COVID-19 Forecast Similarity Analysis

Johannes Bracher, Aaron Gerding, Evan Ray, Nick Reich, Nutcha Wattanachit

07/07/2021

Overview

The diversity of modeling techniques and data sources used by modelers and the variability in forecasting models' performance across time highlight the importance of having a quantitative measure of similarity between short-term COVID-19 forecasts.

Cramer distance

The *Cramer distance* between two predictive distributions F and G is defined as

$$\text{CD}(F,G) = \int_{-\infty}^{\infty} (F(x) - G(x))^2 dx$$

The Cramer distance is the divergence associated with the continuous ranked probability score (Thorarinsdottir 2013, Gneiting and Raftery 2007). We use the following two approximations in the analysis:

Cramer distance approximation for equally-spaced intervals (Approximation 2)

$$\text{CD}(F,G) \approx \frac{1}{(K+1)^2} \times \sum_{i=1}^{2K-1} b_i^2 (q_{i+1} - q_i), \tag{1}$$

- $\mathbf{q}$ is a vector of length $2K$. It is obtained by pooling the $q_k^F, q_k^G, k = 1, \dots, K$ and ordering them in increasing order (ties can be ordered in an arbitrary manner).
- $\mathbf{a}$ is a vector of length $2K$ containing the value 1 wherever $\mathbf{q}$ contains a quantile of F and -1 wherever it contains a value of G.
- $\mathbf{b}$ is a vector of length $2K$ containing the absolute cumulative sums of $\mathbf{a}$, i.e. $b_i = \left|\sum_{j=1}^{i} a_j\right|$.

Cramer distance approximation for unequally-spaced intervals (Trapezoidal riemann sum)

$$\begin{aligned}
\text{CD}(F,G) &\approx \sum_{j=1}^{2K-1} \int_{q_j}^{q_{j+1}} F(x) - G(x)^2 & (2)\\
&\approx \sum_{j=1}^{2K-1} \frac{\{\hat{F}(q_j) - \hat{G}(q_j)\}^2 + \{\hat{F}(q_{j+1}) - \hat{G}(q_{j+1})\}^2}{2} (q_{j+1} - q_j) & (3)
\end{aligned}$$

where $\tau_j^F \in \tau_F$ and $\tau_j^G \in \tau_G$. τ_F and τ_G are vectors of length $2K-1$ with elements

$$\tau_j^F = \begin{cases} I(q_1 = q_1^F) \times \tau_{q_1}^F & \text{for } j = 1 \\ I(q_j \in \{q_1^F, ..., q_K^F\}) \times \tau_{q_j}^F + I(q_j \in \{q_1^G, ..., q_K^G\}) \times \tau_{j-1}^F & \text{for } j > 1 \end{cases}$$

where $\tau_{q_j}^F$ is the probability level corresponding to q_j given q_j in the pooled quantiles comes from F, and τ_{j-1}^F is the $(j-1)^{th}$ probability level in τ_F.

$$\tau_j^G = \begin{cases} I(q_1 = q_1^G) \times \tau_{q_1}^G & \text{for } j = 1 \\ I(q_j \in \{q_1^G, ..., q_K^G\}) \times \tau_{q_j}^G + I(q_j \in \{q_1^F, ..., q_K^F\}) \times \tau_{j-1}^G & \text{for } j > 1 \end{cases}$$

where $\tau_{q_j}^G$ is the probability level corresponding to q_j given q_j in the pooled quantiles comes from G, and τ_{j-1}^G is the $(j-1)^{th}$ probability level in τ_G.

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

Naturally, the differences between the two approximations are larger for further horizons since forecasts are more dissimilar. The differences for the approx. CD between CU-select and the ensemble forecasts seem a bit more pronounced for all horizons - we might want to check how the CDF (built from quantiles) look.

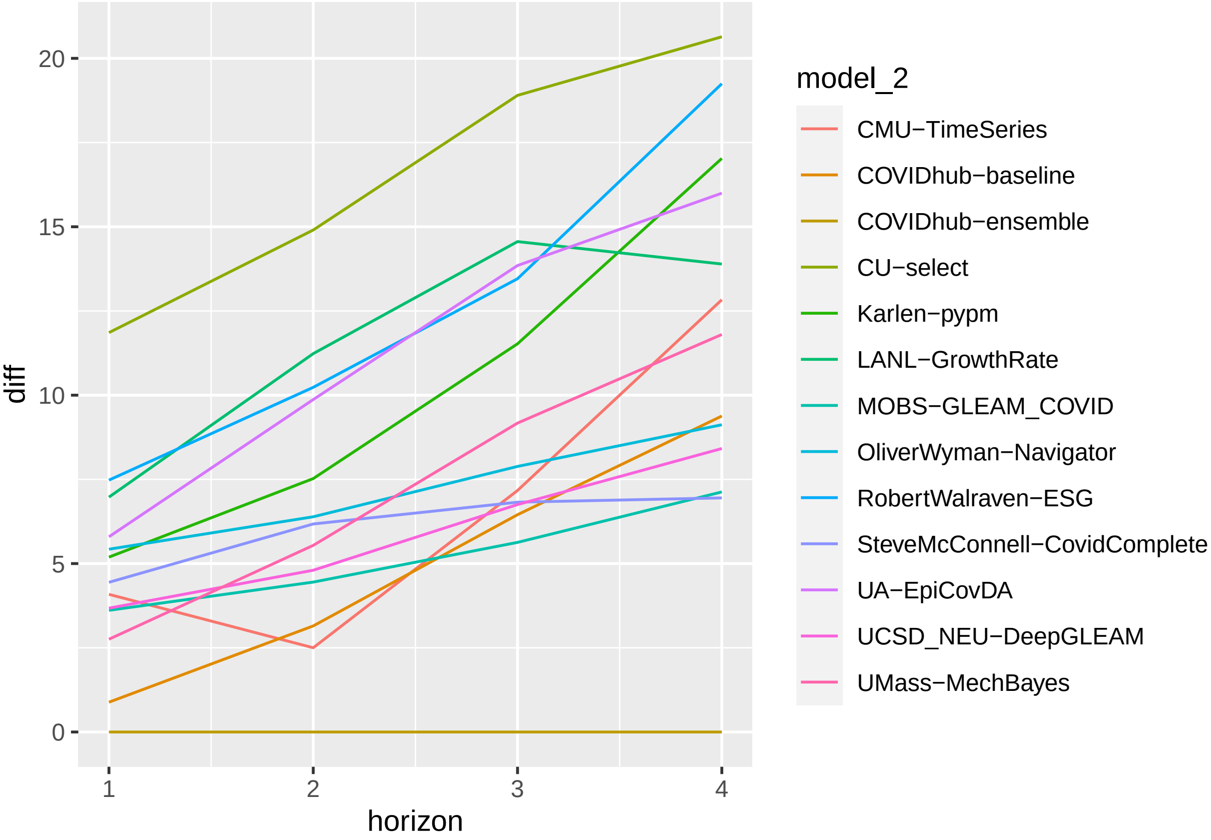
There are 13 models that fulfilled the criteria for the 5 locations with highest cumulative deaths.

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

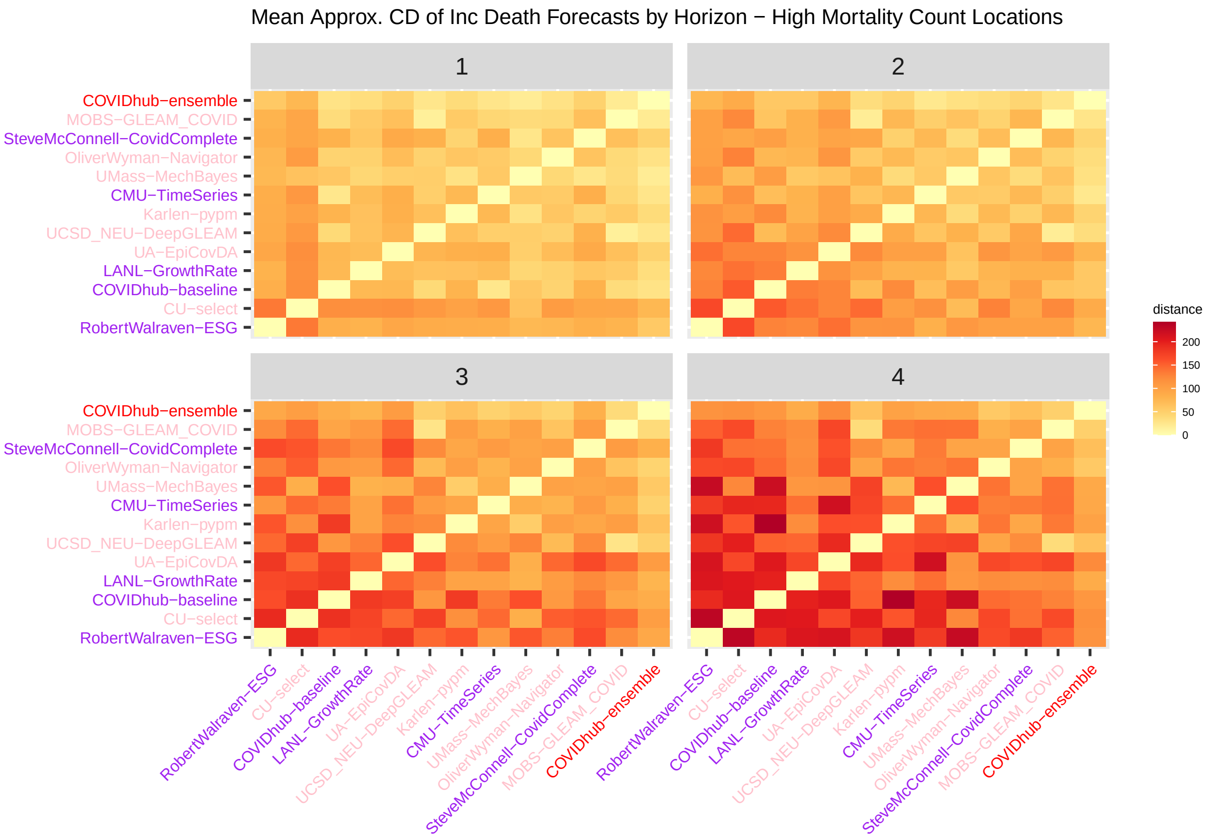
Differences between two approximations (for high count locations only)

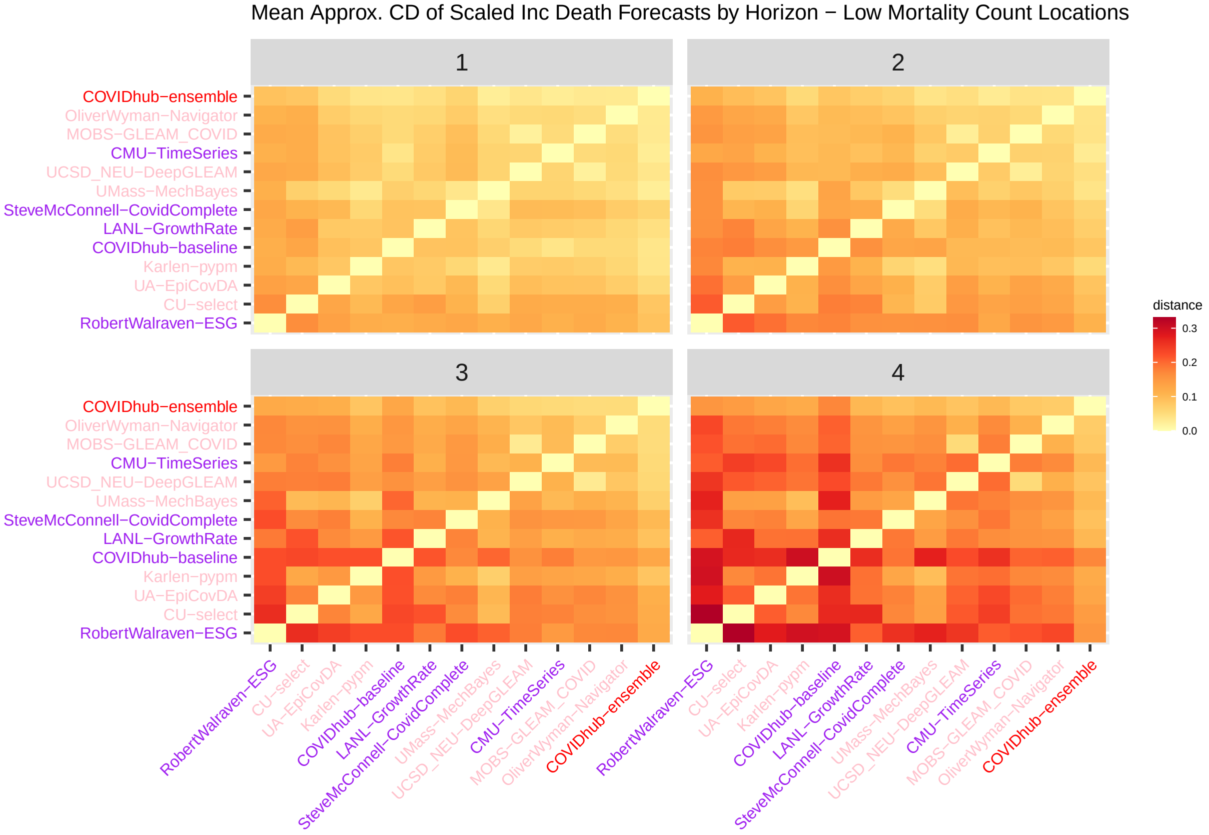
The approximated pairwise Cramer's distances between each forecast and the ensemble are calculated using both types of approximations to check for any large discrepancies between the two methods. The table below shows the averaged approx. CD over all target end dates and all 5 high count locations. Given the differences are positive (unequal-equal) and the equally-spaced formula overweighs the tail quantiles (it actually distort the height of all boxes since it assumes 1/23 (0.043) increment here when most increments are 0.05), the tail differences might be offset by the underestimation of the heights at most quantiles by the equally-weighted formula.



Mean approximated pairwise distances over across 5 high count locations

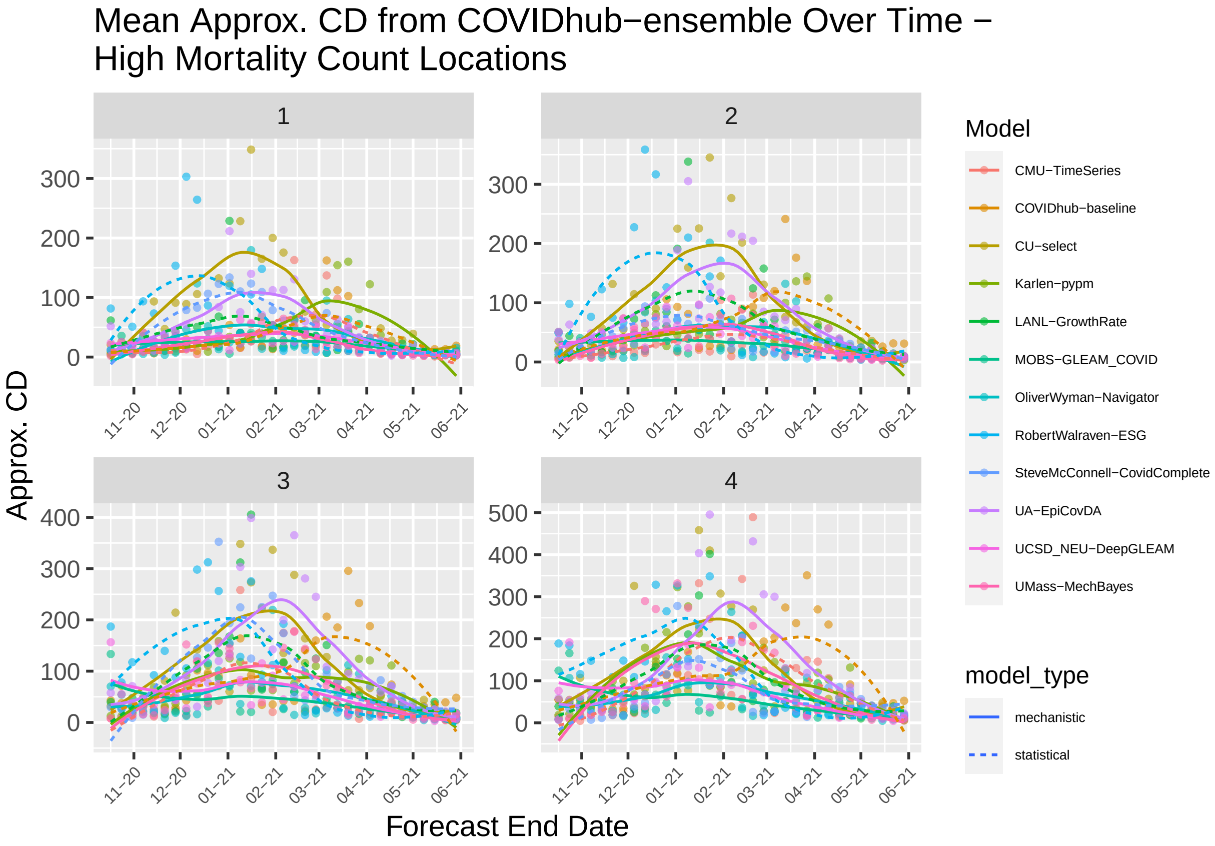
We can visualize the mean approximated pairwise distances across all weeks and locations in heat maps. The distance from the model to itself is zero. The $x-$axis is arranged based in an ascending order of the model's approximate pairwise distance from the COVIDhub-ensemble. So, the first model is the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

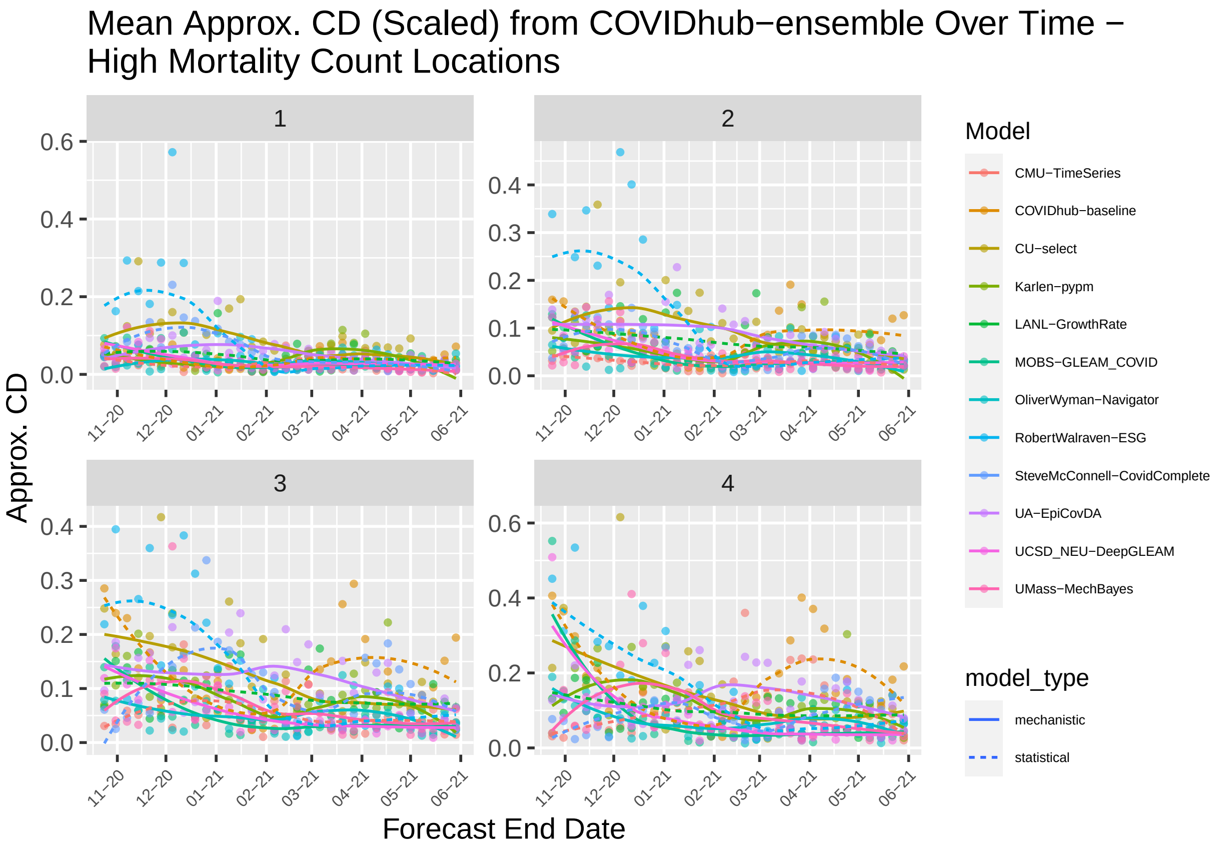




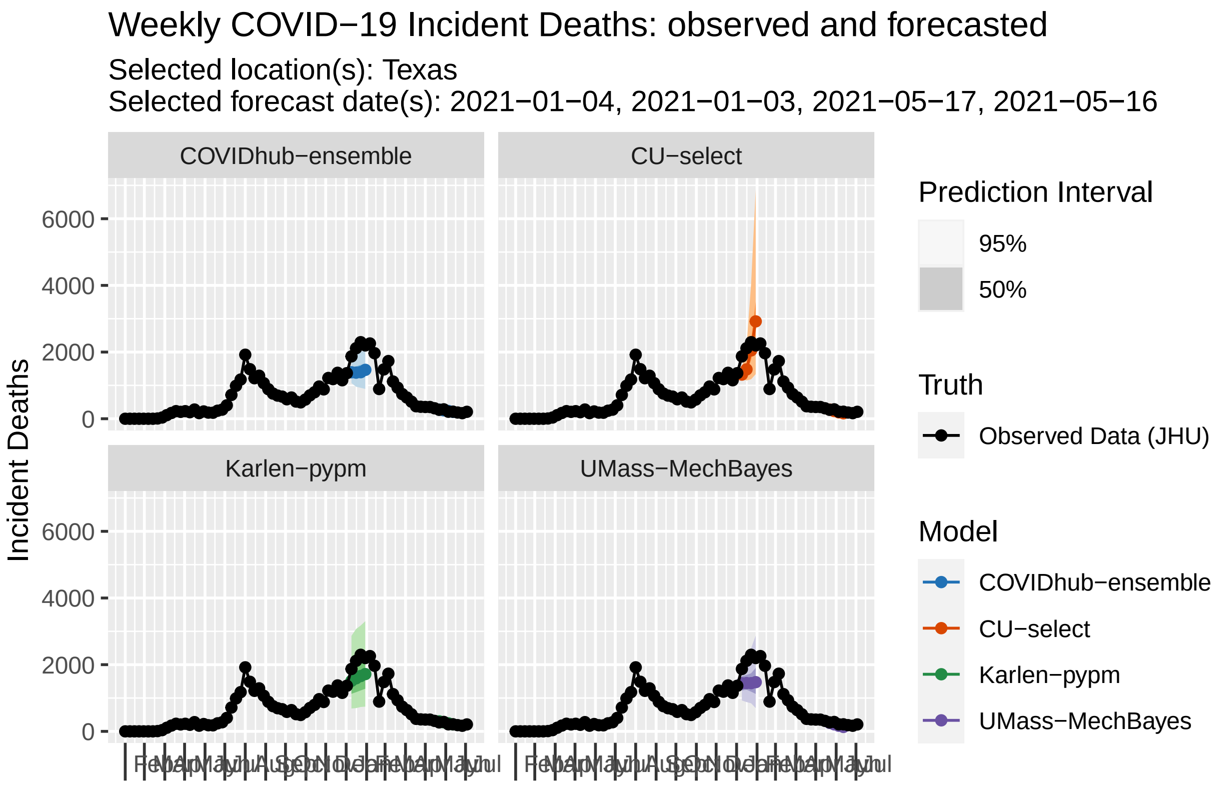
For high morality count locations, the distances between pairs of forecasts are higher for further forecast horizons. TCU-select forecasts for high count locations seem to be more dissimilar to other models on average across all forecast horizons. After scaling by the truth at t-1, some distances get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

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.



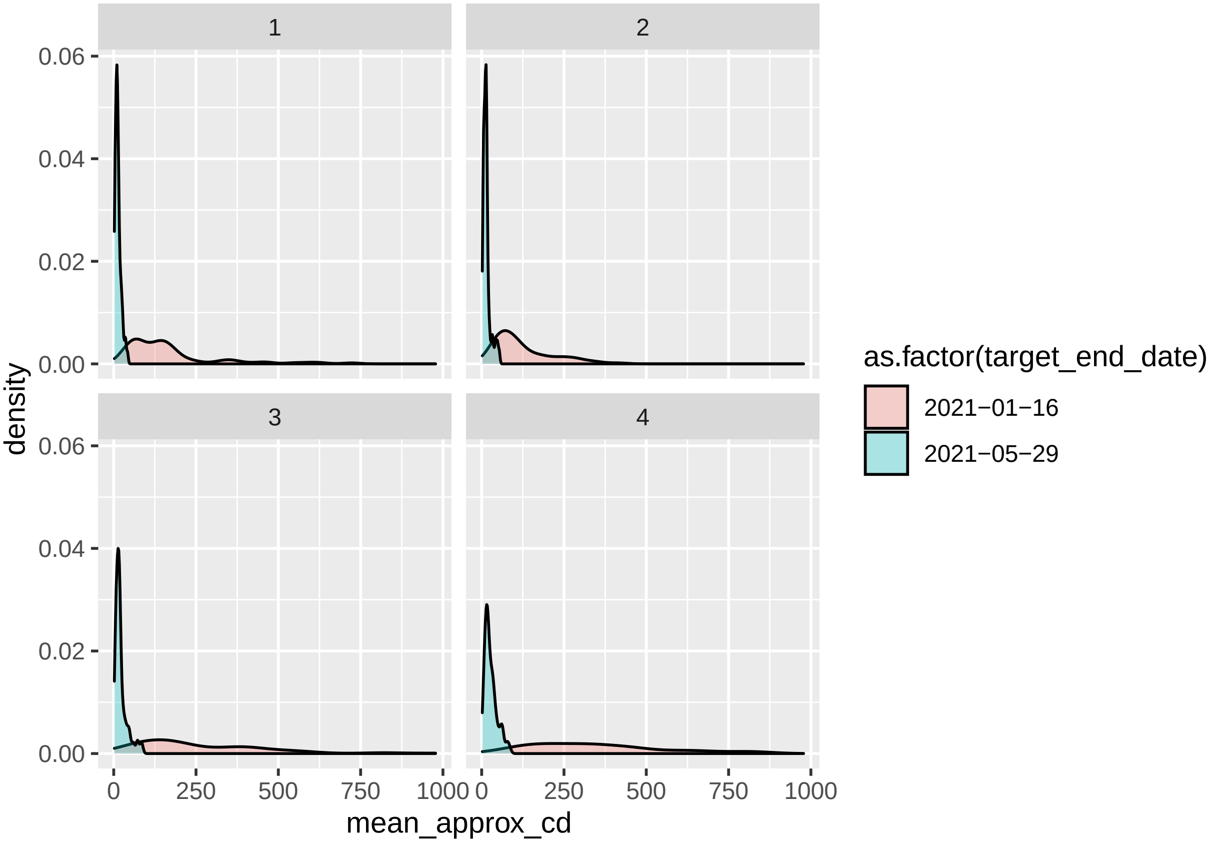


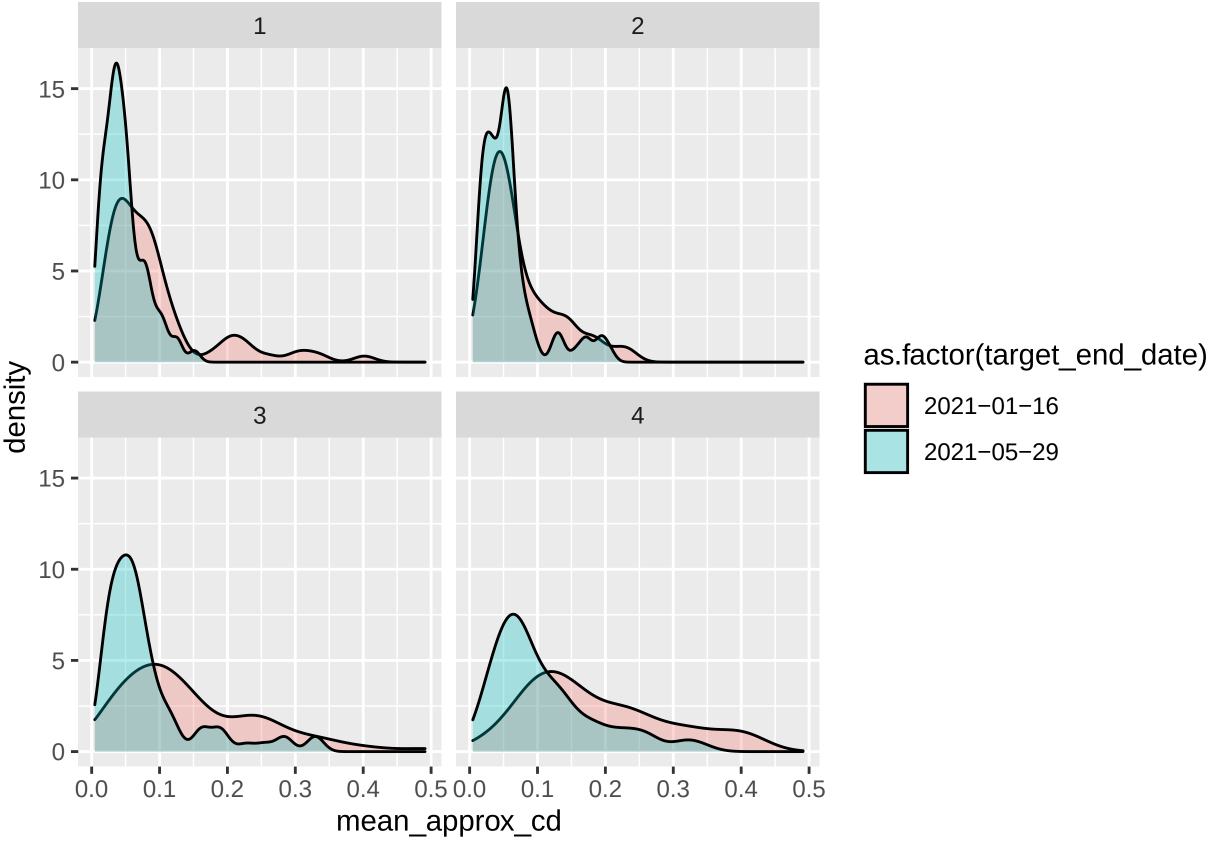
Below is the plots of forecasts in the forecast week where the mean WIS for the ensemble is the highest (for now just Texas). 1-4 wk ahead from this forecast date is from early Jan to early Feb target end dates (which is when, on average across all locations, distance from the ensemble is high).



ι), COVIDhub−ensemble, UMass−MechBayes, CU−select, Karlen−pypm (forecasts)

We can check the distribution of all mean pairwise distances across all locations in that week:


1
2
3
4
0.06
0.04
0.02
0.00
density
0
250
500
750
1000
mean_approx_cd
as.factor(target_end_date)
2021−01−16
2021−05−29


1
2
3
4
15
10
5
0
density
0.0
0.1
0.2
0.3
0.4
0.5
mean_approx_cd
as.factor(target_end_date)
2021−01−16
2021−05−29

Hierarchical clustering based on mean approx. CD across all weeks and locations

We can cluster the distances using hierarchical clustering. Different linkages will result in different clusters - here we use ward linkage.

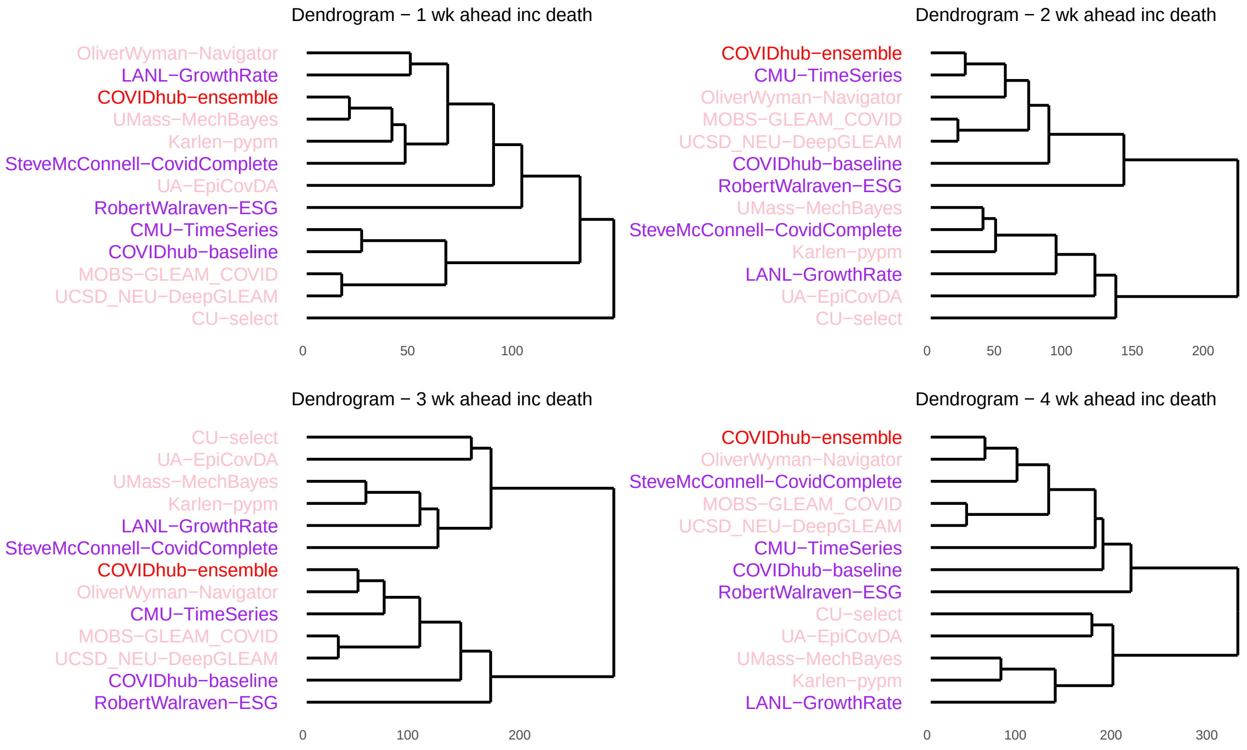
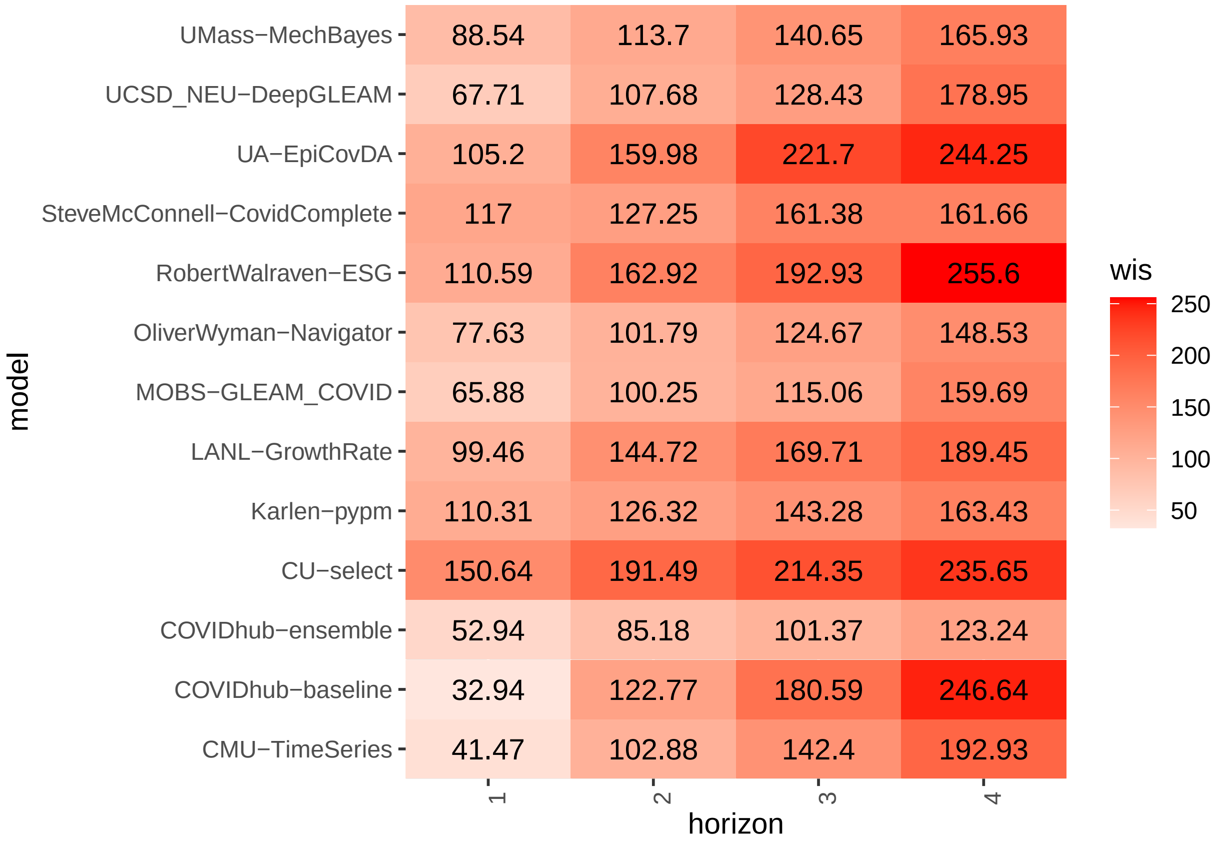


Figure 1: High Mortality Count Locations

We can also check the mean WIS across these location for each horizon. We can see that models with closer WIS are grouped together (CD techically generalizes to WIS in the case where G is a point mass, so it is expected).

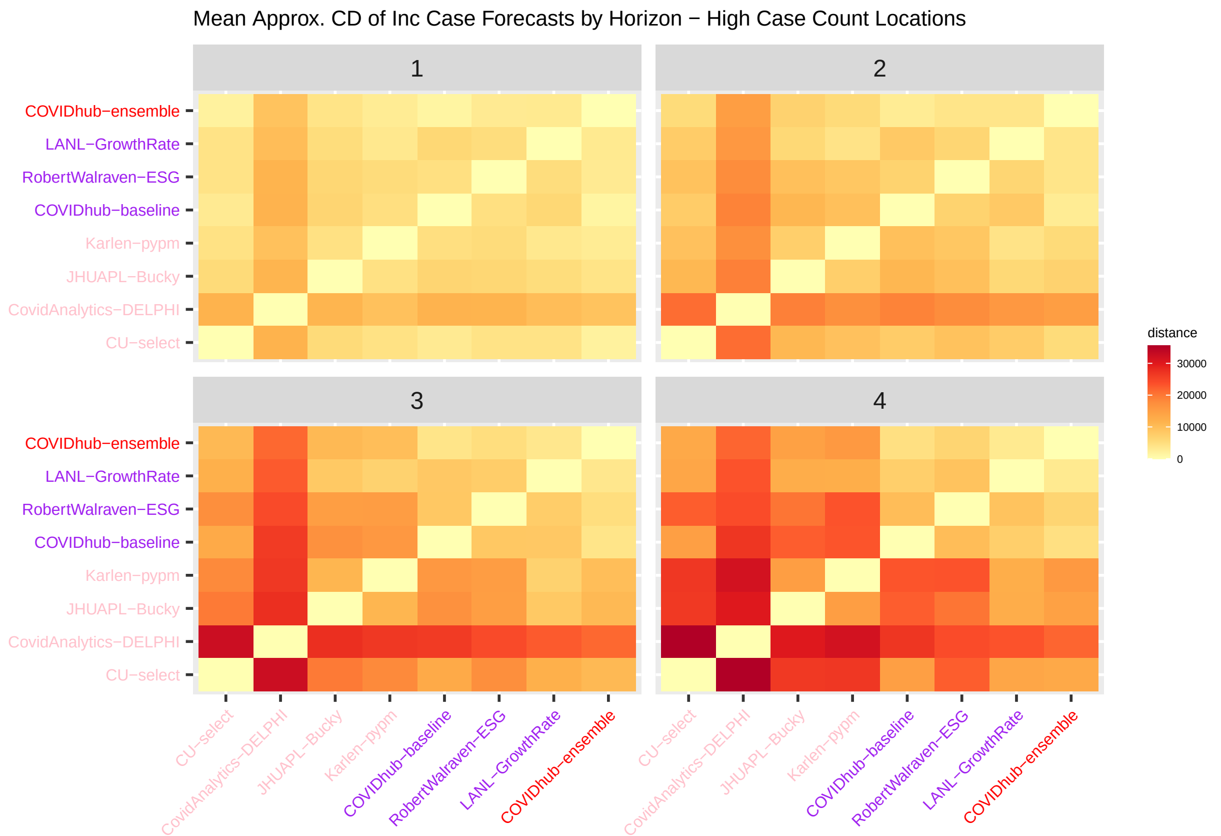


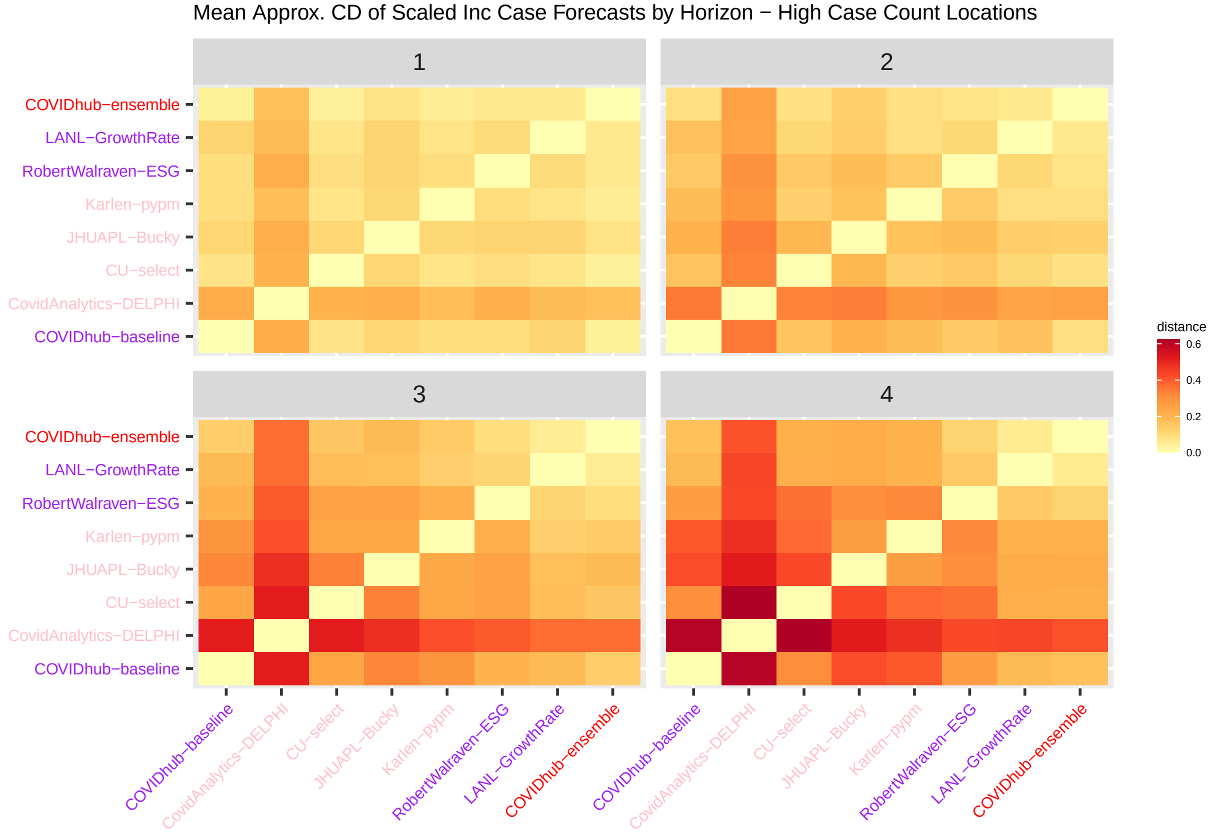
1-4 Week Ahead Incident Case Forecasts

Model types

Model	Type
CovidAnalytics-DELPHI	mechanistic
COVIDhub-baseline	statistical
COVIDhub-ensemble	ensemble
CU-select	mechanistic
JHUAPL-Bucky	mechanistic
Karlen-pypm	mechanistic
LANL-GrowthRate	statistical
RobertWalraven-ESG	statistical

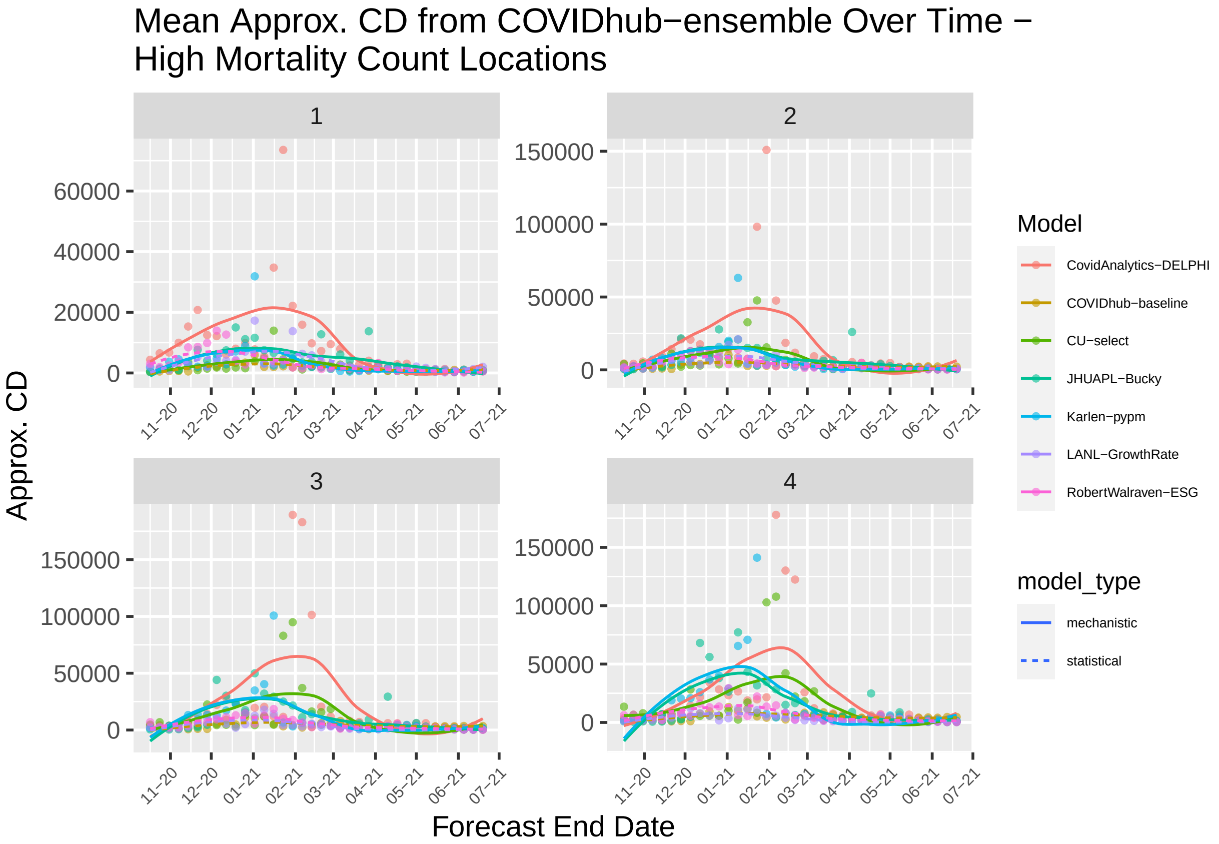
Mean approximated pairwise distances over across 5 high count and 5 low count locations

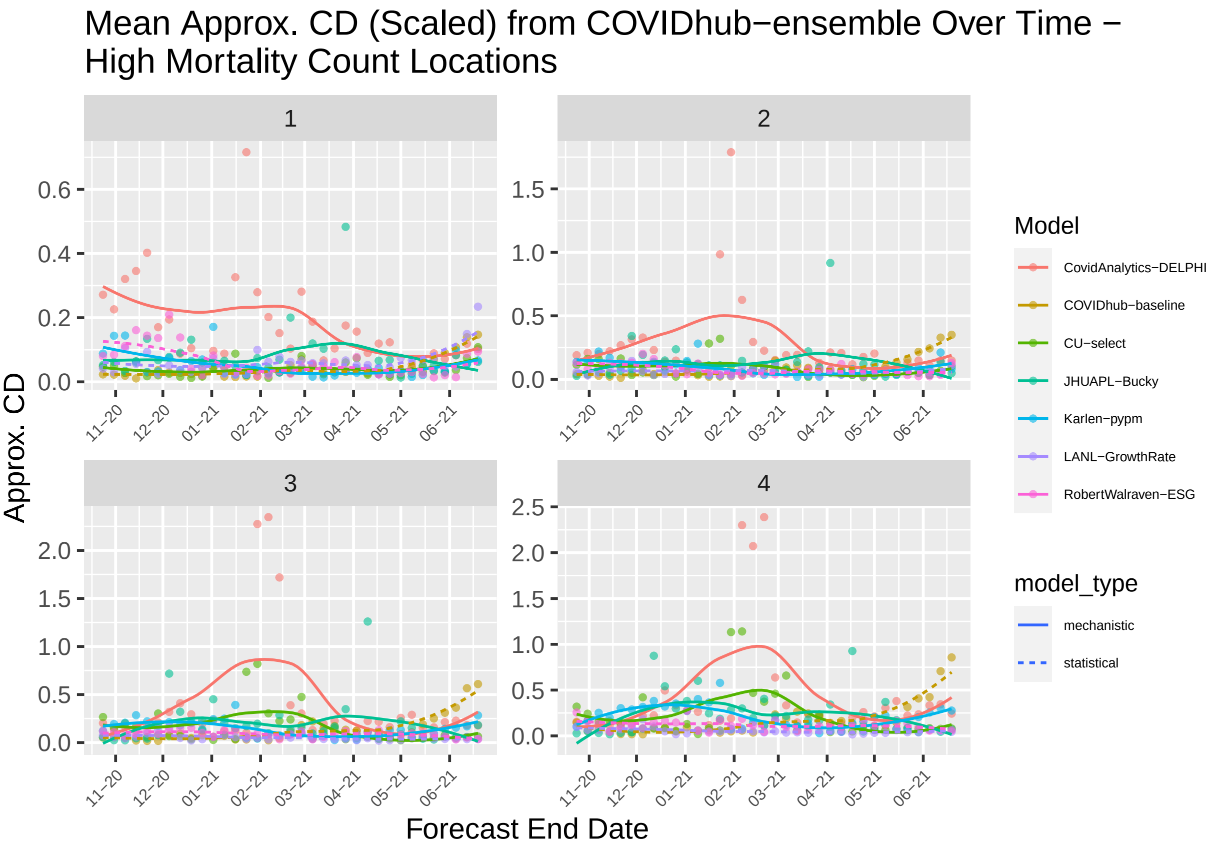

Mean Approx. CD of Inc Case Forecasts by Horizon − High Case Count Locations


Mean Approx. CD of Scaled Inc Case Forecasts by Horizon − High Case Count Locations

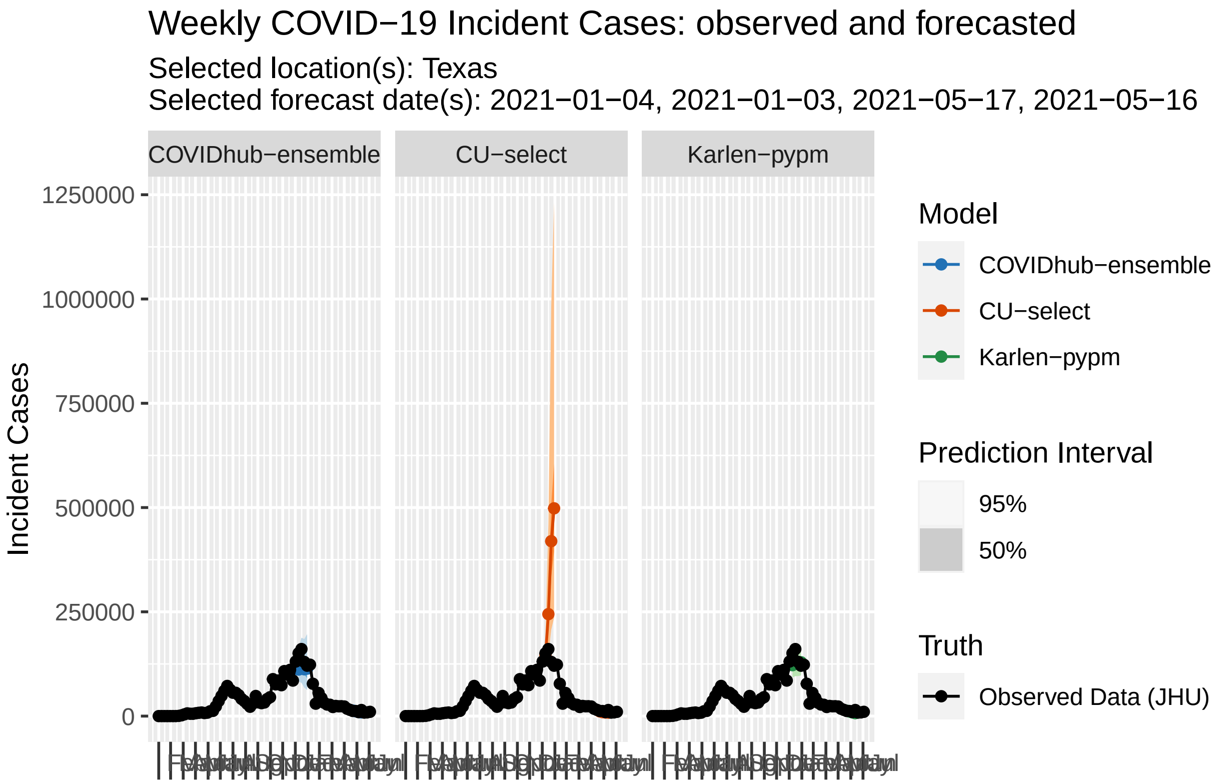
CovidAnalytics−DELPHI forecasts for both high and low count locations seem to be more dissimilar to other models on average across all forecast horizons.

When we look at the approximated pairwise distances over time, we see high distances from the ensemble around Jan-Feb 2021 for high count locations.



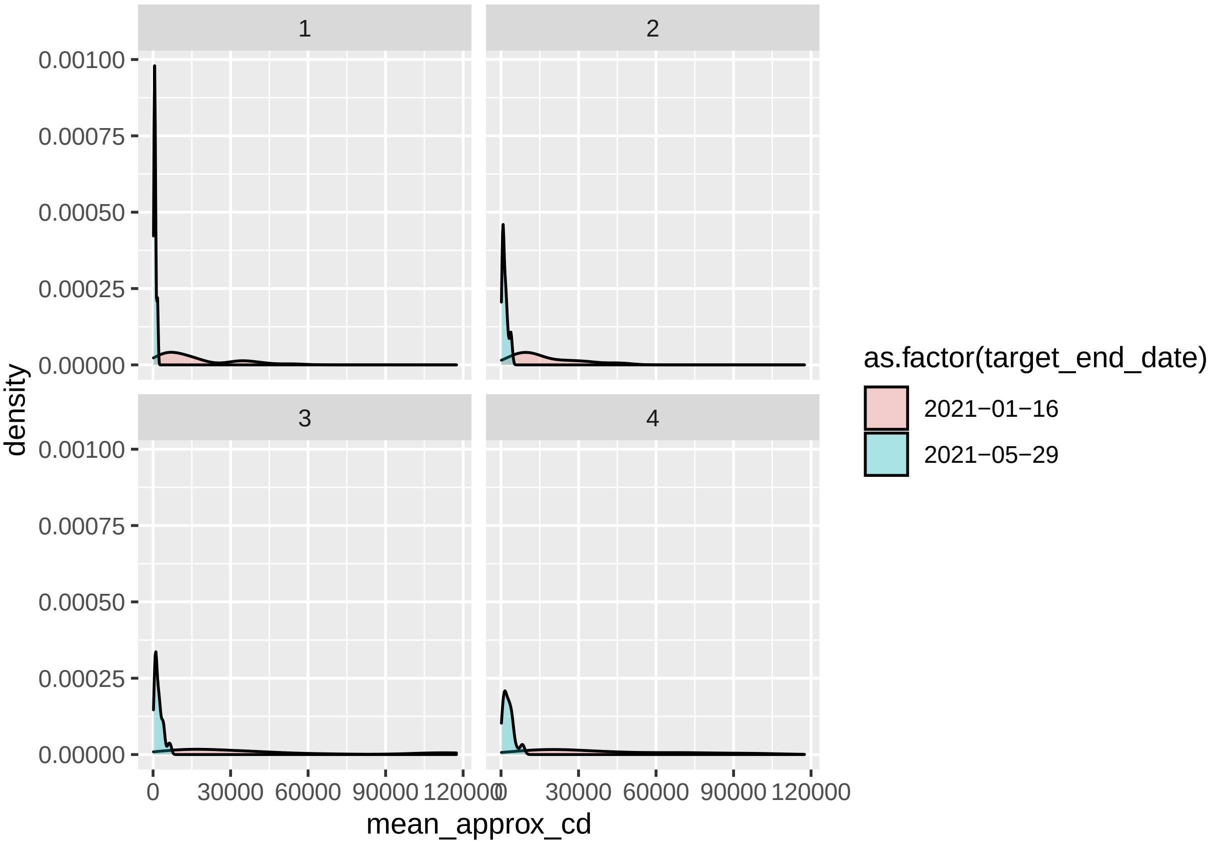


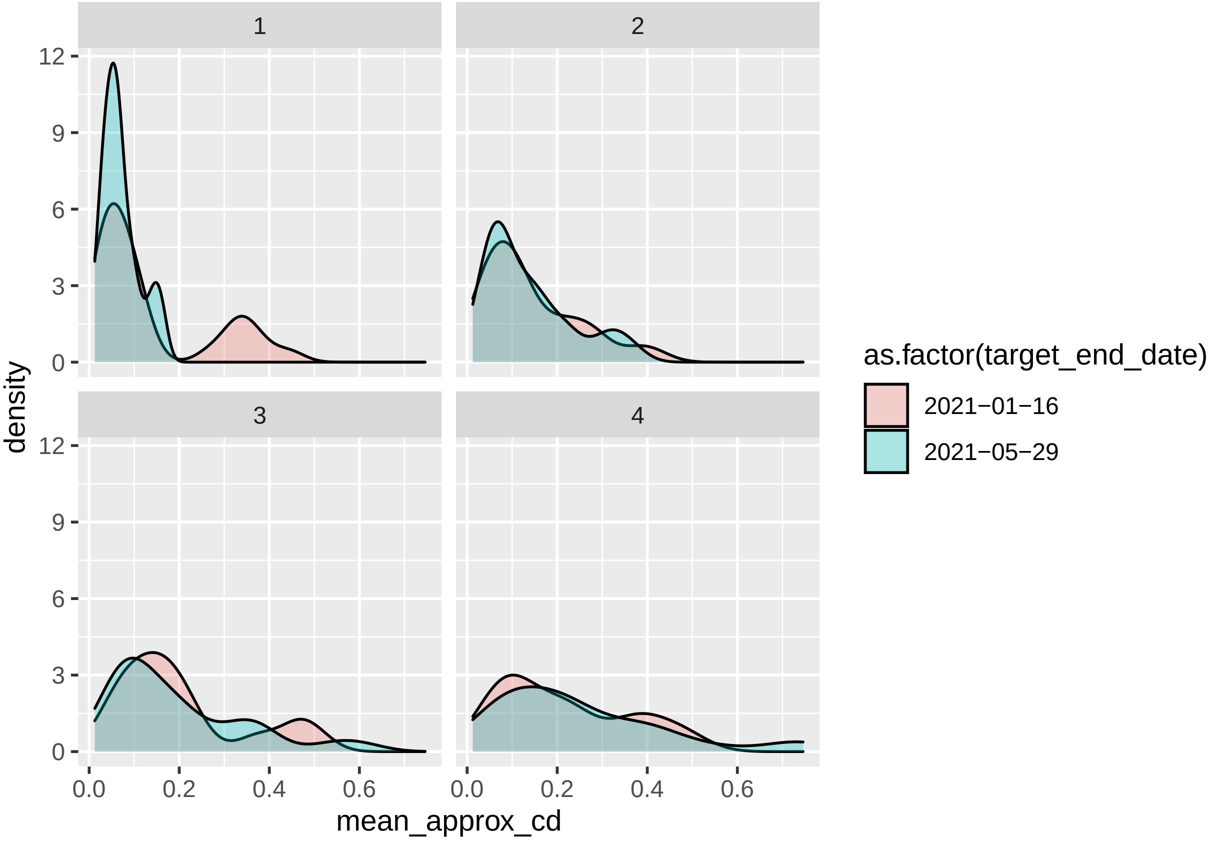
Below is the plots of forecasts in the forecast week where the mean WIS for the ensemble is the highest (for now just Texas). 1-4 wk ahead from this forecast date is from early Jan to early Feb target end dates (which is when, on average across all locations, distance from the ensemble is high).



: JHU (observed data), COVIDhub−ensemble, CU−select, Karlen−pypm (forecasts)

We can check the distribution of all mean pairwise distances across all locations in that week:


1
2
3
4
0.00100
0.00075
0.00050
0.00025
0.00000
density
0
30000
60000
90000
120000
mean_approx_cd
as.factor(target_end_date)
2021−01−16
2021−05−29


1
2
3
4
12
9
6
3
0
density
0.0
0.2
0.4
0.6
mean_approx_cd
as.factor(target_end_date)
2021−01−16
2021−05−29

Hierarchical clustering based on mean approx. CD across all weeks and locations

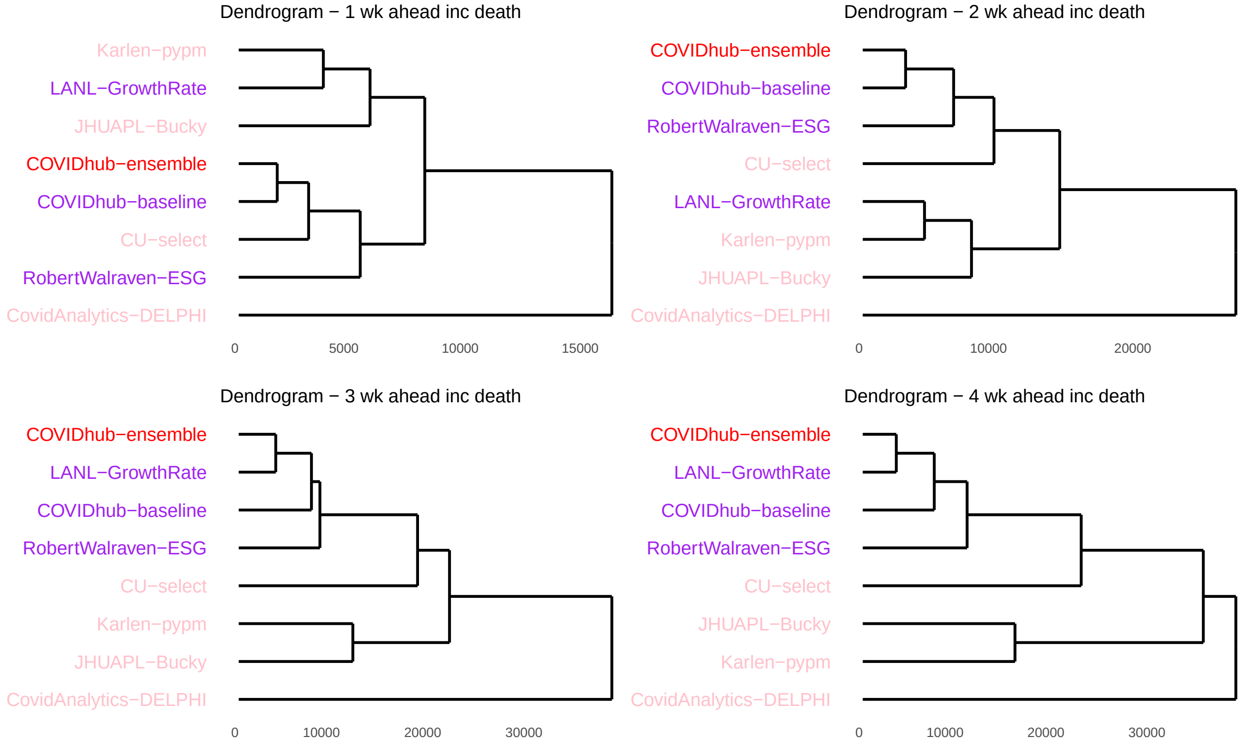
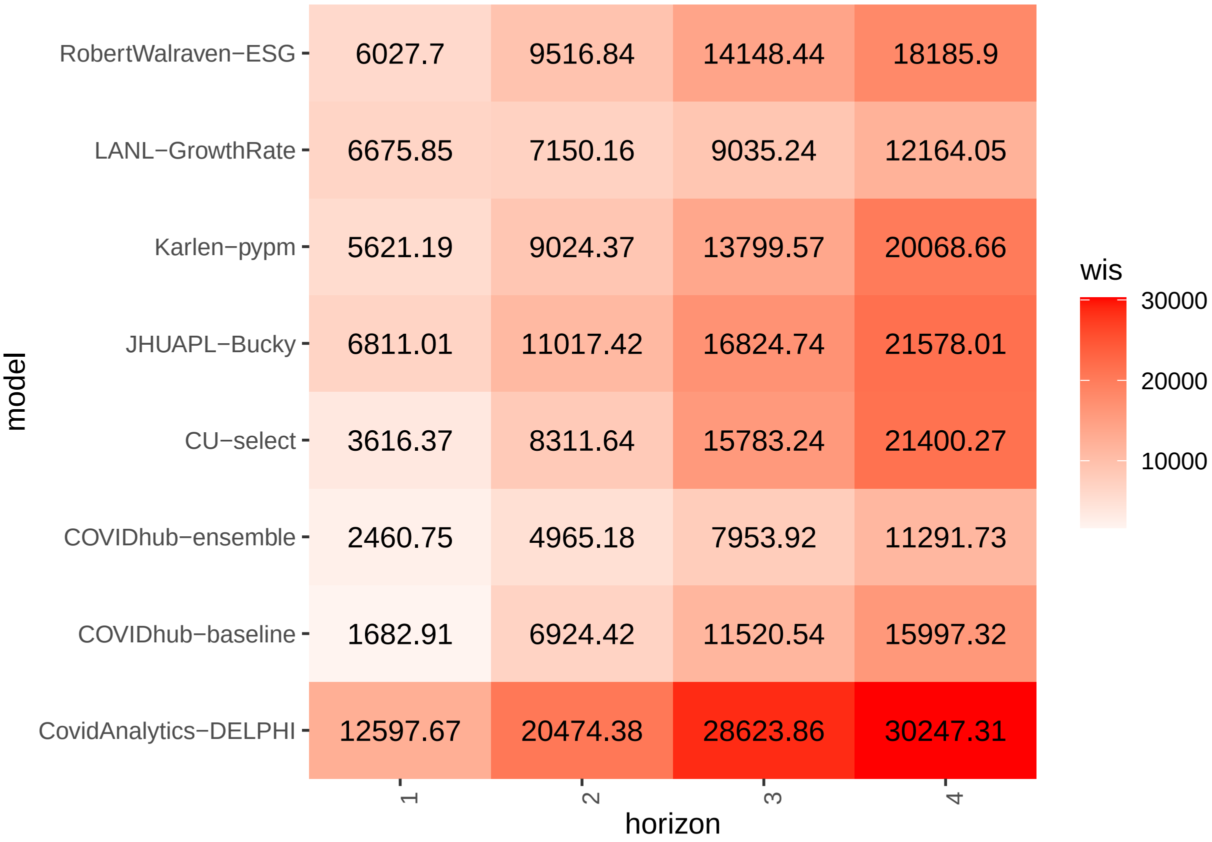


Figure 2: High Case Count Locations

We see similar trend here as we saw in inc death analysis.



model	1	2	3	4
RobertWalraven−ESG	6027.7	9516.84	14148.44	18185.9
LANL−GrowthRate	6675.85	7150.16	9035.24	12164.05
Karlen−pypm	5621.19	9024.37	13799.57	20068.66
JHUAPL−Bucky	6811.01	11017.42	16824.74	21578.01
CU−select	3616.37	8311.64	15783.24	21400.27
COVIDhub−ensemble	2460.75	4965.18	7953.92	11291.73
COVIDhub−baseline	1682.91	6924.42	11520.54	15997.32
CovidAnalytics−DELPHI	12597.67	20474.38	28623.86	30247.31