COVID-19 Forecast Similarity Analysis for Hospitalizations

Johannes Bracher, Evan Ray, Nick Reich, Nutcha Wattanachit, Li Shandross

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.

Extending the work of Bracher et. al, we choose Cramer's Distance as a metric to evaluate the similarity between models. The Cramer'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 Cramer's Distance, and we do so by using Bracher et. al's trapezoidal rule:

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

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

$$\tag{4}$$

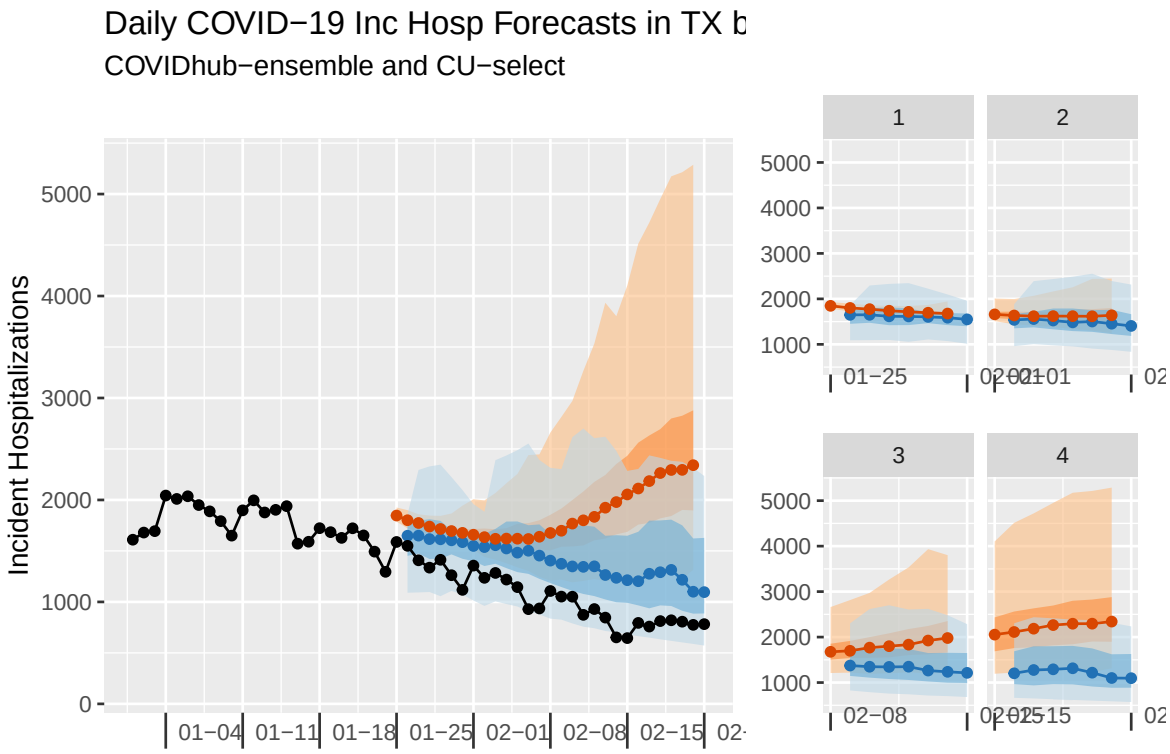
where q_j is the jth quantile.

Cramer'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 Cramer'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 Cramer's Distance calculated separately at each horizon.

Horizon	Approx. CD
1	64.91
2	26.24
3	118.61

Horizon	Approx. CD
4	383.93



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.

<Given that there are so many inc hosp forecasts (since it originally is daily), we choose to focus on only two days of the week for our analysis, one weekday and one weekend (Thursday and Saturday, respectively). We choose these days in the case there is a meaningful difference btwn model similarity on a weekday or weekend.>

Forecast Inclusion Criteria

The pairwise approximated Cramer'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 FL, TX, NY, CA, PA while the low count locations are VT, HI, AK, WY, SD.

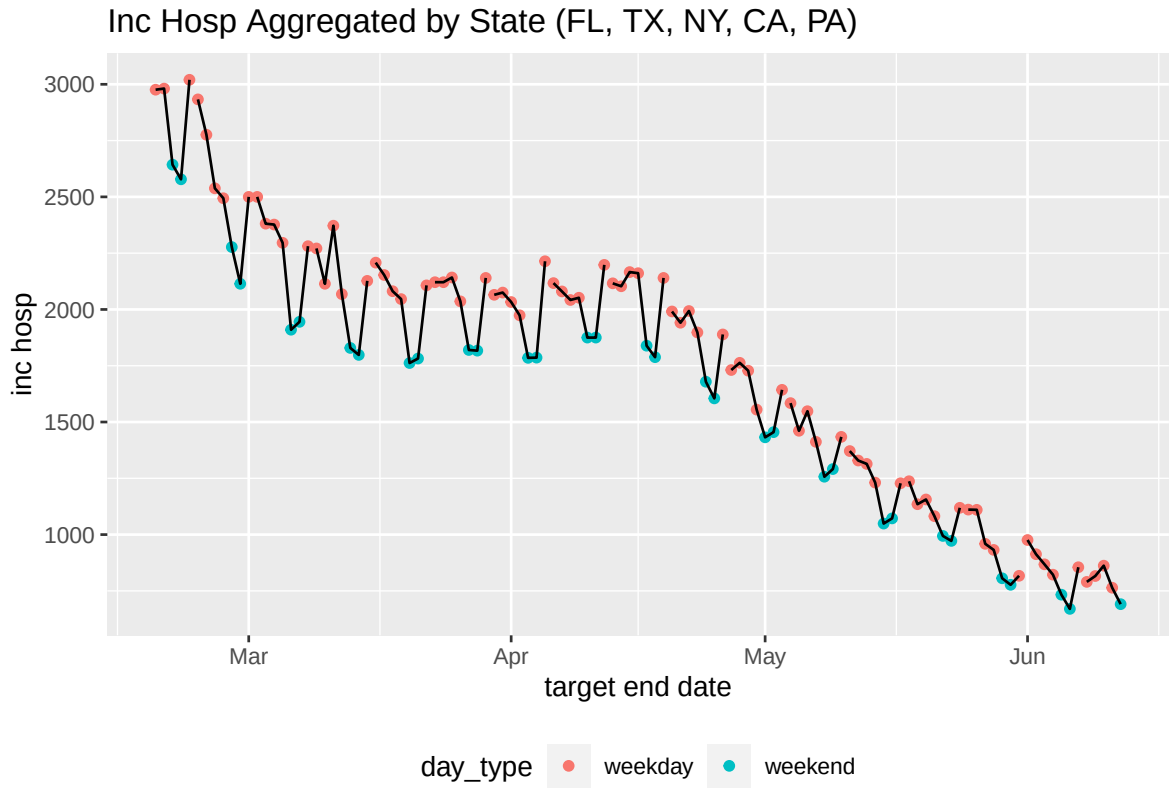
There are nine models that fulfilled the criteria for both the five high count and five low count locations for Thursday forecasts. There are ten that fulfill such criteria for the Saturday forecasts, but we perform the analysis on only the overlapping nine models.

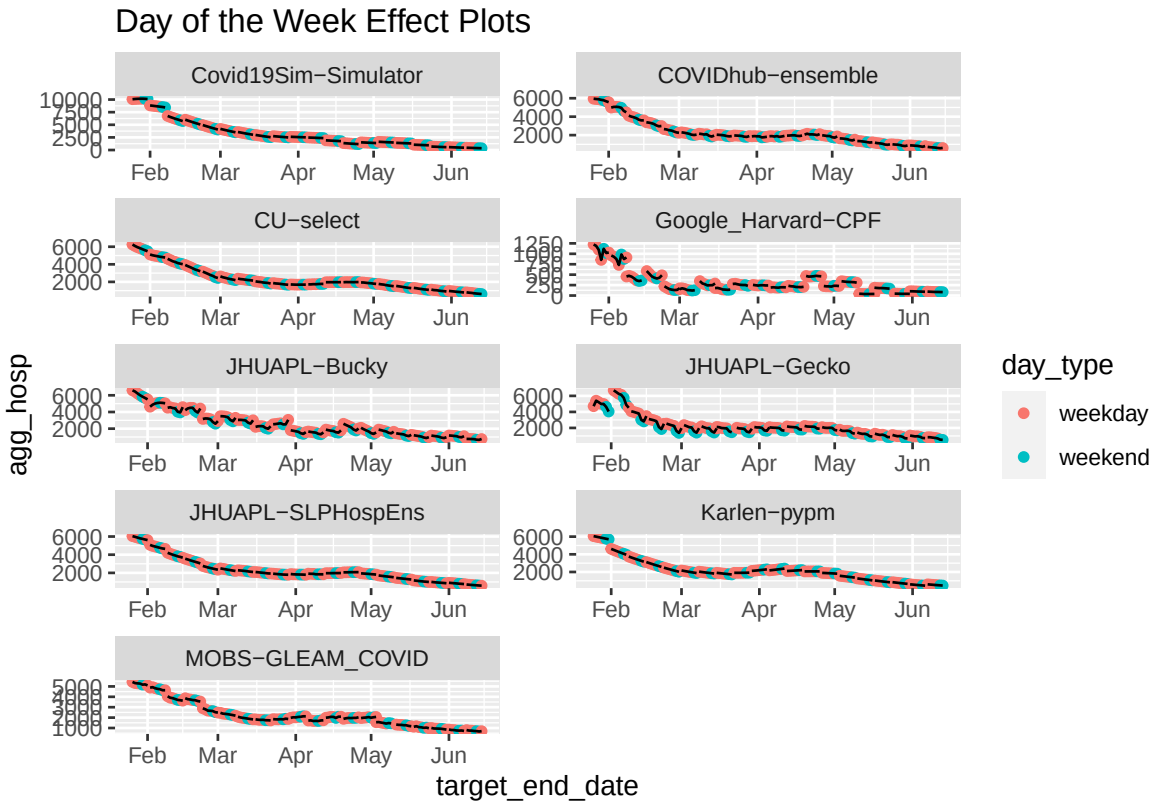
Day of the Week Effects

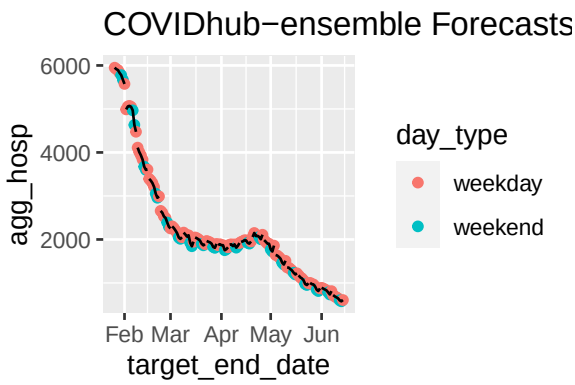
- introduce day of the wk effects, which are cyclic, weekly patterns… -> in this case, we're considering weekdays vs weekends
- left: day of wk effect evident in the truth data - could draw separate lines for weekdays vs weekends
- right: table summarizing if models have the effect, use these colors in our figures
- 2 addtnl day of the week effect plots for 2 models in analysis -> compare model w/o day of wk effects vs model w/ them -> these 2 models chosen specifically for clear illustration purpose

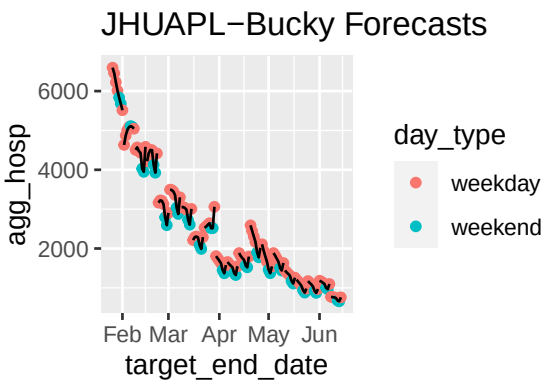
This analysis looks into whether the incorporation of day of the week effects impacts model similarity, e.g. do models that have day of the week effects more similar to each other than those without. We define day of the week effects to be cyclic weekly patterns for which a specific day or group of days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

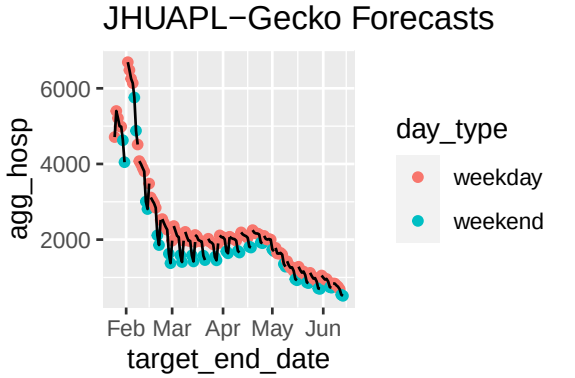
We plot point forecasts with a horizon week of 1 to determine which models include day of the week effects.

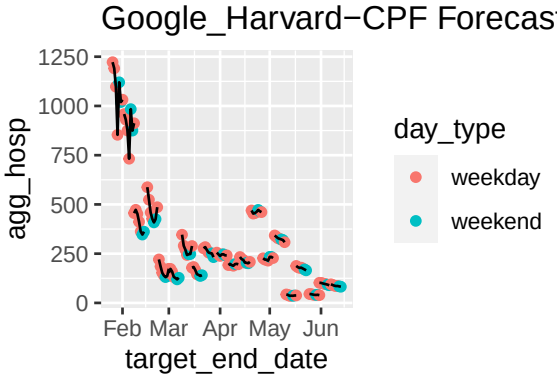


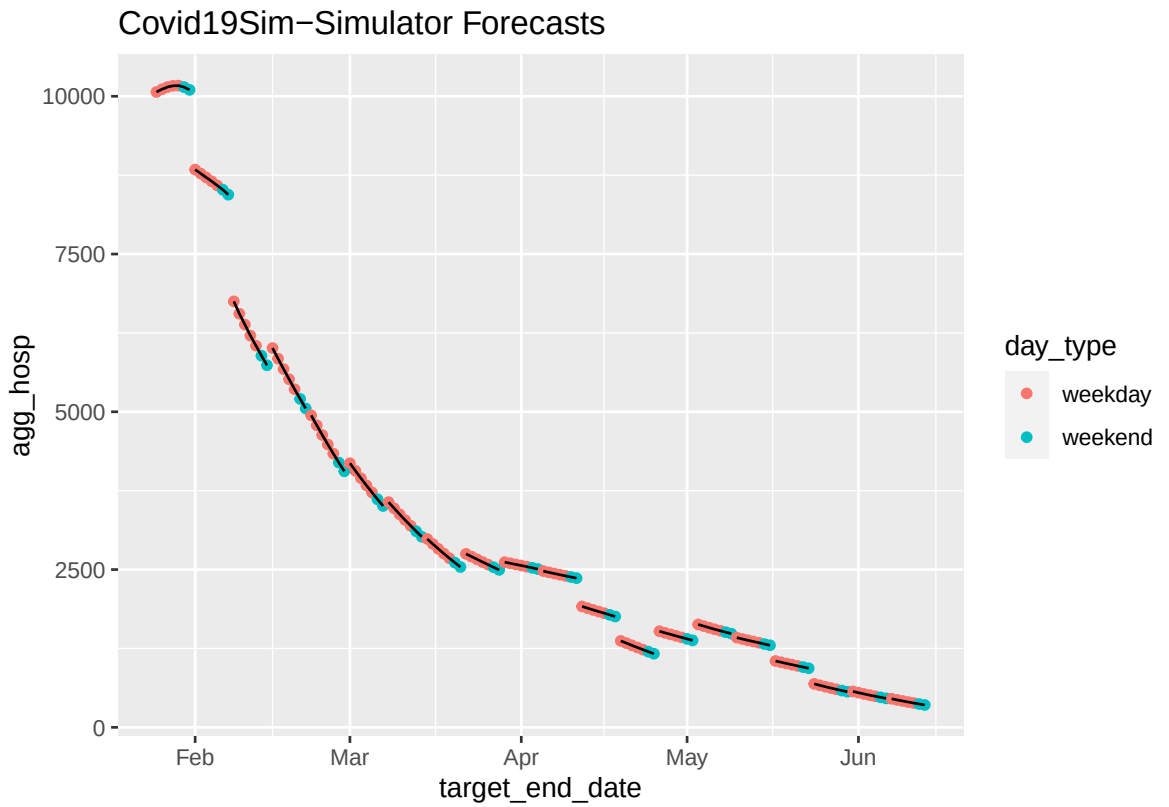

Day of the Week Effect Plots
Covid19Sim−Simulator
COVIDhub−ensemble
CU−select
Google_Harvard−CPF
JHUAPL−Bucky
JHUAPL−Gecko
JHUAPL−SLPHospEns
Karlen−pypm
MOBS−GLEAM_COVID
agg_hosp
target_end_date
Feb
Mar
Apr
May
Jun
day_type
weekday
weekend

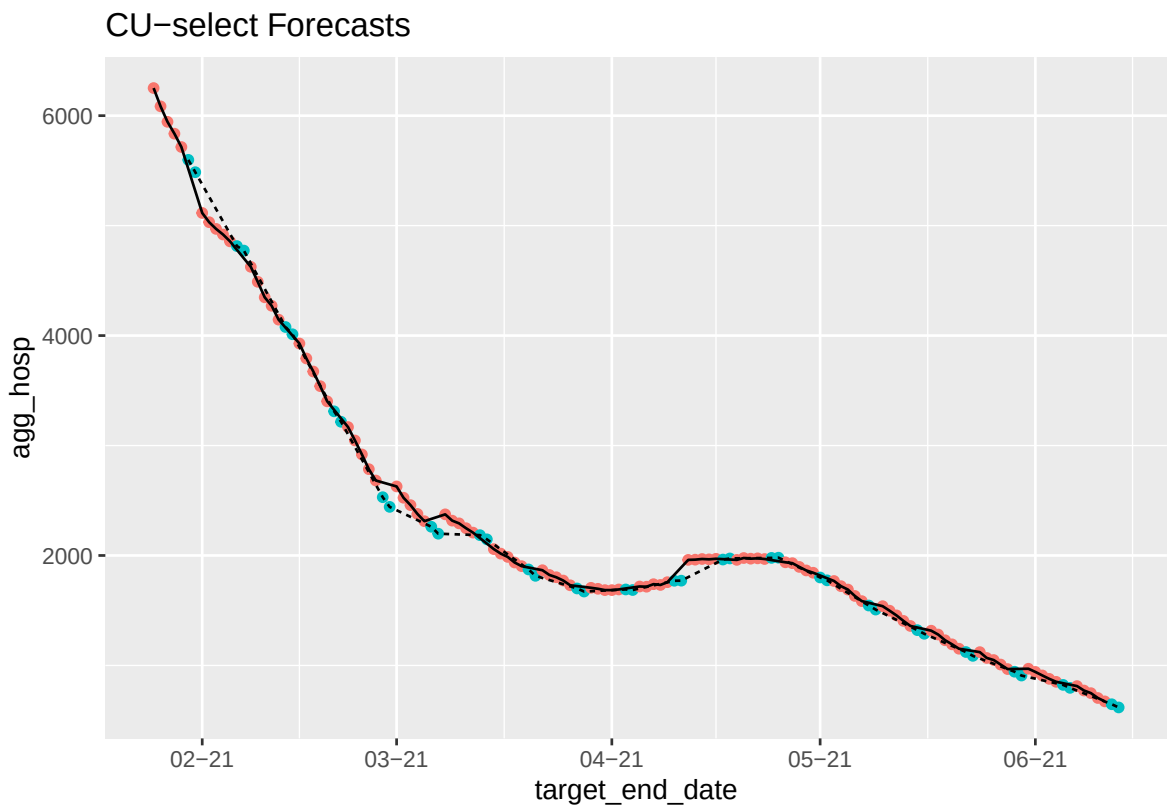

COVIDhub−ensemble Forecasts
agg_hosp
target_end_date
Feb Mar Apr May Jun
day_type
weekday
weekend

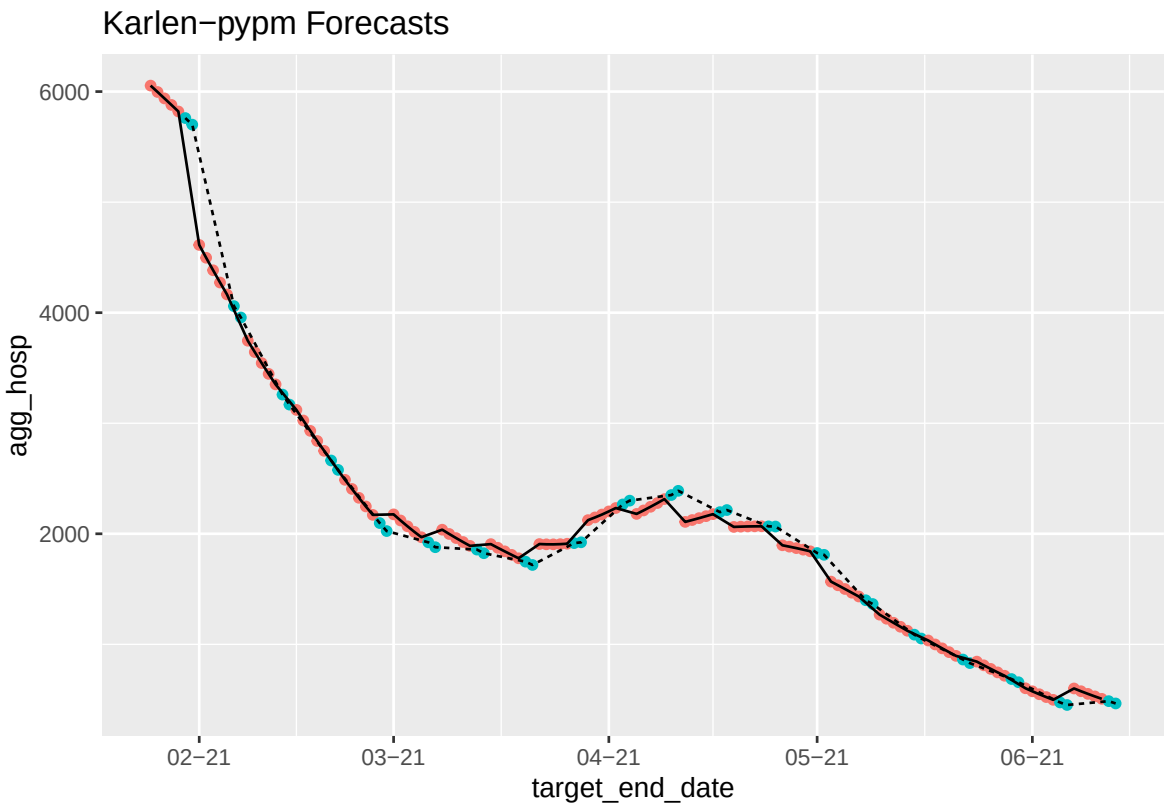

JHUAPL−Bucky Forecasts
agg_hosp
target_end_date
Feb Mar Apr May Jun
day_type
weekday
weekend

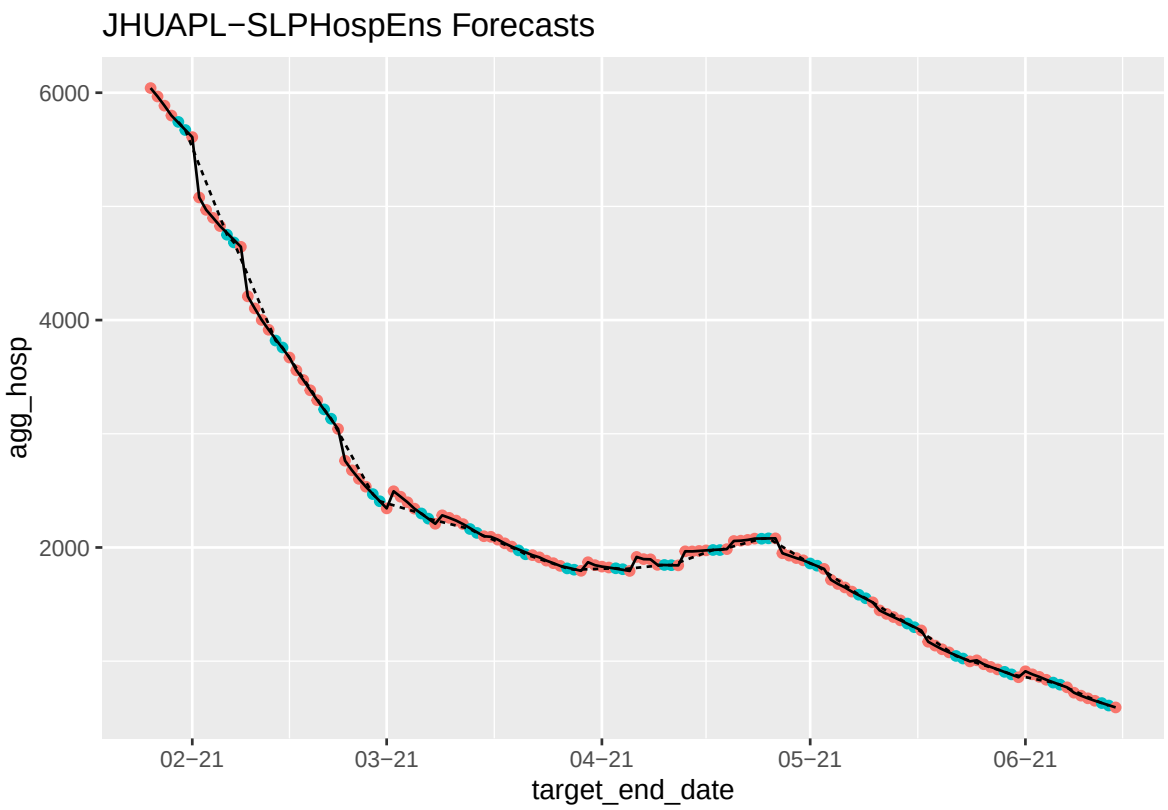

JHUAPL−Gecko Forecasts
agg_hosp
target_end_date
Feb Mar Apr May Jun
day_type
weekday
weekend

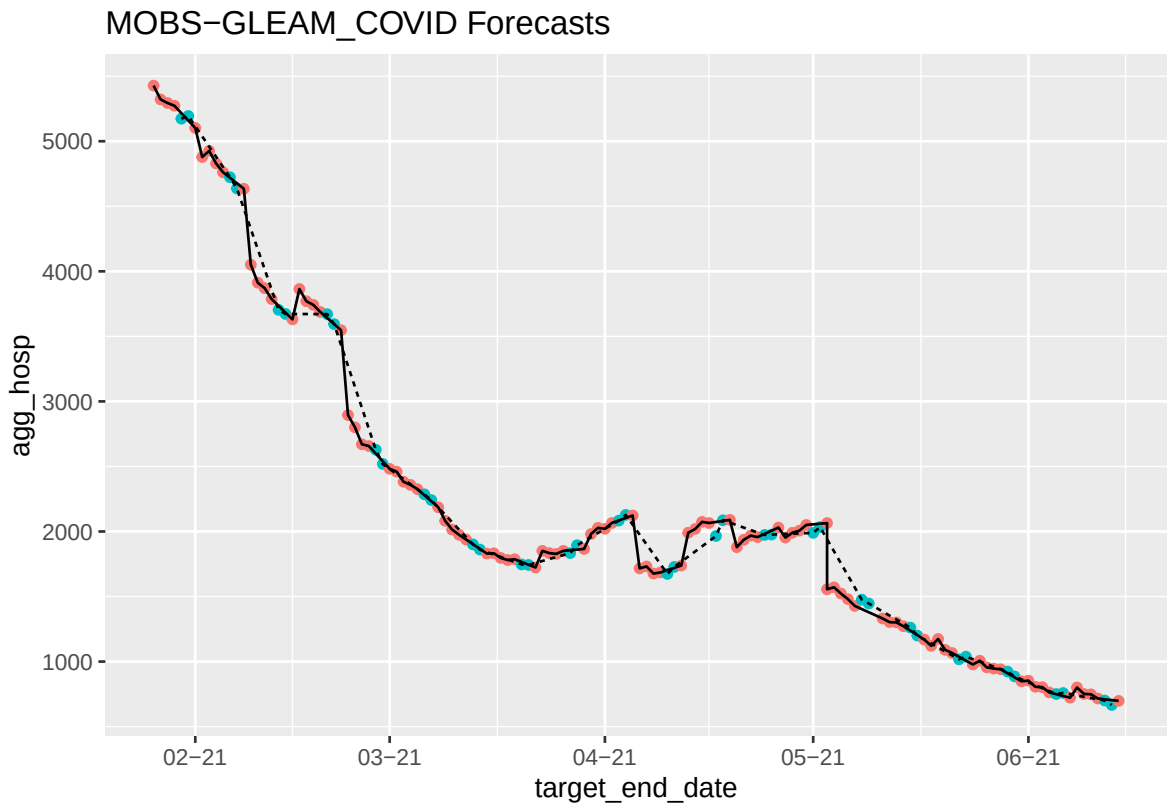

Google_Harvard−CPF Forecasts
agg_hosp
target_end_date
Feb Mar Apr May Jun
day_type
weekday
weekend

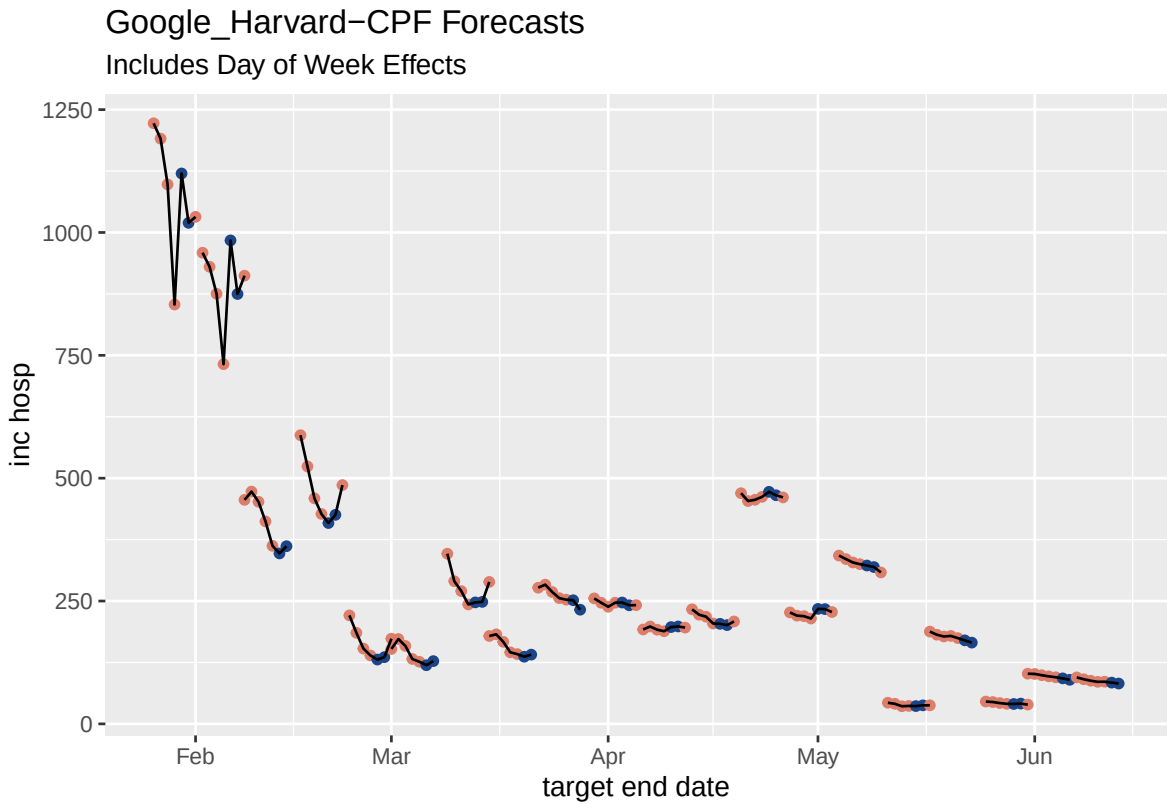

Covid19Sim−Simulator Forecasts
10000
7500
5000
2500
0
agg_hosp
Feb
Mar
Apr
May
Jun
target_end_date
day_type
weekday
weekend


CU−select Forecasts
6000
4000
2000
agg_hosp
02−21
03−21
04−21
05−21
06−21
target_end_date


Karlen−pypm Forecasts
6000
4000
2000
agg_hosp
02−21
03−21
04−21
05−21
06−21
target_end_date

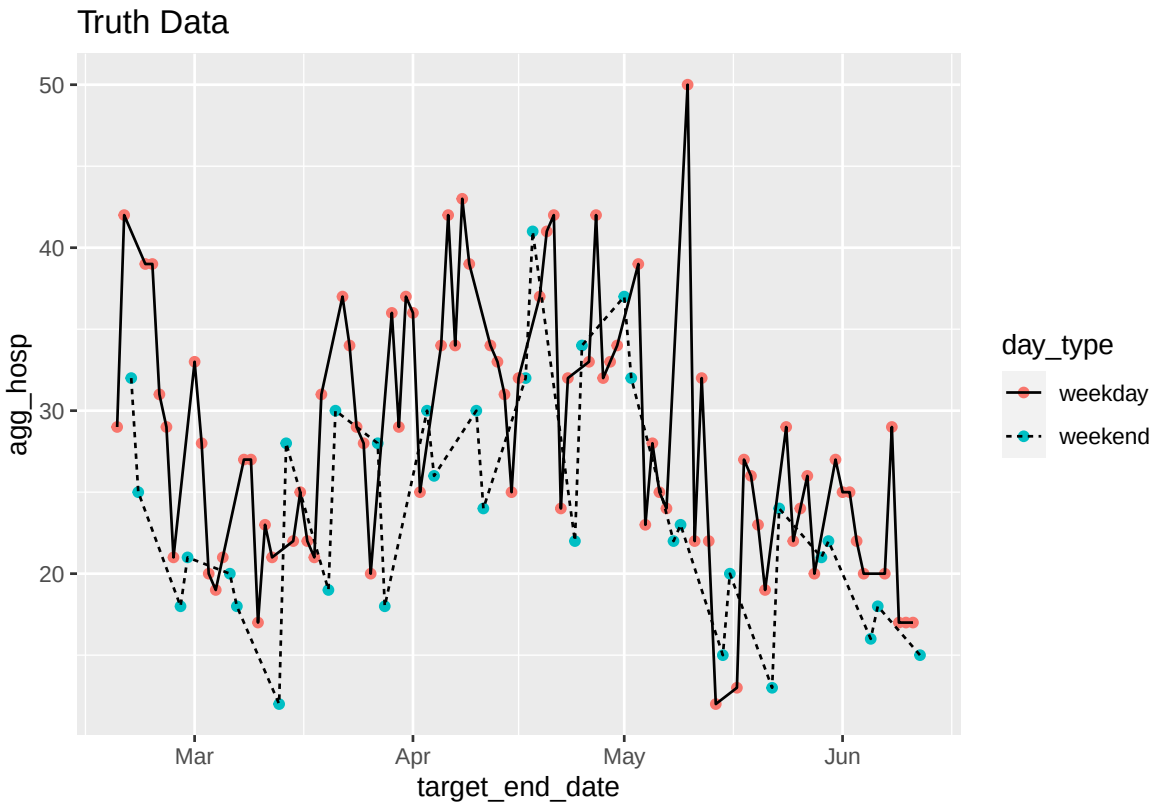

JHUAPL−SLPHospEns Forecasts
6000
4000
2000
agg_hosp
02−21
03−21
04−21
05−21
06−21
target_end_date

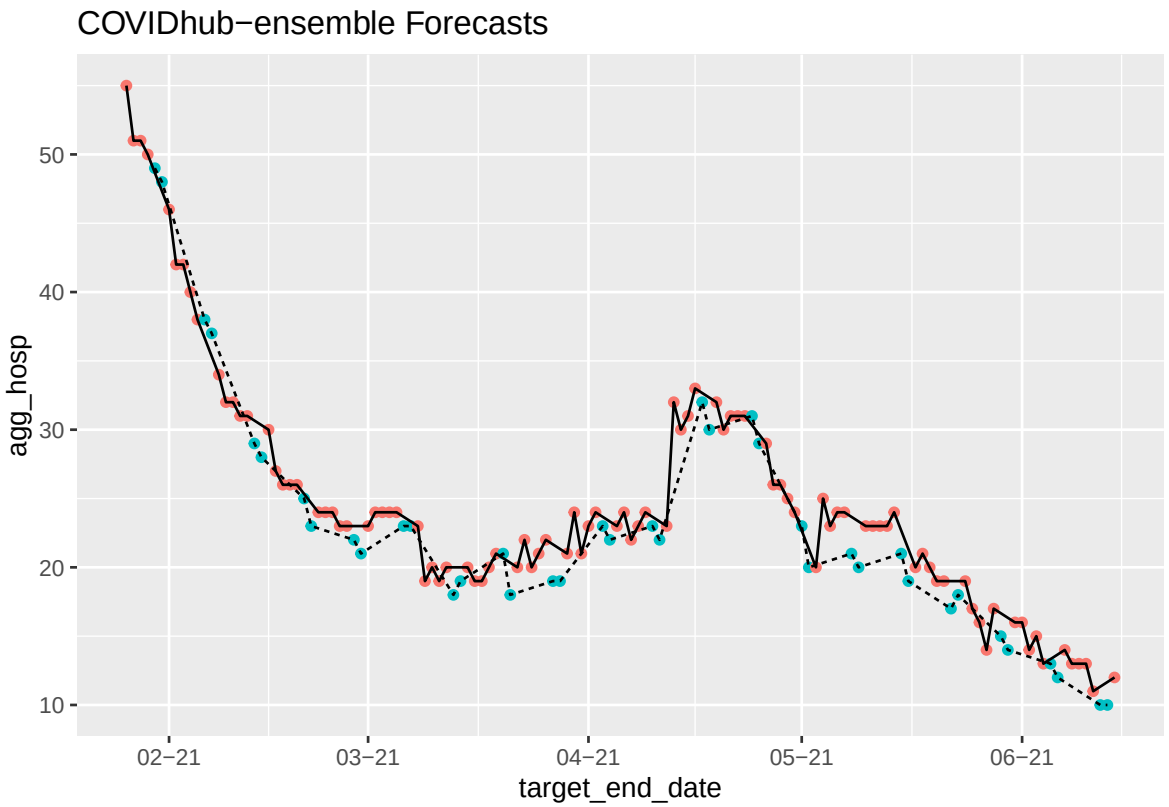


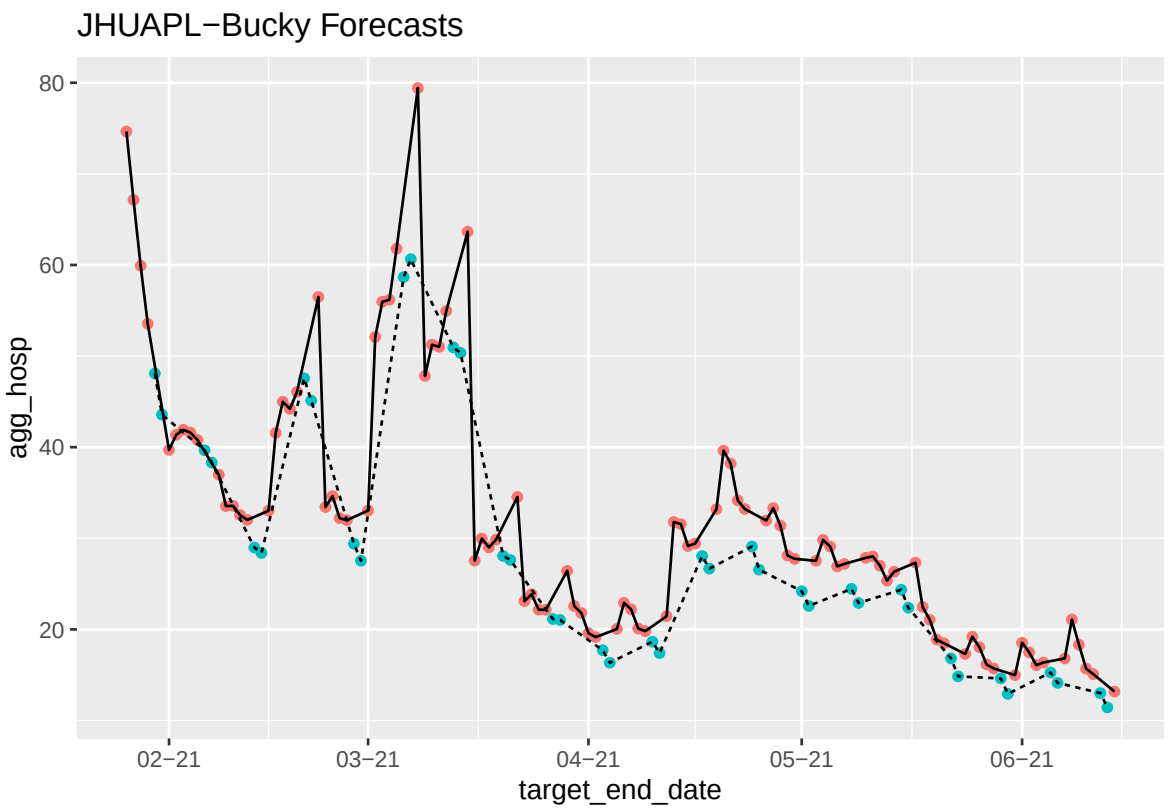


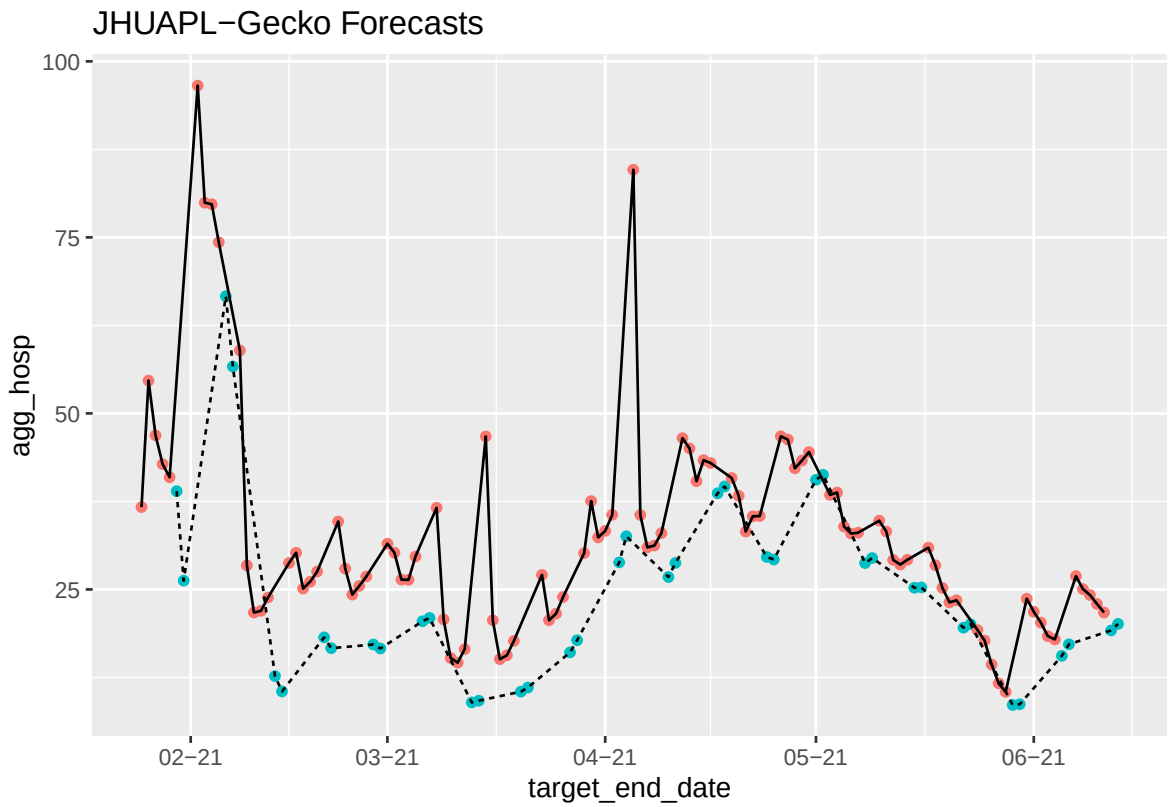
Truth data has always shown day of the week effects (weekends always have lower inc hosp) but the models

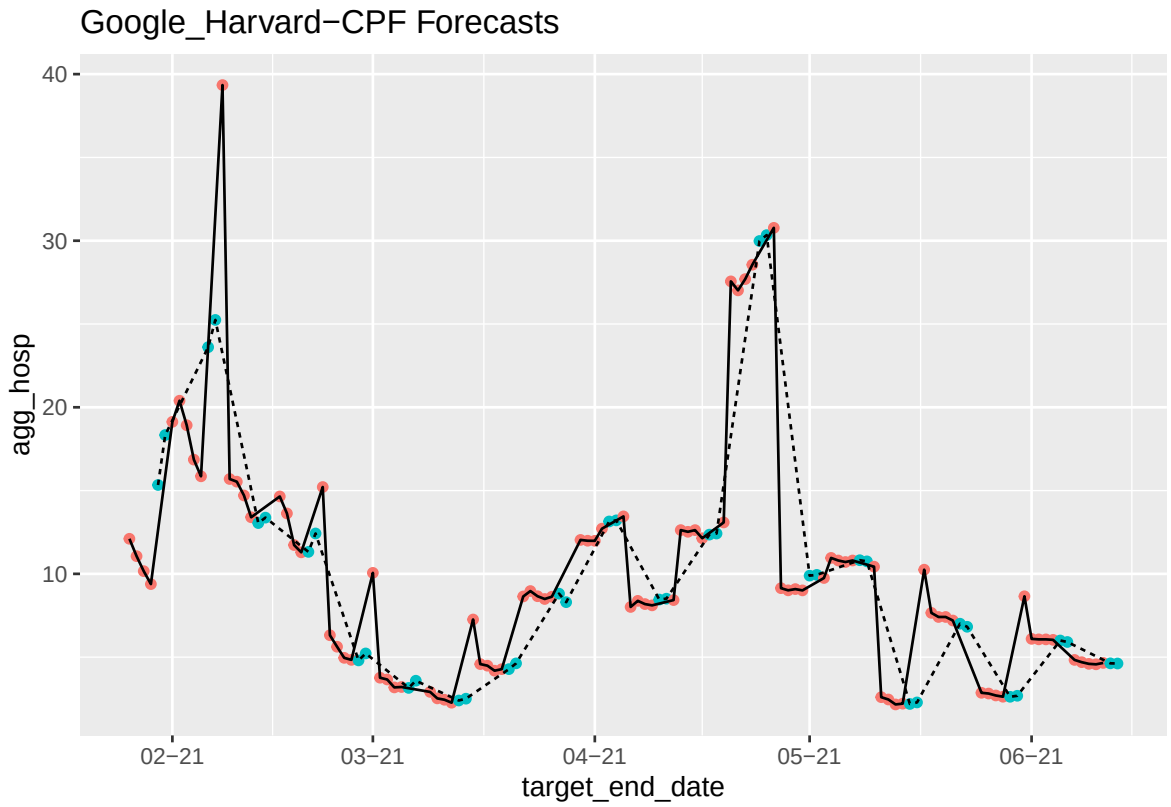
seemed to predict lower inc hosp on Tuesdays for the end of December/beginning of January, then most seem to switch to lower weekend end inc hosp sometime during 2021.

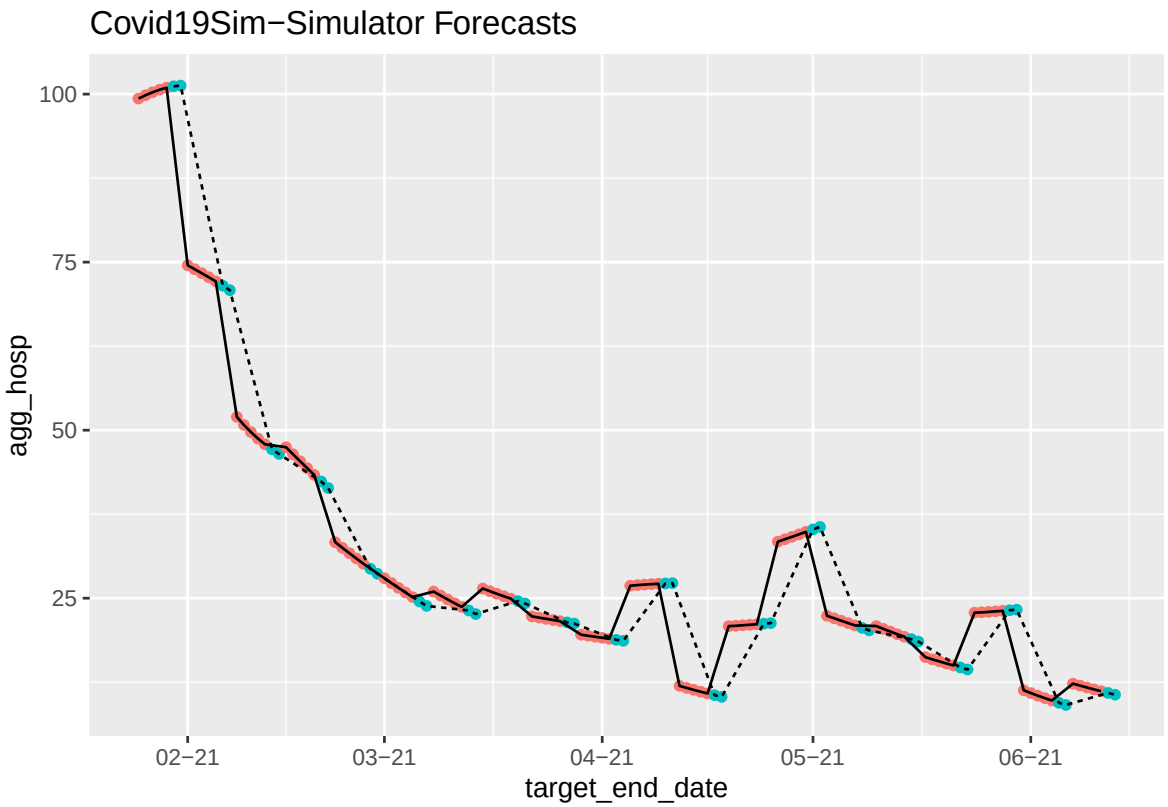


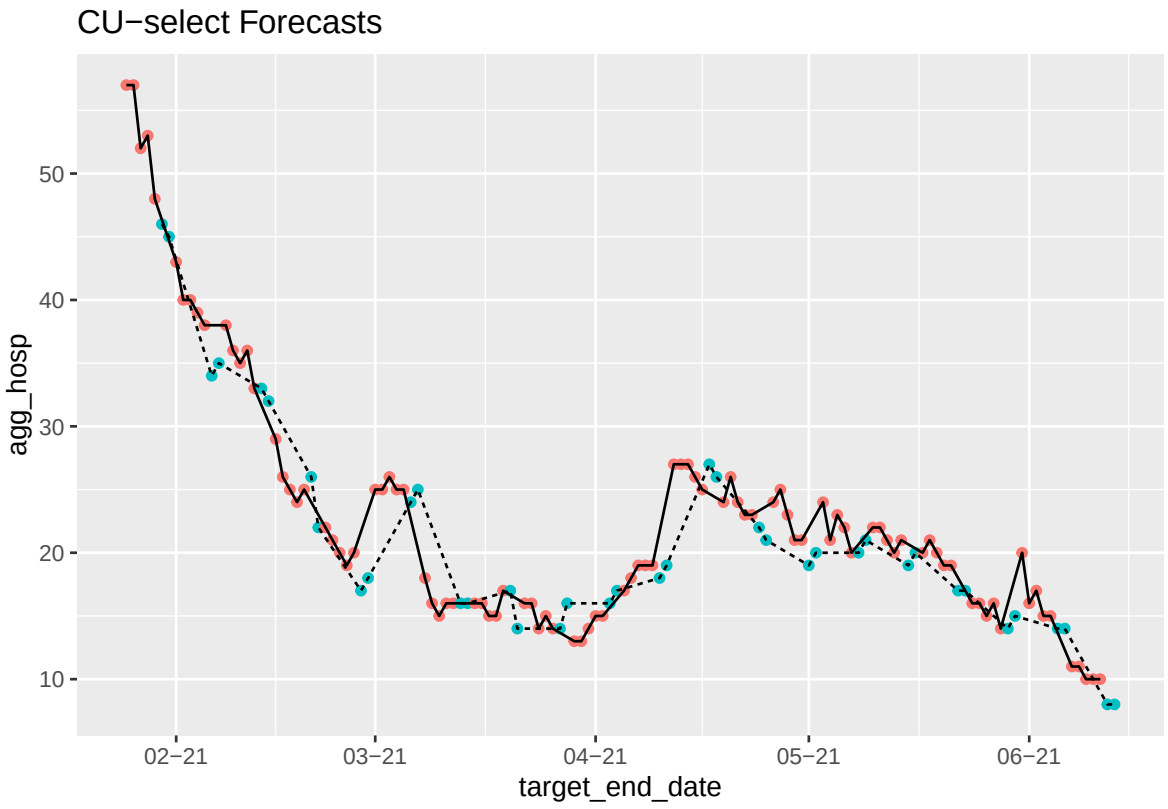

COVIDhub−ensemble Forecasts
agg_hosp
50
40
30
20
10
02−21
03−21
04−21
05−21
06−21
target_end_date

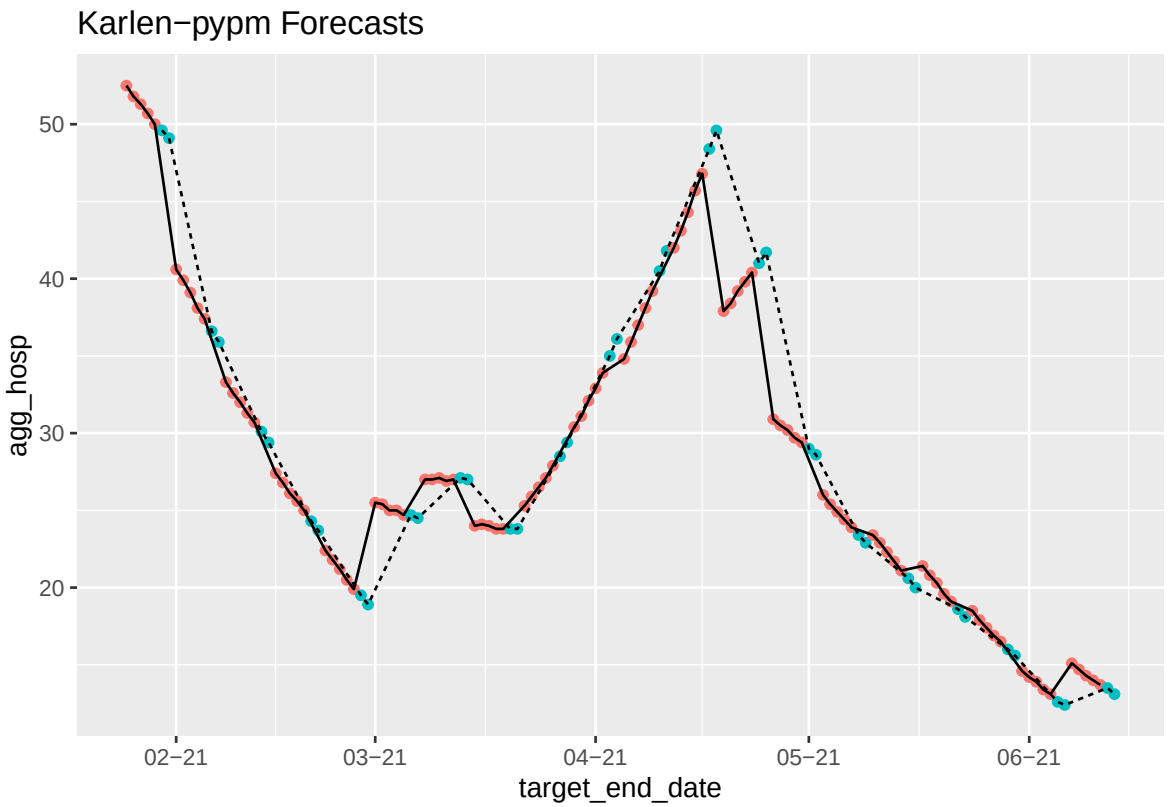

JHUAPL−Bucky Forecasts
agg_hosp
80
60
40
20
02−21
03−21
04−21
05−21
06−21
target_end_date

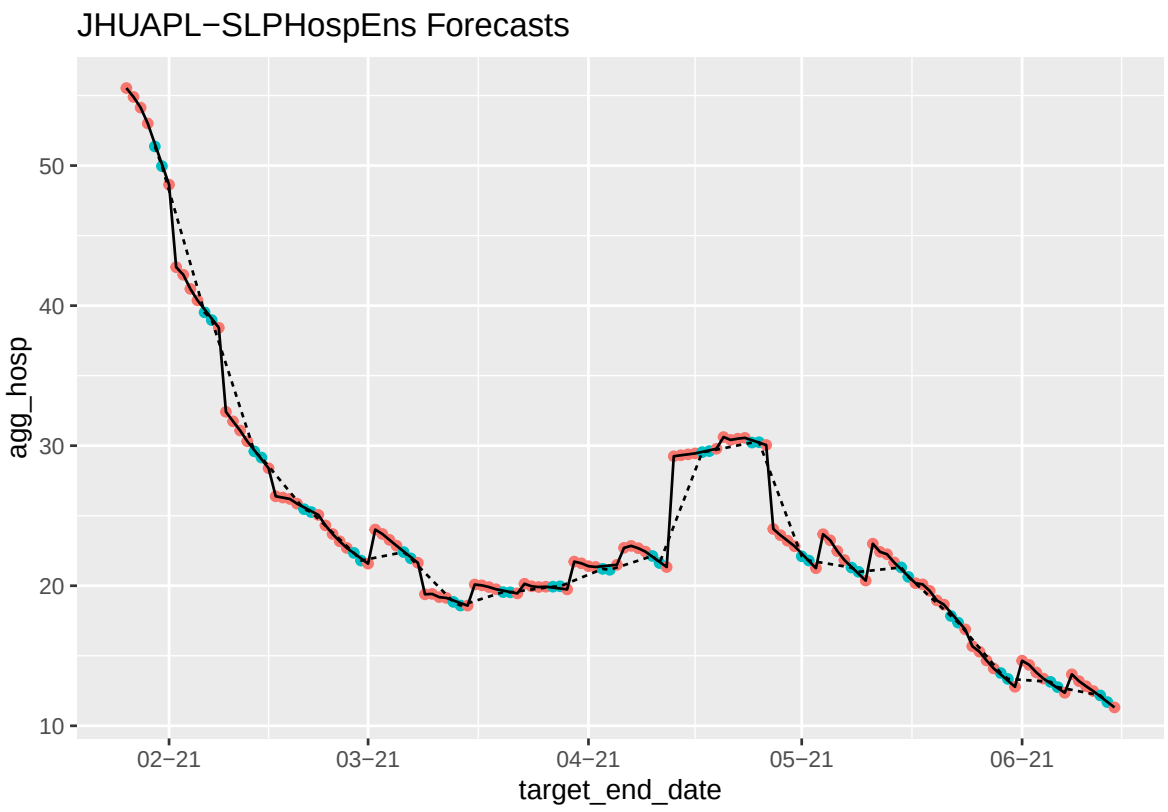

JHUAPL−Gecko Forecasts
100
75
50
25
agg_hosp
02−21
03−21
04−21
05−21
06−21
target_end_date

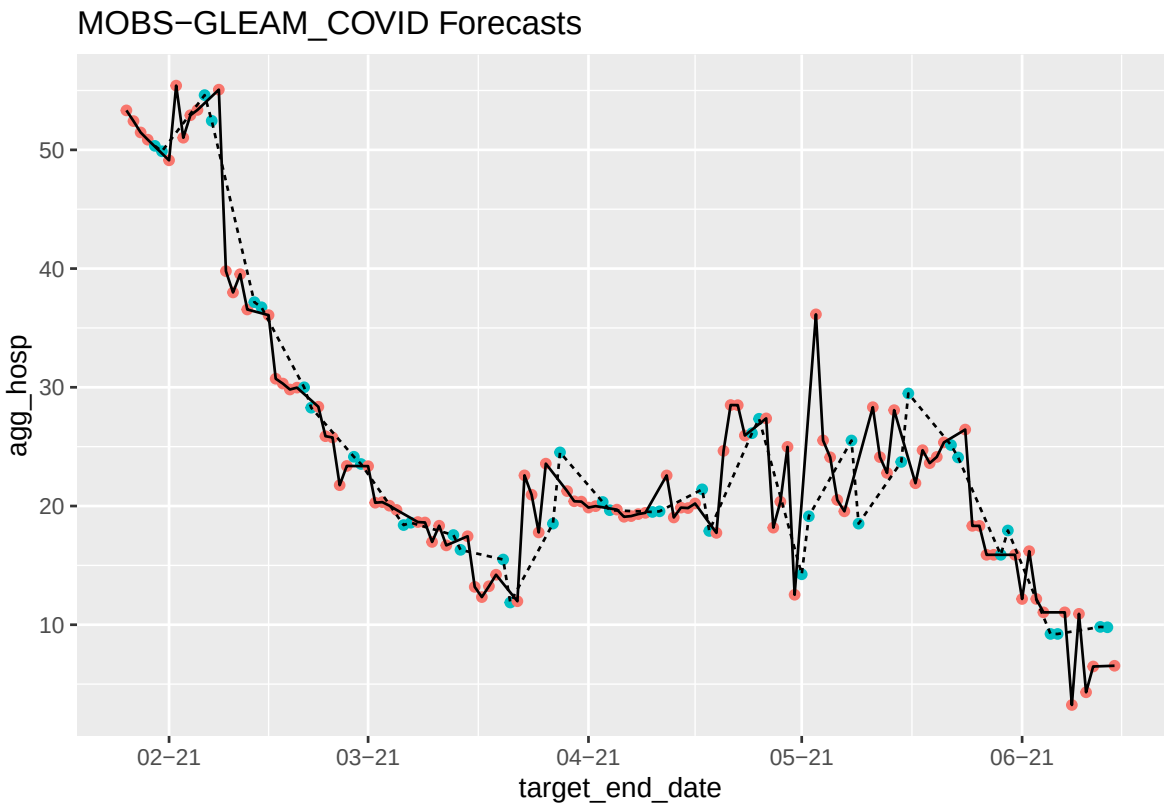

Google_Harvard−CPF Forecasts
40
30
20
10
agg_hosp
02−21
03−21
04−21
05−21
06−21
target_end_date


Covid19Sim−Simulator Forecasts
100
75
50
25
agg_hosp
02−21
03−21
04−21
05−21
06−21
target_end_date


CU−select Forecasts
50
40
30
20
10
agg_hosp
02−21
03−21
04−21
05−21
06−21
target_end_date


Karlen−pypm Forecasts
agg_hosp
50
40
30
20
02−21
03−21
04−21
05−21
06−21
target_end_date


JHUAPL−SLPHospEns Forecasts
agg_hosp
50
40
30
20
10
02−21
03−21
04−21
05−21
06−21
target_end_date



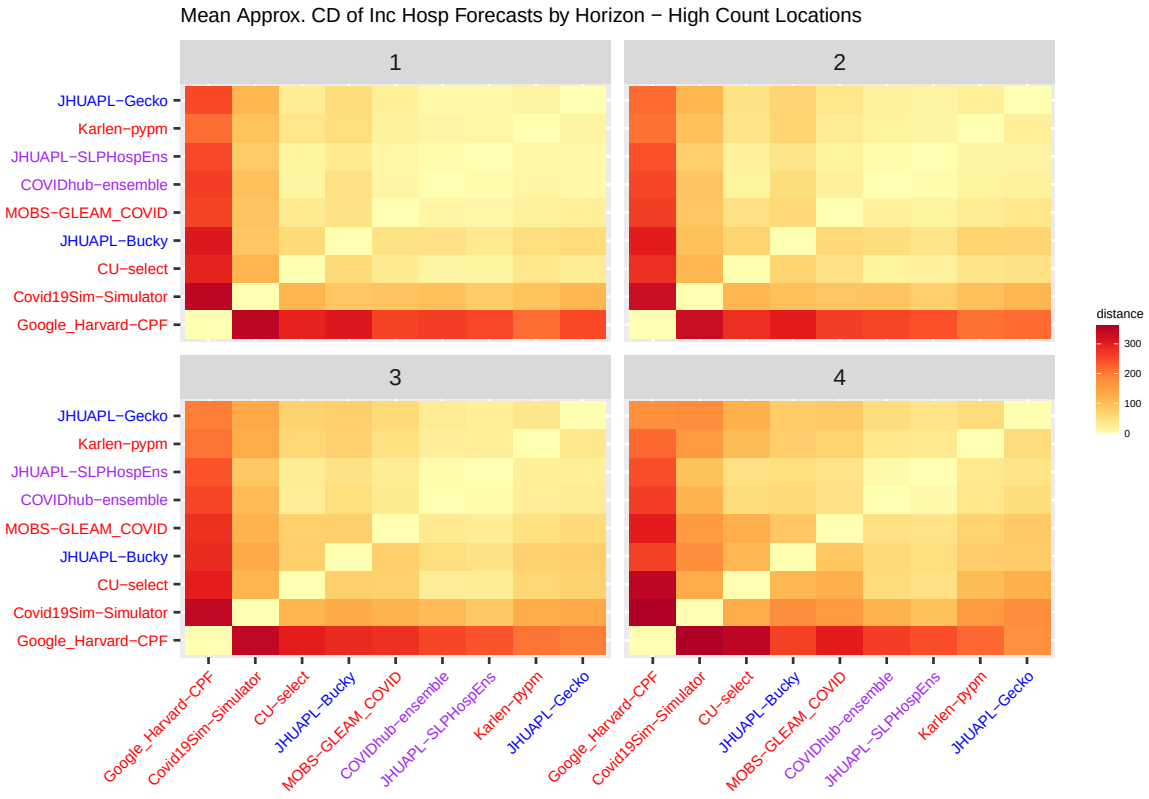
The low hosp location truth data doesn't present such a clear pattern as to day of the week effects that suggest that weekends or weekdays generally have less inc hosp values. However, most of the models seem to switch to incorporating a day of the week effect at about the same time as that for the high count values.

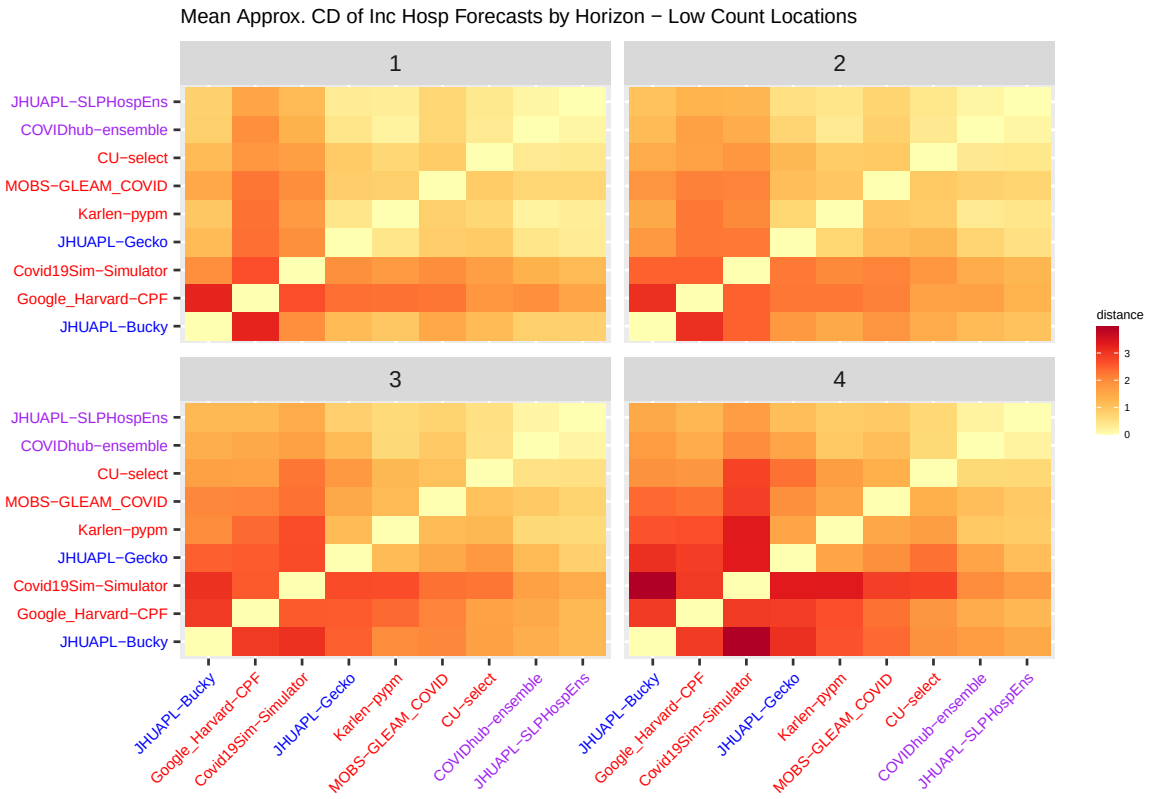
Model	Day of Week Effect
Covid19Sim-Simulator	FALSE
COVIDhub-ensemble	ENS
CU-select	FALSE
Google_Harvard-CPF	FALSE
JHUAPL-Bucky	TRUE
JHUAPL-Gecko	TRUE
JHUAPL-SLPHospEns	ENS
Karlen-pypm	FALSE
MOBS-GLEAM_COVID	FALSE

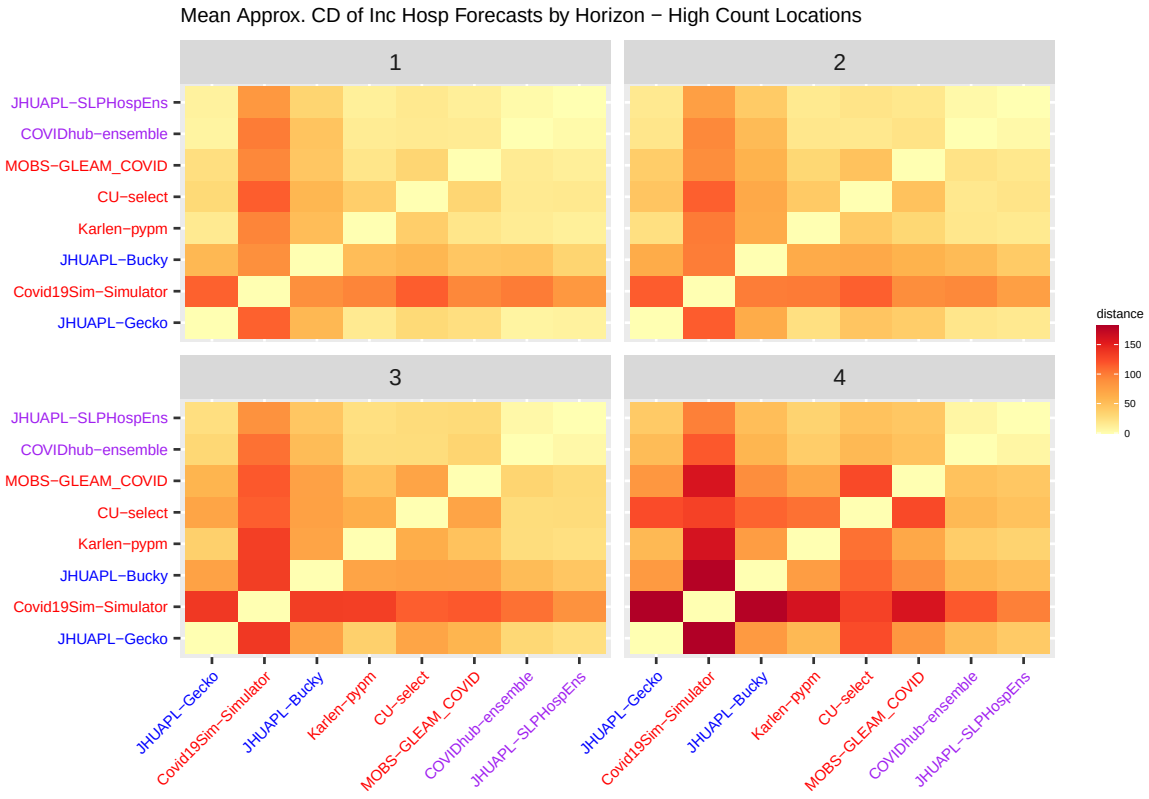
Weekday Analysis

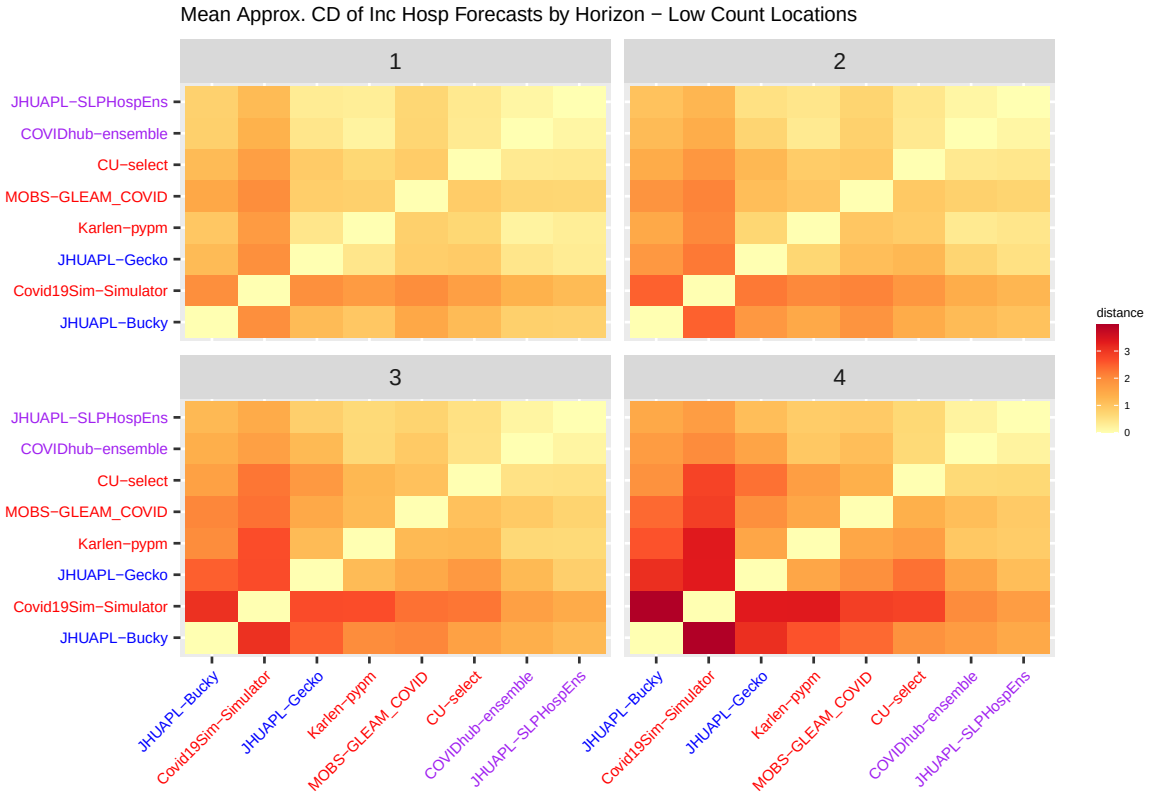
This analysis only examines target end dates for a single day of the week, Thursday, to account for models that include day of the week effects.

We can visualize the mean approximated pairwise distances across all time points in a heat map shown below. 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.


Mean Approx. CD of Inc Hosp Forecasts by Horizon − High Count Locations
1
2
3
4
JHUAPL−Gecko
Karlen−pypm
JHUAPL−SLPHospEns
COVIDhub−ensemble
MOBS−GLEAM_COVID
JHUAPL−Bucky
CU−select
Covid19Sim−Simulator
Google_Harvard−CPF
distance
300
200
100
0


Mean Approx. CD of Inc Hosp Forecasts by Horizon − Low Count Locations
1
2
3
4
JHUAPL−SLPHospEns
COVIDhub−ensemble
CU−select
MOBS−GLEAM_COVID
Karlen−pypm
JHUAPL−Gecko
Covid19Sim−Simulator
Google_Harvard−CPF
JHUAPL−Bucky
distance
3
2
1
0





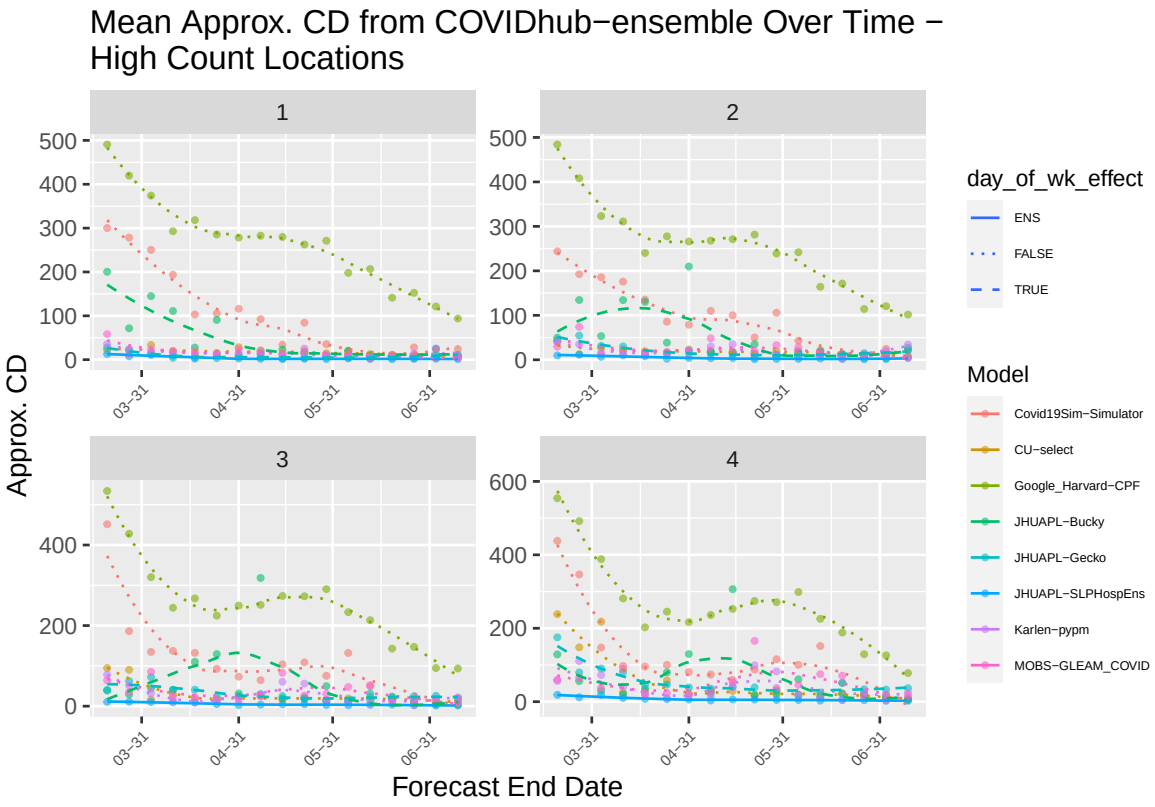
Google_Harvard-CPF is generally the least similar to the other models for both the high count and low

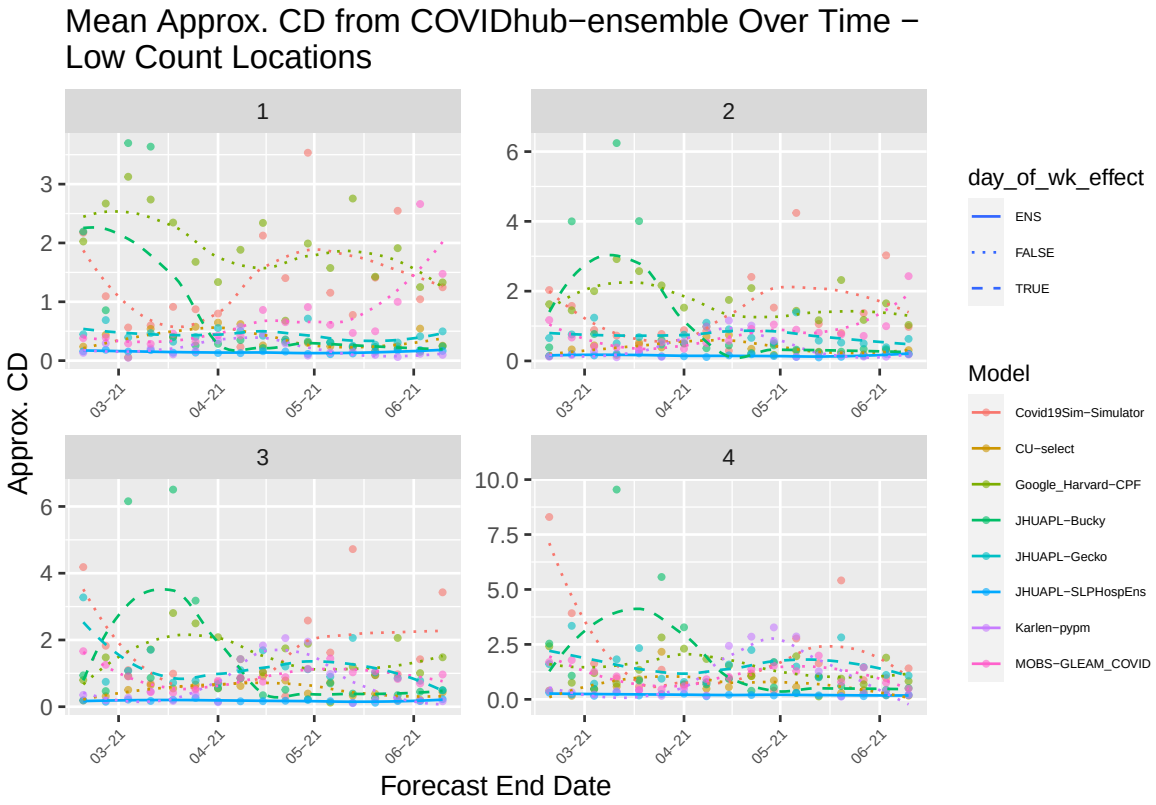
count locations. This is true across all horizons for the high count locations, but only true for the one and two week horizons of the low count locations. Covid19Sim-Simulator, JHUAPL-Bucky, and JHUAPL-Gecko show similar Cramer's Distances at three and four week horizons at the low count locations.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

Covid19Sim-Simulator is the second least similar model for the high count locations, but its Cramer's Distance is much smaller than that of Google_Harvard-CPF. Of note is similar differences between models at all horizons for high count locations, unlike the results shown for inc cases and inc deaths which show substantial differences between models as horizon length increases.

For low count locations, models with day of the week effect tend to have higher Cramer's Distances, but there doesn't seem to be much of a pattern for high count locations.

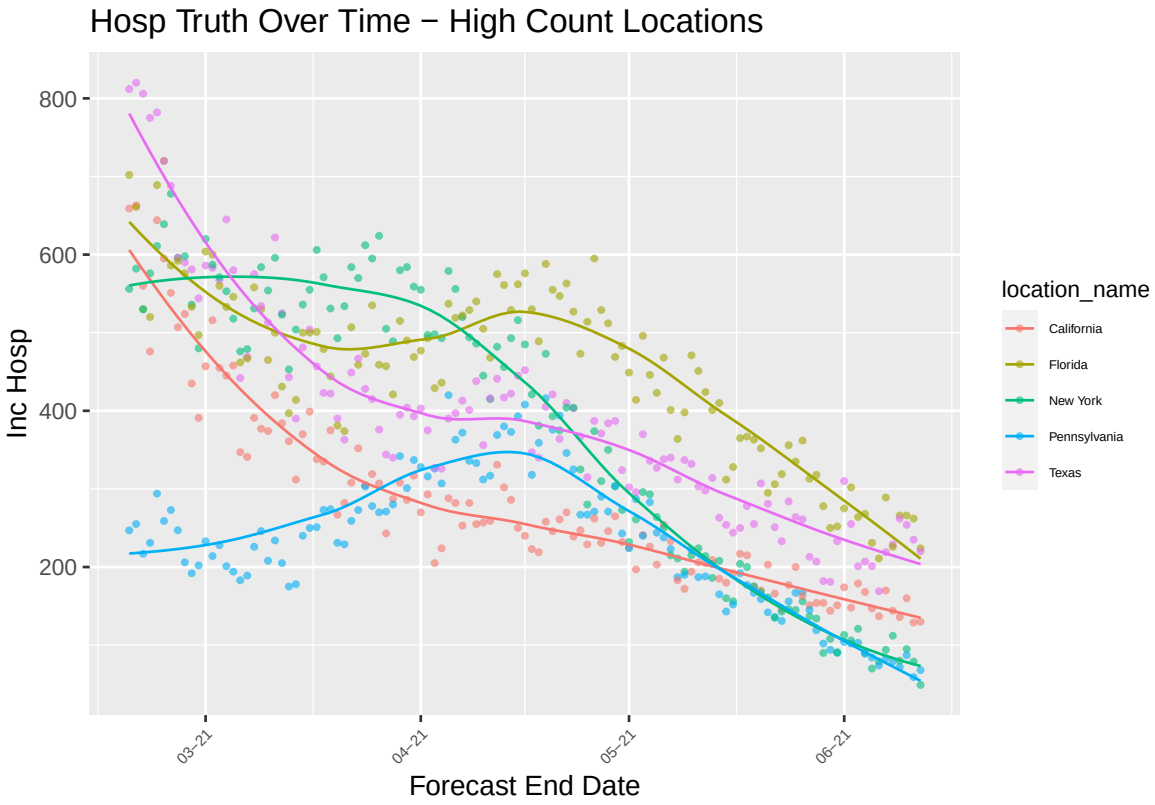
We can also look at the approximated pairwise distances to see how the models become more similar or dissimilar over time.

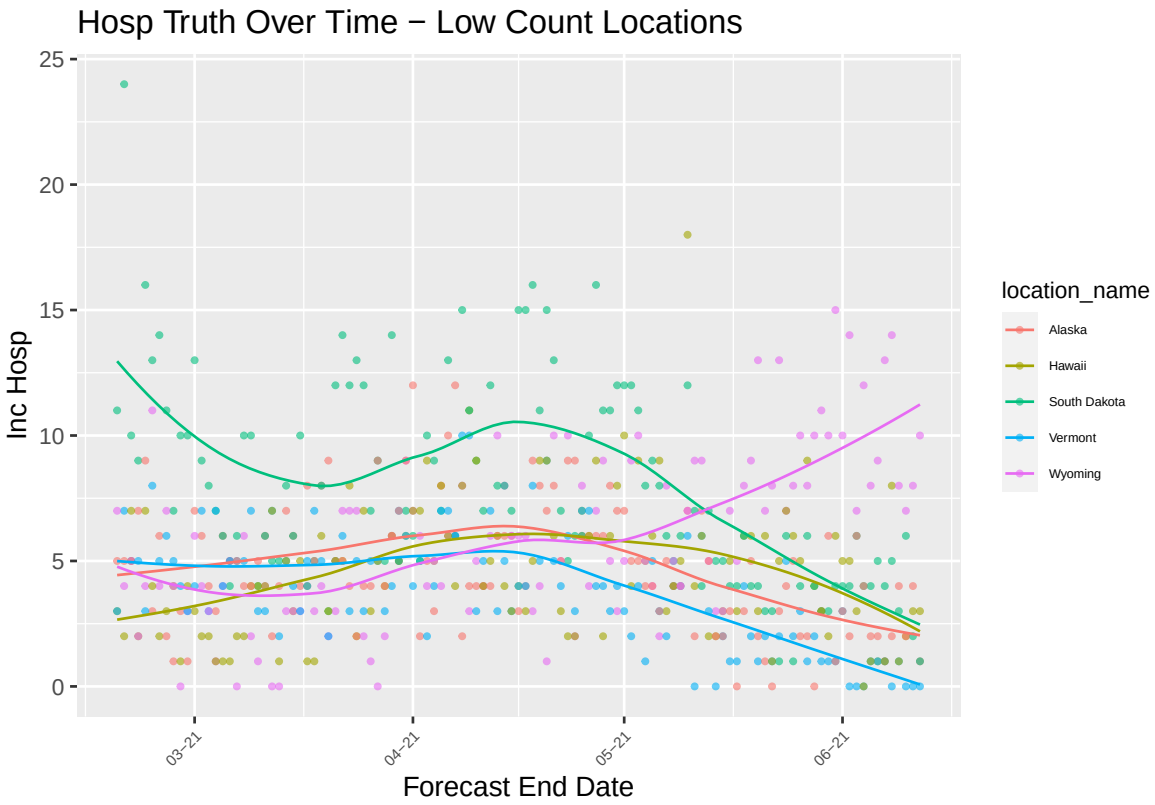




The scatterplots show that the Google_Harvard-CPF, Covid19Sim-Simulator, and JHUAPL-Bucky models tend to differ from the Covidhub-ensemble model compared to the other models. This seems to align with the results shown in the heat maps above that show that Google_Harvard-CPF, Covid19Sim-Simulator, and JHUAPL-Bucky tend to have the highest mean Cramer's Distance from the other models. In high count locations, Google_Harvard-CPF is very different from the ensemble model from February until April. However, in low count locations, JHUAPL-Bucky shows a peak in around March, although this peak is not largely different, as the scale is pretty small.

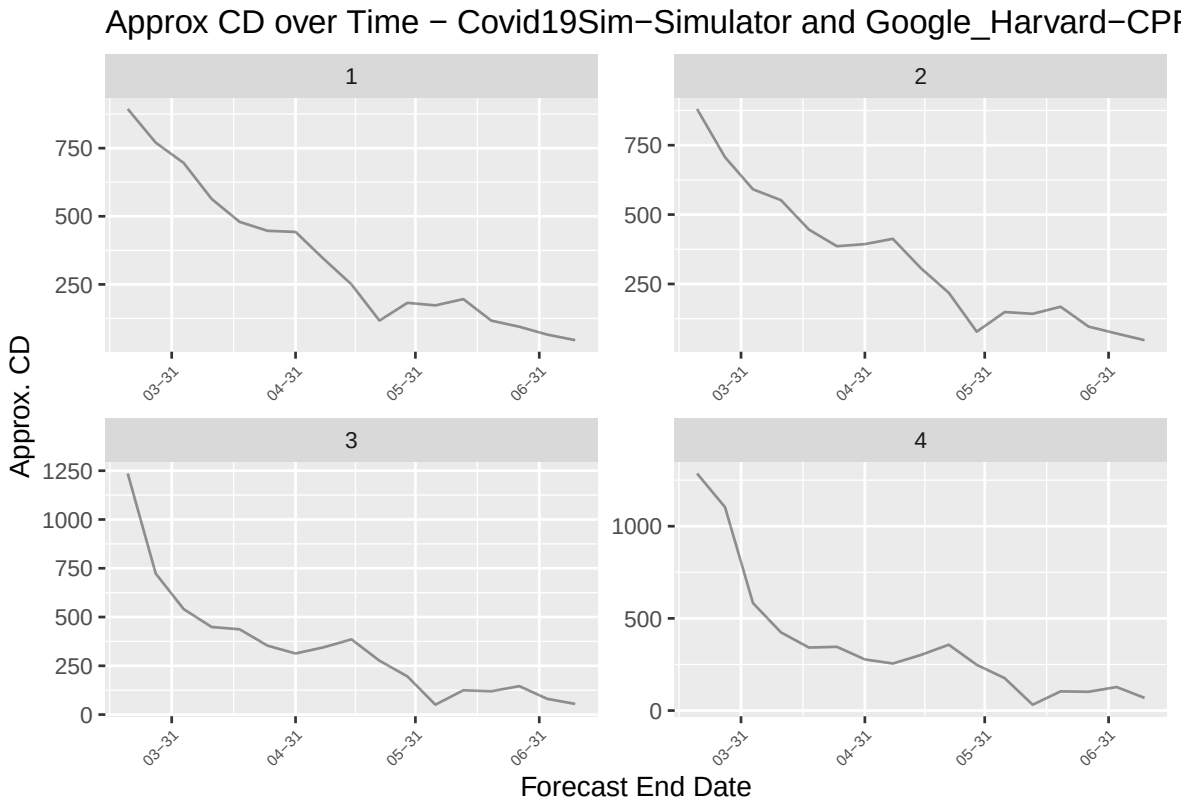
Whether models incorporate a day of the week effect does not seem to have an impact on how much the model differs from the ensemble, nor as to when it differs greatly.

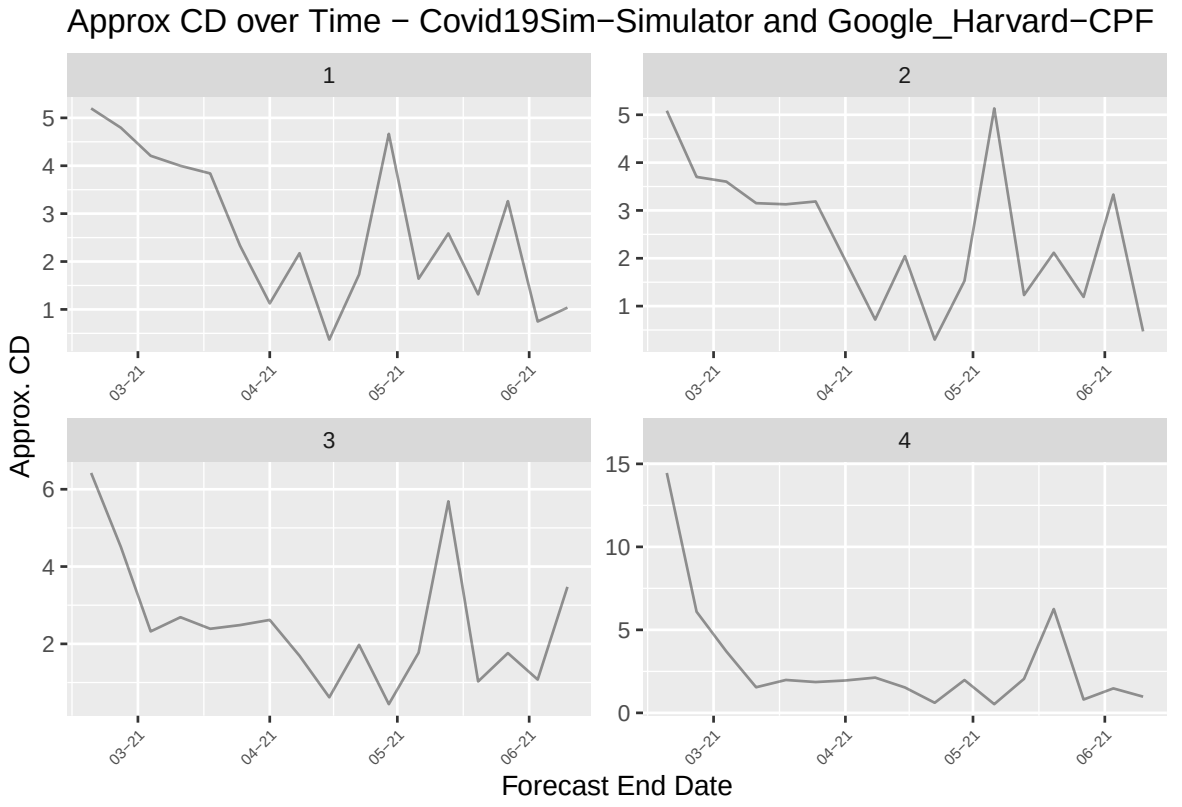




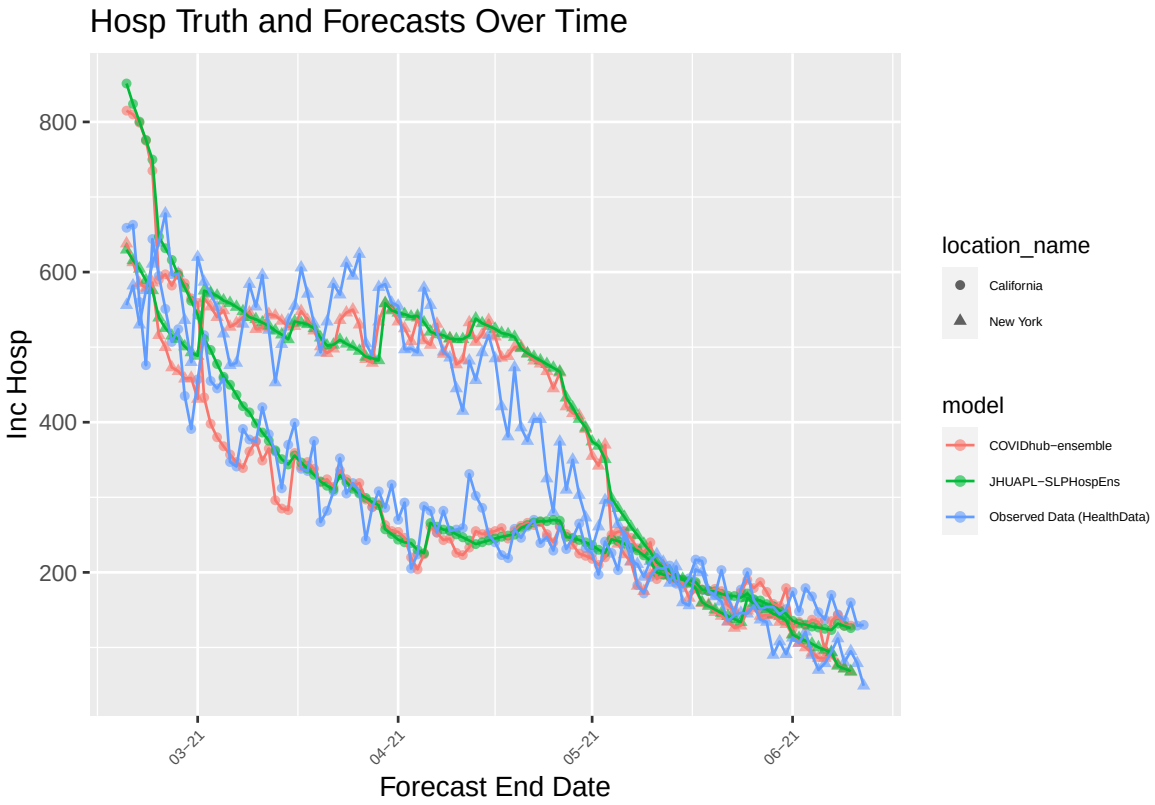
It seems that Google_Harvard-CPF and Covid19Sim-Simulator's differences from the ensemble model follow

the trends shown by the truth data.





These plots indicate that the difference between the models at the high count locations seems to shrink over time. Since the low count locations approx CD is so low, it is difficult to draw conclusions from the plots.



We can also cluster the distances using hierarchical clustering.

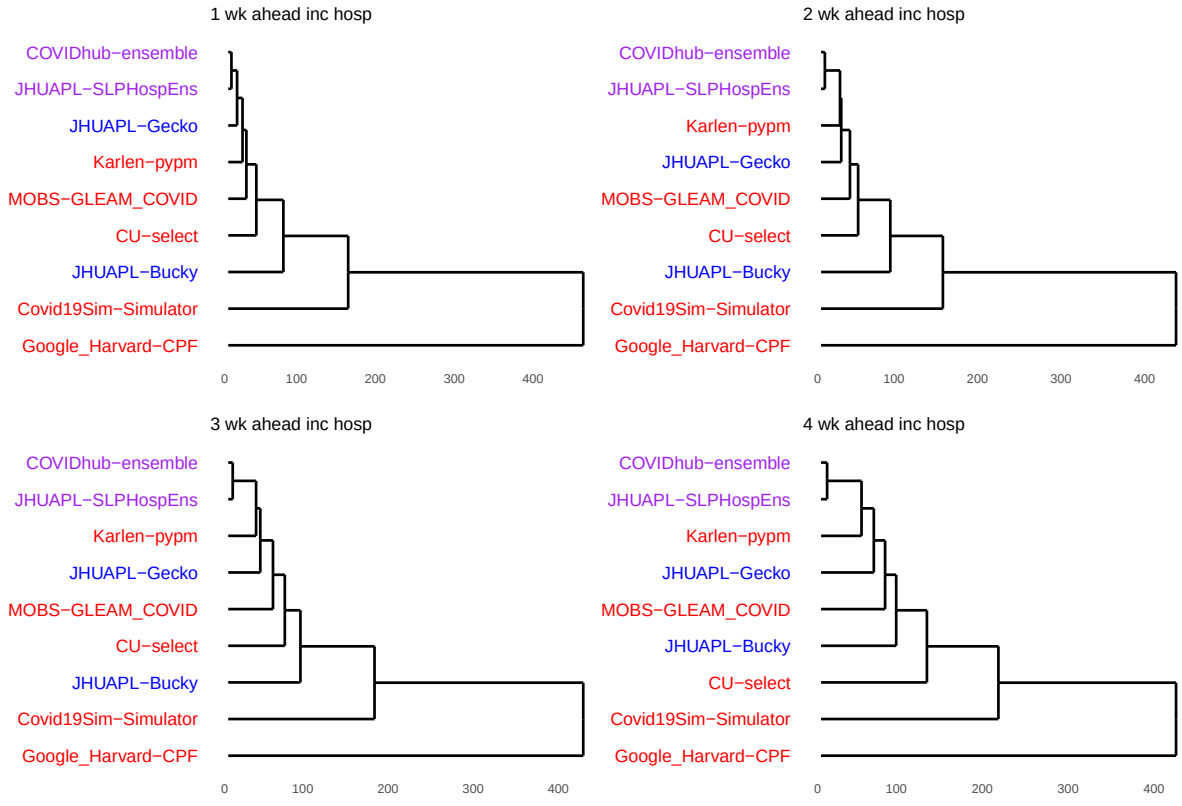


Figure 1: High Hospitalization Count Locations

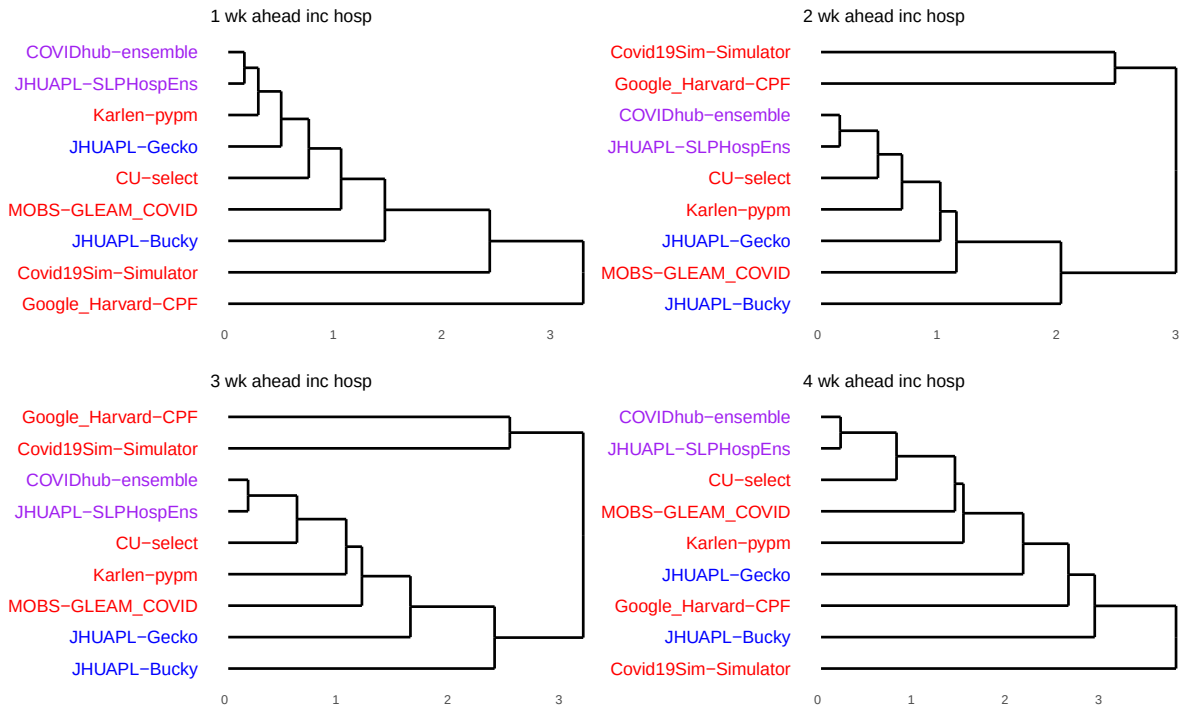


Figure 2: Low Hospitalization Count Locations

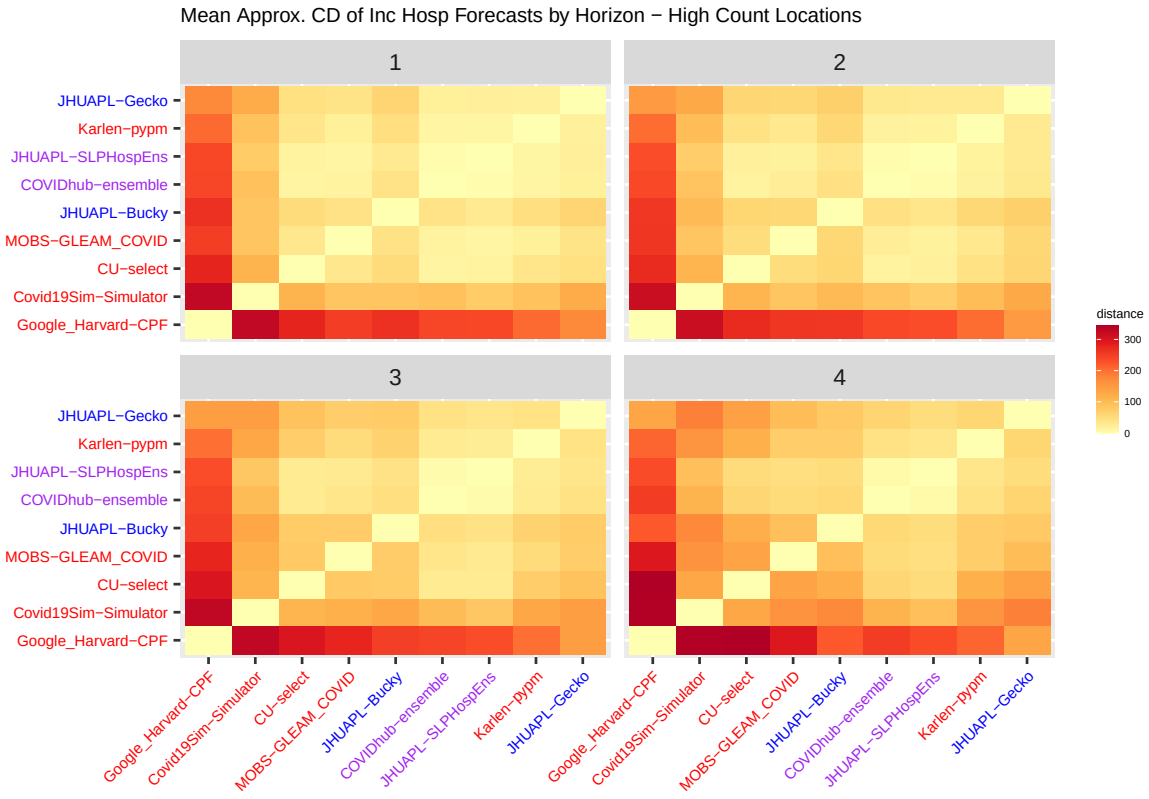
For both high count and low count locations, all of the dendrograms show that Google_Harvard-CPF is generally the most dissimilar to the other models, and substantially so for the high count locations, followed by Covid19Sim-Simulator. However, the scale for differences in Cramer Distance for the low count locations is very small, which may be a result of low hospitalizations, which may explain why Google_Harvard-CPF and Covid19Sim-Simulator show similar Cramer's Distances for low count locations. We can also see that the ensembles are the most similar among models.

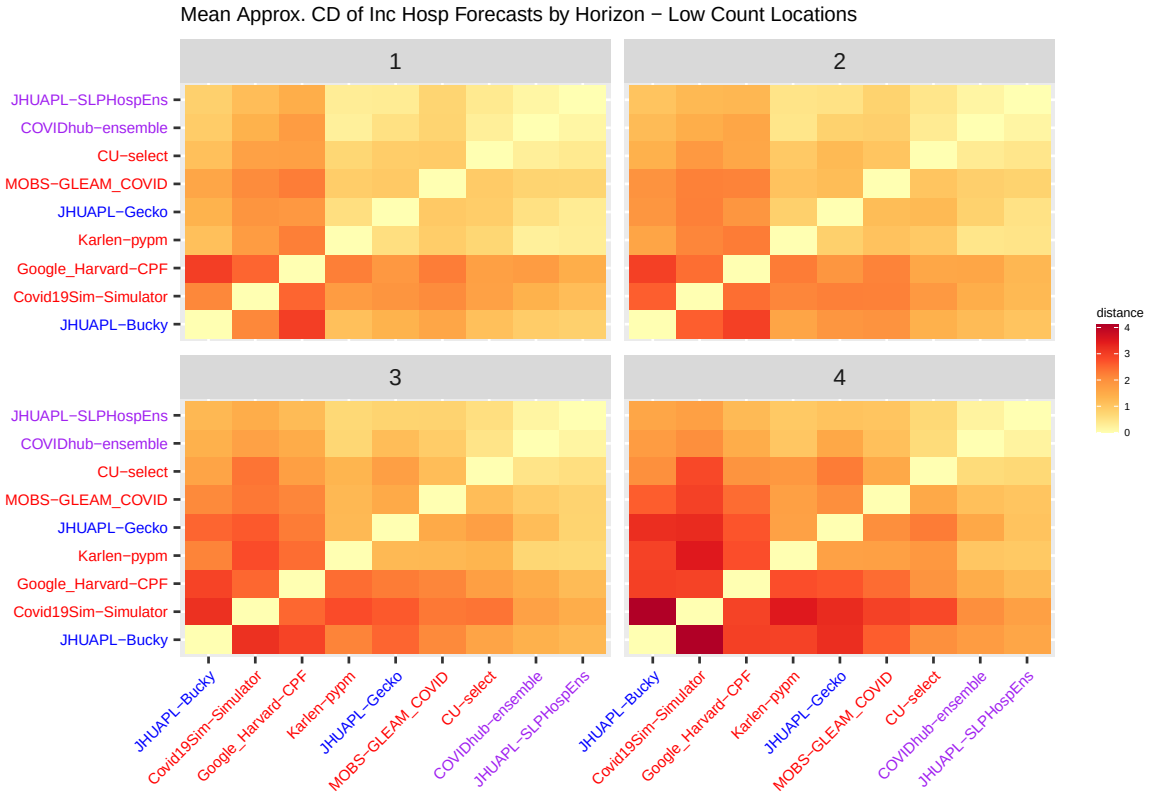
Overall, it seems that Google_Harvard-CPF is consistently the most dissimilar from the other models by a substantial amount, followed by Covid19Sim-Simulator, across almost all horizons for both high-count and low count regions. For Thursday forecasts, it seems models without day of the week effects tend to be more dissimilar from the COVIDhub-ensemble model than models without day of the week effects.

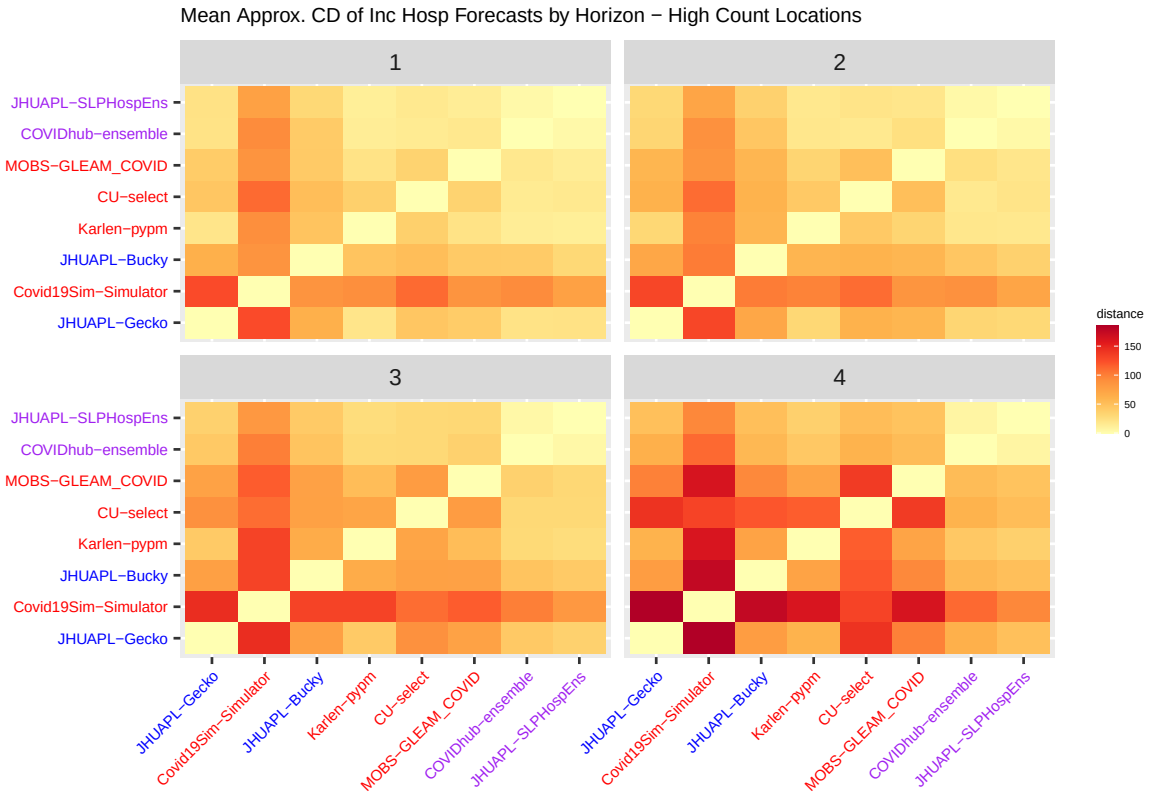
Weekend Analysis

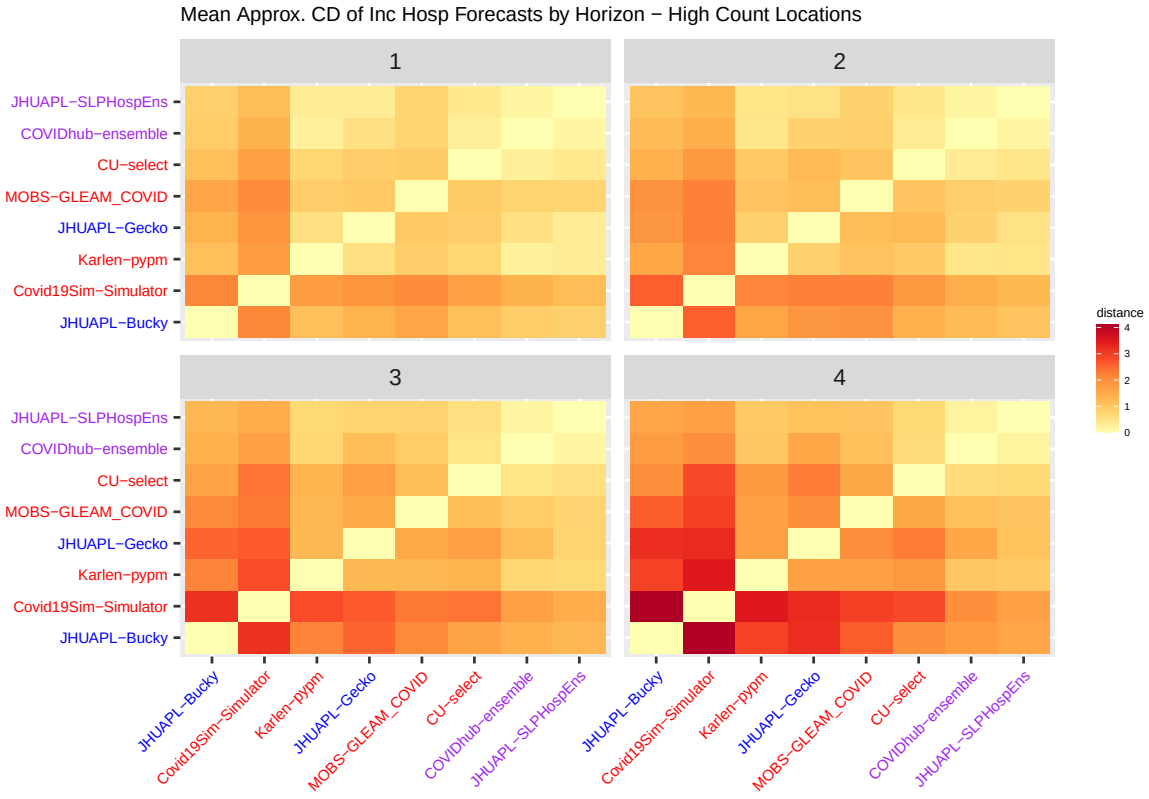
Now we look at forecasts with a target end date on a Saturday to see if day of the week effects change model similarity. We expect forecasts for weekends to show the impact of day of the week effects more strongly than weekdays.

We can visualize the mean approximated pairwise distances across all time points in a heat map shown below. 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.


Mean Approx. CD of Inc Hosp Forecasts by Horizon − High Count Locations
1
2
3
4
JHUAPL−Gecko
Karlen−pypm
JHUAPL−SLPHospEns
COVIDhub−ensemble
JHUAPL−Bucky
MOBS−GLEAM_COVID
CU−select
Covid19Sim−Simulator
Google_Harvard−CPF
distance
300
200
100
0


Mean Approx. CD of Inc Hosp Forecasts by Horizon − Low Count Locations
1
2
3
4
JHUAPL−SLPHospEns
COVIDhub−ensemble
CU−select
MOBS−GLEAM_COVID
JHUAPL−Gecko
Karlen−pypm
Google_Harvard−CPF
Covid19Sim−Simulator
JHUAPL−Bucky
distance
4
3
2
1
0





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

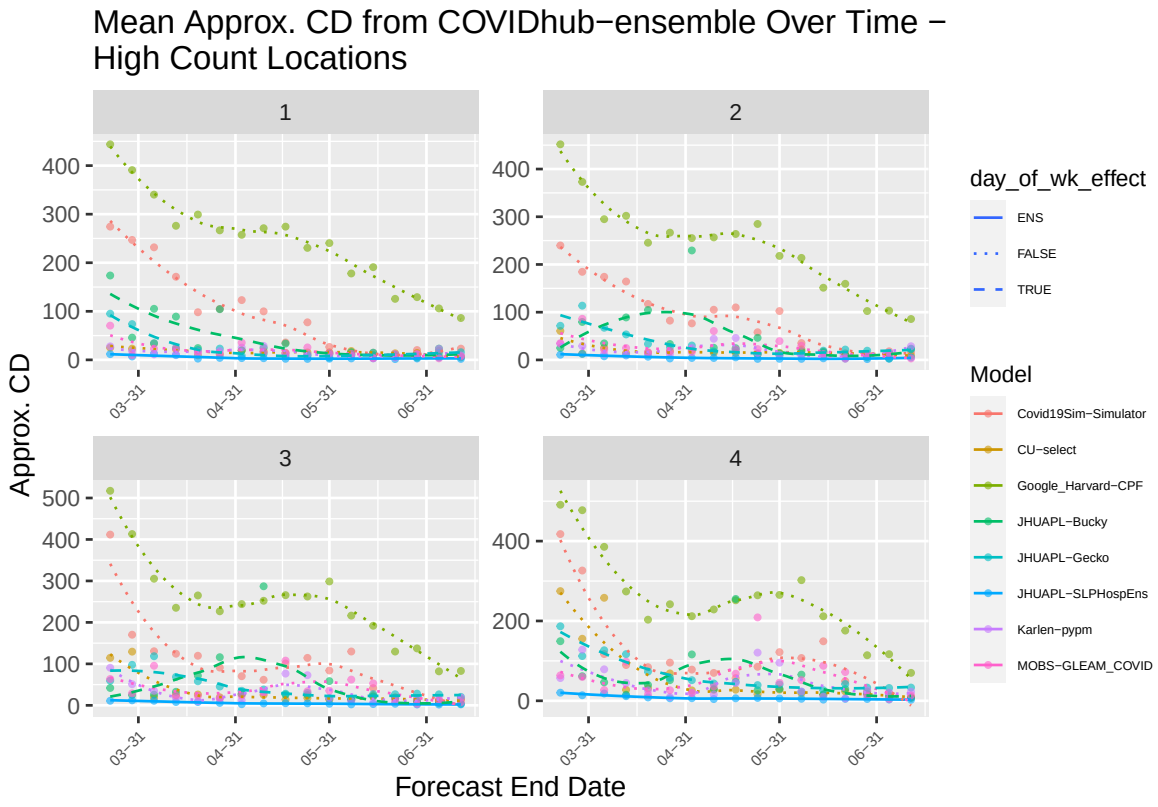
for both the high count and low count locations. This is true across all horizons for the high count locations, but only true for the one and two week horizons of the low count locations. Covid19Sim-Simulator is the most dissimilar for the three and four week horizons of the low count locations. However,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

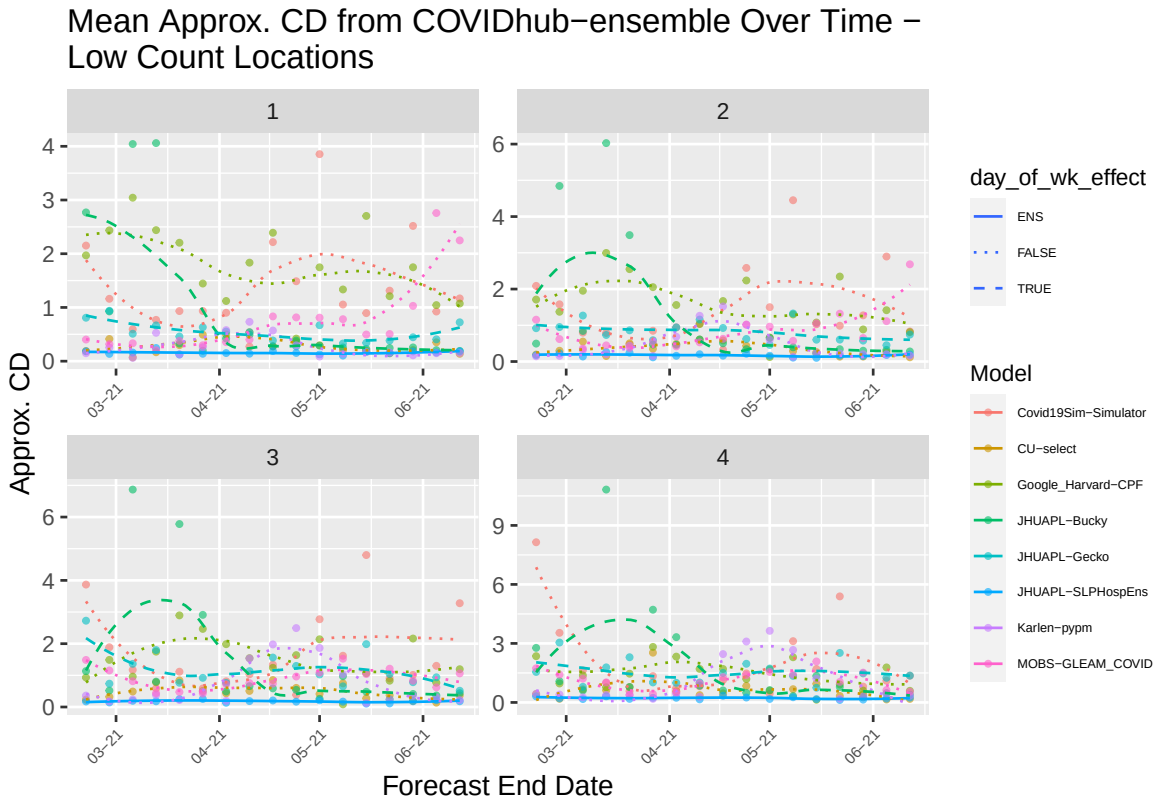
Covid19Sim-Simulator is the second least similar model for the high count locations, but its Cramer's Distance is much smaller than that of Google_Harvard-CPF. Of note is similar differences between models at all horizons for high count locations, unlike the results shown for inc cases and inc deaths which show substantial differences between models as horizon length increases.

For high count locations, models without day of the week effect tend to have higher Cramer's Distances, but it is unclear if there is a pattern for low count locations.

The Saturday forecasts show similar but slightly different heat maps to the Thursday forecasts.

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

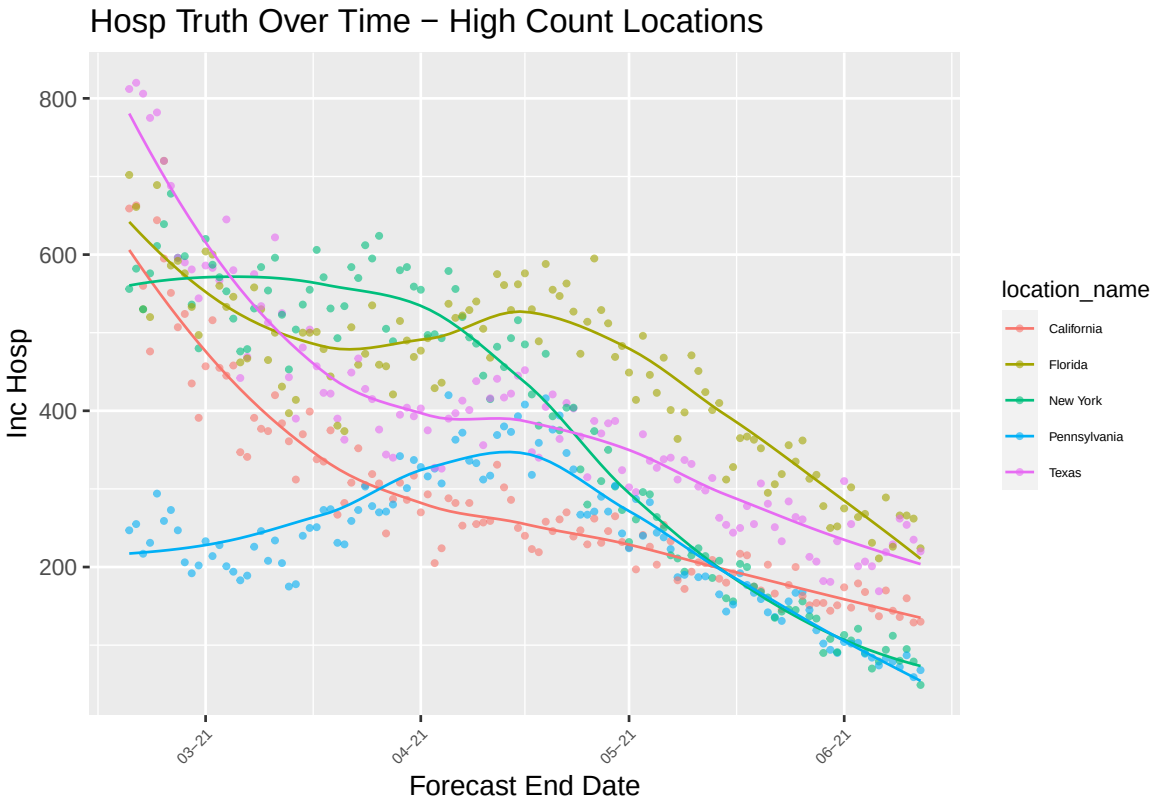


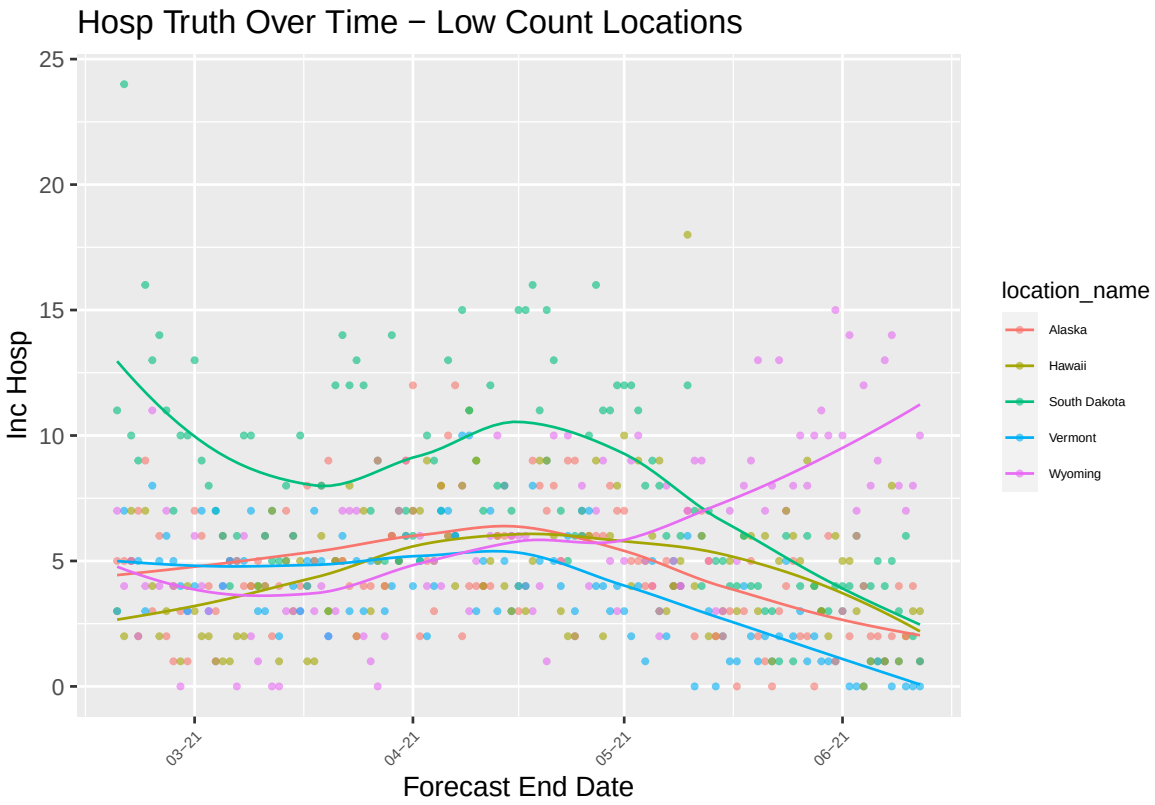


The scatterplots show that the Google_Harvard-CPF, Covid19Sim-Simulator, and JHUAPL-Bucky models tend to differ from the Covidhub-ensemble model compared to the other models. This seems to align with the results shown in the heat maps above that show that Google_Harvard-CPF, Covid19Sim-Simulator, and JHUAPL-Bucky tend to have the highest mean Cramer's Distance from the other models. In high count locations, Google_Harvard-CPF is very different from the ensemble model from February until April. However, in low count locations, JHUAPL-Bucky shows a peak in around March, although this peak is not largely different, as the scale is pretty small.

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.

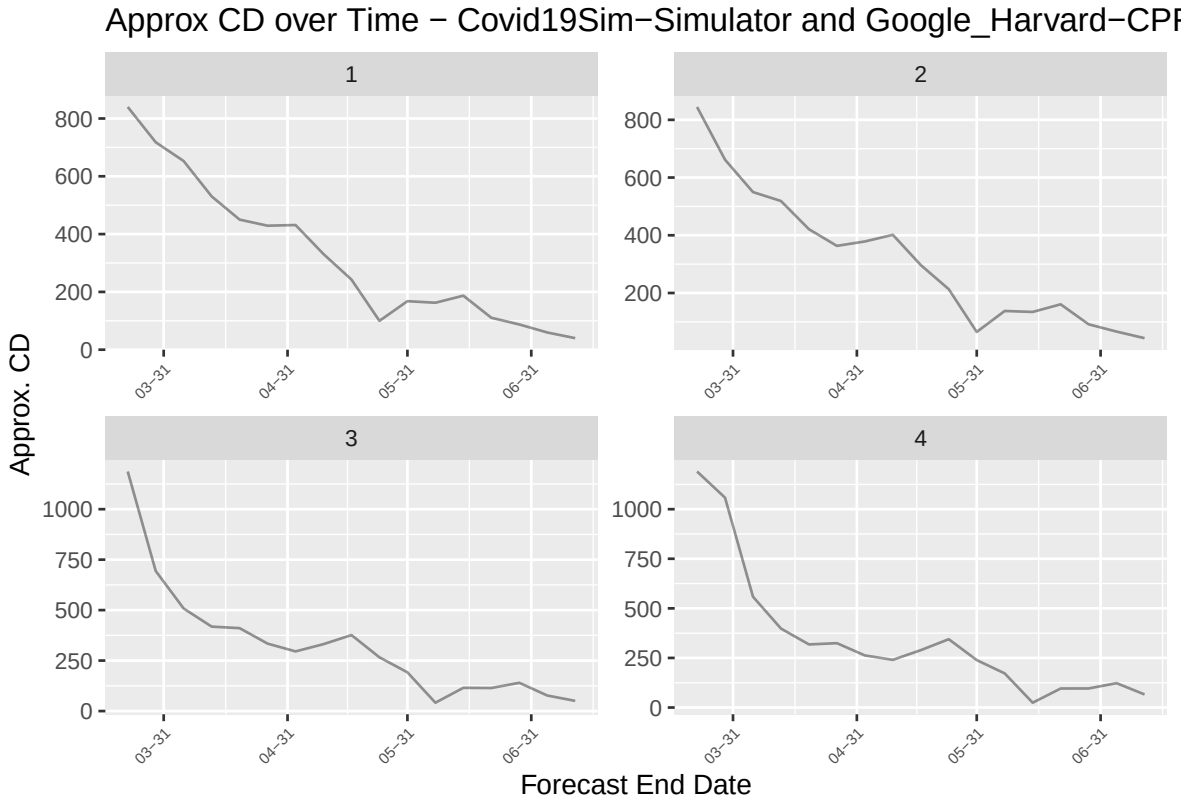
These scatterplots are nearly the same as the ones shown above.

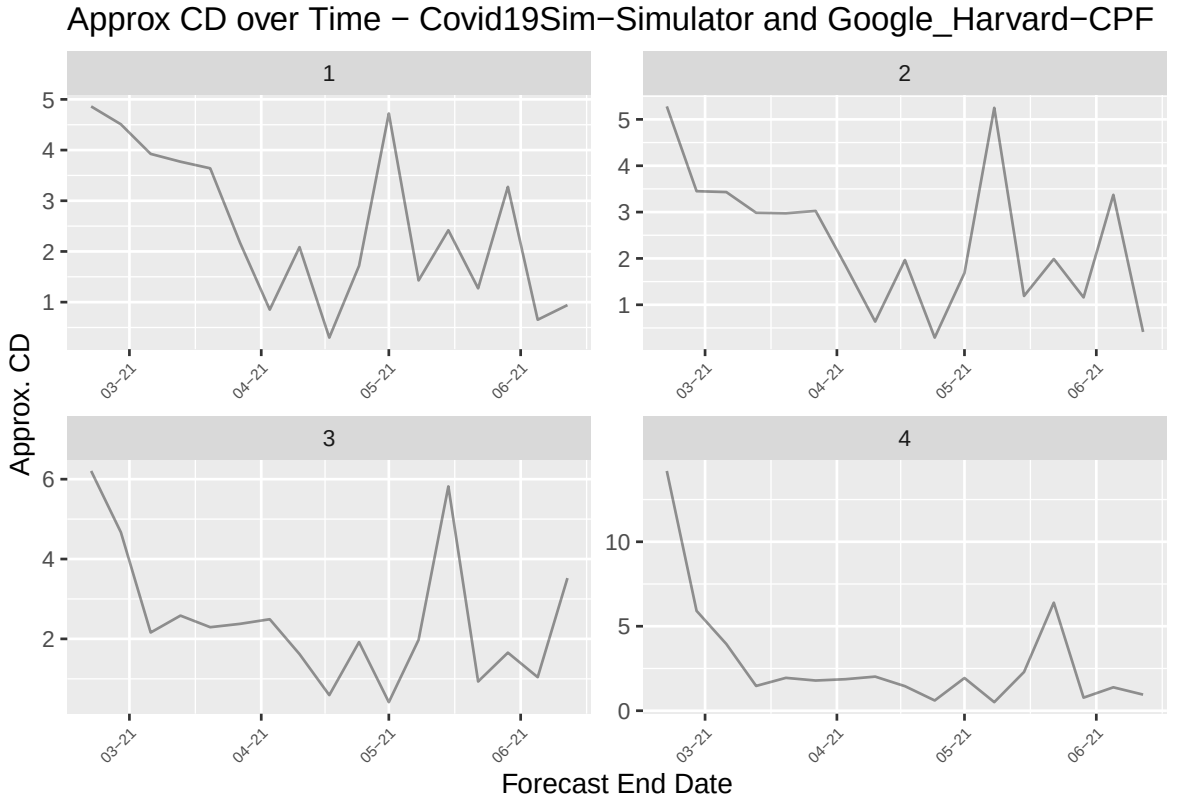




Like with the Thursday forecasts, it seems that Google_Harvard-CPF and Covid19Sim-Simulator's differ-

ences from the ensemble model follow the trends shown by the truth data.





Like with the Thursday forecasts, these plots indicate that the difference between the models at the high count locations seems to shrink over time. Since the low count locations approx CD is so low, it is difficult to draw conclusions from the plots.

We can also cluster the distances using hierarchical clustering.

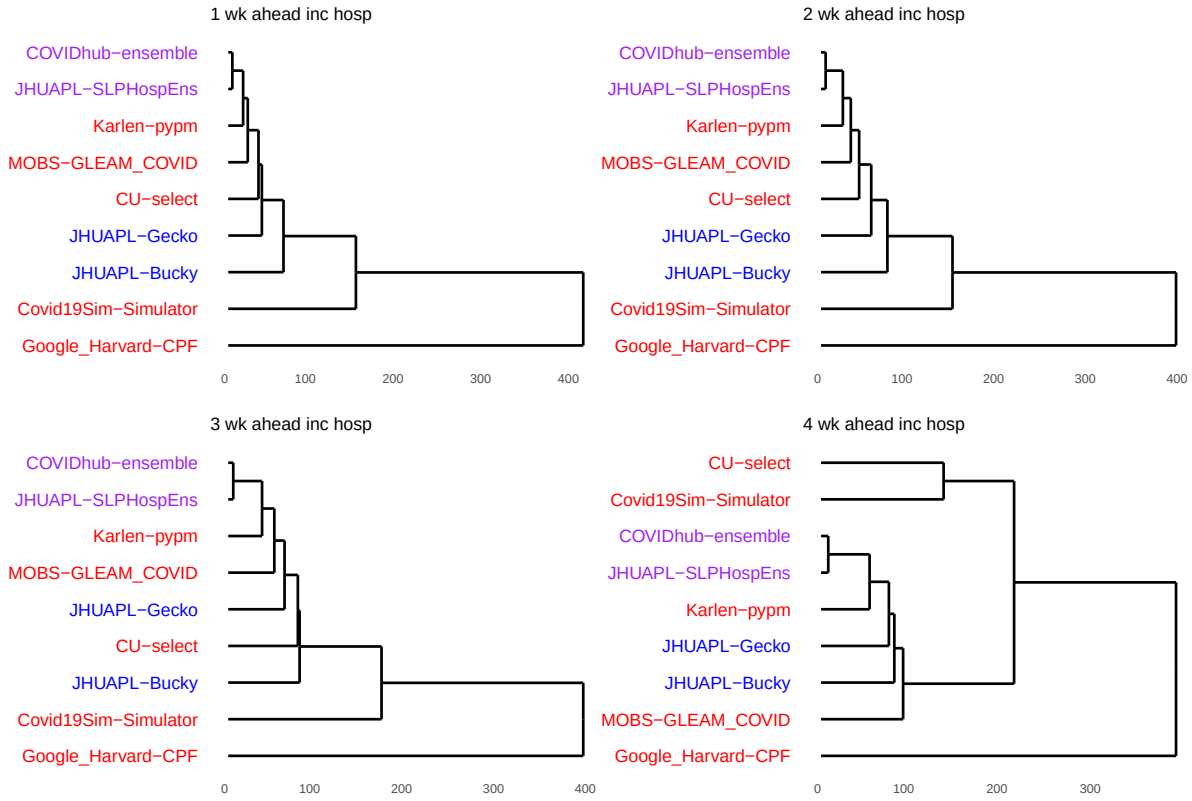


Figure 3: High Hospitalization Count Locations

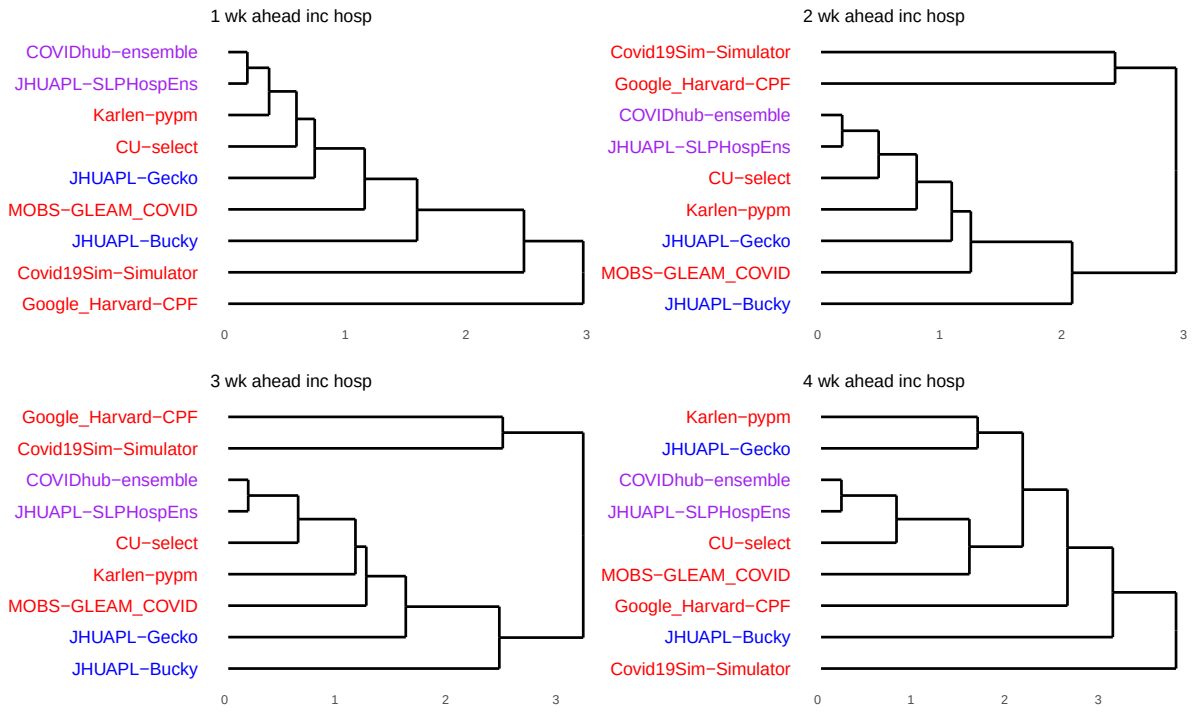


Figure 4: Low Hospitalization Count Locations

For both high count and low count locations, all of the dendrograms show that Google_Harvard-CPF is generally the most dissimilar to the other models, and substantially so for the high count locations, followed by Covid19Sim-Simulator. The one and two week horizons for the high count locations have practically identical dendrograms. The three week horizon one is very similar as well, save for the ordering of CU-select, JHUAPL-Gecko, and JHUAPL-Bucky, but CU-select and JHUAPL-Bucky have almost neglible difference in their Cramer's Distances. However, the shape of the dendrogram changes at the four week horizon for high count locations such that Covid19Sim-Simulator and CU-select are a separate branch together, even though these two models are farther apart for other horizons.

The dendrograms for low count locations are different at each horizon. The one week horizon plot is similar to those for one to three week horizons of high count locations. Meanwhile, the two and three week horizon plots are resemble each other. However, the four week horizon dendrogram is different. At the same time, the scale for differences in Cramer Distance for the low count locations is very small, which may be a result of low hospitalizations, which may explain why there is so much variation in the dendrograms across horizons.

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, the heat maps seem to indicate models without day of the week effects tend to be more dissimilar from the COVIDhub-ensemble model but the dendrograms show less conclusive results.

Day of the week effects may lead to a slight difference between models, but it does not seem to be a significant factor in explaining their differences.