COVID-19 Forecast Similarity Analysis for Hospitalizations

Li Shandross, Nutcha Wattanachit, Nick Reich, Evan Ray

07/30/2021

Over the course of the pandemic, many teams across the United States have worked to create forecasting models for Covid-19 cases, deaths, and hospitalizations. We choose to examine models specifically forecasting incident hospitalizations. But how similar are these models? Are there patterns of similarity, perhaps due to data source, modeling technique, or another incorporated factor? Or do models that tend to perform well tend to be similar? These questions are all reasons why we might be interested in measuring the similarities between covid-19 forecasting models.

Introduction to Cramér's Distance

Extending the work of Bracher et. al, we select Cramér's Distance as a metric to evaluate the similarity between models. The Cramér's Distance of two predictive distributions F and G is defined as follows:

$$\text{CD}(F, G) = \int_{-\infty}^{\infty} (F(x) - G(x))^2 dx \tag{1}$$

where $F(x)$ and $G(x)$ are the two cumulative distribution functions respectively.

However, we actually only know K quantiles from F and G rather than their entire distributions. Thus, we must approximate their Cramér's Distance, and we do so by using Bracher et. al's trapezoidal rule:

$$\text{CD}(F, G) \approx \sum_{j=1}^{2K-1} \int_{q_j}^{q_{j+1}} F(x) - G(x)^2 \tag{2}$$

$$\approx \sum_{j=1}^{2K-1} \frac{\{\hat{F}(q_j) - \hat{G}(q_j)\}^2 + \{\hat{F}(q_{j+1}) - \hat{G}(q_{j+1})\}^2}{2} (q_{j+1} - q_j) \tag{3}$$

where q_j is the jth quantile.

Cramér's Distance (as well as it's approximation) is an ideal quantitative measurement for measuring the similarity of covid-19 forecasting models because it is not only relatively easy to compute but also is compatible with the weighted interval score (WIS) used to score forecasts by the COVID-19 Forecast Hub.

Below, Figure 1 illustrates how Cramér's Distance is calculated between two models in this analysis. Forecasts made on a single date are split into 1 to 4 week horizons, with Cramér's Distance calculated separately at each horizon.

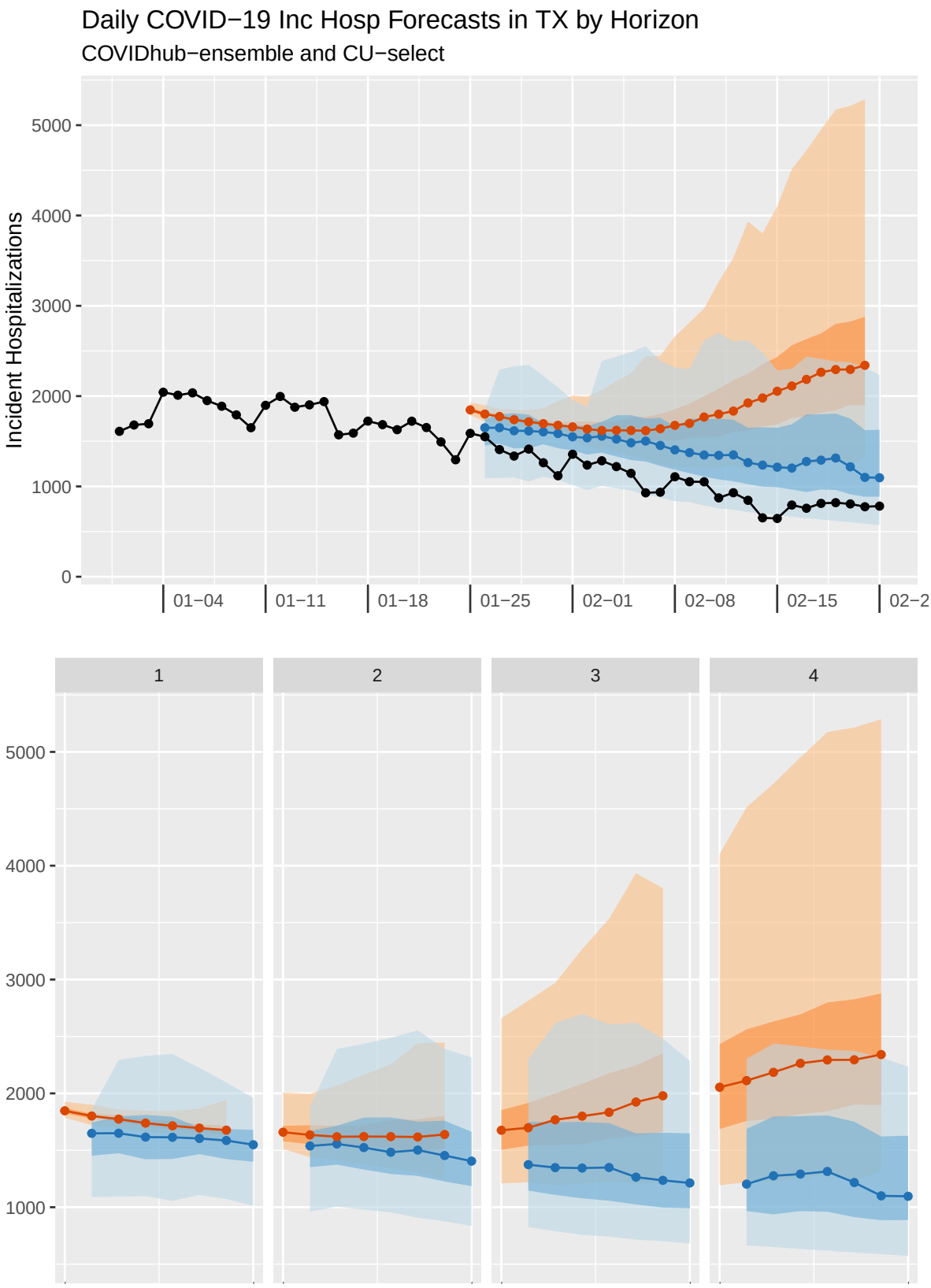


Figure 1: Cramér's Distance Visualized

Horizon	Approx. CD
1	64.91
2	26.24
3	118.61
4	383.93

Aligning Hospitalization Forecasts

Bracher et. al's work on model similarity specifically focused on incident case and death (inc case and inc death, respectively) forecasts, while this analysis is focusing on incident hospitalizations. Incident hospitalization (inc hosp) data has daily targets instead of weekly ones, like inc case and inc death data. This presents an issue with unaligned forecasts because each model only makes predictions on a single day per week but not all models make their forecasts on the same day. That is, forecasts made in the same week but on different days will be predicting for different target end dates, even if they share the horizon. (This is not an issue when the temporal resolution is in terms of weeks, which are defined by epidemiological week, not the number of days between forecast date and target end date.) Thus, we create a new relative horizon variable variable called horizon week to prevent unaligned forecasts. This variable counts horizons between 1 and 7 days to have a horizon week of 1, horizons between 8 and 14 days to have a horizon week of 2, etc. With this new variable, we can easily apply similar analyses from Bracher et. al to inc hosp data.

One potential disadvantage lies in using the horizon week variable. Since models only make forecasts once per week, models that make forecasts later in the week may have an unfair advantage of additional days worth of data informing their forecasts. We will consider and investigate this issue later in the report.

Forecast Inclusion Criteria

The pairwise approximated Cramér's distances are calculated for the models that have complete submissions for all targets, all probability levels, and no missing forecasts between January 28th, 2021 and June 10th 2021. We aggregate results for the five locations that have the highest number of cumulative COVID-19 hospitalizations during this period as well as the five locations with the lowest number, then perform the analysis on both "high count" and "low count" groupings.

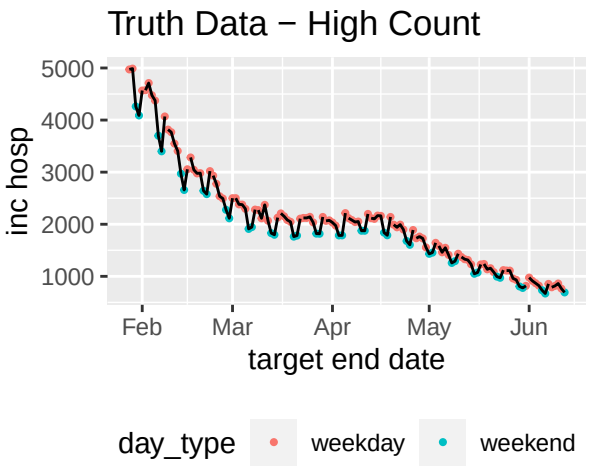
The high count locations are Florida, Texas, New York, California, Pennsylvania while the low count locations are Vermont, Hawaii, Arkansas, Wyoming, South Dakota.

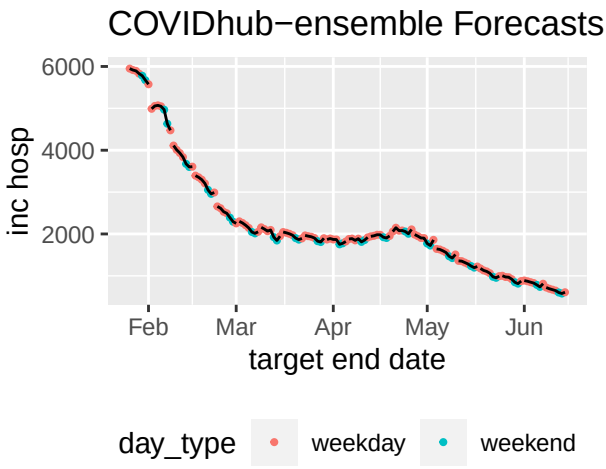
There are nine models that fulfilled the criteria for both the five high count and five low count locations for Thursday forecasts. There are ten that fulfill such criteria for the Saturday forecasts, but we perform the analysis on only the overlapping nine models.

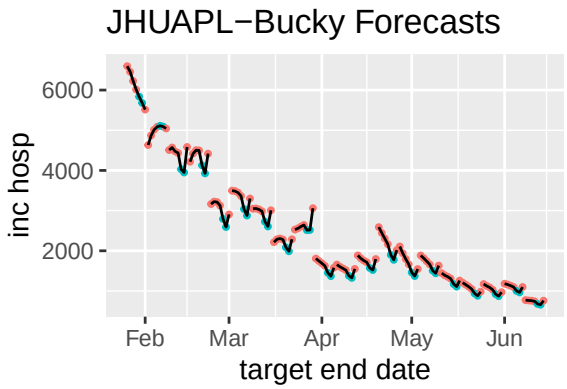
Day of Week Effects

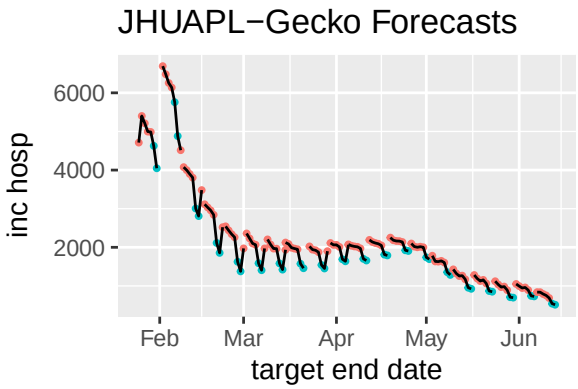
This analysis looks into whether the incorporation of day of the week effects impacts model similarity, e.g. do models that have day of the week effects more similar to each other than those without. We define day of the week effects to be cyclic weekly patterns for which a specific day or group of days specific show higher or lower incident hospitalization values compared to other days. The hospitalization truth data for high count locations clearly shows a day of the week effect in which weekends have noticeably fewer incident hospitalizations in comparison to weekdays. However, hospitalization truth data for low count locations is a little less clear.

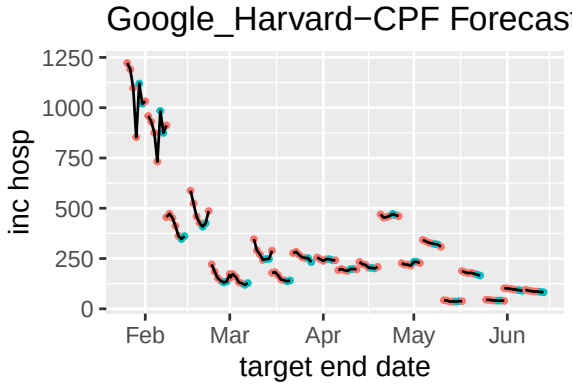
We plot point forecasts with a horizon week of 1 to determine which models include day of the week effects, first looking at high count locations, then low count locations.

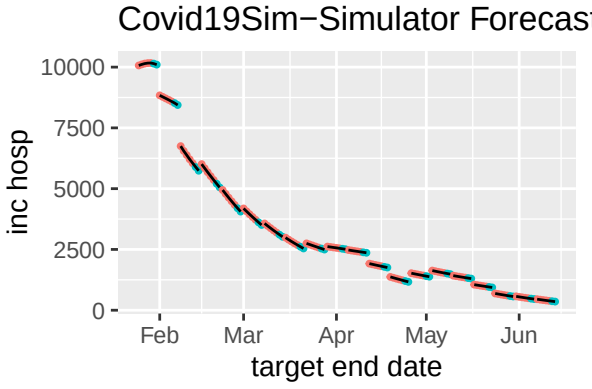

Truth Data − High Count
inc hosp
5000
4000
3000
2000
1000
Feb
Mar
Apr
May
Jun
target end date
day_type
weekday
weekend

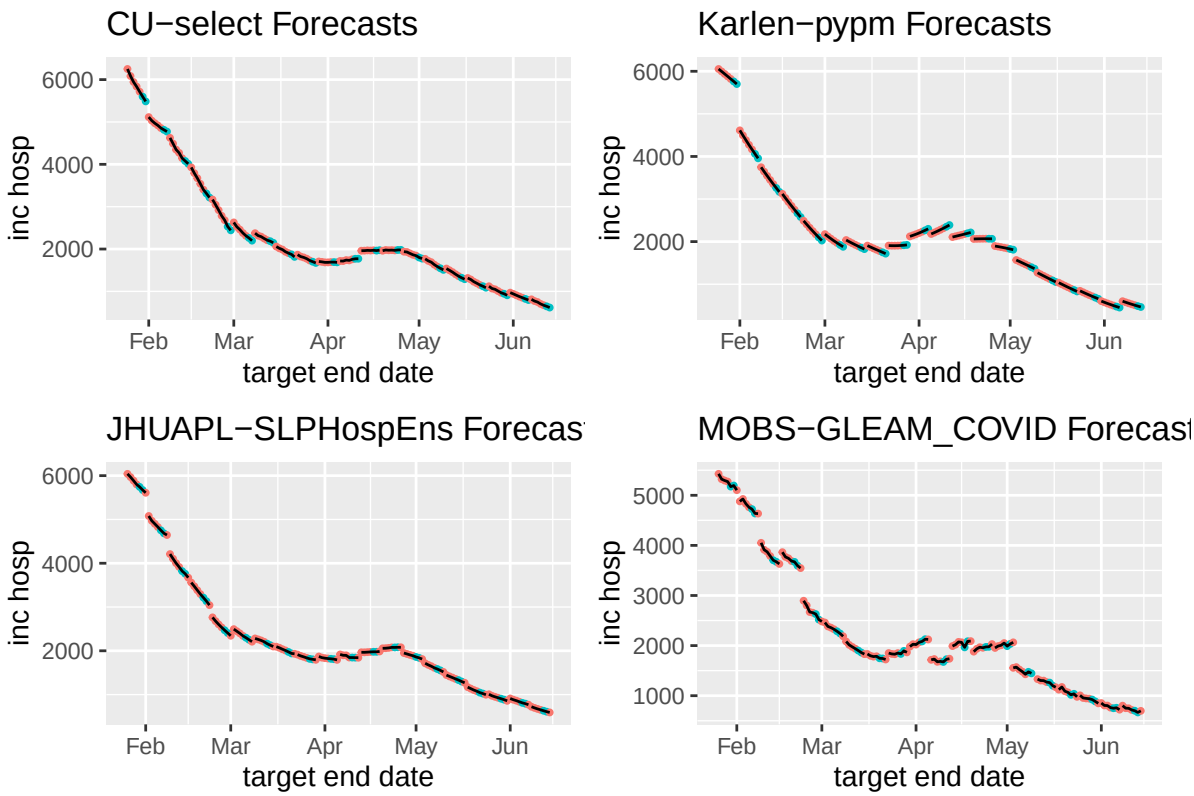

COVIDhub−ensemble Forecasts
inc hosp
6000
4000
2000
Feb
Mar
Apr
May
Jun
target end date
day_type
weekday
weekend


JHUAPL−Bucky Forecasts
inc hosp
6000
4000
2000
Feb
Mar
Apr
May
Jun
target end date

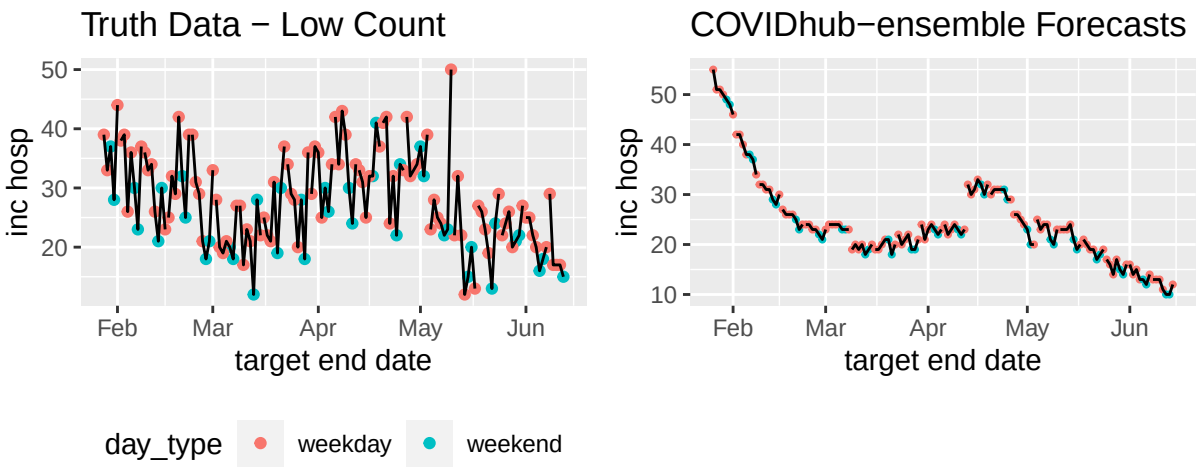

JHUAPL−Gecko Forecasts
inc hosp
6000
4000
2000
Feb
Mar
Apr
May
Jun
target end date

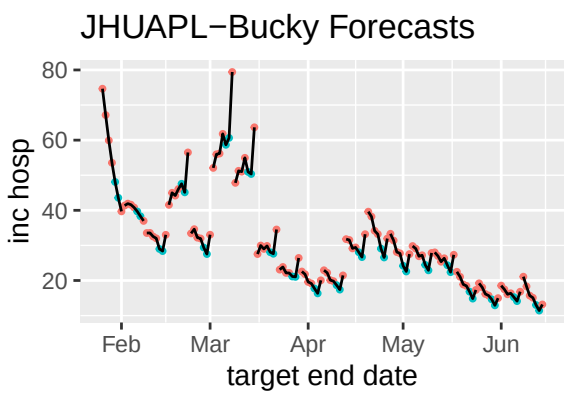

Google_Harvard−CPF Forecast
inc hosp
1250
1000
750
500
250
0
Feb
Mar
Apr
May
Jun
target end date

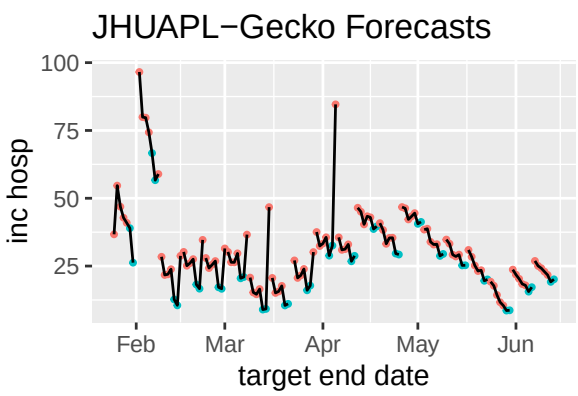

Covid19Sim−Simulator Forecasts
inc hosp
10000
7500
5000
2500
0
Feb
Mar
Apr
May
Jun
target end date

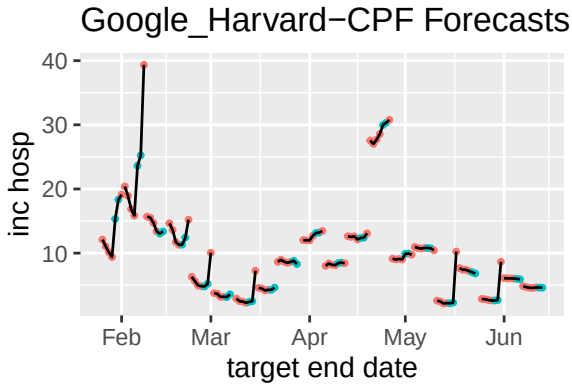


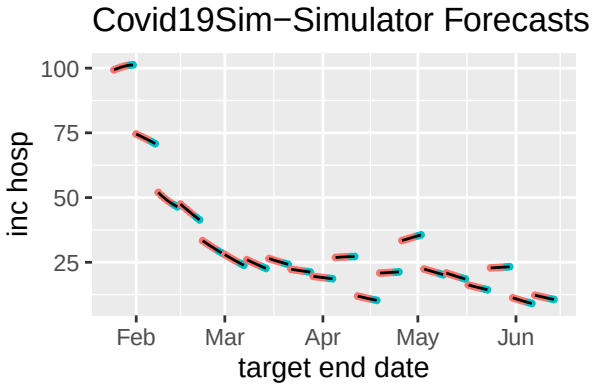
We can see that only two models, JHUAPL-Bucky and JHUAPL-Gecko, incorporate day of week effects based on the high count data.

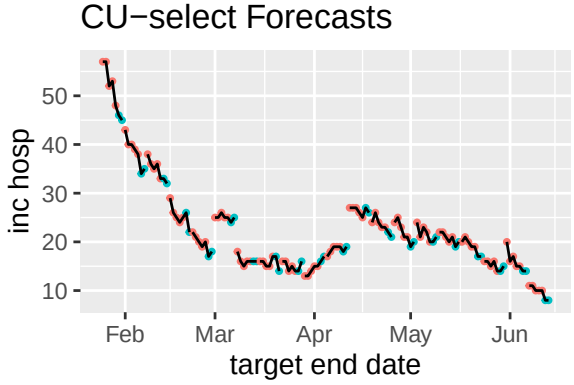


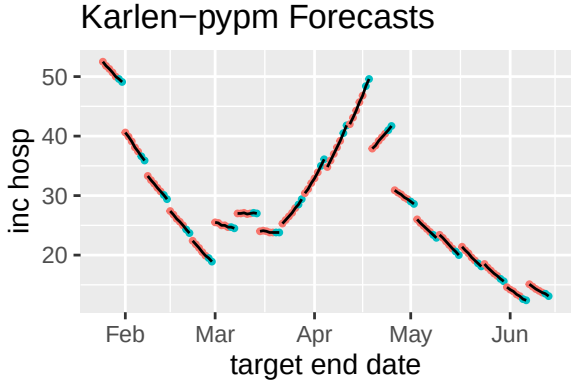

JHUAPL−Bucky Forecasts
inc hosp
80
60
40
20
Feb
Mar
Apr
May
Jun
target end date

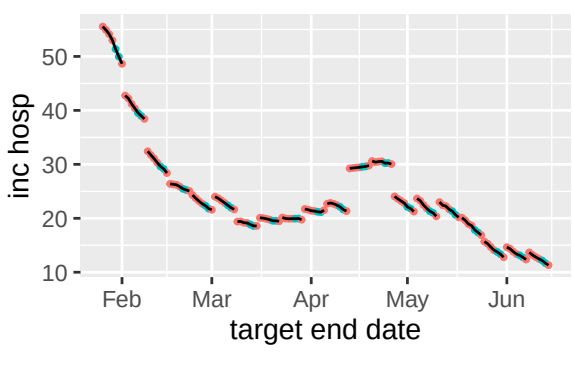

JHUAPL−Gecko Forecasts
inc hosp
100
75
50
25
Feb
Mar
Apr
May
Jun
target end date

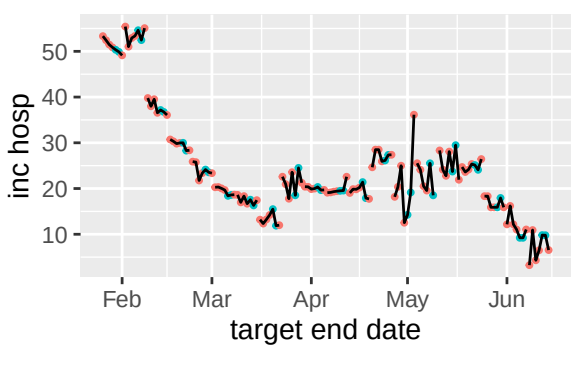

Google_Harvard−CPF Forecasts
inc hosp
40
30
20
10
Feb
Mar
Apr
May
Jun
target end date


Covid19Sim−Simulator Forecasts
inc hosp
100
75
50
25
Feb
Mar
Apr
May
Jun
target end date


CU−select Forecasts
inc hosp
50
40
30
20
10
Feb
Mar
Apr
May
Jun
target end date


Karlen−pypm Forecasts
inc hosp
50
40
30
20
Feb
Mar
Apr
May
Jun
target end date


JHUAPL−SLPHospEns Forecasts
inc hosp
50
40
30
20
10
Feb
Mar
Apr
May
Jun
target end date


MOBS−GLEAM_COVID Forecasts
inc hosp
50
40
30
20
10
Feb
Mar
Apr
May
Jun
target end date

We reiterate that low count location hospitalization truth data doesn't show as clear day of the week effects as high count truth data. However, we can still see that the same two models, JHUAPL-Bucky and JHUAPL-Gecko, do show day of the week effects for low count locations. Thus, we can categorize models into three categories: true (shows a day of the week effect), false (does not show a day of the week effect), or ensemble (a model built from other models that may or may not include day of the week effects). We create a third ensemble category because ensembles themselves do not incorporate day of the week effects, as they are a collection of other models, but the included models may or may not incorporate day of the week effects.

This categorization is shown in the table below.

Model	Day of Week Effect
COVIDhub-ensemble	ENS
JHUAPL-SLPHospEns	ENS
Covid19Sim-Simulator	FALSE
CU-select	FALSE
Google_Harvard-CPF	FALSE
Karlen-pypm	FALSE
MOBS-GLEAM_COVID	FALSE
JHUAPL-Bucky	TRUE
JHUAPL-Gecko	TRUE

Below we investigate the issue of different forecast dates making the horizon week variable not an entirely fair metric for the models. As it turns out, however, this is not a major problem because all of the models make their forecasts on either Sunday or Monday (some models have switched during the period of interest). The following table summarizes which day(s) each of the nine models have made their forecasts on.

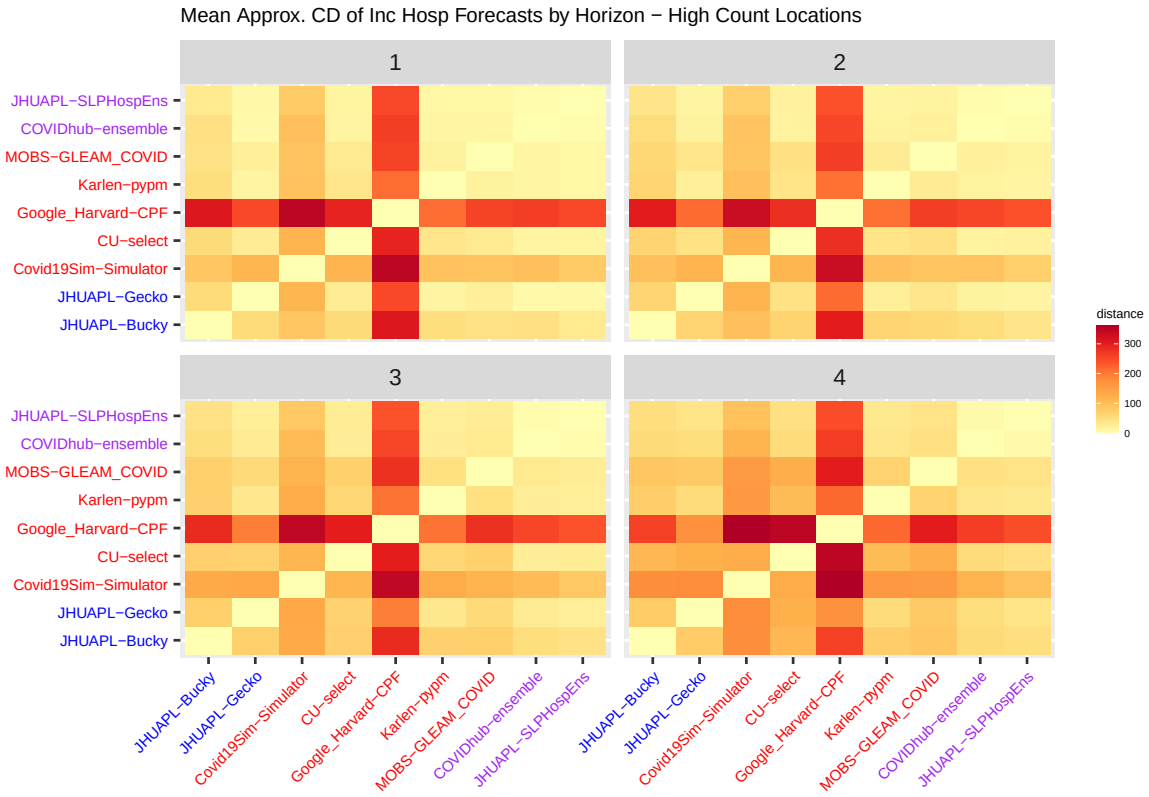
Model	Forecast Day
Covid19Sim-Simulator	Sunday
COVIDhub-ensemble	Monday
CU-select	Sunday
Google_Harvard-CPF	Sunday, Monday
JHUAPL-Bucky	Monday
JHUAPL-Gecko	Sunday, Monday
JHUAPL-SLPHospEns	Monday
Karlen-pypm	Sunday
MOBS-GLEAM_COVID	Monday

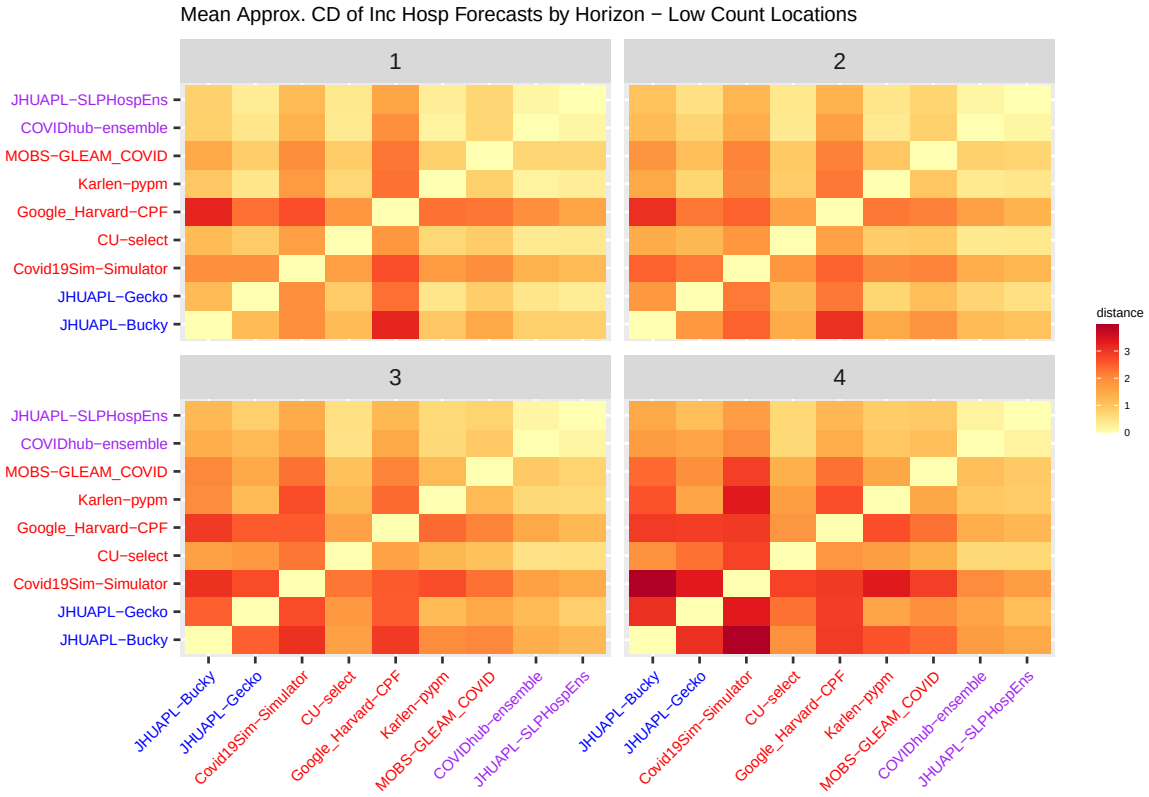
Given that hospitalization forecasts are daily, there are seven times as many forecasts made for incident hospitalizations as compared to incident cases and incident deaths. Because we are looking at day of the week effects, we choose to focus on only two days of the week for our analysis, one weekday and one weekend (Thursday and Saturday, respectively). This way, we can investigate if there seems to be a meaningful difference between model similarity on a weekday versus on a weekend.

Weekday Analysis

This portion of the analysis only examines target end dates for a single weekday, Thursday, to account for models that include day of the week effects.

First, we visualize the mean approximated pairwise Cramér's Distance across the entire period of interest in the heatmaps shown below. The distance from the model to itself is zero. The $x-$axis is arranged based on the categorization of models outlined above: blue is incorporates a day of week effect, red is does not incorporate a day of the week effect, and purple is ensemble model.

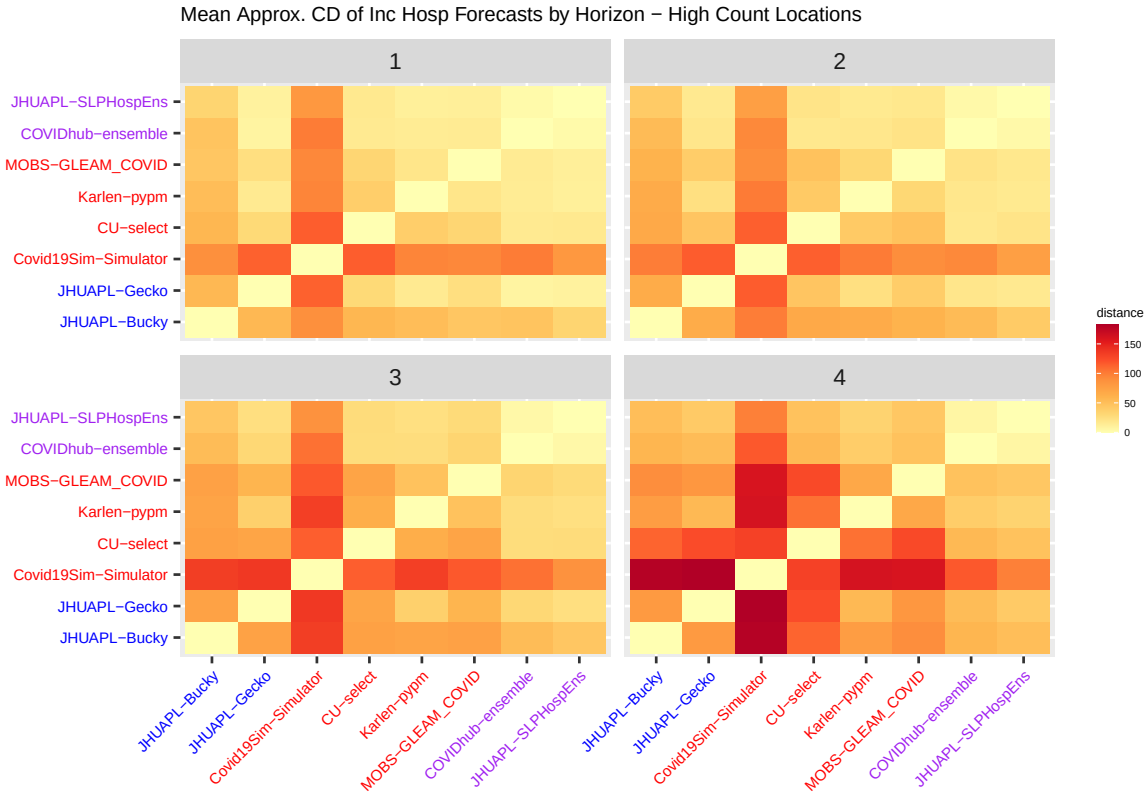


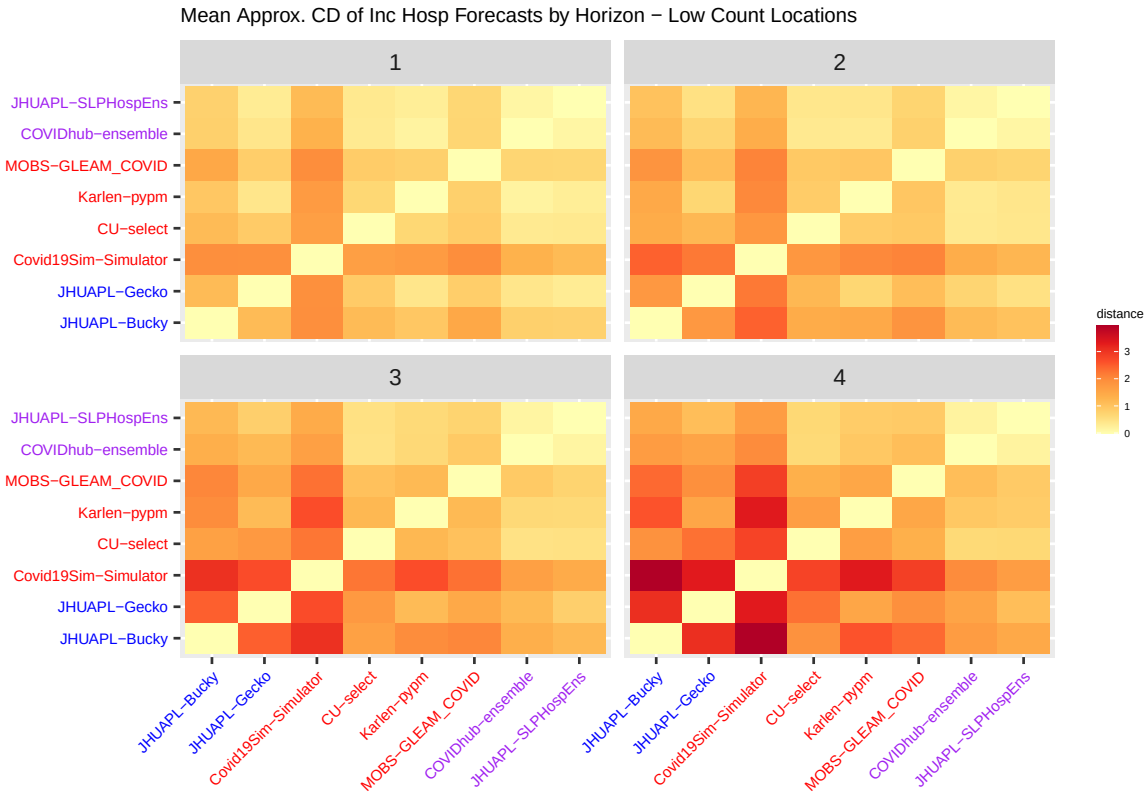


Google_Harvard-CPF is generally the least similar to the other models for both the high count and low

count locations. This is true across all horizons for the high count locations, but only true for the one and two week horizons of the low count locations. Covid19Sim-Simulator, JHUAPL-Bucky, and JHUAPL-Gecko show similar Cramér's Distances at three and four week horizons at the low count locations. However, it is important to note the small scale observed for the low count locations may explain why three models have such similar Cramér's Distances.

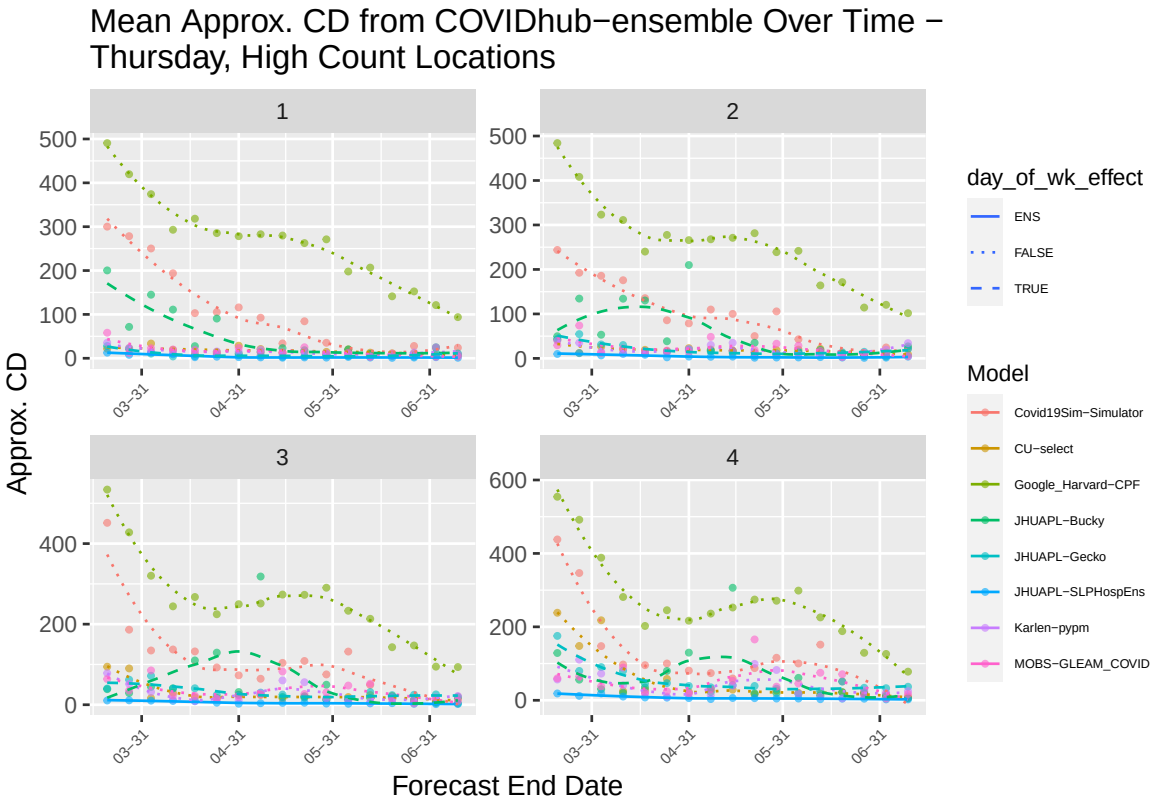
Since Google_Harvard-CPF is substantially dissimilar to the other models, especially for high count locations, it obscures any potential patterns about day of the week effect and model similarity that we might find among the other models. Thus, we plot the heat maps again, this time excluding Google_Harvard-CPF.

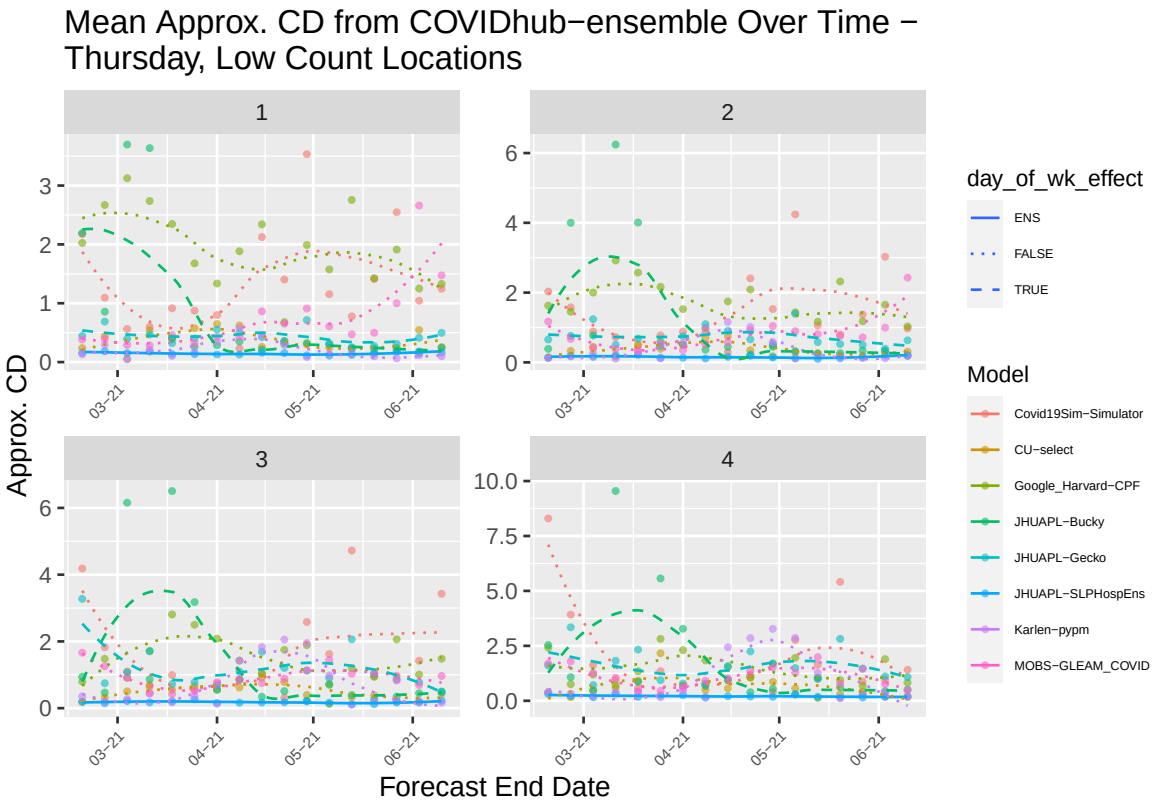




We can see that for both high and low count locations, at all horizons, Covid19Sim-Simulator becomes the most dissimilar model without Google_Harvard-CPF present. JHUAPL-Bucky seems to have the next highest Cramér's Distances at 1 to 3 week horizons, but at a 4-week horizon, the models without a day of the week effect show higher Cramér's Distances. Low count locations show that models with day of the week effect have greater dissimilarity than models without, except for Covid19Sim-Simulator. However, the low count location results can be rather sensitive to small variations due to chance alone given their small Cramér's Distance values.

We can also look at the approximated pairwise distances over time to see how the models become more similar or dissimilar at different points during the period of interest.

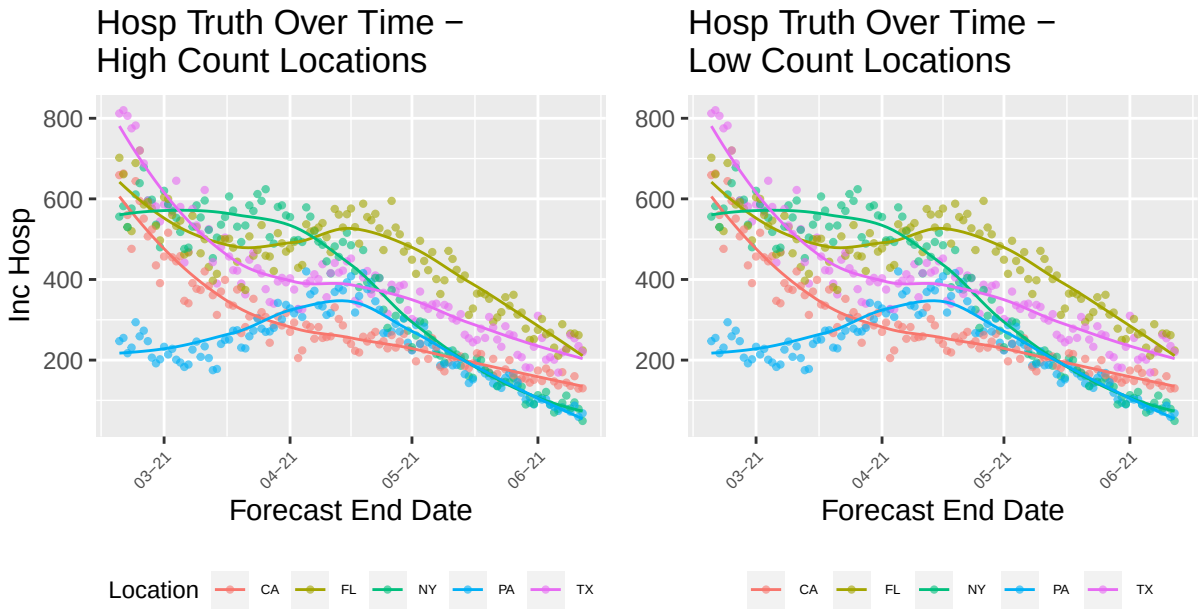




The scatterplots show that the Google_Harvard-CPF, Covid19Sim-Simulator, and JHUAPL-Bucky models

tend to differ from the Covidhub-ensemble model compared to the other models. This seems to align with the results shown in the heat maps above that show that Google_Harvard-CPF, Covid19Sim-Simulator, and JHUAPL-Bucky tend to have the highest mean Cramér's Distance from the other models. In high count locations, Google_Harvard-CPF is substantially different from the ensemble model during the entire period of interest. However, in low count locations, JHUAPL-Bucky shows a peak in March, although this peak is actually rather small, given the scale.

Whether models incorporate a day of the week effect does not seem to have an impact on how much the model differs from the ensemble, nor as to when it differs greatly.



It seems that Google_Harvard-CPF and Covid19Sim-Simulator's differences from the ensemble model follow the trends shown by the truth data.

We can also create dendrograms from the distances using hierarchical clustering.

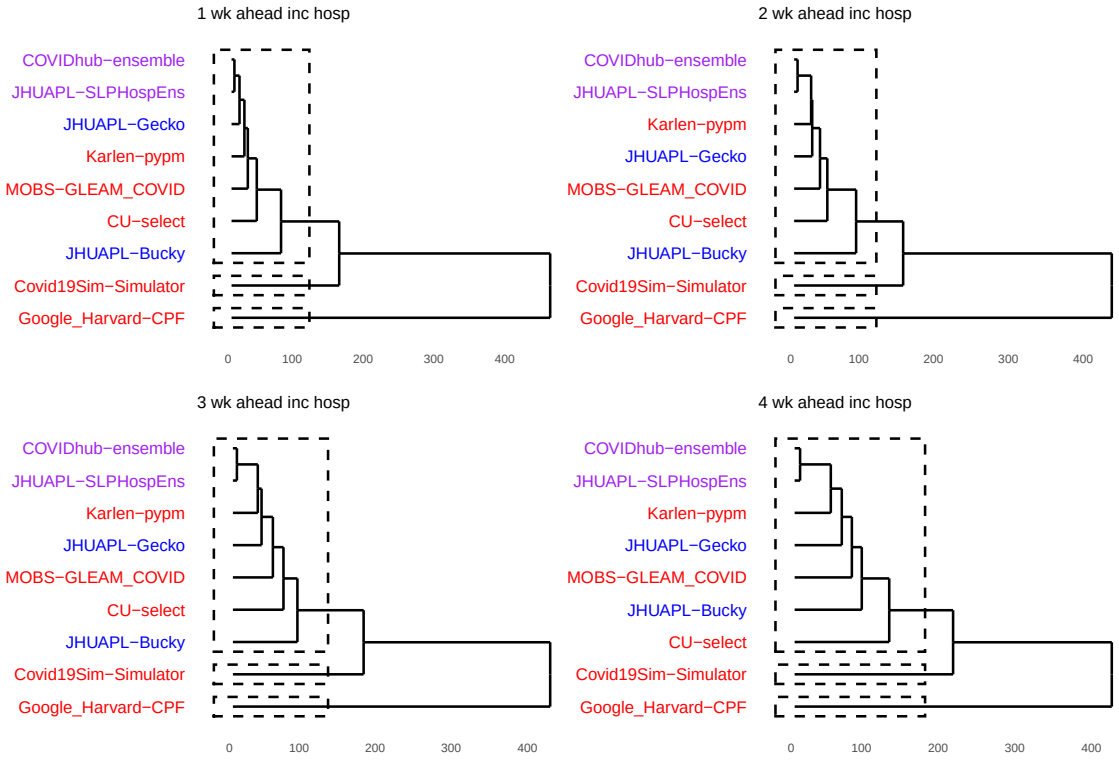


Figure 2: High Hospitalization Count Locations

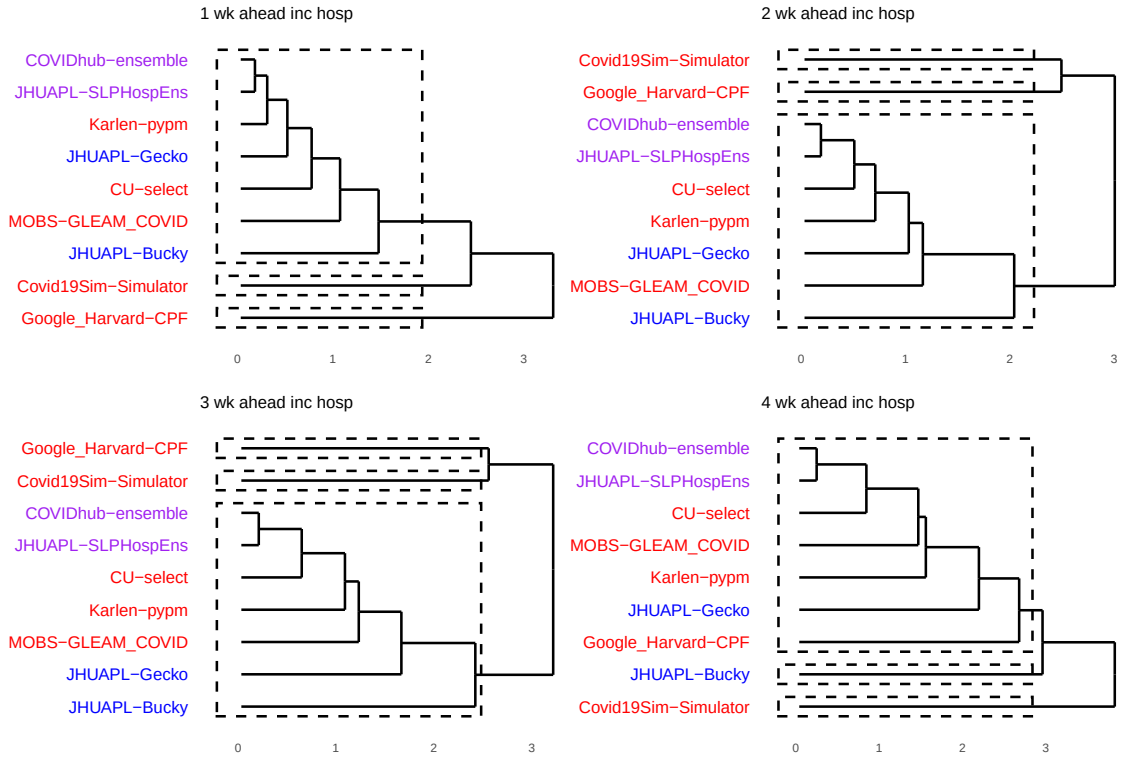


Figure 3: Low Hospitalization Count Locations

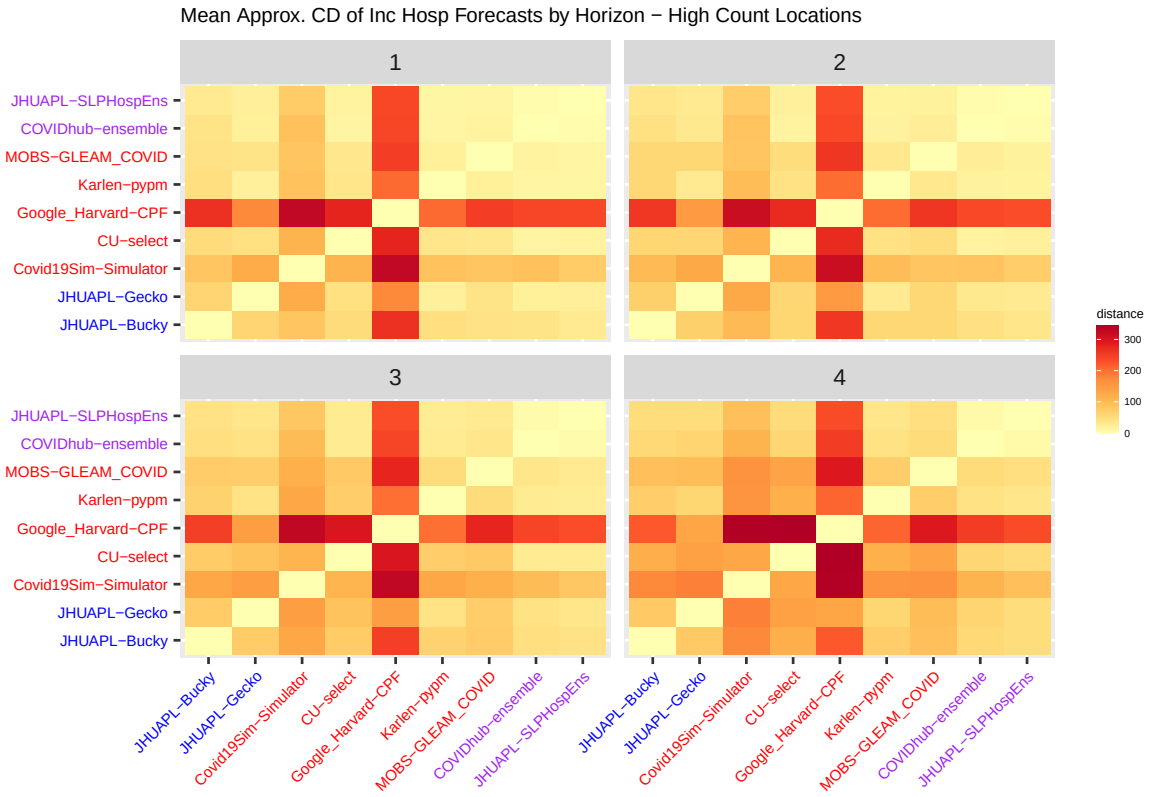
For each dendrogram, we split the models into three groups. Most high count and low count dendrograms include two single-model groups that consist of a model without day of the week effects. However, this does not seem to be indicative of any trends in model similarity related to day of the week effect, more just that Google_Harvard-CPF and Covid19Sim-Simulator tend to be more different from the other models over all. If we were to create the dendrograms without Google_Harvard-CPF, then JHUAPL-Bucky, a model that includes day of the week effects, would become its own group. The low count 3-week horizon dendrogram split into four groups more easily than three groups; however, the scale for differences in Cramér Distance for the low count locations is very small, likely a result of low incident hospitalizations, which may this difference.

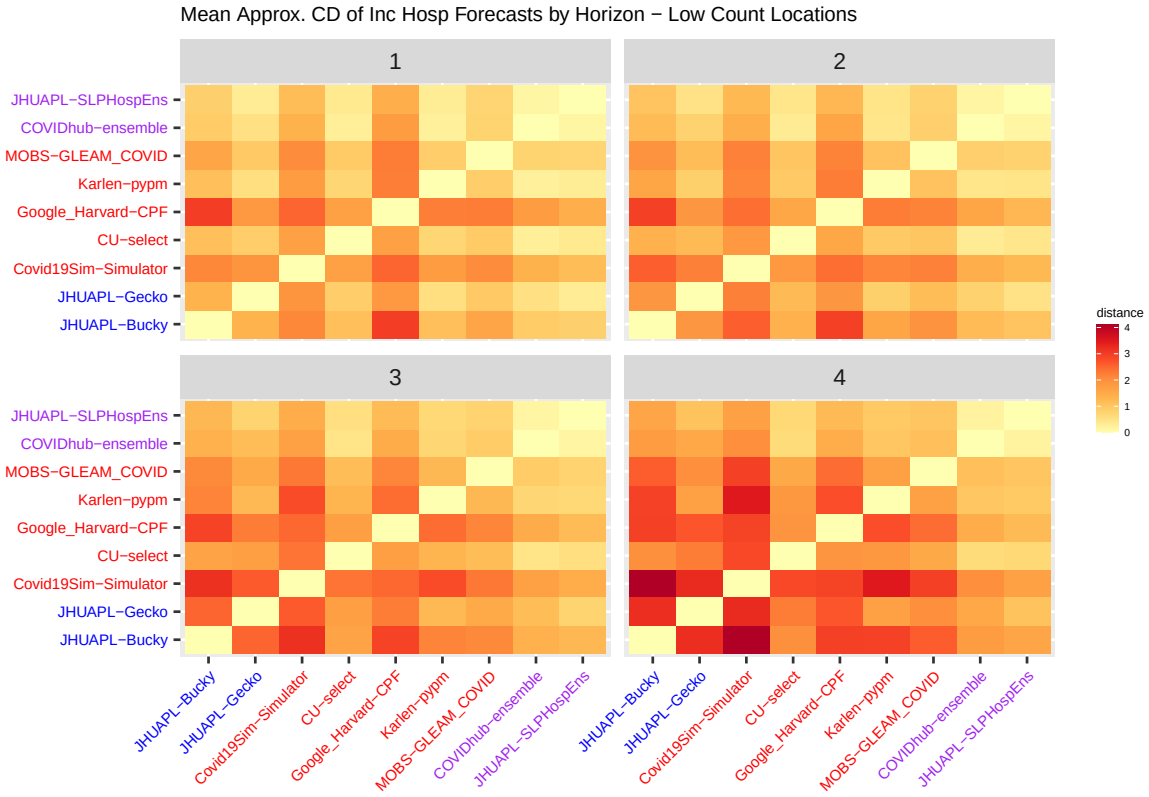
Overall, it seems that Google_Harvard-CPF is consistently the most dissimilar from the other models by a substantial amount, followed by Covid19Sim-Simulator, across almost all horizons for both high-count and low count regions. For Thursday forecasts, there is not a clear conclusion to be drawn about the impact of day of week effects on model similarity.

Weekend Analysis

Now we look at forecasts with a target end date on a Saturday to see if day of the week effects change model similarity. We expect forecasts for weekends to show the impact of day of the week effects more strongly than weekdays.

We again visualize the mean approximated pairwise Cramér's Distance across the entire period of interest in the heatmaps shown below, this time for Saturday forecasts. The distance from the model to itself is zero. The $x-$axis is arranged based on the categorization of models outlined above: blue is incorporates a day of week effect, red is does not incorporate a day of the week effect, and purple is ensemble model.

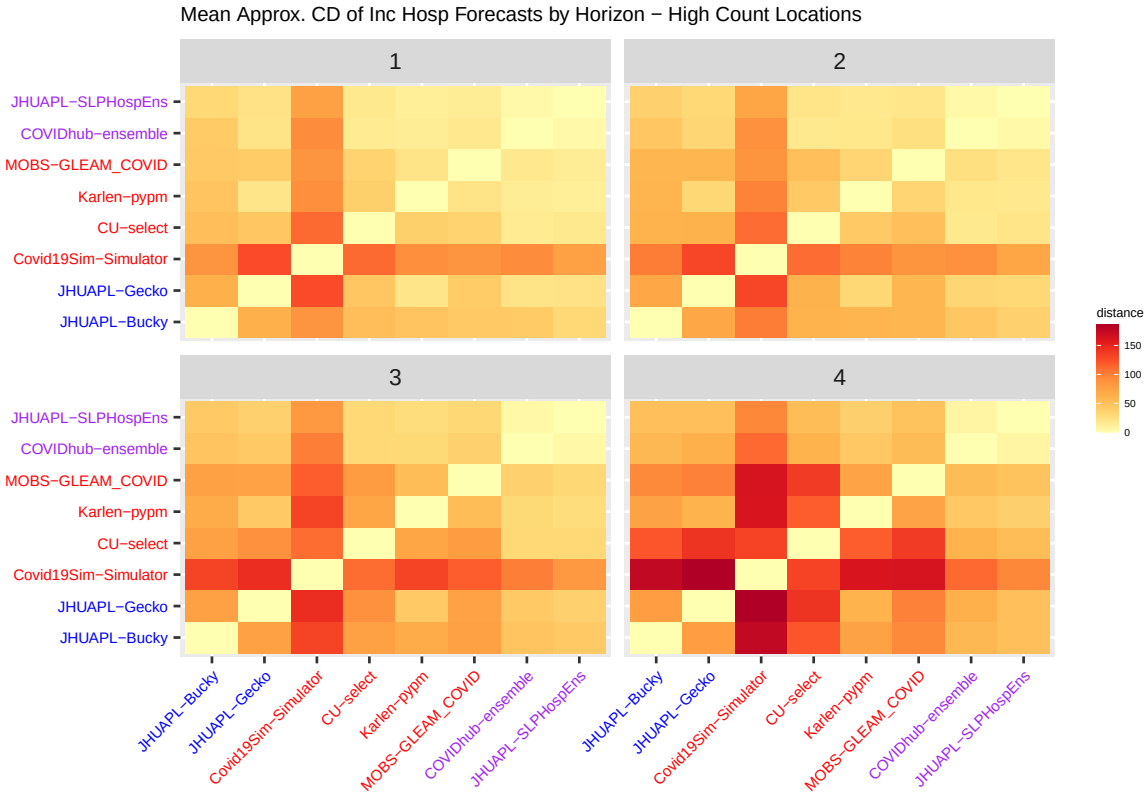


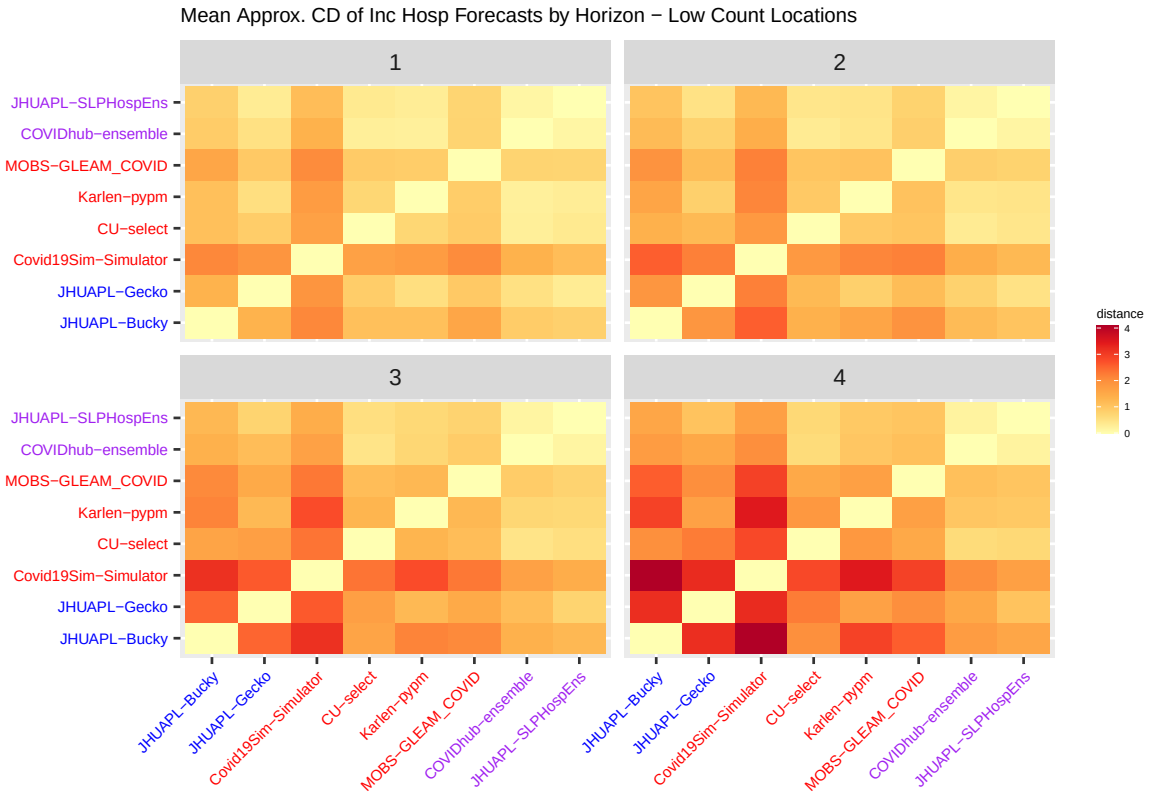


Similarly to the Thursday forecasts, Google_Harvard-CPF is generally the least similar to the other models

for both the high count and low count locations. This is true across all horizons for the high count locations, but only true for the one and two week horizons of the low count locations. Covid19Sim-Simulator is the most dissimilar for the three and four week horizons of the low count locations. However, the small scale observed for the low count locations may explain why this model has a higher Cramér's Distance at longer horizons rather than a true pattern.

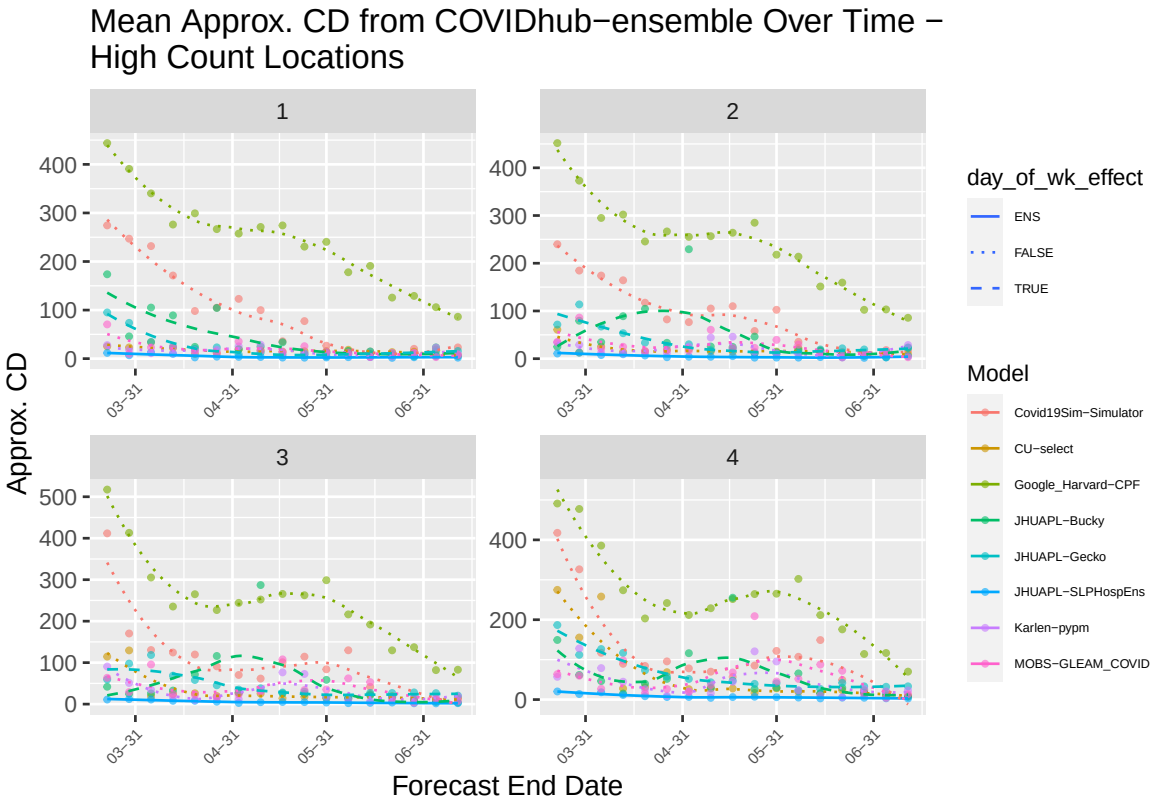
Like before, Google_Harvard-CPF is substantially dissimilar to the other models, especially for high count locations, it obscures any potential patterns about day of the week effect and model similarity that we might find among the other models. Thus, we plot the heat maps again, this time excluding Google_Harvard-CPF.

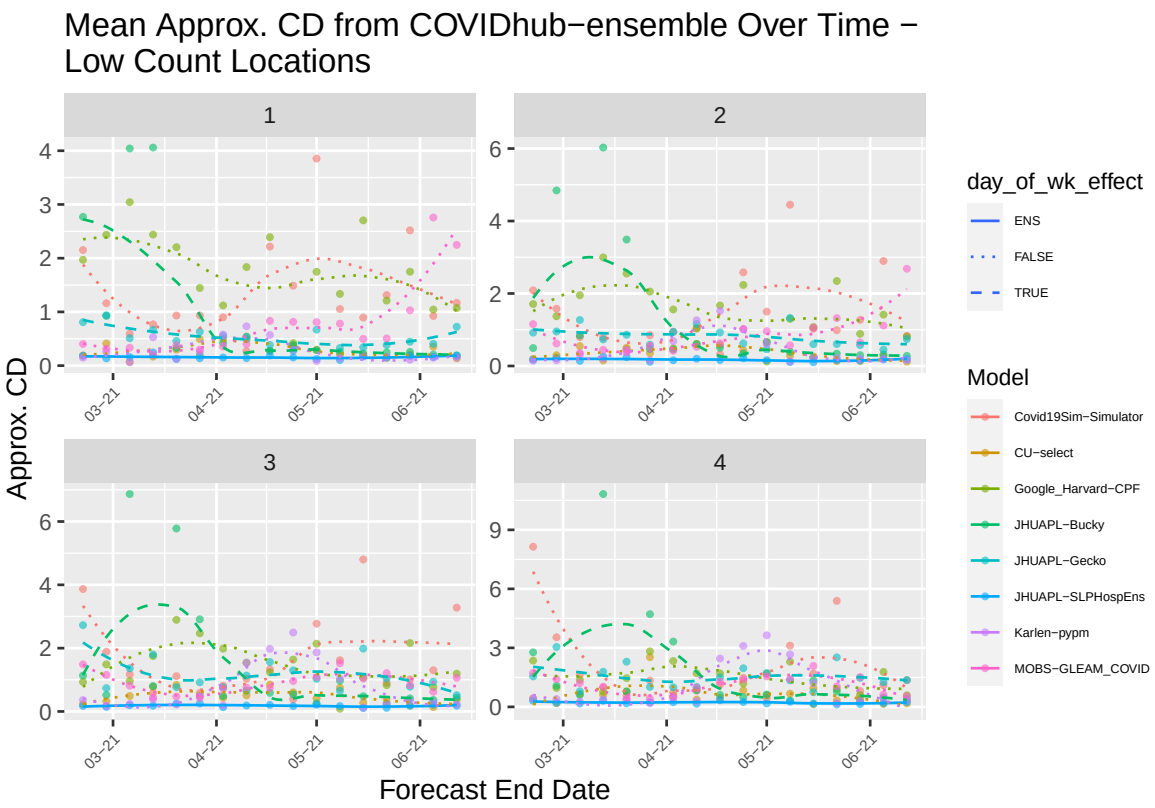




Without Google_Harvard-CPF present, Covid19Sim-Simulator becomes the most dissimilar model for both high and low count locations, at all horizons. JHUAPL-Bucky seems to have the next highest Cramér's Distances at 1 to 3 week horizons, but at a 4-week horizon, the models without a day of the week effect show higher Cramér's Distances. Low count locations show that models with day of the week effect have greater dissimilarity than models without, save for Covid19Sim-Simulator. However, the low count location results can be rather sensitive to small variations due to chance alone given their small Cramér's Distance values.

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



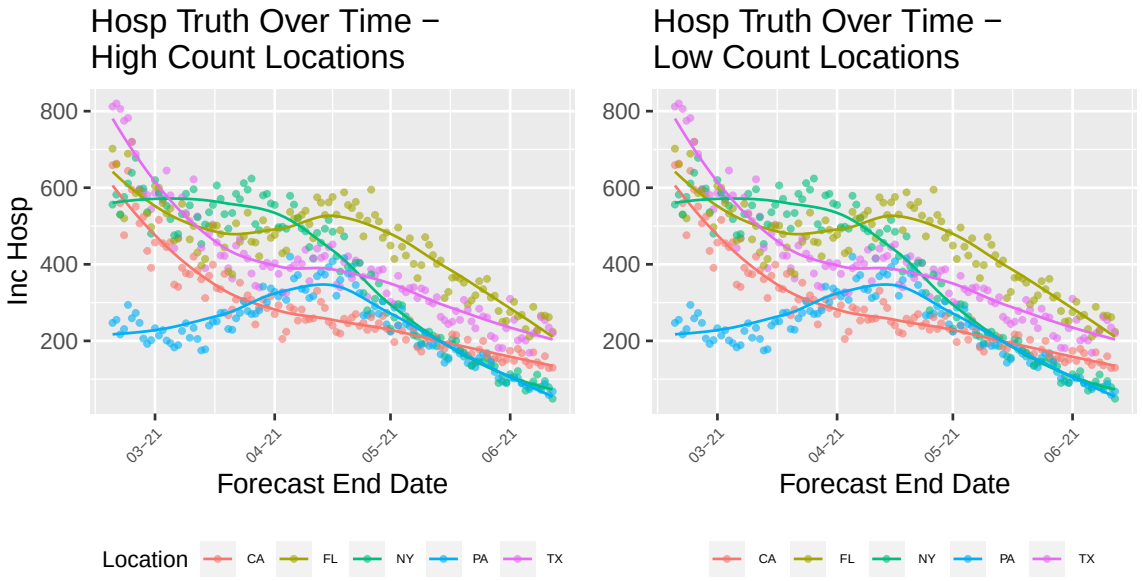


The scatterplots for Saturday forecasts closely resemble those of the Thursday forecasts. Google_Harvard-

CPF, Covid19Sim-Simulator, and JHUAPL-Bucky models generally differ from the Covidhub-ensemble model, which aligns with the results from the heat maps above: Google_Harvard-CPF, Covid19Sim-Simulator, and JHUAPL-Bucky tend to have the highest mean Cramér's Distance from the other models. In high count locations, Google_Harvard-CPF is substantially different from the ensemble model during the entire period of interest. However, in low count locations, JHUAPL-Bucky shows a peak in March, although this peak is actually rather small, given the scale.

Whether models incorporate a day of the week effect does not seem to have an impact on how much the model differs from the ensemble, nor as to when it differs greatly.

These scatterplots are nearly the same as the ones shown above.



Like with the Thursday forecasts, it seems that Google_Harvard-CPF and Covid19Sim-Simulator's differences from the ensemble model follow the trends shown by the truth data.

We can also cluster the distances using hierarchical clustering.

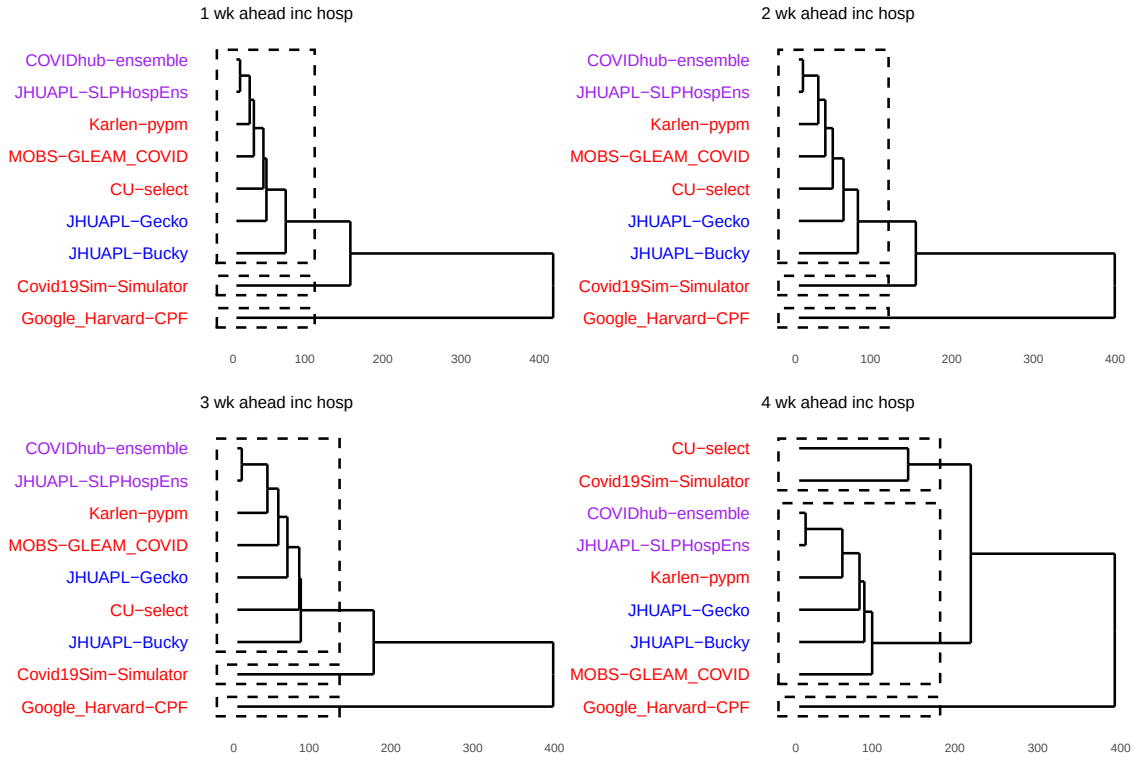


Figure 4: High Hospitalization Count Locations

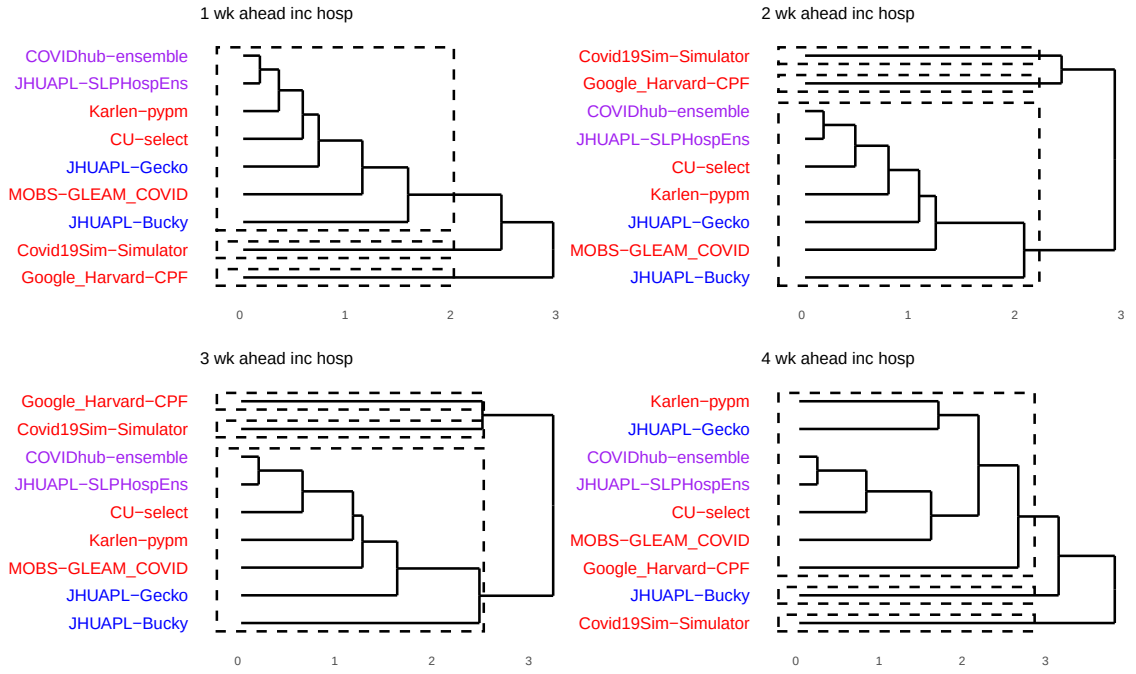


Figure 5: Low Hospitalization Count Locations

For each dendrogram, we split the models into three groups. Most high count and low count dendrograms

include two single-model groups that consist of a model without day of the week effects. However, this does not seem to be indicative of any trends in model similarity related to day of the week effect, more just that Google_Harvard-CPF and Covid19Sim-Simulator tend to be more different from the other models over all. If we were to create the dendrograms without Google_Harvard-CPF, then JHUAPL-Bucky, a model that includes day of the week effects, would become its own group. The low count 3-week horizon dendrogram split into four groups more easily than three groups; however, the scale for differences in Cramér Distance for the low count locations is very small, likely a result of low incident hospitalizations, which may this difference.

For Saturday forecasts, Google_Harvard-CPF is consistently the most dissimilar from other models, followed by Covid19Sim-Simulator, across almost all horizons for both high-count and low count regions. This is the same as for Thursday forecasts. For Saturday forecasts, there is also not an obvious conclusion to be drawn about the impact of day of week effects on model similarity.

Results and Conclusion

Although this investigation of model similarity and day of the week effects for Covid-19 forecasts yielded largely inconclusive results, there are several clear takeaways. First, Cramér's Distance is an attractive metric for measuring model similarity, especially for the models that submit to the COVID-19 Forecast Hub, because it is compatible with WIS and approximations are relatively easy to calculate. Second, the issue of unaligned hospitalization forecasts can be solved fairly easily by creating a new relative horizon variable. In addition, the potential disadvantage of certain models having an unfair forecasting advantage if we use this variable is unfounded for the nine models analyzed (and likely several other incident hospitalization models). Third, incident hospitalization truth data shows day of the week effects in which weekends have noticeably lower counts. Some of the models account for this day of week effect while others do not. Finally, day of week effects may impact model similarity to an extent but results are generally inconclusive, and further research is needed to draw any stronger and/or more definitive conclusions.

There are many directions that future research could take this work, but two stand out at the time of writing this report. One would be to investigate the relationship between the performance of the analyzed models and their similarity by comparing relative WIS. The other would be to explore the decomposition of Cramér's Distance and discover how much shape, spread, and mean of the distributions of forecasts play into the approximate pairwise distance.

Lastly, the repository for this project is available on Github.