Evaluation of incident death forecasts based on pairwise model comparisons

Johannes Bracher (johannes.bracher@kit.edu)

1 What we could write in a paper (still a bit long)

Comparative evaluation of the considered forecasters $1, \dots, M$ is hampered by the fact that not all of them provide forecasts for the same set of locations and time points. One possible approach is to compare each model to the baseline forecaster B and report

$$\theta_{iB} = \frac{\text{mean WIS of model } i}{\text{mean WIS of baseline model } B},$$

evaluated on the set of forecast targets covered by model i. A value of $0 < \theta_{iB} < 1$ means that model i is better than the baseline, a value of $\theta_{iB} > 1$ means that the baseline is better. Models could then be ranked by their performance relative to the baseline. A problem with this approach is that beating the baseline model can be more challenging in some states or weeks than in others; notably, in periods of quick growth or decline, it is much easier to outperform the baseline ("last observation carried forward") than in periods of relative stability, when the latter may actually be doing a good job. To take this into account we adopt the following procedure: For each pair of forecasters i and j we compute the pairwise relative WIS skill

$$\theta_{ij} = \frac{\text{mean WIS of model } i}{\text{mean WIS of model } j}$$

based on the available overlap of forecast targets. Subsequently we compute for each model the geometric mean of the results achieved in the different pairwise comparisons, denoted by

$$\theta_i = \left(\prod_{i=1}^{M} \theta_{ij} \right)^{1/M}. \tag{1}$$

Then θ_i is a measure of relative skill of model i with respect to the set of all other models $1, \dots, M$. The central assumption here is that performing well relative to individual models $1, \dots, M$ (not including the baseline) is similarly difficult for each week and location, so that no model can gain an advantage by focusing on just some of them. We note that the baseline model is not included in these pairwise comparisons, because the difficulty of beating the baseline model can vary considerably over time and space. As is, θ_i is a comparison to a hypothetical "average" model. Because we consider a comparison to the baseline model more straightforward to interpret, we rescaled θ_i and report

$$\theta_i^* = \frac{\theta_i}{\theta_B},$$

where θ_B the geometric mean of the results achieved by the baseline model in pairwise comparisons to all other models. The quantity θ_i^* then describes the relative performance of model i, adjusted for the difficulty of the forecasts model i made, and scaled so the baseline model has relative performance of one. For simplicity, we refer to θ_i^* as the "relative WIS" or "relative MAE" throughout the manuscript. Again, a value of $0 < \theta_i^* < 1$ means that model i is better than the baseline, a value of $\theta_i^* > 1$ means that the baseline is better. Corrected relative WIS skills can also be reported stratified by week or forecast horizon to get more detailed insights on relative performance.

To assess more formally whether the forecast performance between two models i and j differs we apply a permutation test with the relative WIS skill θ_{ij} as the test statistic. To generate one sample from the reference distribution we proceed as follows:

1. Split the available pairs of scores from i and j into blocks by forecast date. This serves to account for dependencies between forecasts made at the same time, but for different states and horizons.

2. For each block independently and with probability 1/2 either flip all pairs of scores between i and j or none.

3. Compute the pairwise relative WIS skill $\theta_{ij}^{\text{perm}}$ from the permuted scores.

We then generate a large number of samples from the reference distribution. To compute a **two-sided** p-value we compute the fraction of samples where $\max(\theta_{ij}^{\text{perm}}, \theta_{ji}^{\text{perm}})$ exceeds the observed $\max(\theta_{ij}, \theta_{ji})$. (Previously I had described a one-sided test here, but the implementation was always two-sided).

2 Longer justification of the approach

I tried to update this to the newest notation and terminology

Consider forecasters $i = 1, \ldots, M$ and denote this set of forecasters by $\mathcal{M}$. In addition there is one baseline forecaster which we denote by B. We would like to summarize the performance of each forecaster over a given set $\mathcal{A}$ of forecast targets (combinations of forecast time points, horizons and locations) into a single number. This number shall enable primarily two types of comparisons:

- comparison to the baseline forecaster to judge whether a model offers at least some additional insight.
- comparison to other available forecast models.

If all forecasters covered the whole set $\mathcal{A}$ of forecast tasks, two possible options are:

- The *relative WIS* with respect to the baseline model:

$$\gamma_{iB} = \frac{\text{mean WIS model } i}{\text{mean WIS of baseline model } B} = \begin{cases} < 1 \text{ if } i \text{ better than baseline} \\ > 1 \text{ if } i \text{ worse than baseline} \end{cases} \tag{2}$$

  This describes the relative (multiplicative) improvement in mean WIS model i offers to the baseline for the set $\mathcal{A}$ of forecast tasks.

- The relative WIS skill with respect to the set of models $\mathcal{M}$:

$$\gamma_i = \frac{\text{mean WIS model } i}{\left(\prod_{m=1}^{M} \text{mean WIS model } m\right)^{1/M}} = \begin{cases} < 1 \text{ if } i \text{ better than average of models in } \mathcal{M} \\ > 1 \text{ if } i \text{ worse than avergage} \end{cases} \tag{3}$$

  This describes the relative (multiplicative) improvement in mean WIS model i offers to the average performance of all considered models.

The summary scores γ_{iB} and γ_i use different references, but both contain the information needed to compare each pair of models as we can recover the relative WIS skill with respect to any model j:

$$\gamma_{ij} = \frac{\text{mean WIS model } i}{\text{mean WIS model } j} = \frac{\gamma_{iB}}{\gamma_{jB}} = \frac{\gamma_i}{\gamma_j} = \begin{cases} < 1 \text{ if } i \text{ better than } j \\ > 1 \text{ if } i \text{ worse than } j \end{cases} \tag{4}$$

In particular we can also recover the γ_{iB} from the γ_i as

$$\gamma_{iB} = \frac{\gamma_i}{\gamma_B}. \tag{5}$$

Things become more complicated if not all forecasters covered all forecast tasks, meaning that we can then no longer compute the γ_{iB}, γ_i and γ_{ij}. We assume that for each pair i, j of forecasters there is at least some overlap of forecast tasks, which we denote by $\mathcal{A}_{ij}$. We can then still compute

$$\theta_{iB} = \frac{\text{mean WIS model } i \text{ on } \mathcal{A}_{iB}}{\text{mean WIS of baseline model } B \text{ on } \mathcal{A}_{iB}}, \tag{6}$$

and all θ_{ij}, which are defined analoguously. A difficulty of using θ_{iB} to summarize relative performance of model i is that beating the baseline model is easier in some periods and states than in others. In situations of relative stability (when one is essentially forecasting a flat line) it can tough to beat the Forecast Hub baseline model. In periods of rapid increase or decline this is much easier. So the θ_{iB} may depend a lot on which targets were covered by a forecaster. This means comparing the relative skills θ_{iB} and θ_{jB} may be misleading, as the direct comparison θ_{ij} and the indirect one θ_{iB}/θ_{jB} can return quite different results.

It is reasonable to assume that the full set $\mathcal{M}$ of models offers a more stable reference to assess performance than the baseline model, and it should be similarly difficult to be good relative to $\mathcal{M}$ irrespective of the exact set of forecast targets covered by a given model. I therefore suggest to take into account all

$$\theta_{ij} = \frac{\text{mean WIS of model } i \text{ on } \mathcal{A}_{ij}}{\text{mean WIS of model } j \text{ on } \mathcal{A}_{ij}}, \tag{7}$$

when computing summary scores. To this end we start by noting that if all forecasts were available for all models,

$$\gamma_i = \left(\prod_{m=1}^{M} \gamma_{im} \right)^{1/M}. \tag{8}$$

would hold. We then analoguously compute

$$\theta_i = \left(\prod_{m=1}^{M} \theta_{im} \right)^{1/M}, \tag{9}$$

which describes how model i is faring on average in all possible pairwise comparisons with other models. This implicitly takes into account the fact that forecasting is more or less difficult in different situations, assuming that this relative difficulty behaves similarly for all "serious" models.

I argue that θ_i is a fair comparison relative to the set of models $\mathcal{M}$. Comparing these relative skills for two models i and j is reasonable as empirically (see next section. Figures 5 and 6)

$$\theta_{ij} \approx \frac{\theta_i}{\theta_j}, \tag{10}$$

with better agreement than for θ_{iB}/θ_{jB}. A downside of θ_i is that the comparison to a hypothetical "average model" may be considered less informative than a comparison to the baseline model. To address this one could follow the lines of relationship (5) and report

$$\theta_i^* = \frac{\theta_i}{\theta_B}. \tag{11}$$

This is a measure of performance relative to the baseline model with a correction for the fact that certain models may have covered prediction tasks for which the baseline is more or less difficult to outperform. Note that if all forecast targets were covered by all forecasters, then $\theta_{iB}^* = \theta_{iB}$ would hold. This reflects the fact that in this case there would not be any need to correct for missing forecasts and varying diffculty of the adressed forecast tasks.

To get a better intuition of what θ_i^* actually measures consider the case where there are only two forecasters. Assume that the baseline forecaster is available for all targets and denote the subset of forecast targets covered by forecaster i by $\mathcal{A}_i$. Then we have

$$\theta_1^* = \left(\frac{\theta_{1,1}\theta_{1,2}}{\theta_{B,1}\theta_{B,2}} \right)^{-1/2} \tag{12}$$

$$= \dots \tag{13}$$

$$= \Bigg(\frac{\text{mean WIS of model 1 on } \mathcal{A}_1}{\text{mean WIS of baseline } B \text{ on } \mathcal{A}_1} \tag{14}$$

$$\times \frac{\text{mean WIS of model 1 on } \mathcal{A}_{1,2} \times \frac{\text{mean WIS of model 2 on } \mathcal{A}_2}{\text{mean WIS of model 2 on } \mathcal{A}_{1,2}}}{\text{mean WIS of baseline } B \text{ on } \mathcal{A}_2} \Bigg)^{-1/2} \tag{15}$$

One can thus interpret θ_1^* as the geometric mean of two quantities:

1. The performance of model 1 relative to the baseline on the set of forecast targets $\mathcal{A}_1$ covered by model 1.

2. An estimate of the performance model 1 would have achieved relative to the baseline on the set of forecasts $\mathcal{A}_2$ covered by model 2. The mean WIS of model 1 on $\mathcal{A}_2$ often cannot be evaluated. As an estimate, we multiply the mean WIS achieved on $\mathcal{A}_{1,2}$ (the subset of $\mathcal{A}_2$ also covered by model 1) with the ratio of the mean WIS model 2 achieved on the full $\mathcal{A}_2$ and the subset $\mathcal{A}_{1,2}$. We thus assume that the"relative difficulty" of forecasting $\mathcal{A}_{1,2}$ and $\mathcal{A}_2 \backslash \mathcal{A}_{1,2}$ would have been the same for model 1 as it was for model 2. Note that this reduces to the observed mean WIS of model 1 on $\in$ if model 1 covered all of $\mathcal{A}_2$.

This also generalizes to a larger number of models M included in the comparison. The summary score θ_i^* can be interpreted as

$$\theta_i^* = \prod_{m=1}^{M} \frac{\text{estimated mean WIS model } i \text{ on } \mathcal{A}_m}{\text{mean WIS of baseline model } B \text{ on } \mathcal{A}_m}. \tag{16}$$

To estimate the mean WIS of model i on $\mathcal{A}_m$, we use the results of model m as a reference to determine the relative difficulty of the forecast targets in $\mathcal{A}_{i,m}$ and $\mathcal{A}_m$. If model i covered all $\mathcal{A}_m$ anyway, no correction is performed and the results obtained by model m have no impact on the summary score θ_i^* of model i. However, equation 16 implies that we give larger weight to forecast targets covered by more models, as these enter in term (16) more often. This means that models covering only part of the targets are allowed to "shift the focus" of the evaluation towards the targets they chose to cover. In a sense, each model is given an equal weight in determining on which set of targets all models are evaluated.

3 Application

The following analysis is restricted to state-level forecasts 1 through 4 wk ahead and models providing forecasts for at least 12 out of 15 weeks between May and August 2020. Weeks with major revisions of truth data have been removed from the evaluation. (This selection has been done by Estee and Nick). As shown in Figure 1, even after this restriction, the number of states covered each week by the different models varies considerably.

We start by displaying the raw pairwise comparisons θ_{ij} as defined in (7). In Figure 2 it can be seen that these pairwise comparisons are *almost* transitive, i.e. if i beats j and j beats k then i usually also beats k. This is encouraging as it seems like the missingness does not cause major incoherences.

```
## Loading required package:  sp
## Loading required package:  xtable
## This is surveillance 1.18.0.  For overview type 'help(surveillance)'.
```

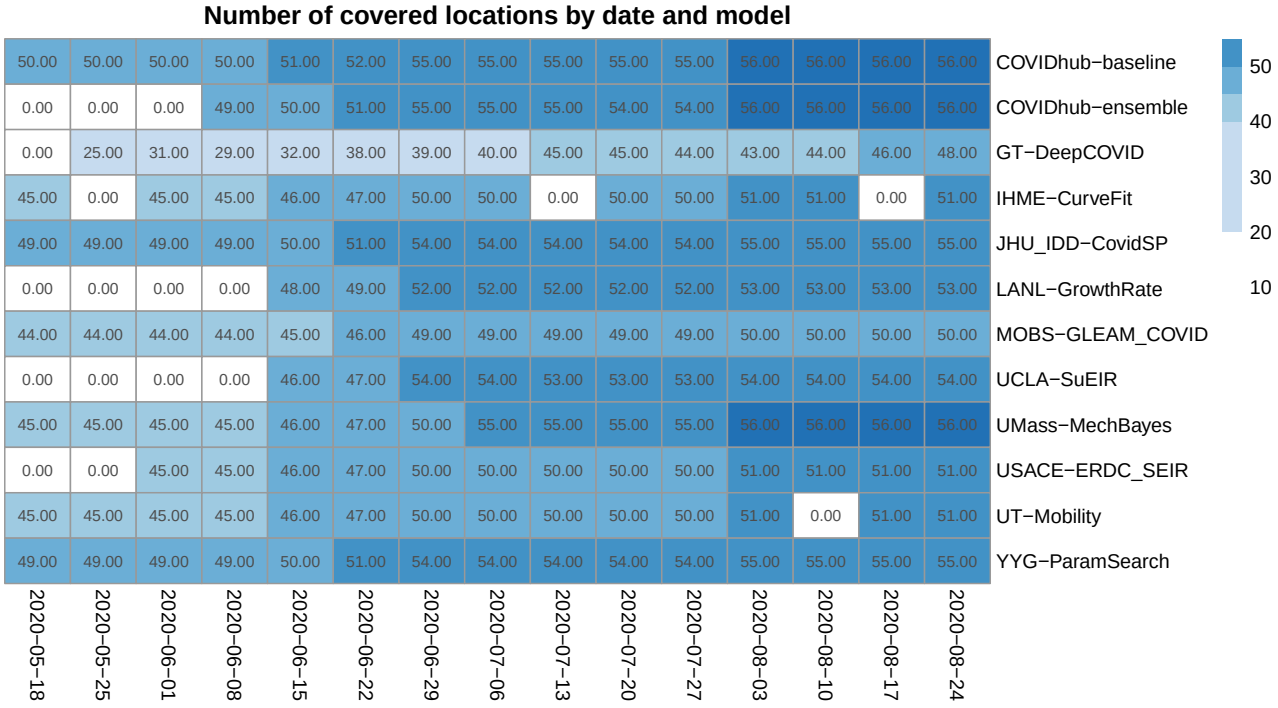

Figure 1: Number of regions covered by each model after restriction of forecasts as described in the text.

In Table 1 we break down this set of pairwise comparisons into the different model-specific summary measures θ_{iB}, θ_i and θ^*_{iB} discussed in the previous section. While θ_i and θ^*_{iB} are obviously proportional, there is some disagreement between the two and θ_{iB}. The models `GT-DeepCOVID` and `JHU IDD-CovidSP` have a better (lower) value after adjusting for how difficult it is to beat the baseline in different states or time periods ($\theta^*_{iB} < \theta_{iB}$). This indicates that `GT-DeepCOVID` and `JHU IDD-CovidSP` have issued more forecasts for targets where beating the baseline was difficult. The opposite holds true for `UT-mobility`.

The θ_{iB}, θ_i and θ^*_{iB} are imperfect ways of aggregating performance as we know that the respective ratios do not always align with the respective direct comparisons, i.e. in general

$$\frac{\theta_{iB}}{\theta_{jB}} \neq \theta_{ij}, \quad \frac{\theta^*_{iB}}{\theta^*_{jB}} = \frac{\theta_i}{\theta_j} \neq \theta_{ij}.$$

However, we can check empirically which one of them allows us to better recover (or preserve) the θ_{ij}, which we know are apples-to-apples comparisons. We can see that overall using $\theta^*_{iB}/\theta^*_{jB}$ leads to less strong deviations from θ_{ij} than θ_{iB}/θ_{jB} (Figures 5 and 6).

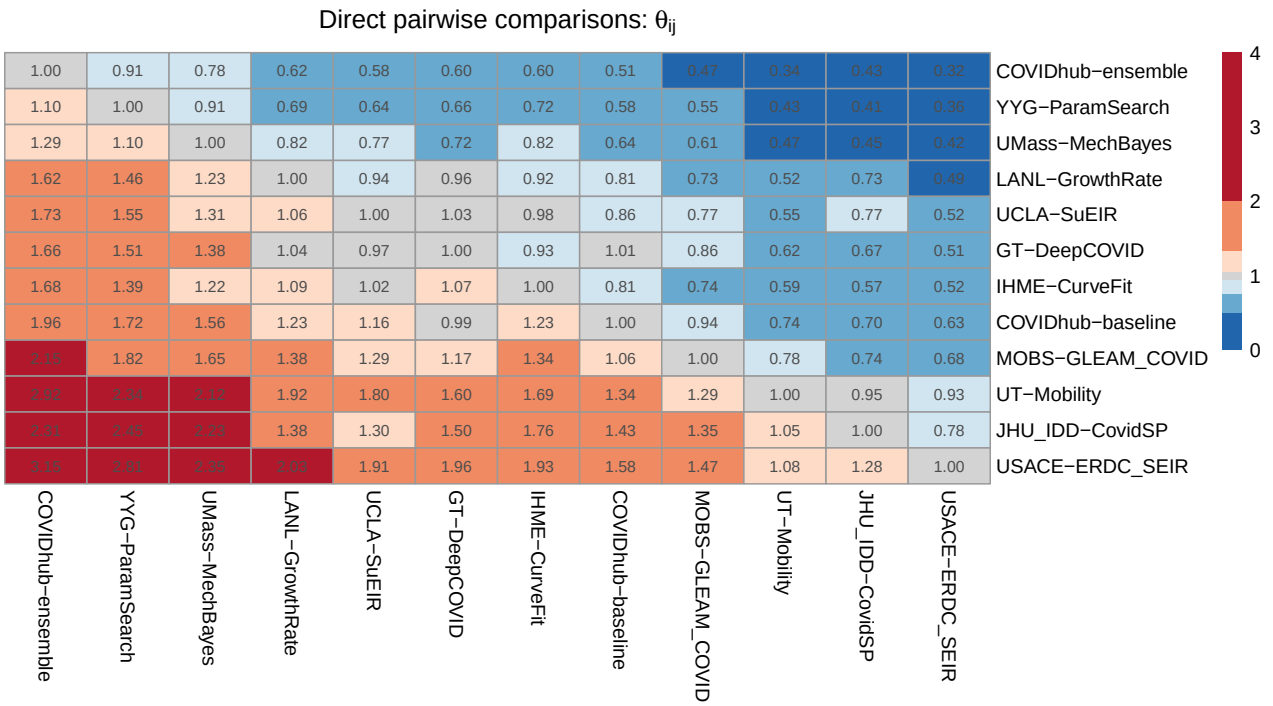

Figure 2: Pairwise relative WIS skills θ_{ij} of twelve forecast models. Note that these are computed from different sets of forecast targets depending on the overlap between time points and states covered by the different models. Numbers below one (blue) indicate that the row model is better, numbers above one (red) indicate that the column model is better.

In the following we therefore focus on θ^*_{iB} as the summary measure to describe performance of model i relative to the baseline. The relative performance measure θ^*_{iB} per model can also be shown over time by computing it for subsets of forecasts made on specific date (Figure 8) or with a specific forecast horizon (Figure 7). Here we can see that the ensemble, `UMass-MechBayesand` and `YYG-ParamSearch` are rather consistently the best models

Lastly we provide the p-values for the pariwise tests for different forecast performance in Figure 9. It can be seen that the ensemble forecast provides indeed significantly better forecasts than any of the member forecasts.

Table 1: Summary scores of included models. θ_{iB} and θ_i/θ_B are constructed such that the baseline model has a value of 1. θ_i and $\theta i/\theta_B$ take into account all pairwise comparisons, while θ_{iB} takes only into account direct comparisons to the baseline.

Model	θ_i	θ_{iB}	θ^*_{iB}
COVIDhub-ensemble	0.57	0.51	0.52
YYG-ParamSearch	0.64	0.58	0.58
UMass-MechBayes	0.72	0.64	0.66
LANL-GrowthRate	0.90	0.81	0.82
IHME-CurveFit	0.93	0.81	0.84
GT-DeepCOVID	0.95	1.01	0.87
UCLA-SuEIR	0.96	0.86	0.87
COVIDhub-baseline	1.10	1.00	1.00
MOBS-GLEAM_COVID	1.19	1.06	1.09
JHU_IDD-CovidSP	1.46	1.43	1.33
UT-Mobility	1.58	1.34	1.44
USACE-ERDC_SEIR	1.79	1.58	1.63

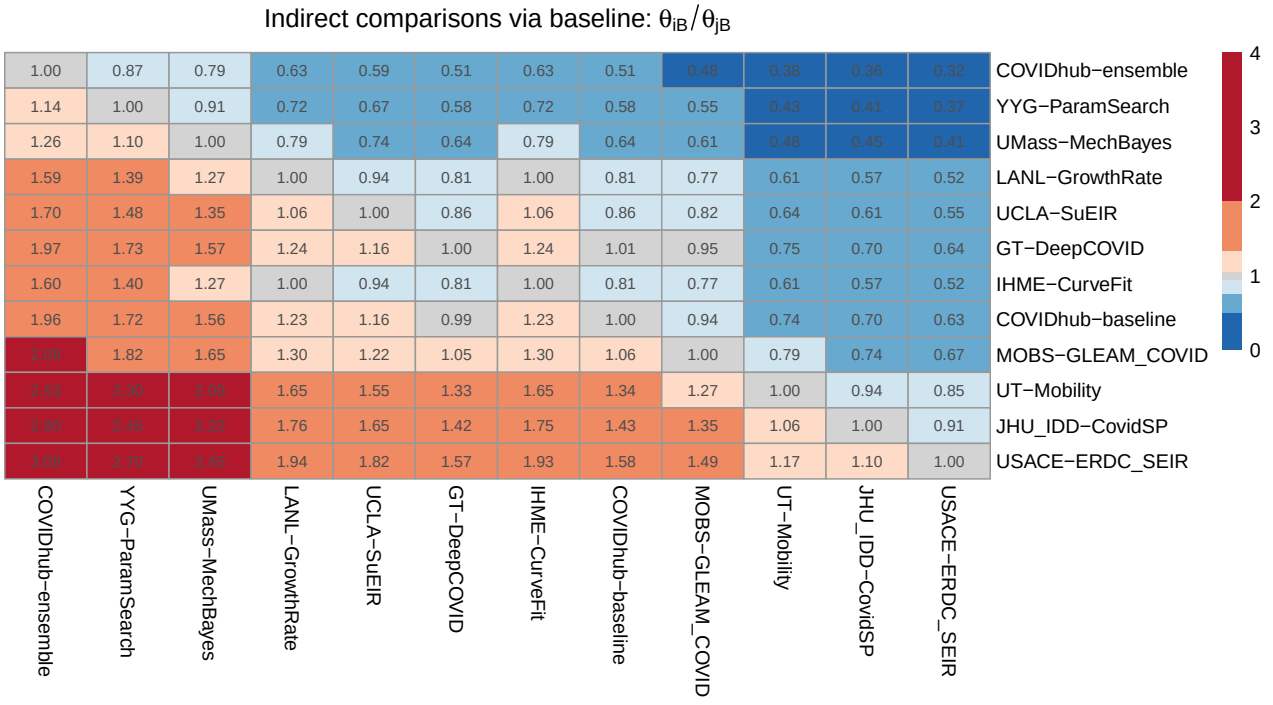

Figure 3: Indirect pairwise comparisons θ_{iB}/θ_{jB}. Here we compare each of the models i and j to the baseline B and then compare the results achieved by both models.

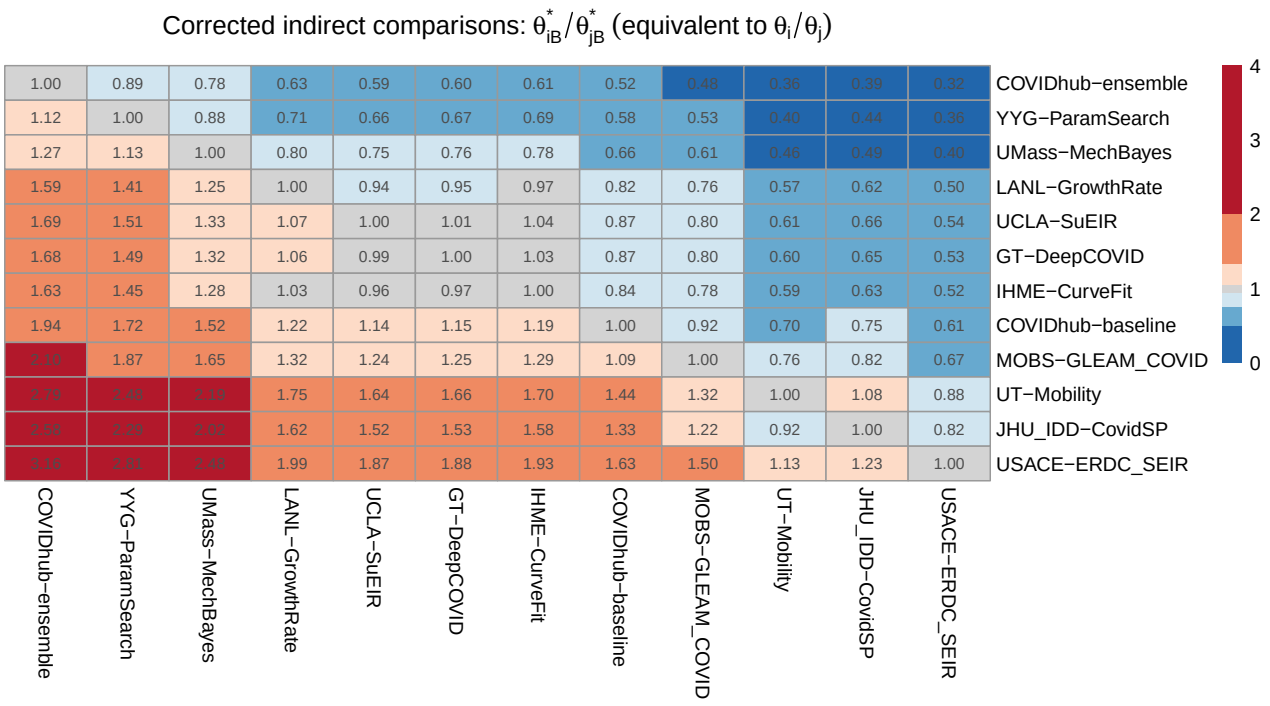

Figure 4: Corrected indirect pairwise comparisons $\theta^*_{iB}/\theta^*_{jB} = \theta_i/\theta_j$. Here we compare each of the models i and j to the entire set $\mathcal{M}$ of models and then compare the results achieved by both models.

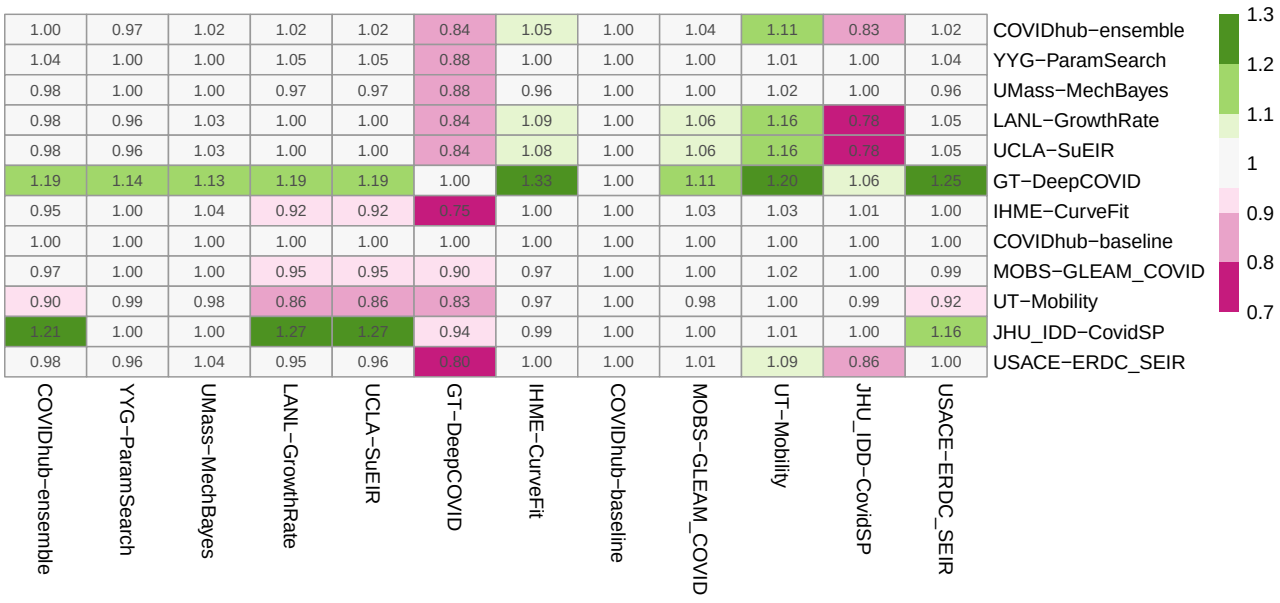

Figure 5: Relative discrepancy between indirect comparisons θ_{iB}/θ_{jB} and direct comparisons θ_{ij}. This describes how much we move away from the original pairwise comparisons when using θ_{iB} as a summary measure.

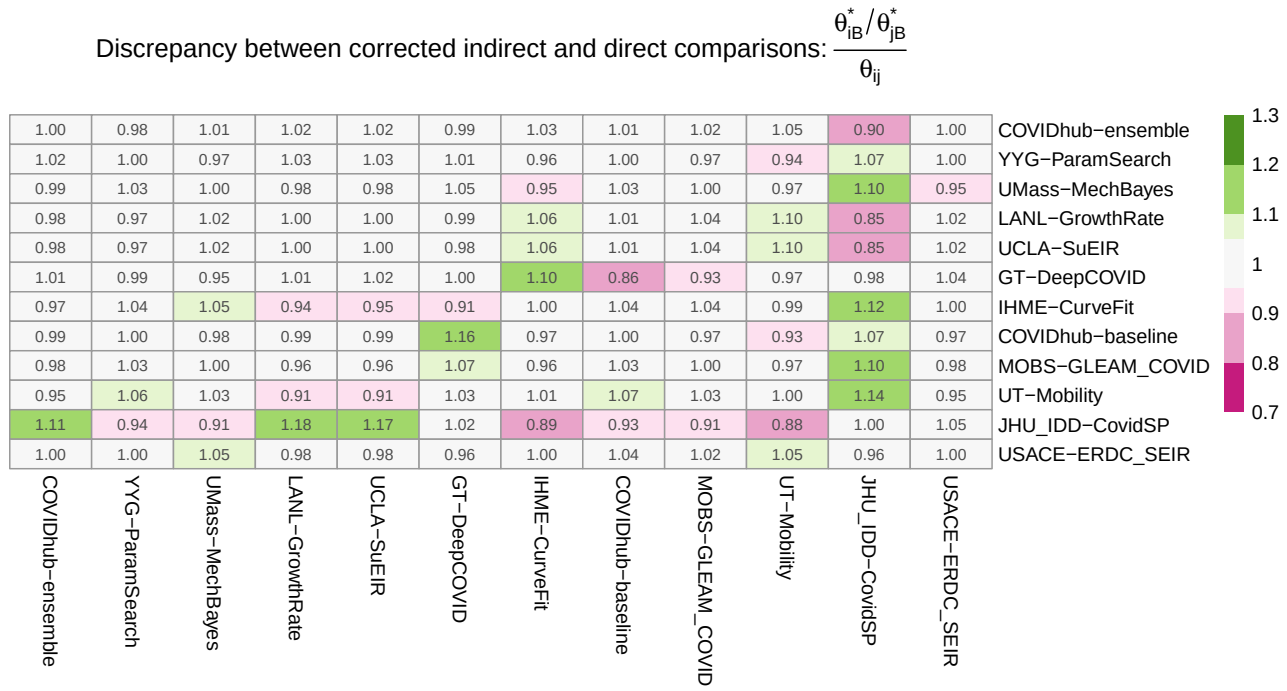

Figure 6: Relative discrepancy between indirect comparisons $\theta^*_{iB}/\theta^*_{jB}$ with correction and direct pairwise comparisons θ_{ij}. This describes how much we move away from the original pairwise comparisons when using the corrected θ^*_{iB} as a summary measure.

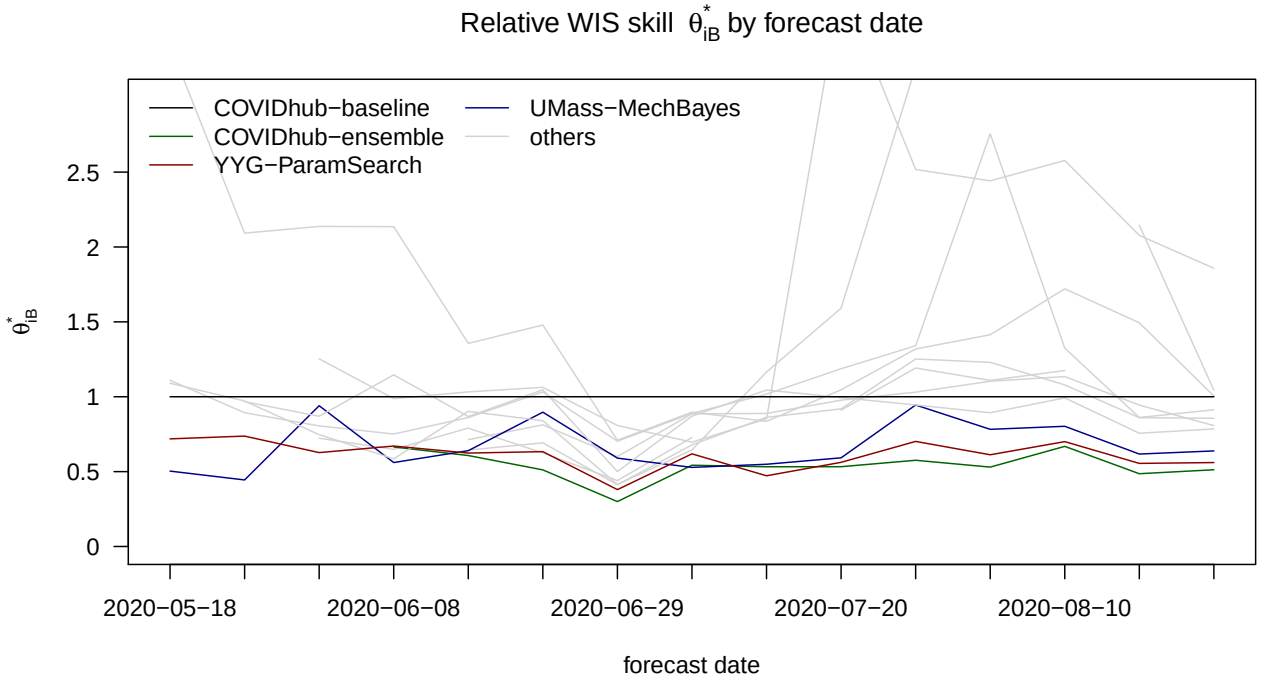

Figure 7: Corrected relative WIS skill θ^*_{iB} with respect to the baseline model, shown per forecast date.

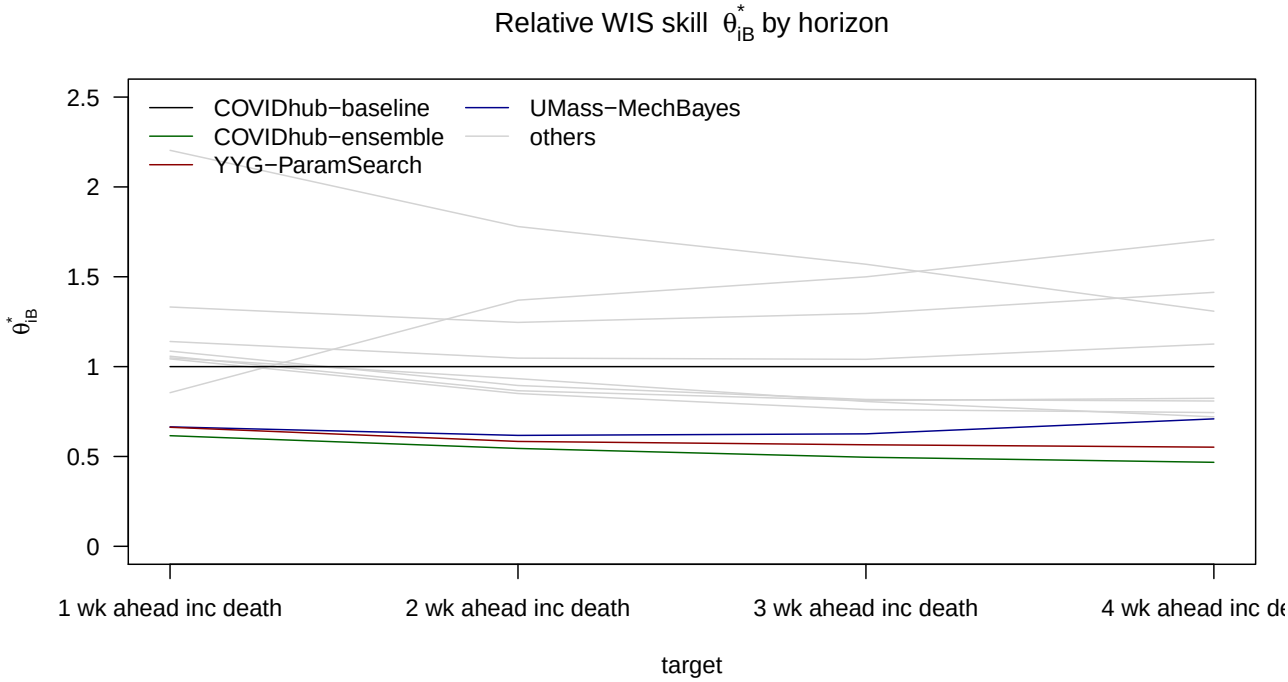

Figure 8: Corrected relative WIS skill θ^*_{iB} with respect to the baseline model, shown per forecast horizon.

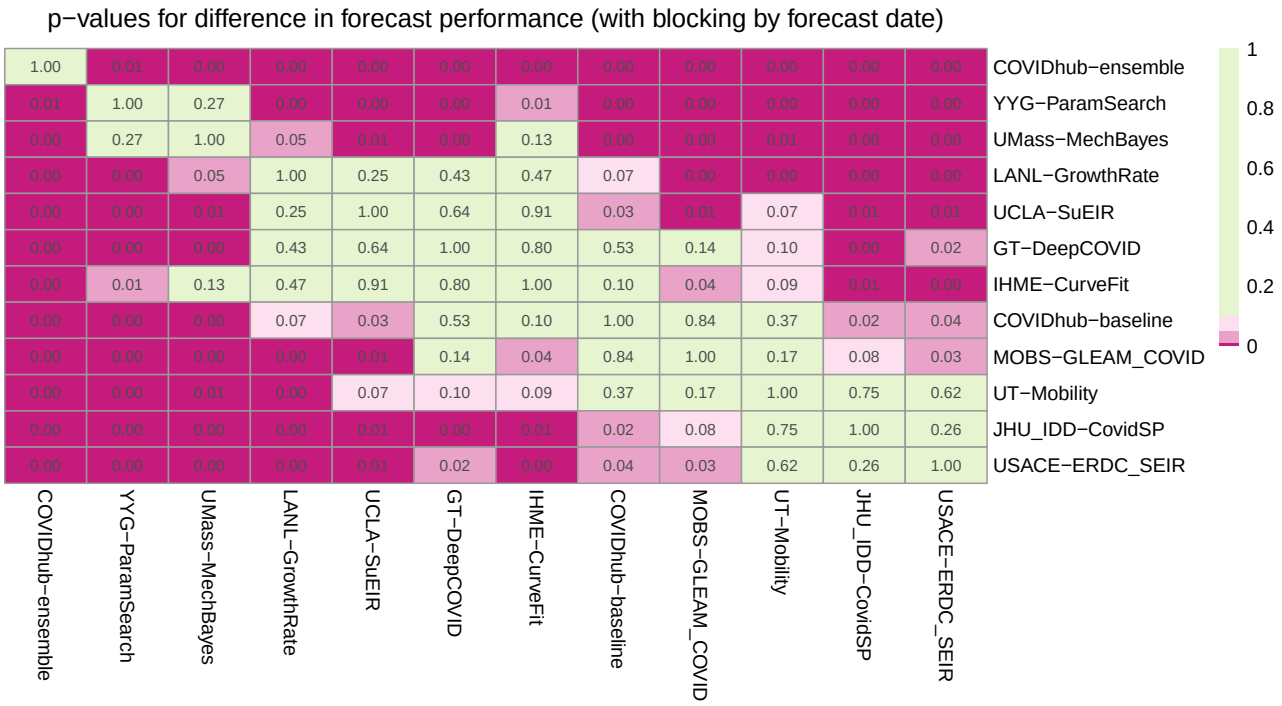

Figure 9: Pairwise p-values for differences in forecast performance, based on permutation test with blocking by forecast date.