COVID-19 Forecast Similarity Analysis

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

06/24/2021

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 $q_k^F, q_k^G, k = 1, \ldots, K$ 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)

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

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 and 12 models for the 5 locations with lowest cumulative deaths.

Model types

Model	Type
CMU-TimeSeries	statistical
COVIDhub-baseline	statistical
COVIDhub-ensemble	ensemble
CU-select	mechanistic
Karlen-pypm	statistical
LANL-GrowthRate	statistical
MOBS-GLEAM_COVID	mechanistic
OliverWyman-Navigator	mechanistic
RobertWalraven-ESG	statistical
SteveMcConnell-CovidComplete	statistical
UA-EpiCovDA	mechanistic
UCSD_NEU-DeepGLEAM	neither stats nor mech
UMass-MechBayes	mechanistic

Model	Type
PSI-DRAFT	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.

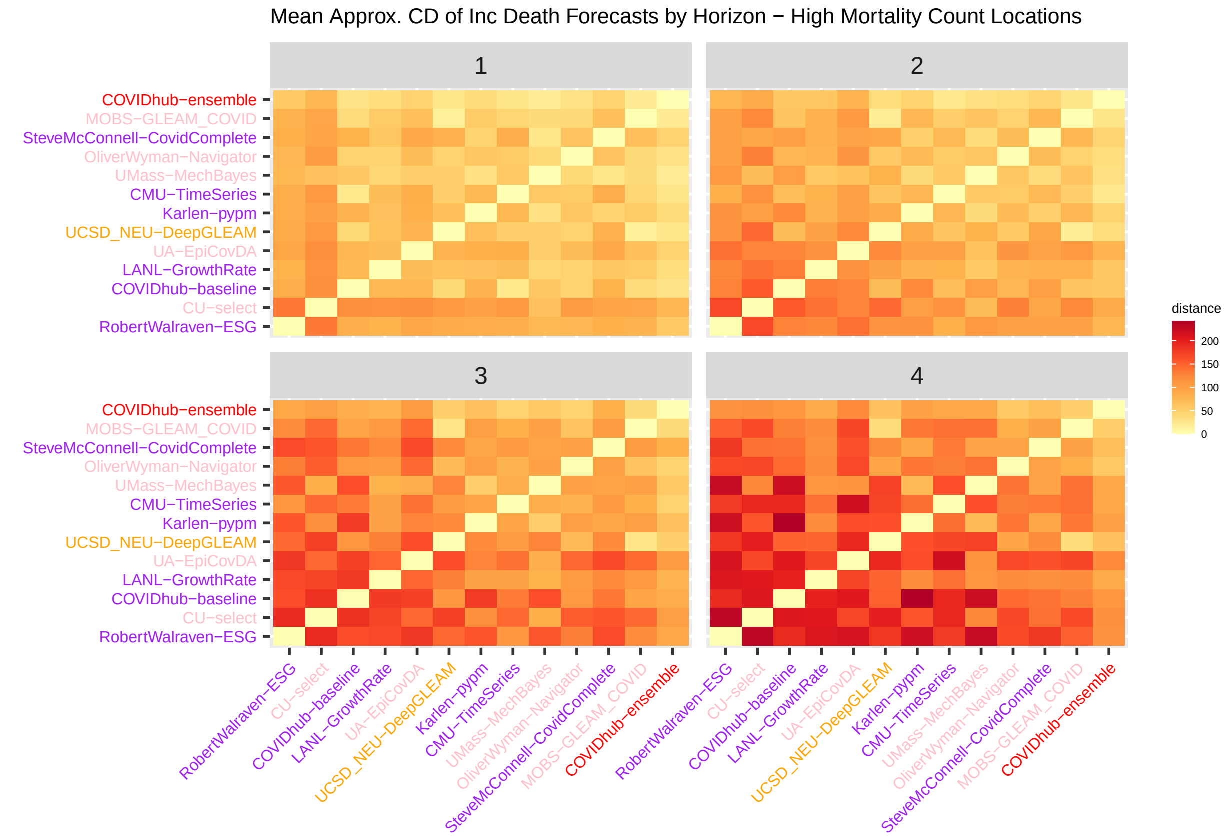
Table 2: Mean approx. CDs relative to the ensemble

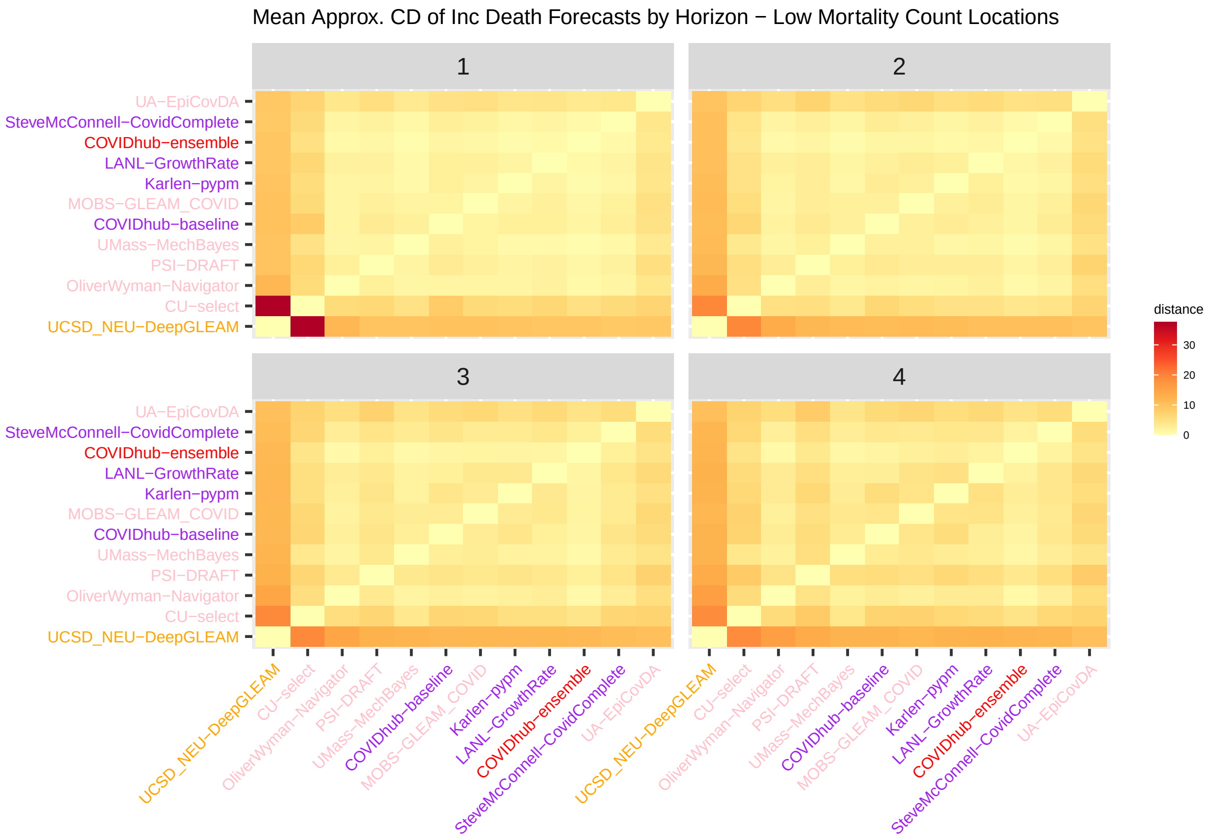
Anchor Model	Model	Horizon	Target	CD (uneq)	CD (eq)	Diff
COVIDhub-ensemble	COVIDhub-ensemble	1	inc death	0.00	0.00	0.00
COVIDhub-ensemble	COVIDhub-baseline	1	inc death	29.76	28.87	0.88
COVIDhub-ensemble	UMass-MechBayes	1	inc death	19.71	16.96	2.75
COVIDhub-ensemble	MOBS-GLEAM_COVID	1	inc death	20.72	17.11	3.61
COVIDhub-ensemble	UCSD_NEU-DeepGLEAM	1	inc death	26.12	22.45	3.68
COVIDhub-ensemble	CMU-TimeSeries	1	inc death	27.17	23.08	4.09
COVIDhub-ensemble	SteveMcConnell-CovidComplete	1	inc death	47.81	43.36	4.44
COVIDhub-ensemble	Karlen-pypm	1	inc death	37.69	32.50	5.19
COVIDhub-ensemble	OliverWyman-Navigator	1	inc death	30.52	25.09	5.43
COVIDhub-ensemble	UA-EpiCovDA	1	inc death	48.19	42.39	5.79
COVIDhub-ensemble	LANL-GrowthRate	1	inc death	35.45	28.48	6.97
COVIDhub-ensemble	RobertWalraven-ESG	1	inc death	56.77	49.29	7.48
COVIDhub-ensemble	CU-select	1	inc death	74.72	62.86	11.86
COVIDhub-ensemble	COVIDhub-ensemble	2	inc death	0.00	0.00	0.00
COVIDhub-ensemble	CMU-TimeSeries	2	inc death	24.36	21.86	2.50
COVIDhub-ensemble	COVIDhub-baseline	2	inc death	57.73	54.58	3.15
COVIDhub-ensemble	MOBS-GLEAM_COVID	2	inc death	27.38	22.93	4.45
COVIDhub-ensemble	UCSD_NEU-DeepGLEAM	2	inc death	35.83	31.03	4.80
COVIDhub-ensemble	UMass-MechBayes	2	inc death	32.26	26.72	5.54
COVIDhub-ensemble	SteveMcConnell-CovidComplete	2	inc death	44.27	38.10	6.18
COVIDhub-ensemble	OliverWyman-Navigator	2	inc death	36.35	29.96	6.39
COVIDhub-ensemble	Karlen-pypm	2	inc death	45.50	37.98	7.52
COVIDhub-ensemble	UA-EpiCovDA	2	inc death	77.26	67.39	9.87
COVIDhub-ensemble	RobertWalraven-ESG	2	inc death	75.93	65.70	10.23
COVIDhub-ensemble	LANL-GrowthRate	2	inc death	57.90	46.67	11.23
COVIDhub-ensemble	CU-select	2	inc death	88.41	73.51	14.90
COVIDhub-ensemble	COVIDhub-ensemble	3	inc death	0.00	0.00	0.00
COVIDhub-ensemble	MOBS-GLEAM_COVID	3	inc death	37.79	32.16	5.63
COVIDhub-ensemble	COVIDhub-baseline	3	inc death	86.58	80.13	6.45
COVIDhub-ensemble	UCSD_NEU-DeepGLEAM	3	inc death	49.79	43.04	6.75
COVIDhub-ensemble	SteveMcConnell-CovidComplete	3	inc death	83.16	76.33	6.82
COVIDhub-ensemble	CMU-TimeSeries	3	inc death	47.64	40.47	7.17
COVIDhub-ensemble	OliverWyman-Navigator	3	inc death	45.76	37.88	7.88
COVIDhub-ensemble	UMass-MechBayes	3	inc death	56.83	47.66	9.17
COVIDhub-ensemble	Karlen-pypm	3	inc death	64.99	53.46	11.53
COVIDhub-ensemble	RobertWalraven-ESG	3	inc death	91.75	78.29	13.46
COVIDhub-ensemble	UA-EpiCovDA	3	inc death	105.13	91.28	13.85
COVIDhub-ensemble	LANL-GrowthRate	3	inc death	77.83	63.27	14.56
COVIDhub-ensemble	CU-select	3	inc death	102.98	84.08	18.90
COVIDhub-ensemble	COVIDhub-ensemble	4	inc death	0.00	0.00	0.00

Anchor Model	Model	Horizon	Target	CD (uneq)	CD (eq)	Diff
COVIDhub-ensemble	SteveMcConnell-CovidComplete	4	inc death	67.09	60.14	6.95
COVIDhub-ensemble	MOBS-GLEAM_COVID	4	inc death	49.71	42.58	7.13
COVIDhub-ensemble	UCSD_NEU-DeepGLEAM	4	inc death	63.79	55.38	8.42
COVIDhub-ensemble	OliverWyman-Navigator	4	inc death	56.61	47.49	9.12
COVIDhub-ensemble	COVIDhub-baseline	4	inc death	110.09	100.71	9.38
COVIDhub-ensemble	UMass-MechBayes	4	inc death	91.02	79.22	11.80
COVIDhub-ensemble	CMU-TimeSeries	4	inc death	91.66	78.83	12.83
COVIDhub-ensemble	LANL-GrowthRate	4	inc death	87.28	73.39	13.89
COVIDhub-ensemble	UA-EpiCovDA	4	inc death	122.55	106.56	16.00
COVIDhub-ensemble	Karlen-pypm	4	inc death	98.05	81.03	17.03
COVIDhub-ensemble	RobertWalraven-ESG	4	inc death	114.35	95.11	19.25
COVIDhub-ensemble	CU-select	4	inc death	118.20	97.56	20.64

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

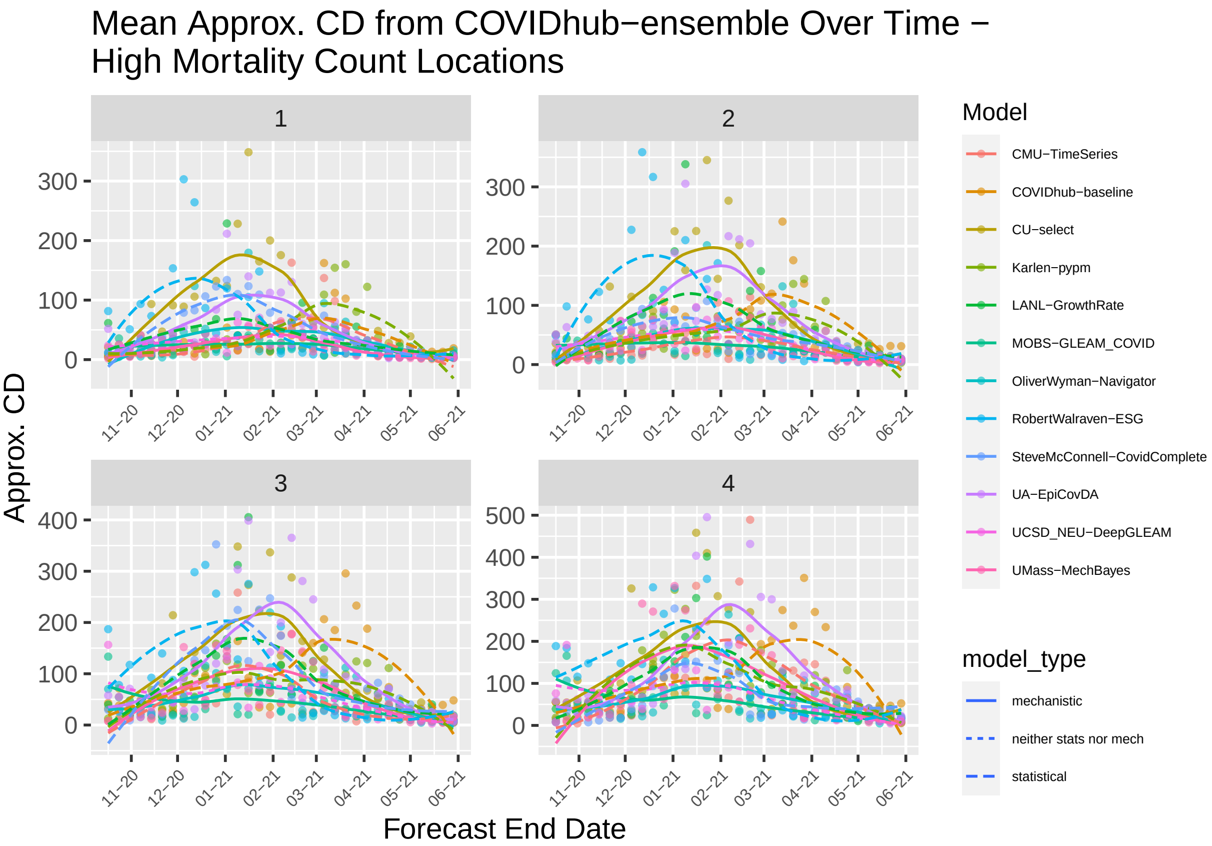
We can visualize the mean approximated pairwise distances across all weeks and locations in heat maps. The distance from the model to itself is zero. The $x-$axis is arranged based in an ascending order of the model's approximate pairwise distance from the COVIDhub-ensemble. So, the first model is the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

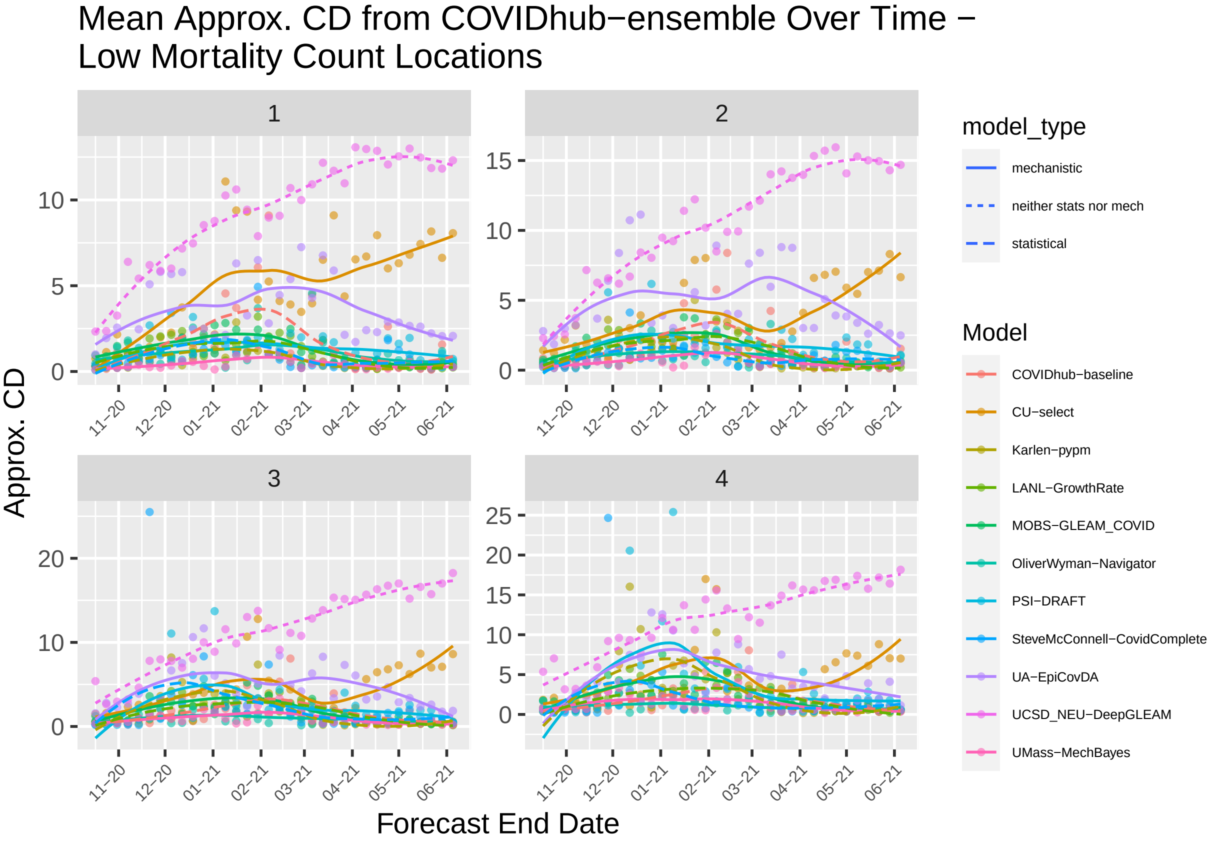




For high morality count locations, the distances between pairs of forecasts are higher for further forecast horizons. This is less less pronounced for low count locations. CU-select forecasts for high count locations and UCSD_NEU−DeepGLEAM forecasts for low count locations seem to be more dissimilar to other models on average across all forecast horizons.

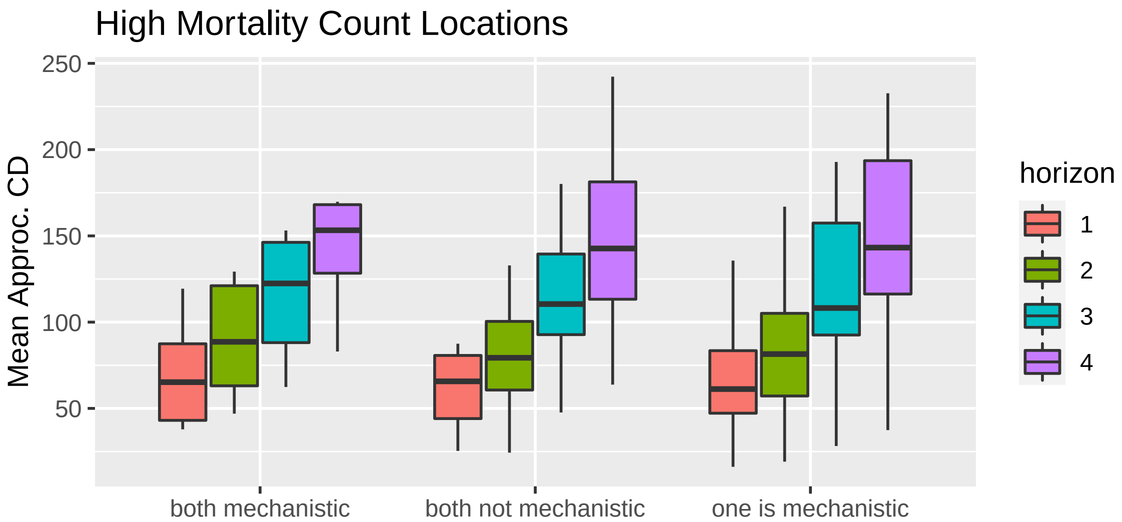
We can also look at the mean approximated pairwise distances across locations only to see how the models become more similar or dissimilar over time.

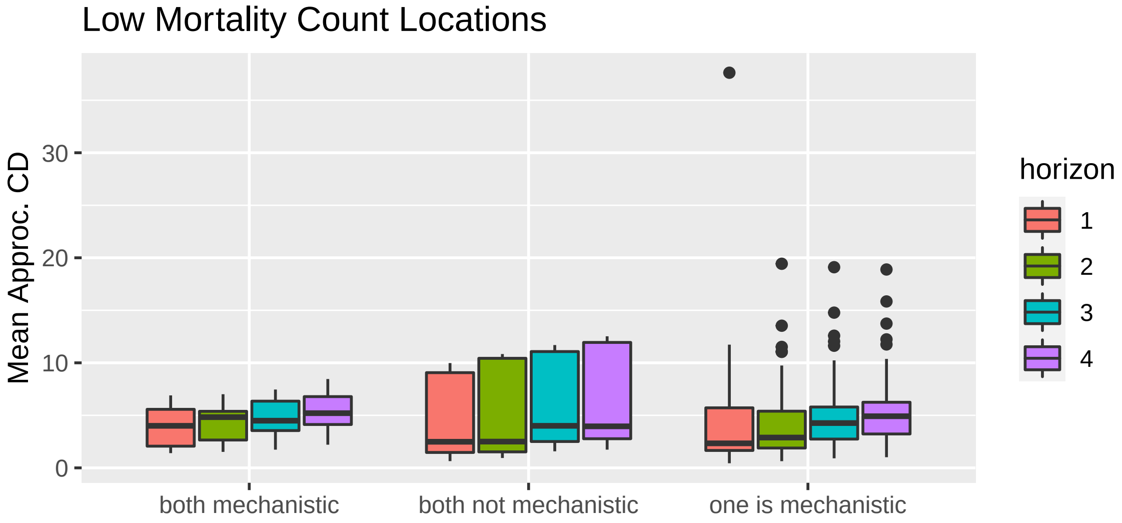

Mean Approx. CD from COVIDhub−ensemble Over Time − High Mortality Count Locations


Mean Approx. CD from COVIDhub−ensemble Over Time − Low Mortality Count Locations

Relationship between a mechanistic model type and similarity

Here we created categorical variable with 3 levels for each pair of models in the analysis: 1) both models are mechanistic 2) only one of the two models is mechanistic 3) neither of the two models are mechanistic. The approx. distances shown in the plots are averaged across locations and weeks. The distance from the model to itself and any duplicated pairs are excluded.


High Mortality Count Locations


Low Mortality Count Locations

For low count locations, there are noticeably more outliers between forecasts when one model of a pair is mechanistic and we also see larger range of distances between forecasts when both models are not mechanistic.

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.

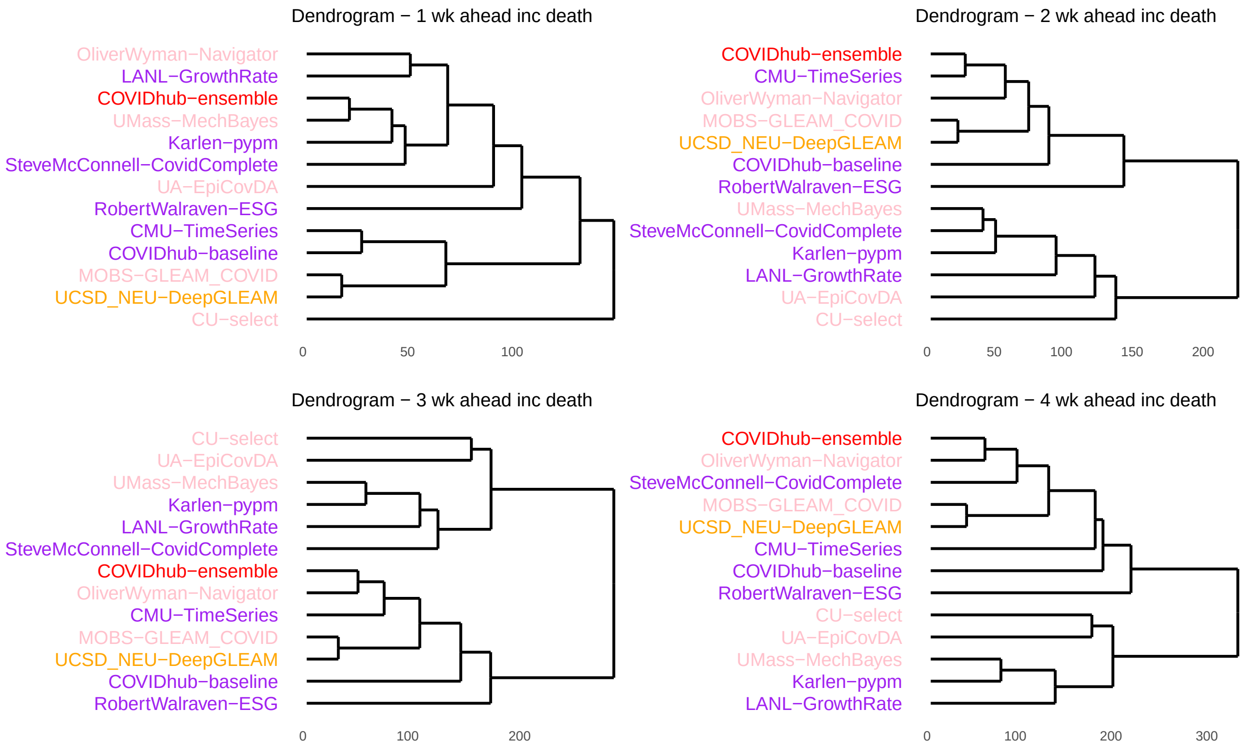


Figure 1: High Mortality Count Locations

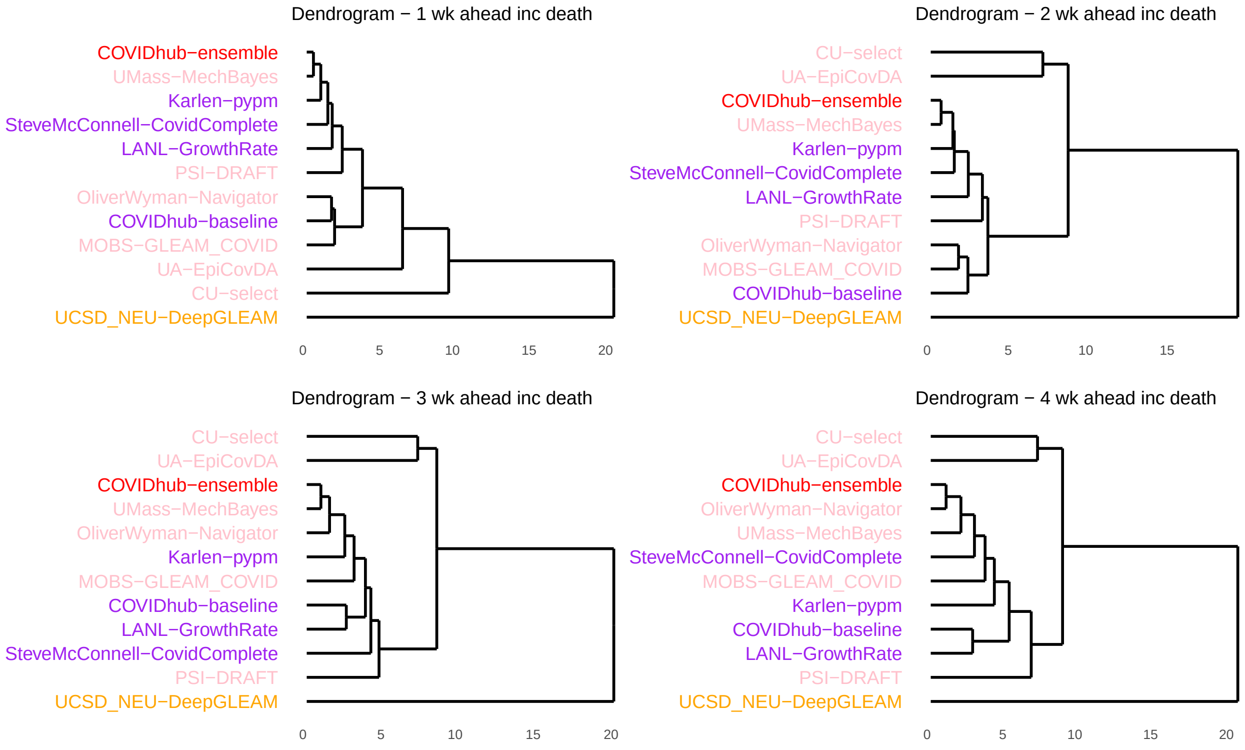


Figure 2: Low Mortality Count Locations

1-4 Week Ahead Incident Case Forecasts

There are 8 models for both the 5 locations with lowest cumulative cases.

Model types

Model	Type
CovidAnalytics-DELPHI	mechanistic
COVIDhub-baseline	statistical
COVIDhub-ensemble	ensemble
CU-select	mechanistic
JHUAPL-Bucky	neither stats nor mech
Karlen-pypm	statistical
LANL-GrowthRate	statistical
RobertWalraven-ESG	statistical
LNQ-ens1	ensemble

Differences between two approximations (for high count locations only)

Similar to Table 1, this table below shows the averaged approx. CD over all target end dates and all 5 high count locations.

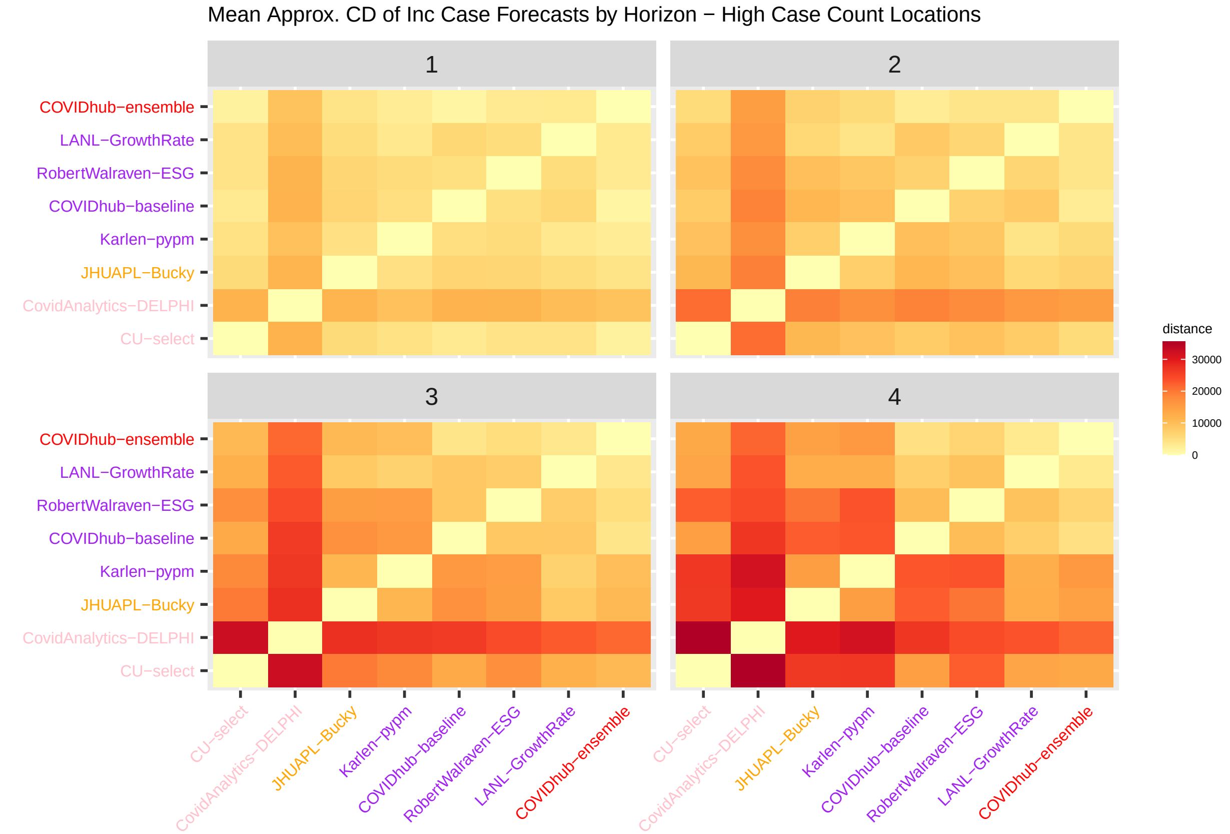
Table 4: Mean approx. CDs relative to the ensemble

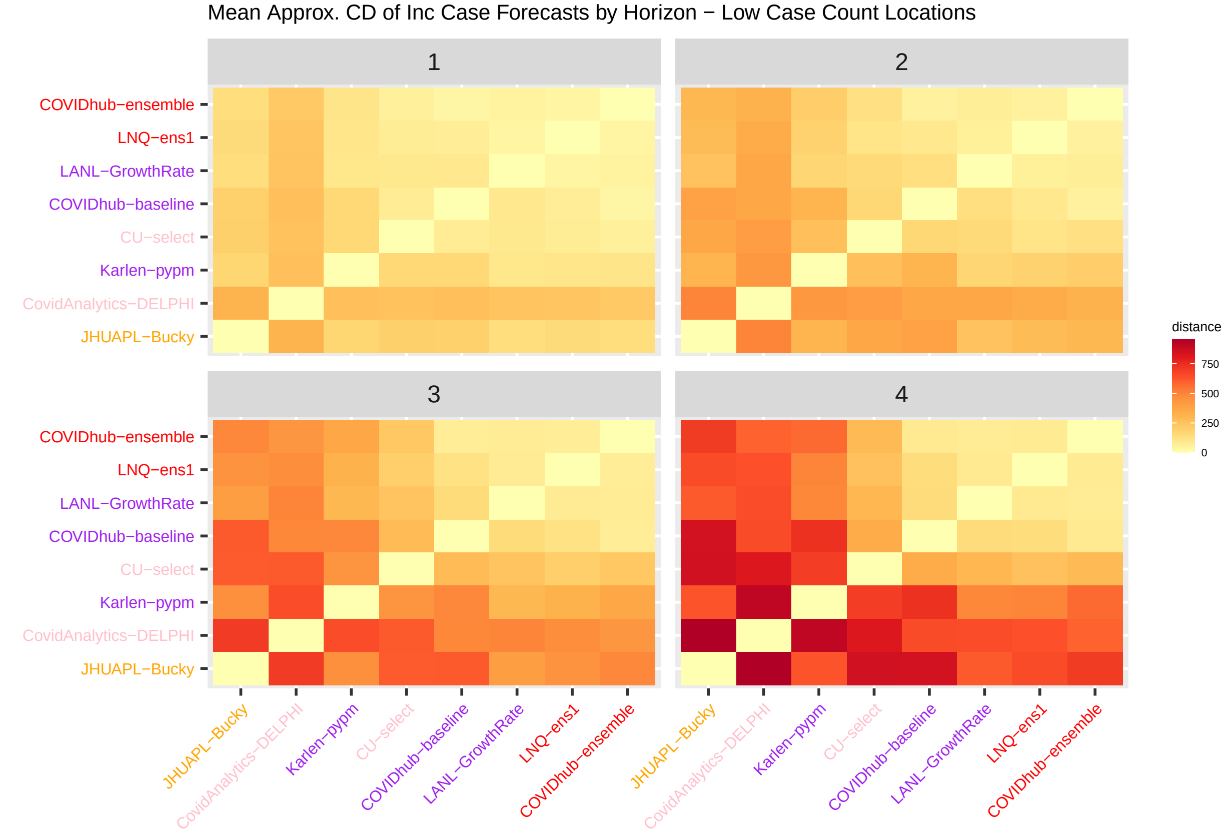
Anchor Model	Model	Horizon	Target	CD (uneq)	CD (eq)	Diff
COVIDhub-ensemble	COVIDhub-ensemble	1	inc case	0.00	0.00	0.00
COVIDhub-ensemble	COVIDhub-baseline	1	inc case	1604.19	1211.62	392.57
COVIDhub-ensemble	CU-select	1	inc case	1973.13	1276.39	696.74
COVIDhub-ensemble	RobertWalraven-ESG	1	inc case	3187.32	2418.77	768.55
COVIDhub-ensemble	Karlen-pypm	1	inc case	2976.27	2079.54	896.73
COVIDhub-ensemble	LANL-GrowthRate	1	inc case	3313.36	2008.50	1304.86
COVIDhub-ensemble	JHUAPL-Bucky	1	inc case	4220.34	2634.11	1586.23
COVIDhub-ensemble	CovidAnalytics-DELPHI	1	inc case	9276.59	6401.22	2875.38
COVIDhub-ensemble	COVIDhub-ensemble	2	inc case	0.00	0.00	0.00
COVIDhub-ensemble	RobertWalraven-ESG	2	inc case	4037.05	3149.83	887.22
COVIDhub-ensemble	COVIDhub-baseline	2	inc case	2969.32	2026.07	943.25
COVIDhub-ensemble	LANL-GrowthRate	2	inc case	4007.72	2243.03	1764.69
COVIDhub-ensemble	Karlen-pypm	2	inc case	5632.78	3825.92	1806.86
COVIDhub-ensemble	CU-select	2	inc case	5512.79	3424.65	2088.15
COVIDhub-ensemble	JHUAPL-Bucky	2	inc case	6926.10	4350.33	2575.77
COVIDhub-ensemble	CovidAnalytics-DELPHI	2	inc case	15097.39	10785.48	4311.92
COVIDhub-ensemble	COVIDhub-ensemble	3	inc case	0.00	0.00	0.00
COVIDhub-ensemble	RobertWalraven-ESG	3	inc case	5135.75	4242.86	892.89
COVIDhub-ensemble	COVIDhub-baseline	3	inc case	4066.76	2660.77	1405.99
COVIDhub-ensemble	LANL-GrowthRate	3	inc case	3655.31	2098.86	1556.45
COVIDhub-ensemble	Karlen-pypm	3	inc case	9972.79	6699.65	3273.14
COVIDhub-ensemble	CU-select	3	inc case	10781.08	7135.68	3645.40
COVIDhub-ensemble	JHUAPL-Bucky	3	inc case	10744.90	6857.48	3887.42
COVIDhub-ensemble	CovidAnalytics-DELPHI	3	inc case	21483.36	15699.46	5783.90
COVIDhub-ensemble	COVIDhub-ensemble	4	inc case	0.00	0.00	0.00
COVIDhub-ensemble	RobertWalraven-ESG	4	inc case	6492.18	5347.47	1144.71

Anchor Model	Model	Horizon	Target	CD (uneq)	CD (eq)	Diff
COVIDhub-ensemble	LANL-GrowthRate	4	inc case	3325.29	2070.15	1255.14
COVIDhub-ensemble	COVIDhub-baseline	4	inc case	4776.39	3028.06	1748.33
COVIDhub-ensemble	CU-select	4	inc case	13292.09	8902.87	4389.22
COVIDhub-ensemble	Karlen-pypm	4	inc case	15819.79	10907.20	4912.59
COVIDhub-ensemble	JHUAPL-Bucky	4	inc case	14631.70	9629.66	5002.04
COVIDhub-ensemble	CovidAnalytics-DELPHI	4	inc case	21771.17	15571.58	6199.59

Again, the differences for the approx. CD between CU-select and the ensemble case forecasts seem more pronounced compared to other models.

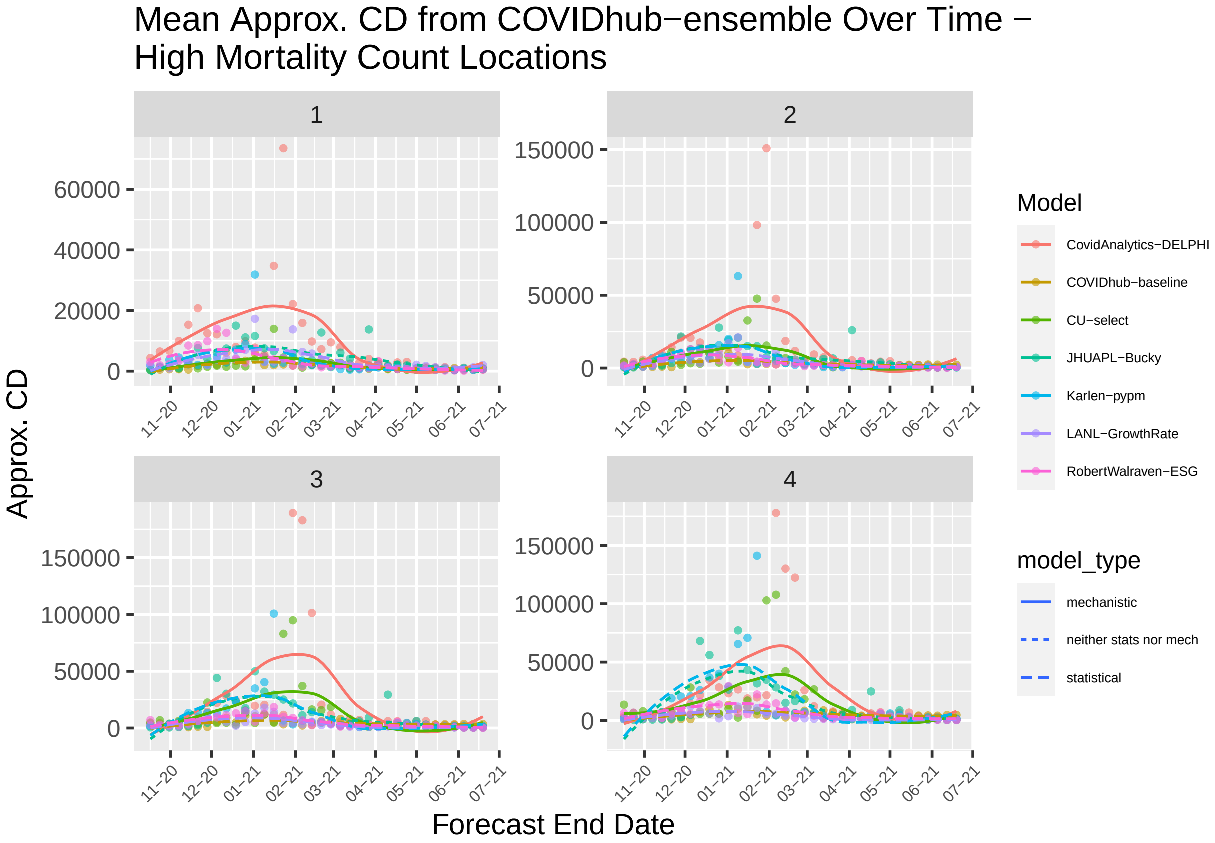
Mean approximated pairwise distances over across 5 high count and 5 low count locations

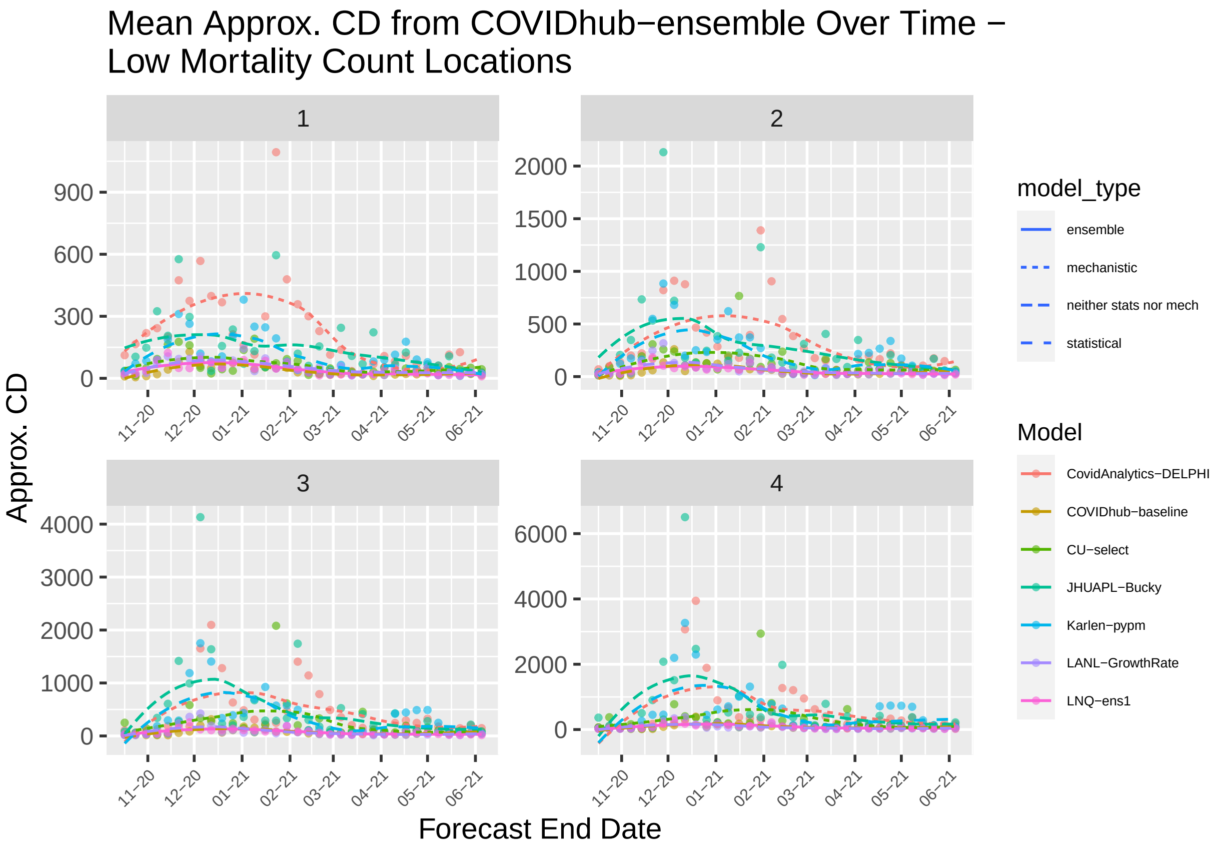




Interestingly, if we put the scale aside, forecasts are not significantly more dissimilar for high count locations compared to low count locations here (which is the case for inc death target). CovidAnalytics−DELPHI forecasts for both high and low count locations seem to be more dissimilar to other models on average across all forecast horizons.For low count locations, JHUAPL−Bucky and Karlen-pypm (for 3-4 wk ahead) are also more dissimilar.

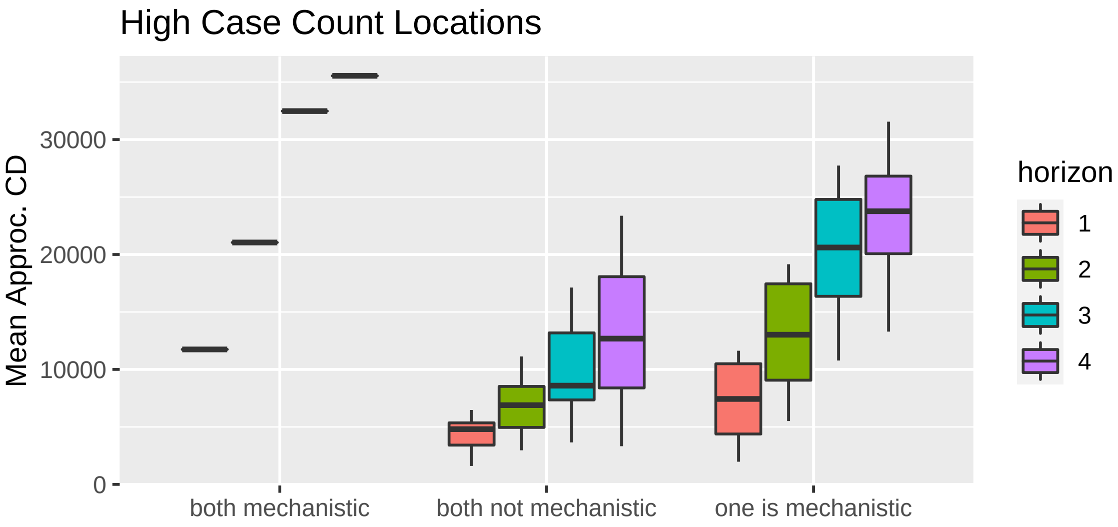
When we look at the approximated pairwise distances over time, we see high distances from the ensemble around Jan-Feb 2021 for high count locations, while we see that about a month earlier for low count locations.

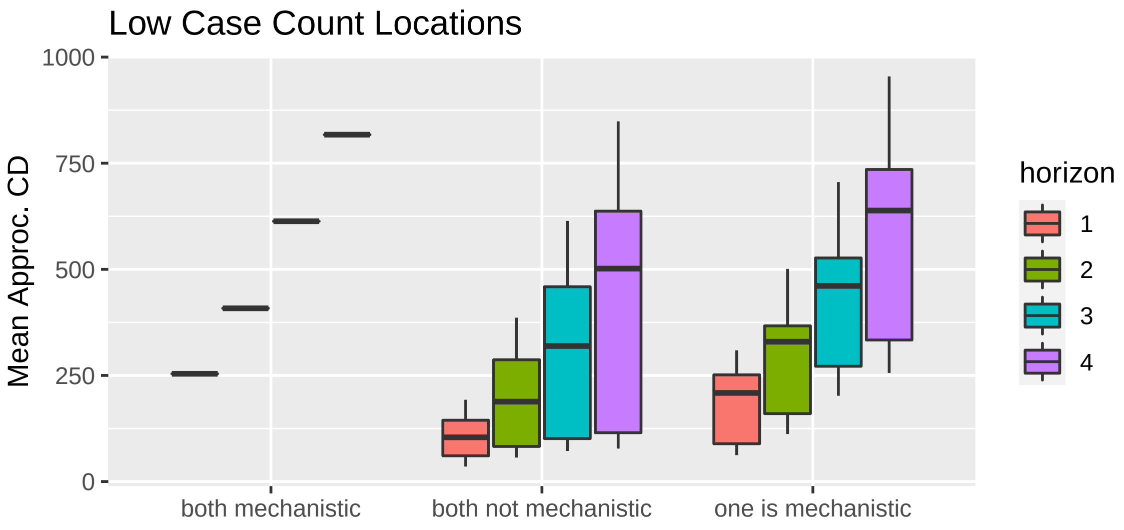




Relationship between a mechanistic model type and similarity

These are the same plots as in the previous section for inc death forecasts, but for inc case forecasts. It seems there is only one pair of model that are both mechanistic. We see higher medians of the mean approx. cd when one of the models in a pair is mechanistic for both high and low count locations.





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

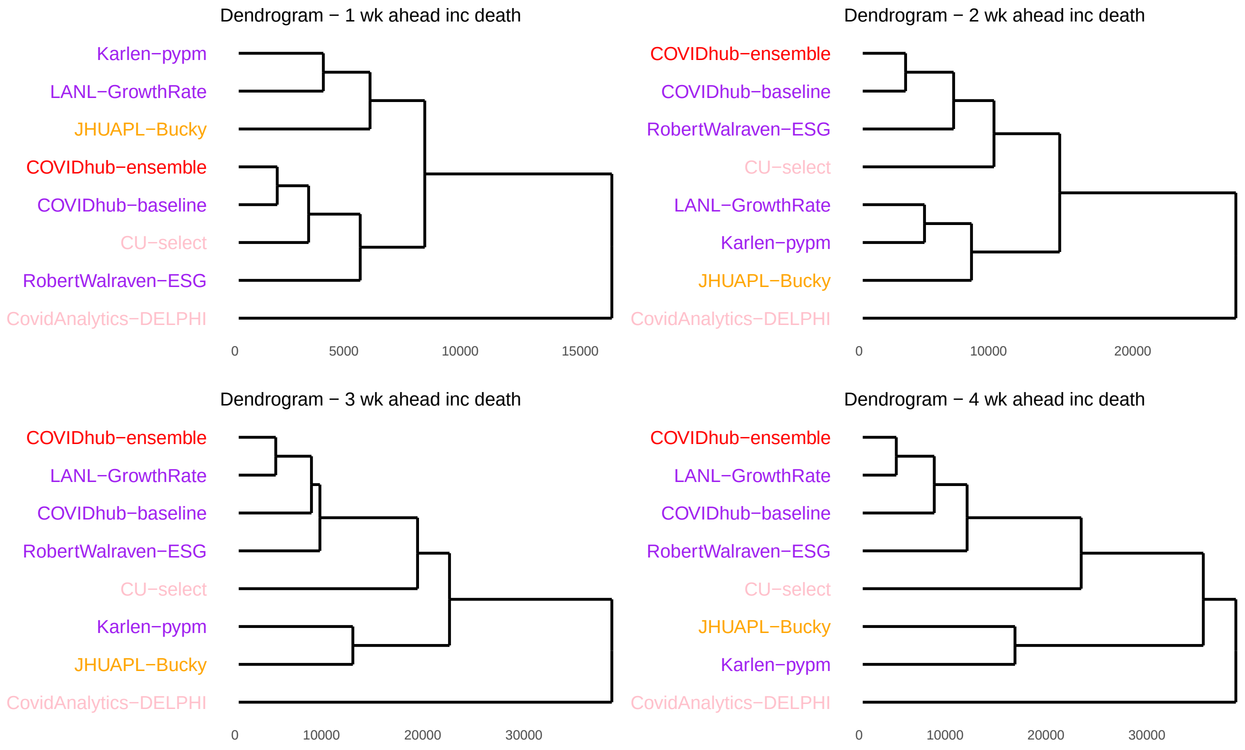


Figure 3: High Case Count Locations

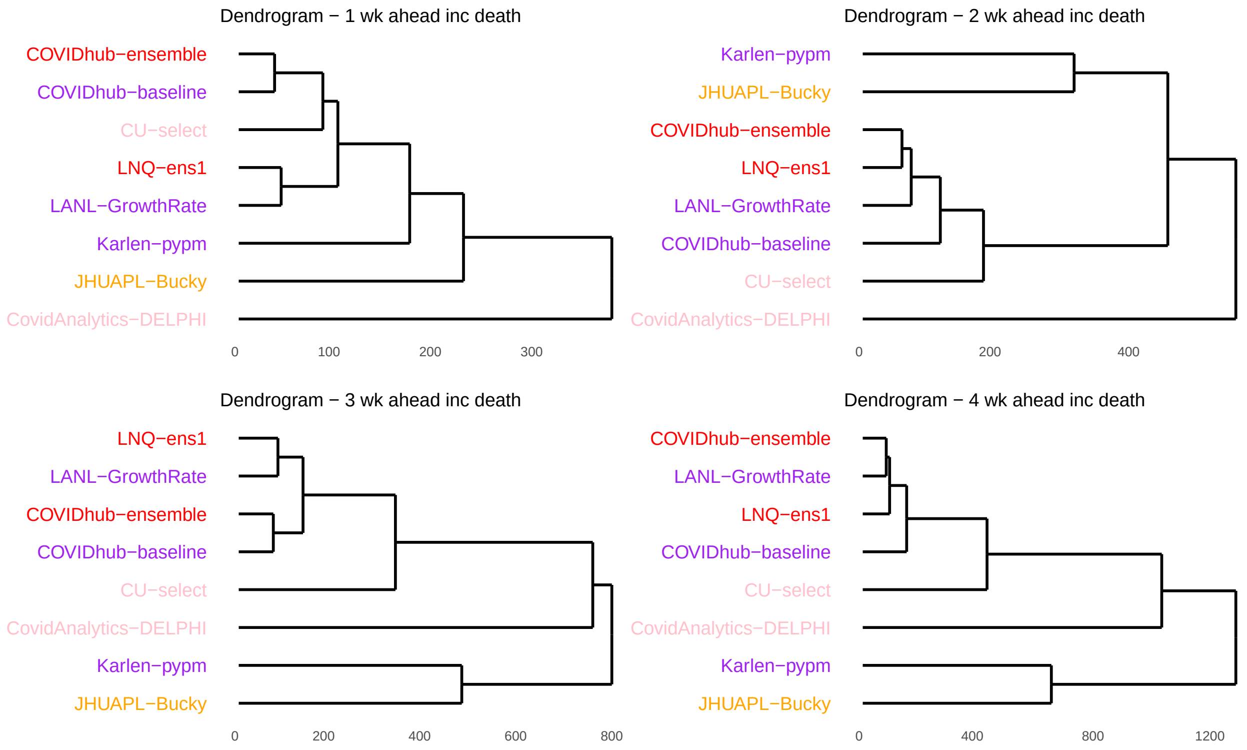


Figure 4: Low Case Count Locations