COVID-19 Forecast Similarity Analysis

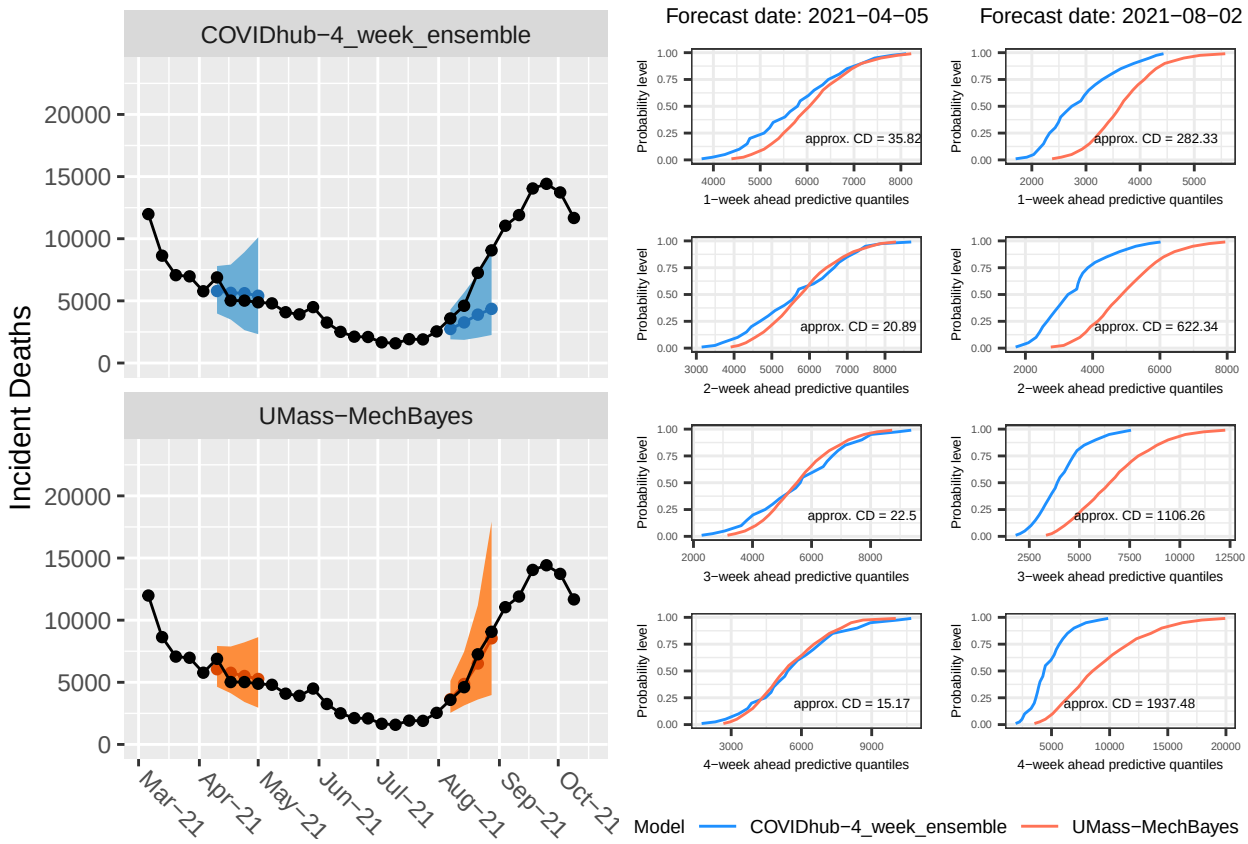
Nutcha Wattanachit, Johannes Bracher, Evan Ray, Nick Reich

04/28/2022

Forecast inclusion criteria

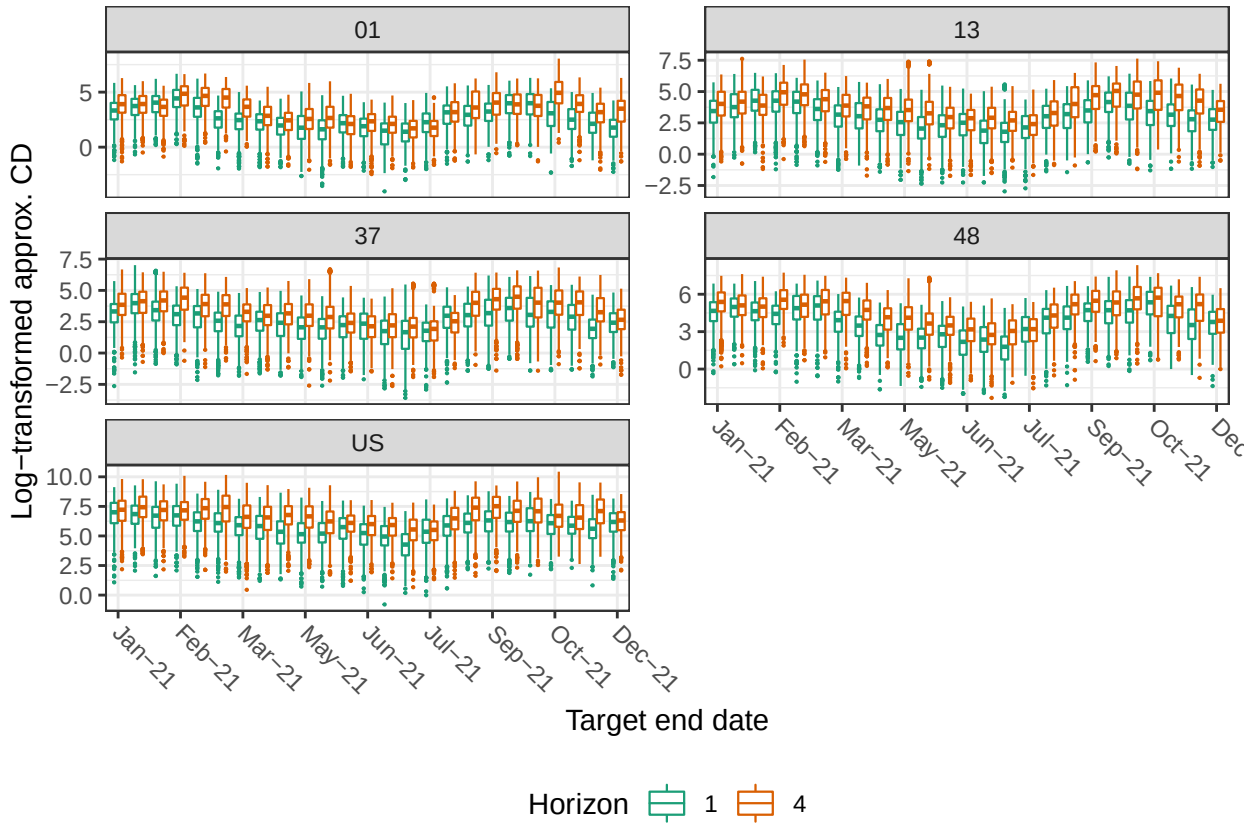
- Targets: 1-4 wk ahead inc death
- Target End Dates: Jan-Dec 2021 for general analyses and summer/fall 2021 for the model type comparison.
- Probability levels: All
- Locations: Varies, depending on the analysis

Examples of forecasts and Cramer distances - US National



Similarity - 1 and 4 week-ahead horizons

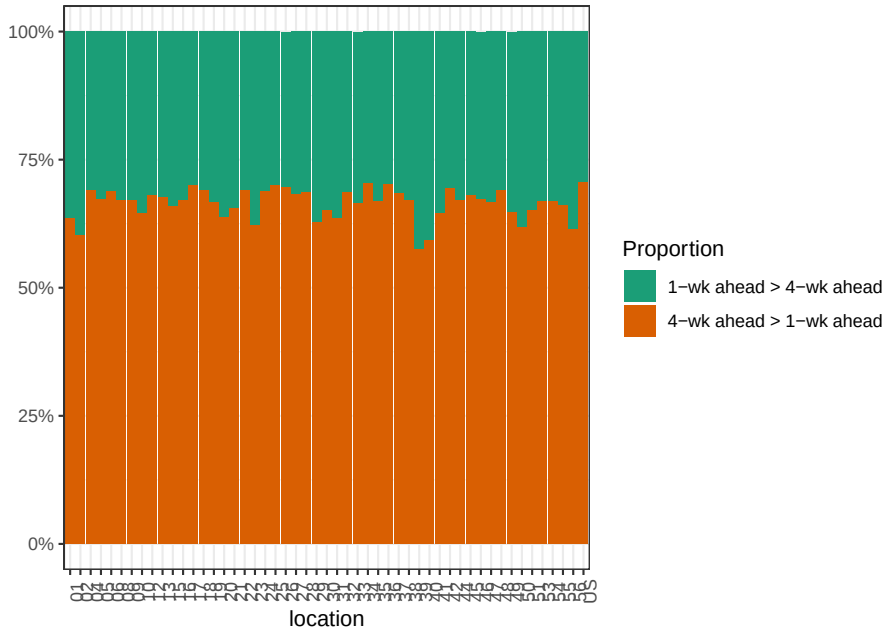
The figure shows log-transformed approximated CD for all 1 and 4 week-ahead forecasts submitted for each target end date in 2021. To avoid overcrowding, the plot only shows every other week across the date range for 4 states and US national.



- From eyeballing these panels, we can see that generally forecasts made 4 weeks from the target end date exhibit more dissimilarity in predictive quantiles compared to forecasts made 1 week from the target end dates.

Proportion of median distances among 4-wk ahead forecasts being higher than median distances among 1-wk ahead forecasts

For each target end date and location, we can compare the median of approximated CD of 4-wk ahead forecasts to that of 1-wk ahead forecasts. For each location, the proportions are calculated based on medians of each end date. All locations are included here.



- The proportion of the median of 4-wk ahead CDs being higher than the median of 1-wk ahead CDs (for forecasts of incident death for the same target end date) is higher than vice versa for all locations, which supports the previous observation made from the previous boxplot. Similar to what we observed with WIS degrading at further horizons, models produced less similar forecasts at further horizons.

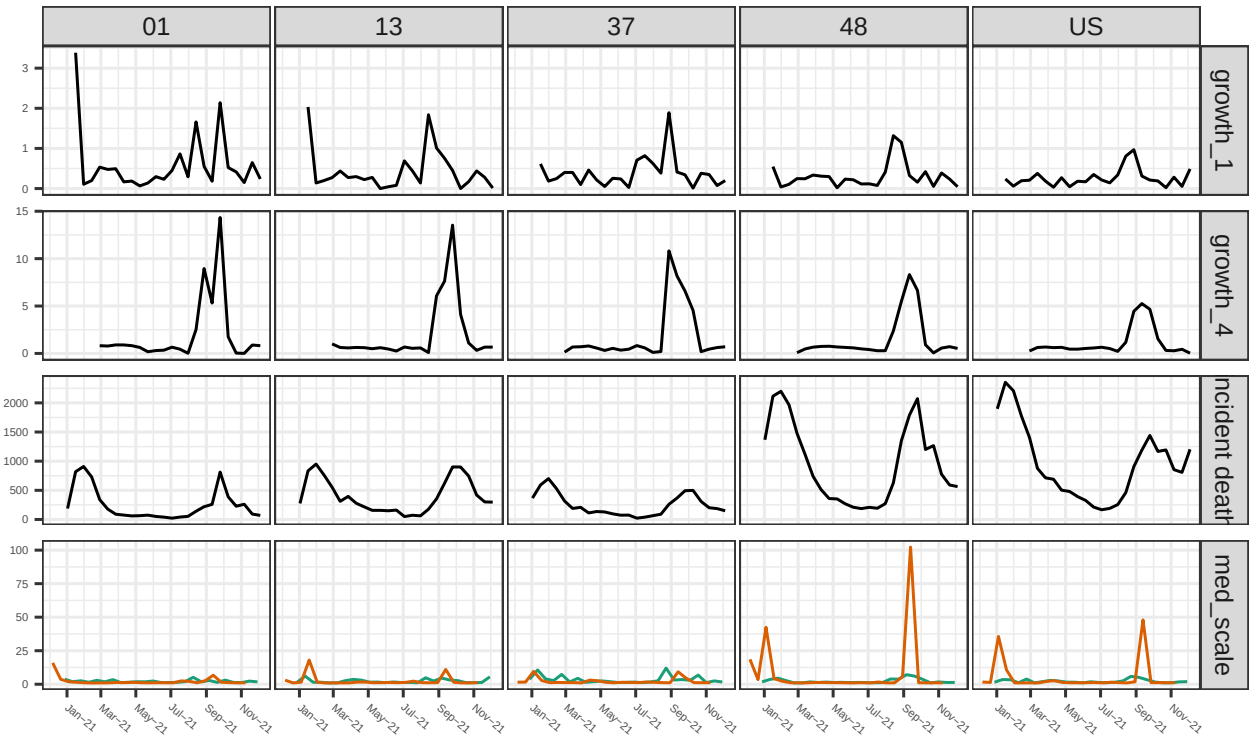
Distance/dissimilarity as a signal of a change in trends

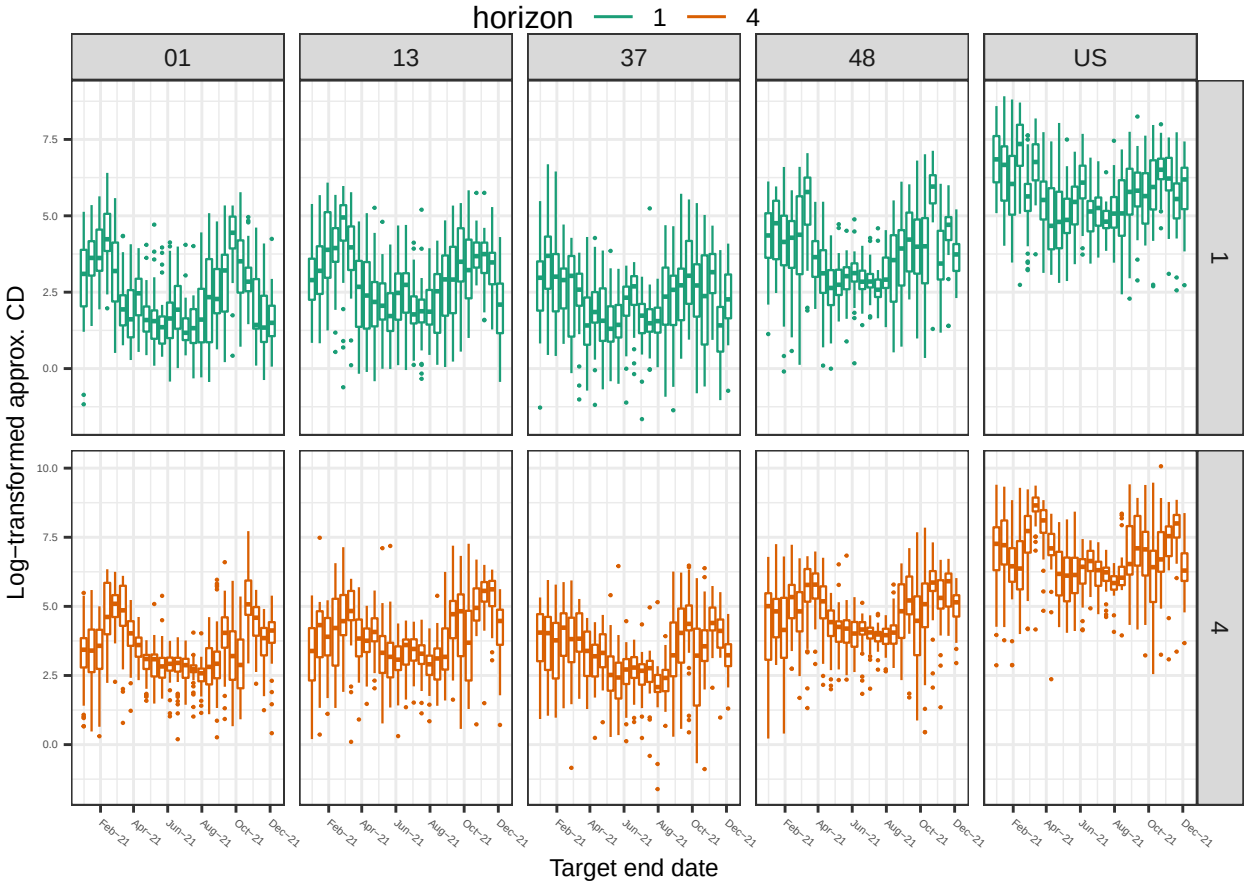
We are interested in looking at dissimilarity measure across dates to see if we see any trends prior to a change in observed incident deaths. On the plot are:

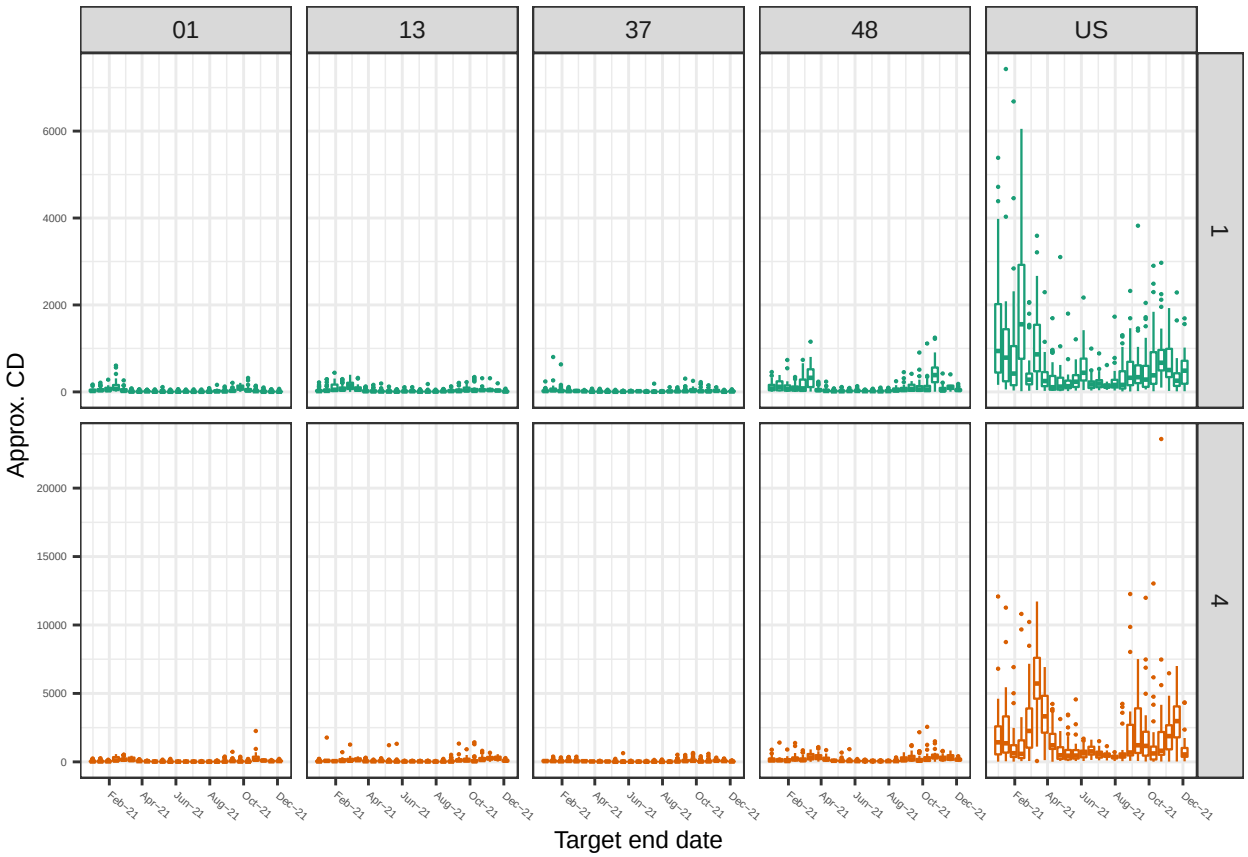
- The `growth_1` row is the 1-week absolute growth rate (absolute increase or decrease from previous week)
- The `growth_4` row is the 4-week absolute growth rate (absolute increase or decrease from previous 4 weeks)
- Scaled CD is defined as $\frac{\text{CD(baseline,model)}}{\text{CD(baseline,ensemble)}}$. This is a measure of how dissimilar a forecast from a model is to the baseline forecast, relative to how dissimilar a baseline forecast is to the ensemble forecast (the median/central forecast). To state simply, it measures the similarity/dissimilarity of a model's forecast relative to the central forecast (ensemble).
- The median scaled CD rows are now plotted by forecast date.
- The last row is the median of scaled CD at a given target end date. Each point in a line is the CD of forecasts made 1 or 4 weeks prior to the date on the x axis.

Note that the truth for US is divided by a factor of 10 so it can be plotted together with other states.

- We see increasing scaled dissimilarity among forecasts made for target end dates during the wave. There seems to be some relationship between 1-wk absolute changes and scaled CD among 1-wk ahead forecasts, and the same can be said about 4-wk change and 4-wk ahead forecasts.
- This could be useful when we are a few weeks away from the beginning of a wave because we can anticipate a change in trends of incident deaths from the dissimilarity among 1-4 wk ahead forecasts.
- Should we calculate some correlation between two time series here?


01
13
37
48
US
growth_1
growth_4
ncident death
med_scale
Jan-21
Mar-21
May-21
Jul-21
Sep-21
Nov-21

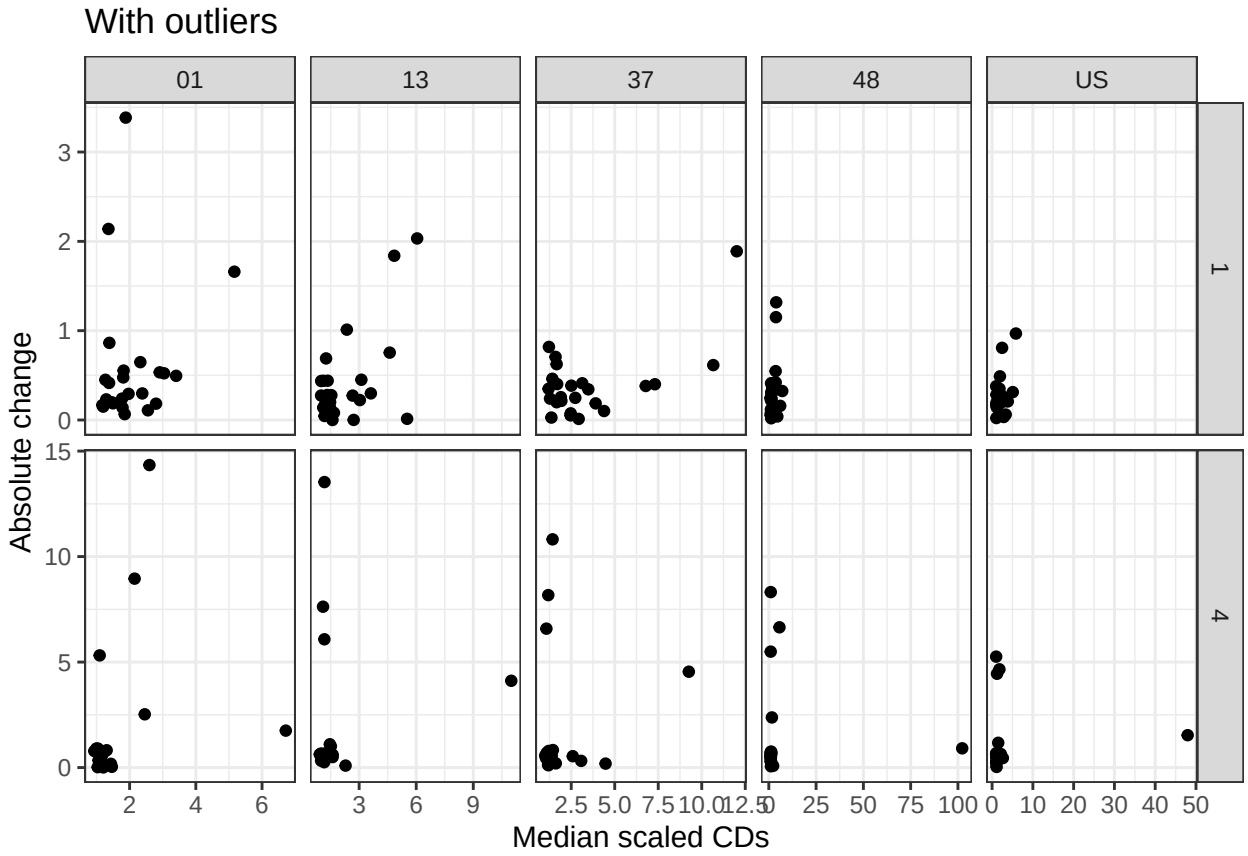

horizon
1
4
01
13
37
48
US
Log−transformed approx. CD
Target end date
Feb-21
Apr-21
Jun-21
Aug-21
Oct-21
Dec-21

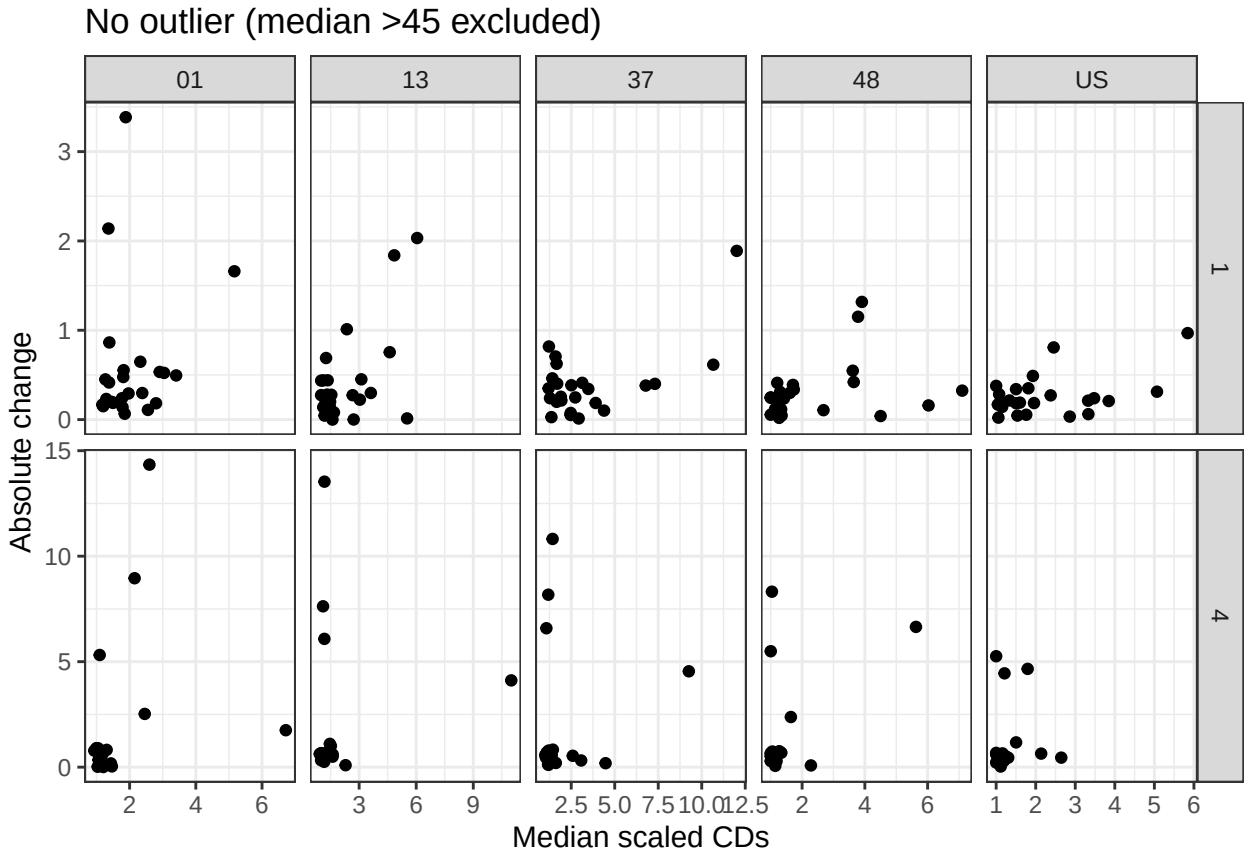


- This plot compares the raw CD and the transformed scaled CD to give us more information how the CDs change post transformation. Raw CDs roughly follow the observed incident deaths due to scale.

Scatterplots of median scaled CDs vs absolute change

I excluded a couple of outliers (median scaled cd »> 45).


With outliers
01
13
37
48
US
1
4
Absolute change
Median scaled CDs


No outlier (median >45 excluded)
01
13
37
48
US
1
4
Absolute change
Median scaled CDs

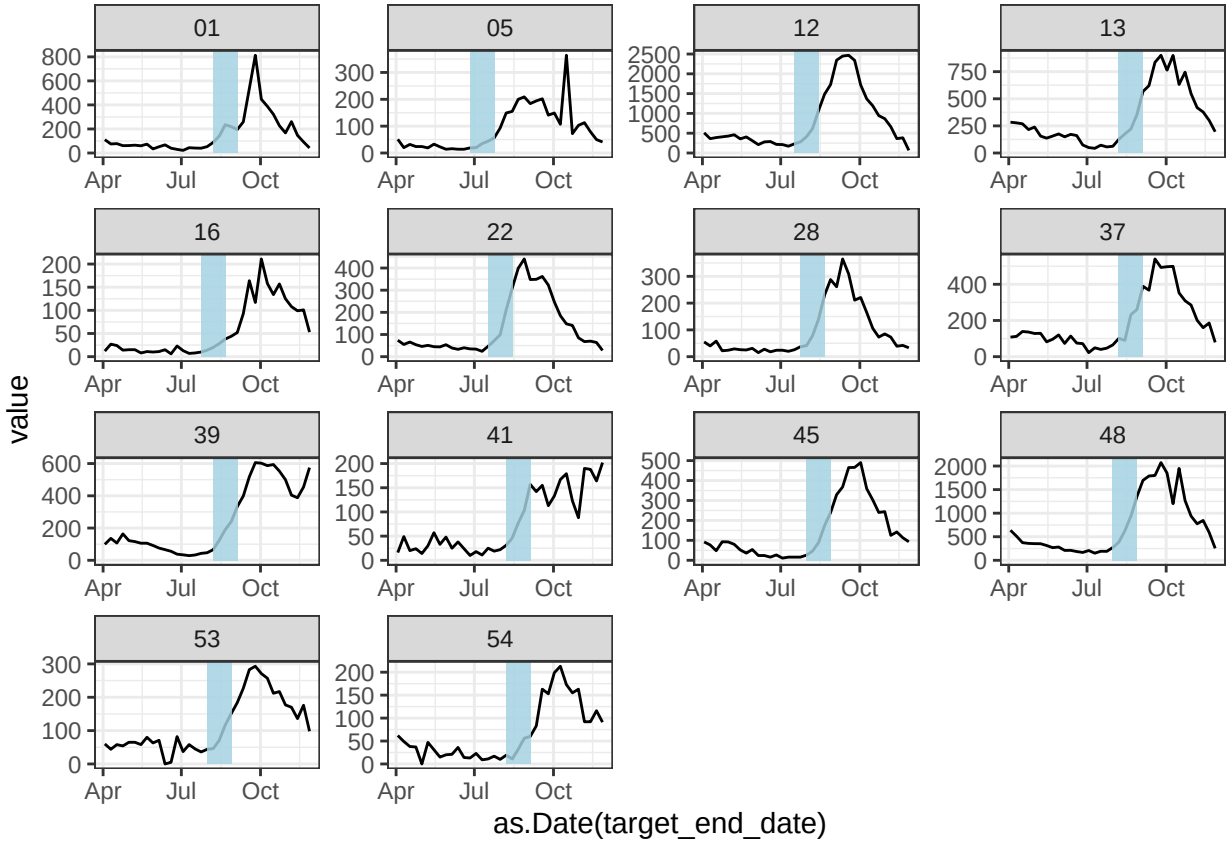
Similarity by type

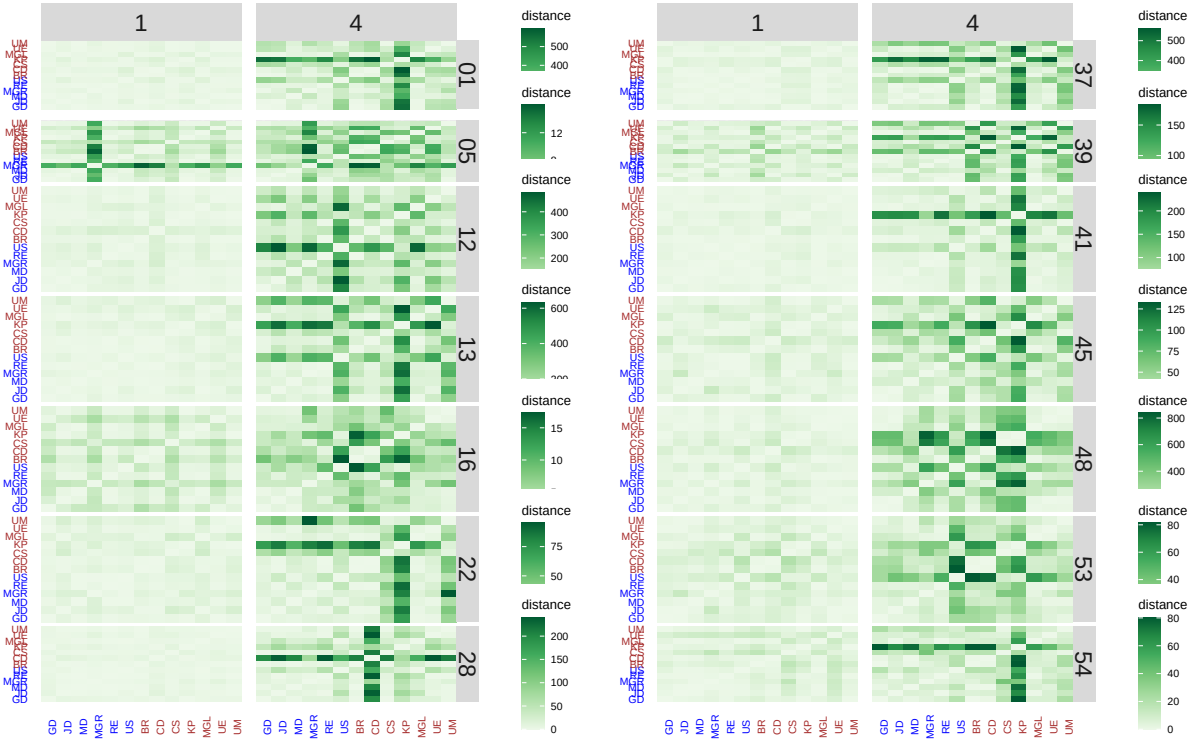
The models are selected based on their submission on the date designated as the moment of change dates and the number of submissions overall during 2021. We have 6 stats/ML models and 7 mechanistic models. Model included in the analysis are:

```
FALSE               model_abbr             type
FALSE 1       BPagano.RtDriven      mechanistic
FALSE 2  CovidAnalytics.DELPHI      mechanistic
FALSE 3              CU.select      mechanistic
FALSE 4          GT.DeepCOVID data_adaptive/ML
FALSE 5         JHU_CSSE.DECOM data_adaptive/ML
FALSE 6            Karlen.pypm      mechanistic
FALSE 7     Microsoft.DeepSTIA data_adaptive/ML
FALSE 8      MIT_CritData.GBCF data_adaptive/ML
FALSE 9       MOBS.GLEAM_COVID      mechanistic
FALSE 10    RobertWalraven.ESG data_adaptive/ML
FALSE 11           UA.EpiCovDA      mechanistic
FALSE 12       UMass.MechBayes      mechanistic
FALSE 13        USC.SI_kJalpha data_adaptive/ML
```

Moment of change before the summer/fall 2021 wave of incident deaths

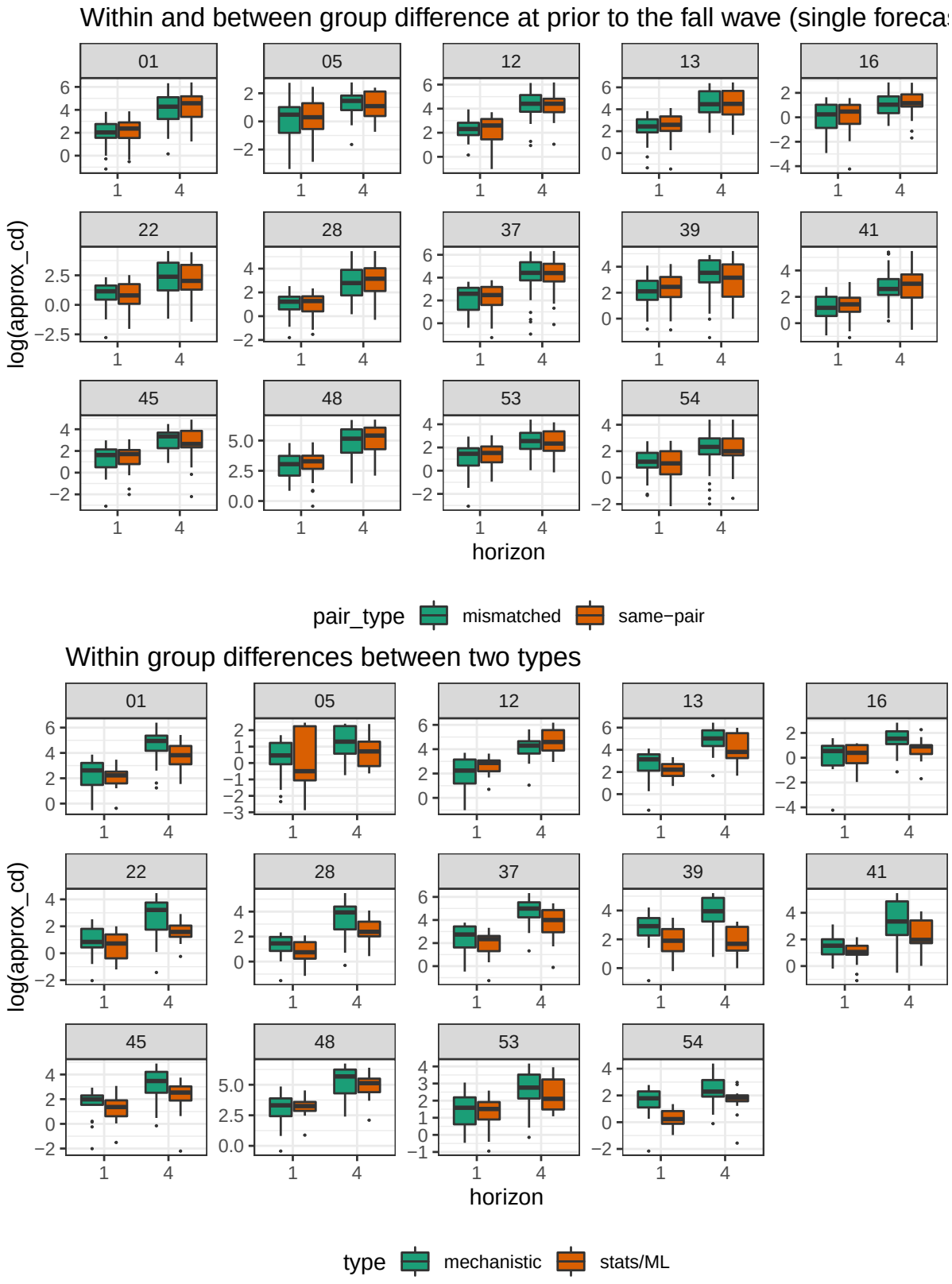
Below are the plots of incident deaths in 14 locations that experienced a noticeable summer wave in 2021 (they are selected based on objective criteria). The shaded portions show the target end dates (we only test the first and the last dates of the portion) selected for testing.





- The plots are by model acronyms. Blue are stats/ML and Orange are mechanistic models. Karlenpypm seems to be dissimilar from other models for all 14 locations.
- Looking at this by quadrants, there seems to be more dissimilarity among mechanistic models than among stats/ML models. There are some dissimilarity between groups, but not as strong as the first relationship.

Boxplots of log-transformed approx. CDs by categories



- Plotted on a log scale since these are right-skewed.

- Despite variations, the median of distances between different groups and the median of distances within group are not that different across all locations at the beginning of a wave. This agrees with the heatmap.
- We see more distinct differences between distances among stats/ML models anddistances among mechanistic models. Mechanistic models are dissimilar among one another at the beginning of a wave.

Permutation test

To check if the dissimilarity between groups is significant prior to observing a wave, we did a permutation test. The hypotheses are

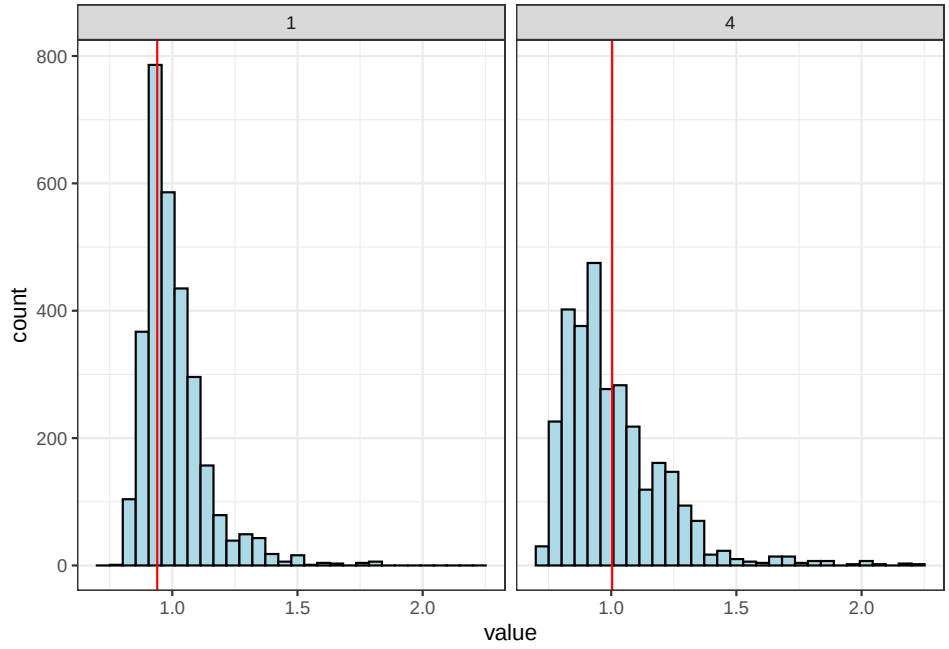
- H_0 : Prior to an increase in deaths, forecasts from different model types are as similar as forecasts from same model types.
- H_A : Prior to an increase in deaths, forecasts from different model types are more dissimilar compared to forecasts from the same model types.

This simple two-way design permute forecasts within each location and keep the permutation order constant across location. This will keep the location-specific variability constant across all permutations. The test statistic is

$$T_{cd} = \frac{\text{median(between-group approx. CD)}}{\text{median(within-group approx. CD)}}.$$

The ratio of medians can alleviate an issue of outliers in the distances that would effect the mean. We have 4 sets off hypotheses for each horizon. All possible number of permutations for each test is 14!/(6!7!). For now I shuffle 3000 times.

```
FALSE [1] "P-values: 0.670333333333333,0.414333333333333"
```

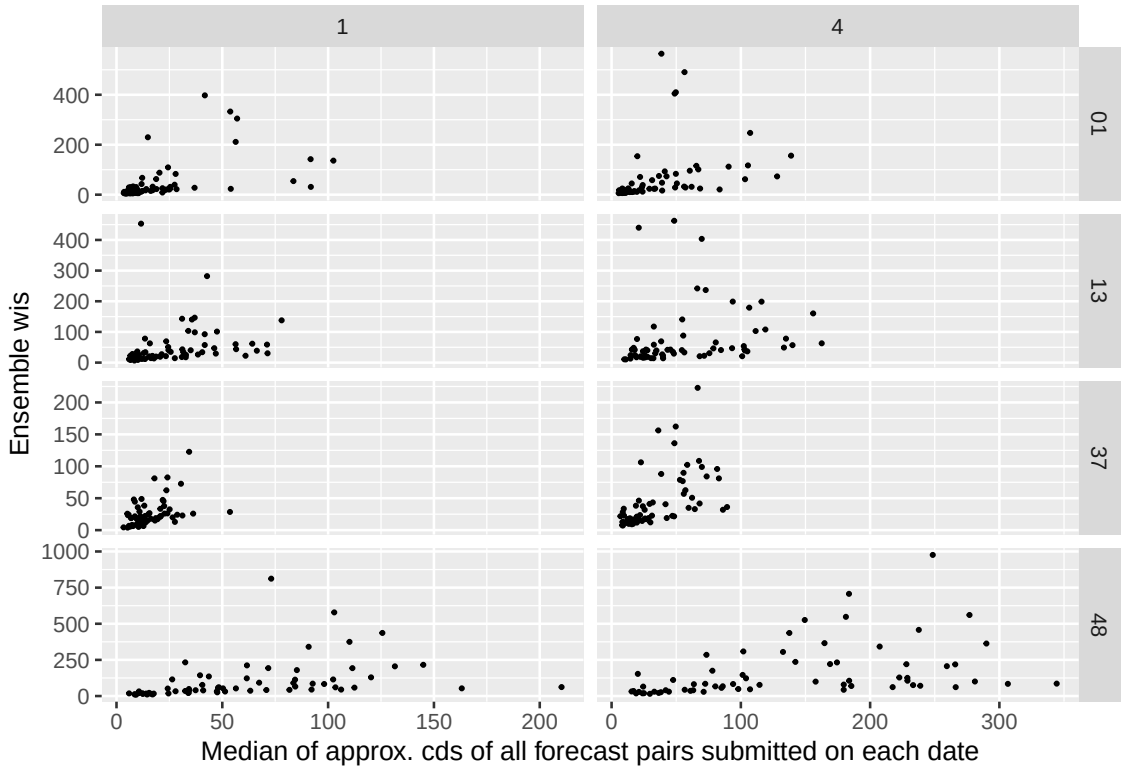


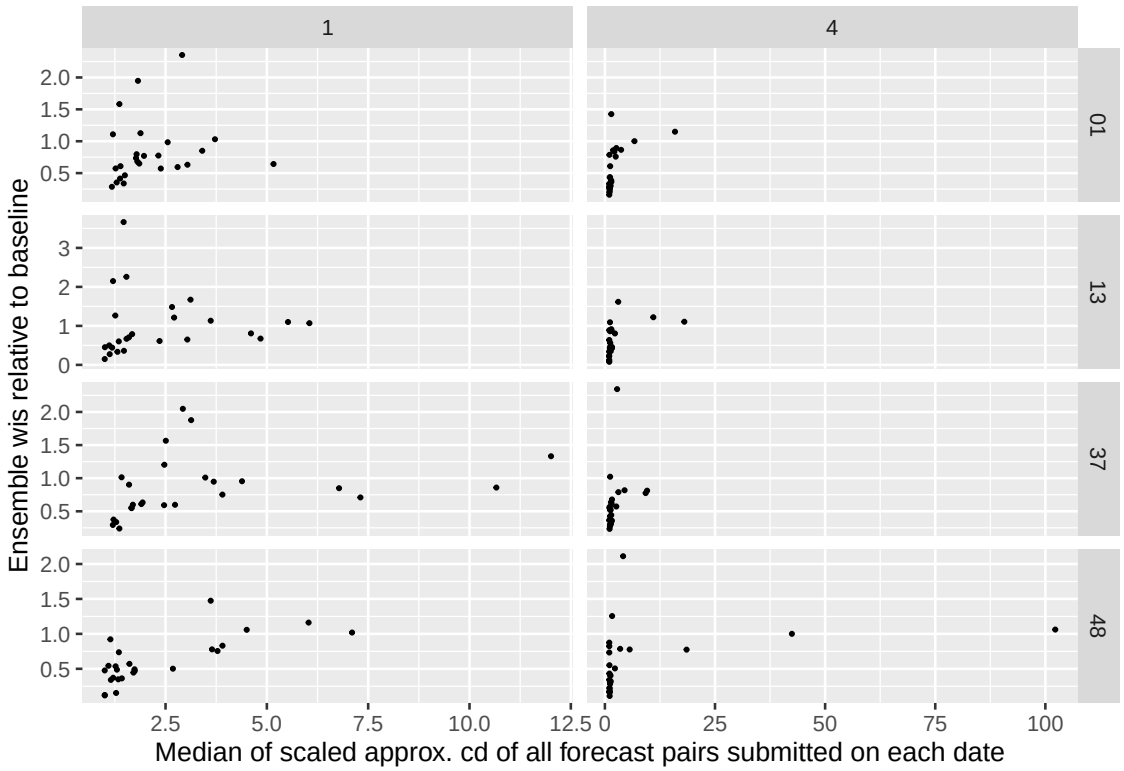
- None of the p-values are significant. We probably don't have to adjust for correlated tests given that we didn't do a lot of tests for it to be concerning and the results won't change. We fail to reject H_0 that prior to an increase in deaths in the summer/fall of 2021, forecasts from different model types are as similar as forecasts from same model types

- I think we have enough power given the total possible number of permutations are not that small, but we will need simulation if this testing framework make sense.
- It's possible that ML models' forecasts are similar things at the beginning of the wave, and mechanistic models are all over the place. We can test the hypothesis that (median) dissimilarity among stats/ML models is lower compared to (median) dissimilarity among mechanistic models right before we see the wave.

Median dissimilarity vs ensemble WIS

The scaled approx. CDs here are the same as the ones in the earlier section.





- Maybe a bit of a positive correlation here?