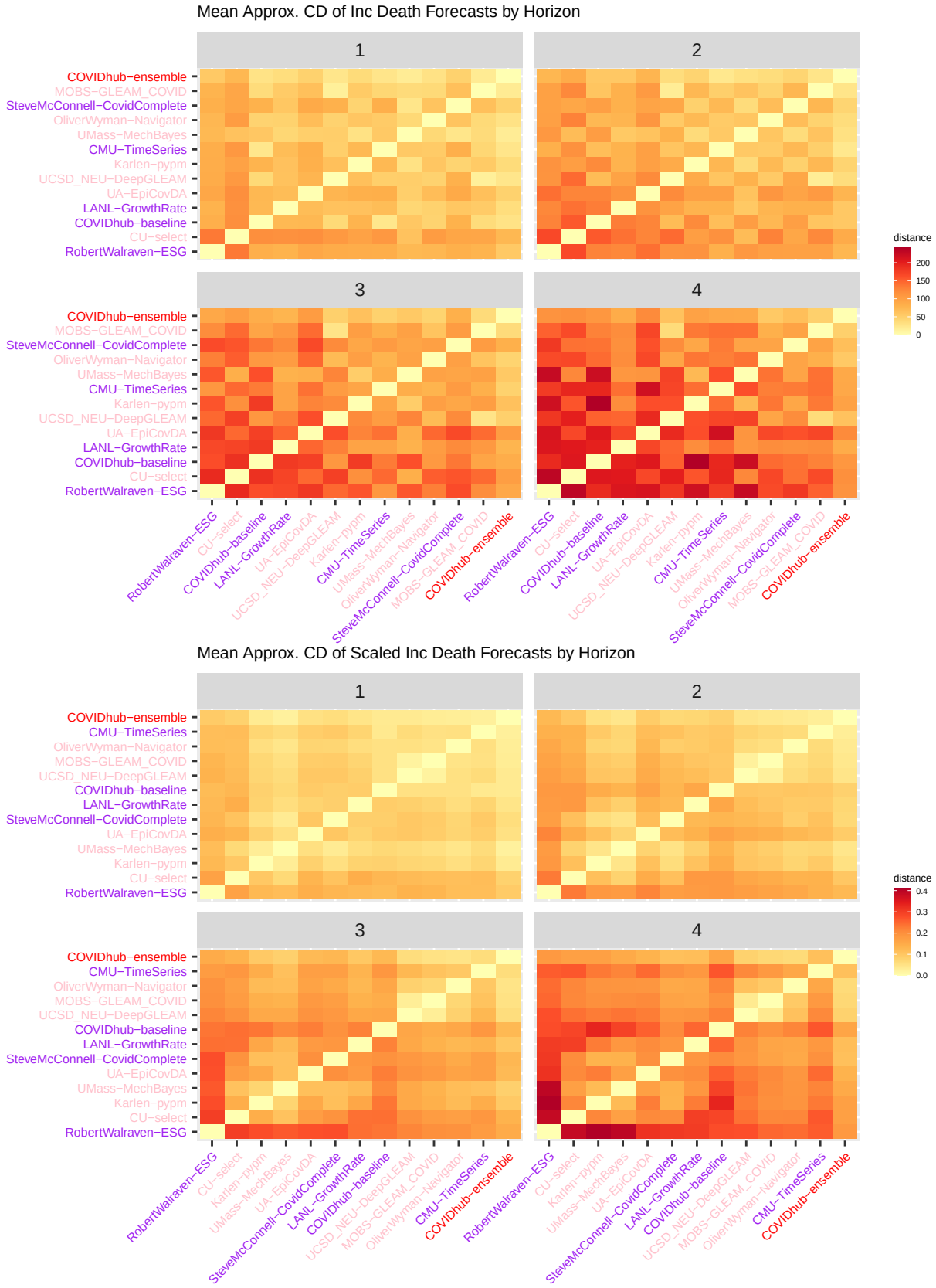
There are 13 models that fulfilled the criteria for the 5 locations with highest cumulative deaths. Below is an example of approximated Cramér distances between COVIDhub-ensemble and UMass-Mechbayes for 1-4 week ahead forecasts of incident deaths in CA in the week of 01/11/2021. It is evident from the approx. CD and the 95% prediction interval that Mechbayes forecasts for this peak week are more dispersive compared to the ensemble forecasts (not surprising considering "stacking" lacks dispersion?).



Mean approximated pairwise distances across 5 high count locations

We can visualize the mean approximated pairwise distances across all weeks and locations in heatmaps. The distance from the model to itself is zero. The $x-$axis is arranged based in an ascending order of the model's approximate pairwise distance from the COVIDhub-ensemble. So, the first model is the model that is most dissimilar (on average) to the ensemble in this time frame.

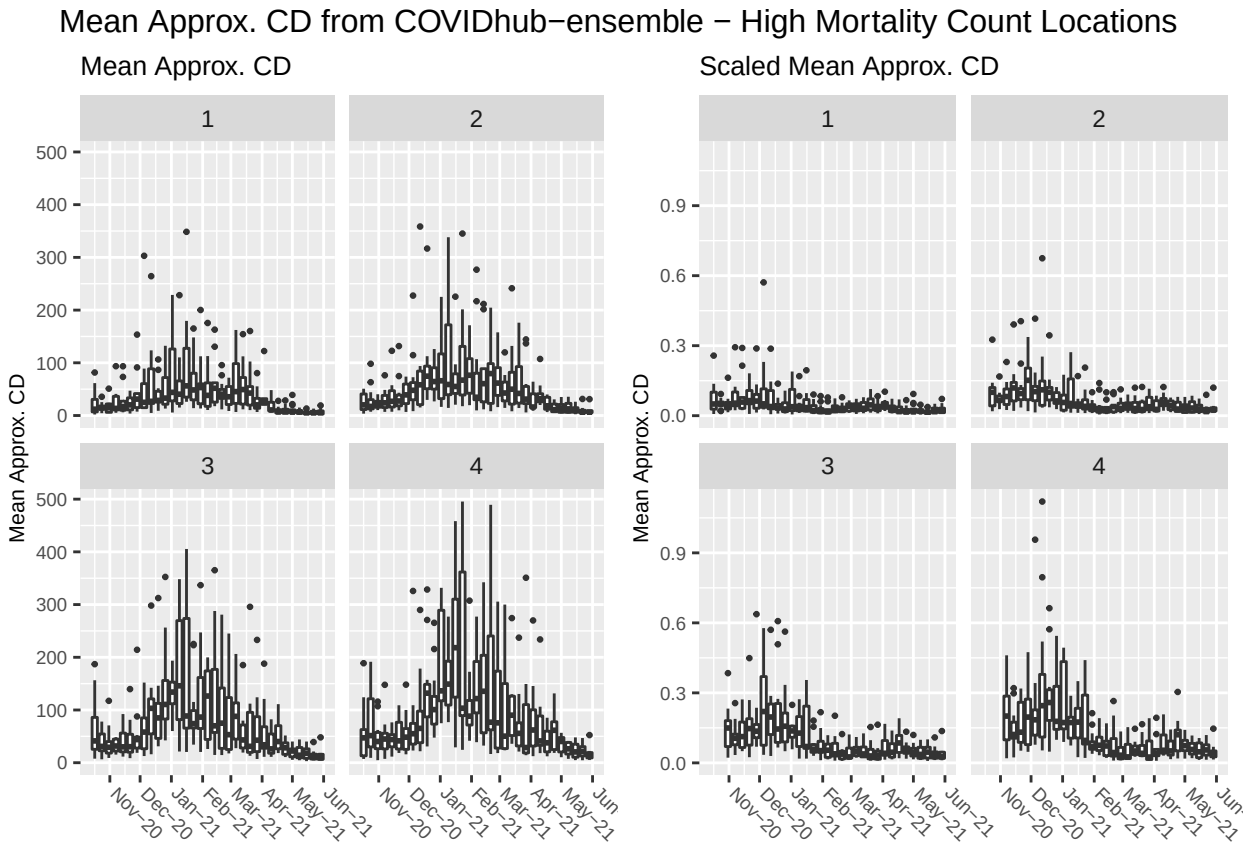
For high morality count locations, the distances between pairs of forecasts are higher for further forecast horizons. CU-select and RobertWalraven-ESG forecasts for high count locations seem to be more dissimilar to other models on average across all forecast horizons. After scaling by the truth at $t-1$ from the forecast week, some distances get "smoothed out", but the observed trends persist.


Mean Approx. CD of Inc Death Forecasts by Horizon
1
2
3
4
COVIDhub−ensemble
MOBS−GLEAM_COVID
SteveMcConnell−CovidComplete
OliverWyman−Navigator
UMass−MechBayes
CMU−TimeSeries
Karlen−pypm
UCSD_NEU−DeepGLEAM
UA−EpiCovDA
LANL−GrowthRate
COVIDhub−baseline
CU−select
RobertWalraven−ESG
distance
200
150
100
50
0
Mean Approx. CD of Scaled Inc Death Forecasts by Horizon
1
2
3
4
COVIDhub−ensemble
CMU−TimeSeries
OliverWyman−Navigator
MOBS−GLEAM_COVID
UCSD_NEU−DeepGLEAM
COVIDhub−baseline
LANL−GrowthRate
SteveMcConnell−CovidComplete
UA−EpiCovDA
UMass−MechBayes
Karlen−pypm
CU−select
RobertWalraven−ESG
distance
0.4
0.3
0.2
0.1
0.0

Approximated pairwise distances over time, across 5 high count locations

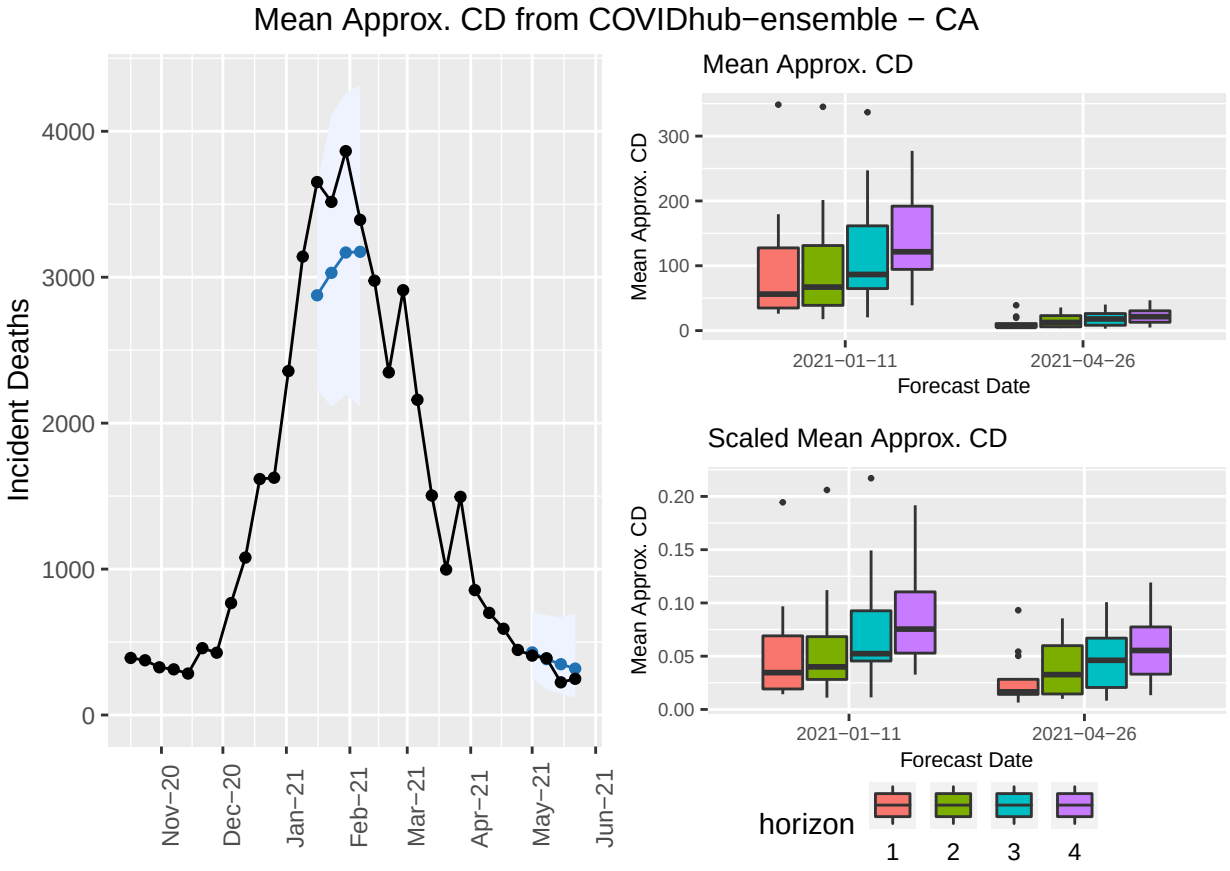
Dissimilarity to COVIDhub-ensemble

We can also look at the mean approximated pairwise distances across locations only to see how the models become more similar or dissimilar over time. Across all horizons, we see a large range of distances from the ensemble forecasts around the second week of January 2021. After scaling, the dissimilarity is attenuated. Specifically, the peak of dissimilarity is shifted to early/mid December across all horizons. The large range of dissimilarity is also observed for about a month after for the 4 week ahead horizon. This might be due to the higher uncertainty exhibited in longer horizon forecasts as we were about to observe the peak month?



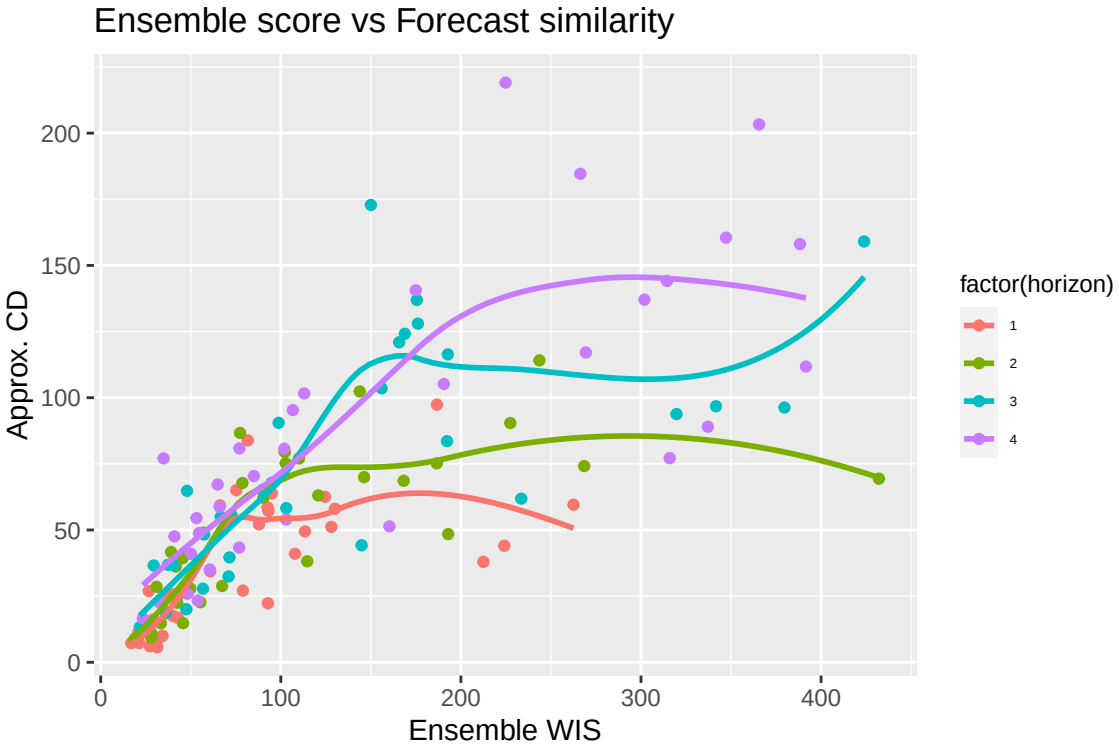
The week of high incident deaths and forecast similarities to the COVIDhub-ensemble forecasts for CA

This boxplot shows that we see more dissimilarity from the ensemble forecasts during the week that the actual death number is high compared to the week that we observed lower deaths. I assume this is probably the case for all high count locations, but I have yet to generate the plots for the combined locations. However, this is attenuated after scaling. The median CDs are more similar after scaling, but the range of scaled CDs in the week with higher deaths are still larger. We might want to check the standard deviation here?



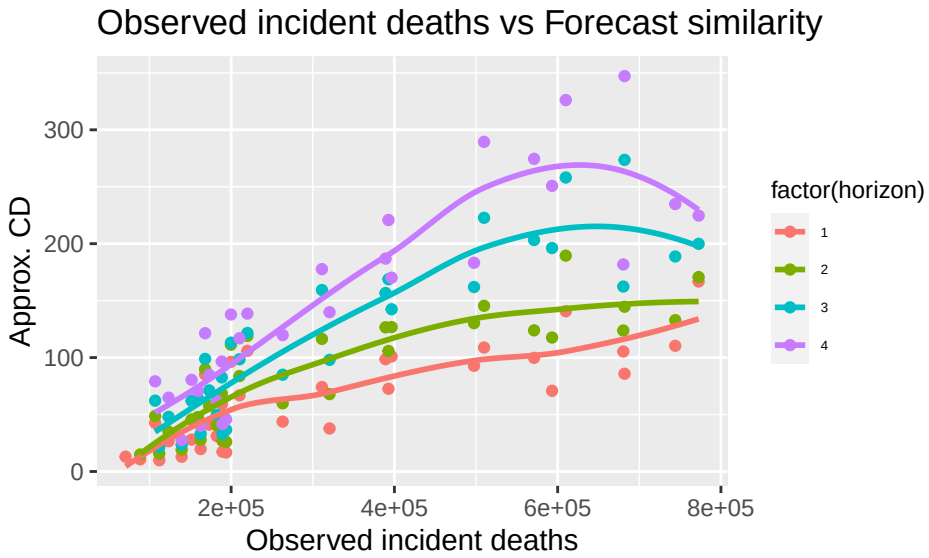
Disagreement among models and ensemble scores

The y-axis is the approximated Cramer distances from the ensemble for each target end dates in the analysis. Maybe it's better to plot Cramer distances every pairs and not just the model to the ensemble? I see something like "a higher disagreement is associated with a higher ensemble score"?. If forecasts are dissimilar, the ensemble forecasts tend to be less accurate? But both x- and y-axis are affected by the scale of observed values...

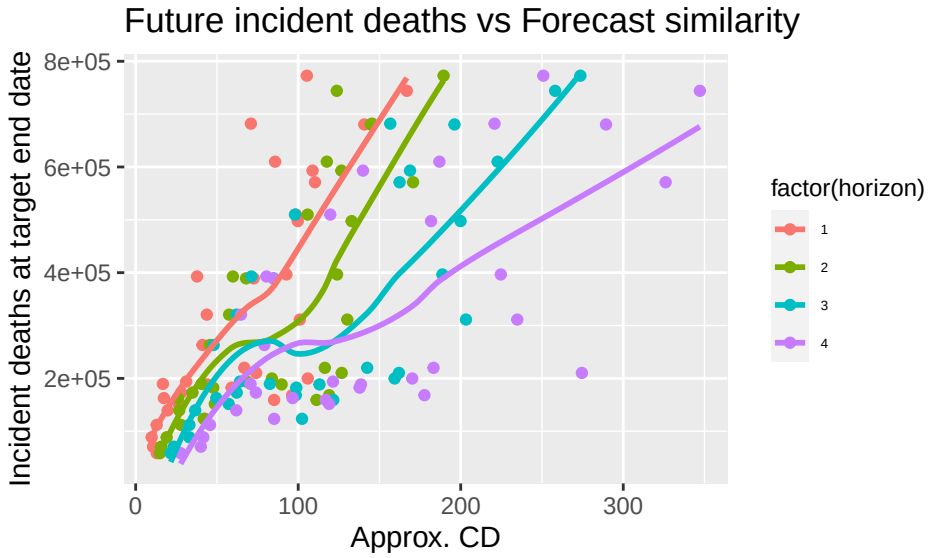


Relationship between observed values and similarities

Observed incident deaths at $t-1$ seem to be associated with more dissimilarity among forecasts made at time t. But we know the scale of observed values impacts the distances.



Future incident deaths at target end dates seem to be associated with more dissimilarity among forecasts made at time t.



Model types

Model	Type
CMU-TimeSeries	statistical
COVIDhub-baseline	statistical
COVIDhub-ensemble	ensemble
CU-select	mechanistic
Karlen-pypm	mechanistic
LANL-GrowthRate	statistical
MOBS-GLEAM_COVID	mechanistic
OliverWyman-Navigator	mechanistic
RobertWalraven-ESG	statistical
SteveMcConnell-CovidComplete	statistical
UA-EpiCovDA	mechanistic
UCSD_NEU-DeepGLEAM	mechanistic
UMass-MechBayes	mechanistic

Proportions of each model's nearest neighbor being the same type

We have 12 models, not including the ensemble. Each model's nearest neighbor is the model that is most similar to it (not including the model itself and duplicated pairs). Based on the proportions doesn't look like model types impact the similarity of forecasts overall.

Horizon	Target	Proportion of same-type nearest neighbors
1	inc death	0.6666667
2	inc death	0.6000000
3	inc death	0.6000000
4	inc death	0.5000000