PNAS
www.pnas.org

Supplementary Information for

Can Auxiliary Indicators Improve COVID-19 Forecasting and Hotspot Prediction?

Daniel J. McDonald, Jacob Bien, Alden Green, Addison J. Hu, Nat DeFries, Sangwon Hyun, Natalia L. Oliveira, James Sharpnack, Jingjing Tang, Robert Tibshirani, Valerie Ventura, Larry Wasserman, and Ryan J. Tibshirani

Daniel J. McDonald.
E-mail: daniel@stat.ubc.ca

This PDF file includes:

Supplementary text
Figs. S1 to S23 (not allowed for Brief Reports)
Table S1 (not allowed for Brief Reports)
Legends for Dataset S1 to S4
SI References

Other supplementary materials for this manuscript include the following:

Datasets S1 to S4

Supporting Information Text

1. Finalized Versus Vintage Data

The goal of this section is to quantify the effect of not properly accounting for the question of "what was known when" in performing retrospective evaluations of forecasters. Figures S1 and S2 show what Figures 3 and 4 in the main paper would have looked like if we had simply trained all models using the finalized data rather than using vintage data. This comparison can be seen more straightforwardly in Figures S3 and S4, which show the ratio in performance between the vintage and finalized versions. When methods are given the finalized version of the data rather than the version available at the time that the forecast would have been made, all methods appear (misleadingly) to have better performance than they would have had if run prospectively. For example, for forecasting case rates 7-days ahead, the WIS of all methods is at least 8% larger than what would have been achieved using finalized data. This effect diminishes as the forecasting horizon increases, reflecting the fact that longer-horizon forecasters rely less heavily on recent data than very short-horizon forecasters. Crucially, some methods are "helped" more than others by the less scrupulous retrospective evaluation, underscoring the difficulty of avoiding misleading conclusions when performing retrospective evaluations of forecasters.

CHNG-CLI (and, to a lesser extent, the other claims-based signals) is the most affected by this distinction, reflecting the latency in claims-based reporting. This underscores the importance of efforts to provide "nowcasts" for claims signals (which corresponds to a 0-ahead forecast of what the claims signal's value will be once all data has been collected). Looking at the CHNG-CLI and DV-CLI curves in Figure S1, we can see that they perform very similarly when trained on the finalized data. This is reassuring because they are, in principle, measuring the same thing (namely, the percentage of outpatient visits that are primarily about COVID-related symptoms). The substantial difference in their curves in Figure 3 of the main paper must therefore reflect their having very different backfill profiles.

While using finalized rather than vintage data affects DV-CLI the least for forecasting, it is one of the most affected methods for the hotspot problem. This is a reminder that the forecasting and hotspot problems are fundamentally distinct. For example, the hotspot problem does not measure the ability to distinguish between flat and downward trends.

Even the AR model is affected by this distinction, reflecting the fact that the case rates themselves (i.e., the response values) are also subject to revision. The forecasters based on indicators are thus affected both by revisions to the indicators and by revisions to the case rates. In the case of the Google-AA model, in which we only used finalized values for the Google-AA indicator, the difference in performance can be wholly attributed to revisions of case rates.

2. Aggregation with Geometric Mean

In this section, we consider using the geometric mean instead of the arithmetic mean when aggregating the weighted interval score (WIS) across location-time pairs. There are three reasons why using the geometric mean may be desirable.

1. WIS is right-skewed, being bounded below by zero and having occasional very large values. Figure S5 illustrates that the densities appear roughly log-Gaussian. The geometric mean is a natural choice in such a context since the relative ordering of forecasters is determined by the arithmetic mean of the *logarithm* of their WIS values.

2. In the main paper, we report the ratio of the mean WIS of a forecaster to the mean WIS of the baseline forecaster. Another choice could be to take the mean of the ratio of WIS values for the two methods. This latter choice would penalize a method less for doing poorly where the baseline forecaster also does poorly. Using instead the geometric mean makes the order of aggregation and scaling immaterial since the ratio of geometric means is the same as the geometric mean of ratios.

3. If one imagines that a forecaster's WIS is composed of multiplicative space-time effects $S_{\ell,t}$ shared across all forecasters, i.e., $\text{WIS}(F_{\ell,t,f}, Y_{\ell,t}) = S_{\ell,t}E_{f,t}$ with $E_{f,t}$ a forecaster-specific error, then taking the ratio of two forecasters' geometric mean WIS values will effectively cancel these space-time effects.

Figure S6 uses the geometric mean for aggregation. Comparing this with Figure 3 of the main paper, we see that the main conclusions are largely unchanged; however, CHNG-CLI now appears better than AR. This behavior would be expected if CHNG-CLI's poor performance is attributable to a relatively small number of large errors (as opposed to a large number of moderate errors). Indeed, Figure 5 of the main paper further corroborates this, in which we see the heaviest left tails occuring for CHNG-CLI.

3. Comparing with the COVID-19 Forecast Hub

Since July of 2020, teams have been submitting real time forecasts to the COVID-19 Forecast Hub (1). While a number of forecast targets can be submitted, the most related to that used in this paper—the 7-day trailing average of COVID-19 case incidence rates at the Hospital Referral Region—are state-level forecasts for total case incidence over an epiweek. Essentially, for each state, teams are asked to predict the total number of cases that will be recorded over the week from Monday to Sunday. Forecasts must be submitted by Tuesday for the current Sunday as well as the next three Sundays. This corresponds to producing forecasts for 5, 12, 19, and 26 days ahead. Some teams choose to submit their forecasts a few days earlier, but always for the same targets.

To properly compare the models presented in this paper with those submitted to the Hub, we limit ourselves to the same set of forecast dates and states, and we predict case incidence over a week by: (1) predicting the 7-day trailing average case rate (as in the manuscript); (2) multiplying by 7 to give the total over a week; (3) multiplying by the state population as reported in the 2019 US Census Population Estimate (available in the Covidcast R package (2)). Finally, we perform the same evaluations as in the manuscript Figure 3 and also using geometric mean as described above. For the purpose of the comparison, we show only the AR model along with submissions collected by the Forecast Hub in Figure S7. Since not all teams submitted forecasts for the entire period, we only display those which submitted for at least 6 weeks. By both metrics, the AR model examined in this paper is competitive with the submitted by top teams, even beating the Ensemble for smaller target horizons.

We also are undertook the same analysis shown in Figure 3 in the manuscript and Figure S6 above for all models (the AR, as well as the indicator-assisted models). This comparison is shown in Figure S8. TODO: some text about why indicators don't help as much.

4. Statistical Significance

In the Introduction of the main paper, we give some reasons that we avoid making formal statements about statistical significance, preferring instead to examine the stability of our results in different contexts. There are strong reasons to avoid model-based significance tests because the necessary assumptions about stationarity, independence, and the model being true (or at least approximately true) are certainly violated. With those caveats in mind, we undertake two relatively assumption-lean investigations in this section. The first is a sign-test for whether the difference between the AR model's relative WIS and each other model's relative WIS is centered at zero. By "relative WIS" we mean scaled by the strawman as displayed in Figure 3. To mitigate the dependence across time (which intuitively seems to matter more than that across space), we computed these tests in a stratified way, where for each forecast date we run a sign test on the scaled errors between two models over all 440 counties. The results are plotted as histograms in Figure S9. In this case, we use the total relative WIS over all aheads, but the histograms are largely similar for individual target horizons. If there were no difference, we would expect to see a uniform distribution. However, for all indicator-assisted models, we see many more small p-values than would be expected if the null hypothesis (that the AR model is better) were true.

Another relatively assumption-lean method of testing for differences in forecast accuracy is the Diebold-Mariano test (3–5). Essentially, the differences between forecast losses are assumed to have a constant mean and a covariance that depends on time. Under these conditions, the asymptotic distribution for the standardized mean of the differences is limiting normal provided that a heteroskedasticity and autocorrelation robust estimate of the variance is used. Using the loss as weighted interval score across all HRRs and horizons (7 to 21 days ahead), we perform the DM test using both the mean relative to the strawman (as reported in the manuscript) and the geometric mean relative to the strawman as described above. The first two rows of Table S1 displays p-values for the test that the indicator-assisted model is no-better than the AR model. Only the CHNG-CLI model's p-value exceeds conventional statistical significance thresholds.

5. Bootstrap Results

As explained in Section 2.B. of the main paper, a (somewhat cynical) hypothesis for why we see benefits in forecasting and hotspot prediction is that the indicators are not actually providing useful information but they are instead acting as a sort of "implicit regularization," leading to shrinkage on the autoregressive coefficients and therefore to less volatile predictions. To investigate this hypothesis, we consider fitting "noise features" that in truth should have zero coefficients. Recall (from the main paper) that at each forecast date, we train a model on 6,426 location-time pairs. Indicator models are based on six features, corresponding to the three autoregressive terms and the three lagged indicator values. To form noise indicator features, we replace their values with those from a randomly chosen time-space pair (while keeping the autoregressive features fixed). In particular, at each location ℓ and time t, for the forecasting task we replace the triplet $(X_{\ell,t}, X_{\ell,t-7}, X_{\ell,t-14})$ in Eq. (3) of the main paper with the triplet $(X_{\ell^*,t^*}, X_{\ell^*,t^*-7}, X_{\ell^*,t^*-14})$, where (ℓ^*, t^*) is a location-time pair sampled with replacement from the 6,426 location-time pairs. Likewise in the hotspot prediction task, we replace the triplet $(X^{\Delta}_{\ell,t}, X^{\Delta}_{\ell,t-7}, X^{\Delta}_{\ell,t-14})$ in Eq. (5) of the main paper with $(X^{\Delta}_{\ell^*,t^*}, X^{\Delta}_{\ell^*,t^*-7}, X^{\Delta}_{\ell^*,t^*-14})$. Figures S10–S12 show the results. No method exhibits a noticeable performance gain over the AR method, leading us to dismiss the implicit regularization hypothesis.

6. Upswings and Downswings

In this section we provide extra details about the upswing / flat / downswing analysis described in the main text. Figure S13 shows the overall results, examining the average difference $\mathrm{WIS}(AR) - \mathrm{WIS}(F)$ in period. Figure S14 shows the same information for the hotspot task. On average, during downswings and flat periods, the indicator-assisted models have lower classification error and higher log likelihood than the AR model. For hotspots, both Google-AA and CTIS-CLIIC perform better than the AR model during upswings, in contrast to the forecasting task, where only Google-AA improves. For a related analysis, Figure S15 shows histograms of the Spearman correlation (Spearman's ρ, a rank-based measure of association) between the $\mathrm{WIS}(F)/\mathrm{WIS}(AR)$ and the magnitude of the swing. Again we see that case rate increases are positively related to diminished performance of the indicator models.

One hypothesis for diminished relative performance during upswings is that the AR model tends to overpredict downswings and underpredict upswings. Adding indicators appears to help avoid this behavior on the downswing but not as much on upswings. Figure S16 shows the correlation between between $\mathrm{WIS}(AR) - \mathrm{WIS}(F)$ and the difference of their median forecasts.

During downswings, this correlation is large, implying that improved relative performance of F is related to making lower forecasts than the AR model. The opposite is true during upswings. This is largely to be expected. However, the relationship attenuates in flat periods and during upswings. That is, when performance is better in those cases, it may be due to other factors than simply making predictions in the correct direction, for example, narrower confidence intervals.

It is important to note, that even though some indicators, notably CHNG-COVID and CHNG-CLI underperform relative to the AR model during upswings, all models dramatically outperform the baseline in such periods. Figure S17 shows the performance of all forecasters relative to the null model. All forecasters suffer relative to the baseline during down periods, but the AR is the worst. In contrast, all models beat the baseline during up periods, even CHNG-COVID and CHNG-CLI, though not by quite as much as the AR does.

7. Leadingness and Laggingness

In Section 2.D of the main text, we discuss the extent to which the indicators are leading or lagging case rates during different periods. To define the amount of leadingness or laggingness at location ℓ, we use the cross correlation function (CCF) between the two time series. The $\mathrm{CCF}_\ell(a)$ of an indicator X_ℓ and case rates Y_ℓ is defined as their Pearson correlation where X_ℓ has been aligned with the values of Y_ℓ that occurred a days earlier. Thus, for any $a > 0$, $\mathrm{CCF}_\ell(a) > 0$ indicates that $Y_{\ell,t}$ is moving together with $X_{\ell,t+a}$. In this case we say that X_ℓ is lagging Y_ℓ. For $a < 0$, $\mathrm{CCF}_\ell(a) > 0$ means that $Y_{\ell,t}$ is positively correlated with $X_{\ell,t-a}$, so we say that X_ℓ leads Y_ℓ.

Figure S18 shows the standardized signals for the HRR containing Charlotte, North Carolina, from August 1, 2020 until the end of September. These are the same signals shown in Figure 1 in the manuscript but using finalized data. To define "leadingness" we compute $\mathrm{CCF}_\ell(a)$ (as implemented with the R function `ccf()`) for each $a \in \{-15, \ldots, 15\}$ using the 56 days leading up to the target date. This is the same amount of data used to train the forecasters: 21 days of training data, 21 days to get the response at $a = 21$, and 14 days for the longest lagged value. The orange dashed horizontal line represents the 95% significance threshold for correlations based on 56 observations. Any correlations larger in magnitude than this value are considered statistically significant under the null hypothesis of no relationship. We define leadingness to be the sum of the significant correlations that are leading (those above the dashed line with $a < 0$) while laggingness is the same but for $a > 0$. In the figure, there are three significant correlations on the "leading" side (at $a = -5, -4, -3$), so leadingness will be the sum of those values while laggingness is 0: on September 28 in Charlotte, DV-CLI is leading cases leading but not lagging.

Figure S19 shows the correlation between laggingness and the difference in indicator WIS and AR WIS. Unlike leadingness (Figure 5 in the manuscript) there is no obvious relationship that holds consistently across indicators. This is heartening as laggingness should not aid forecasting performance. On the other hand, if an indicator is more lagging than it is leading, this may suggest diminished performance. Figure S20 shows the correlation of the difference in leadingness and laggingness with the difference in WIS. The pattern here is largely similar to the pattern in leadingness described in the manuscript: the relationship is strongest in down periods and weakest in up periods with the strength diminishing as we move from down to flat to up for all indicators.

In calculating the CCF and the associated leadingness and laggingness scores, we have used the finalized data, and we look at the behavior at the target date of the forecast. That is we are using the same data to evaluate predictive accuracy as to determine leadingness and laggingness. It should be noted that the leadingness of the indicator at the time the model is trained may also be important. Thus, we could calculate separate leadingness and laggingness scores for the trained model and for the evaluation data and examine their combination in some way. We do not pursue this combination further and leave this investigation for future work.

8. Aggregation Over Time and Space

Following the suggestion of an anonymous reviewer, we investigate two other disaggregated versions of the main forecasting result shown in Figure 3 on the manuscript. The first (Figure S21) displays the cumulative error (summed over all horizons) of each forecaster divided by the cumulative error of the baseline. This perspective should illustrate how the models perform over time, drawing attention to any nonstationary behavior. During the initial increase in cases in July 2020, CTIS-CLIIC and Google-AA gain a lot relative to the AR model. All the indicators do much better than the AR model during the following downturn (the ebb of the second wave). The AR model actually improves over the indicators in October 2020, before losing a bit in late November. The second (Figure S22) repeats this analysis but aggregating by Geometric mean as described above and displays a similar pattern.

Figure S23 examines the spatial behavior over the entire period of the indicator-assisted models relative to the AR. For ease of comparison, we show the percent improvement in each HRR. Negative numbers (blue) mean that the indicator helped while positives (red) mean that the indicator hurt forecasting performance. The clear pattern is that in most HRRs, the indicators improved performance, though usually by small amounts (2.5%–10%). In some isolated HRRs, performance was markedly worse, though there does not appear to be any particular pattern to these locations.

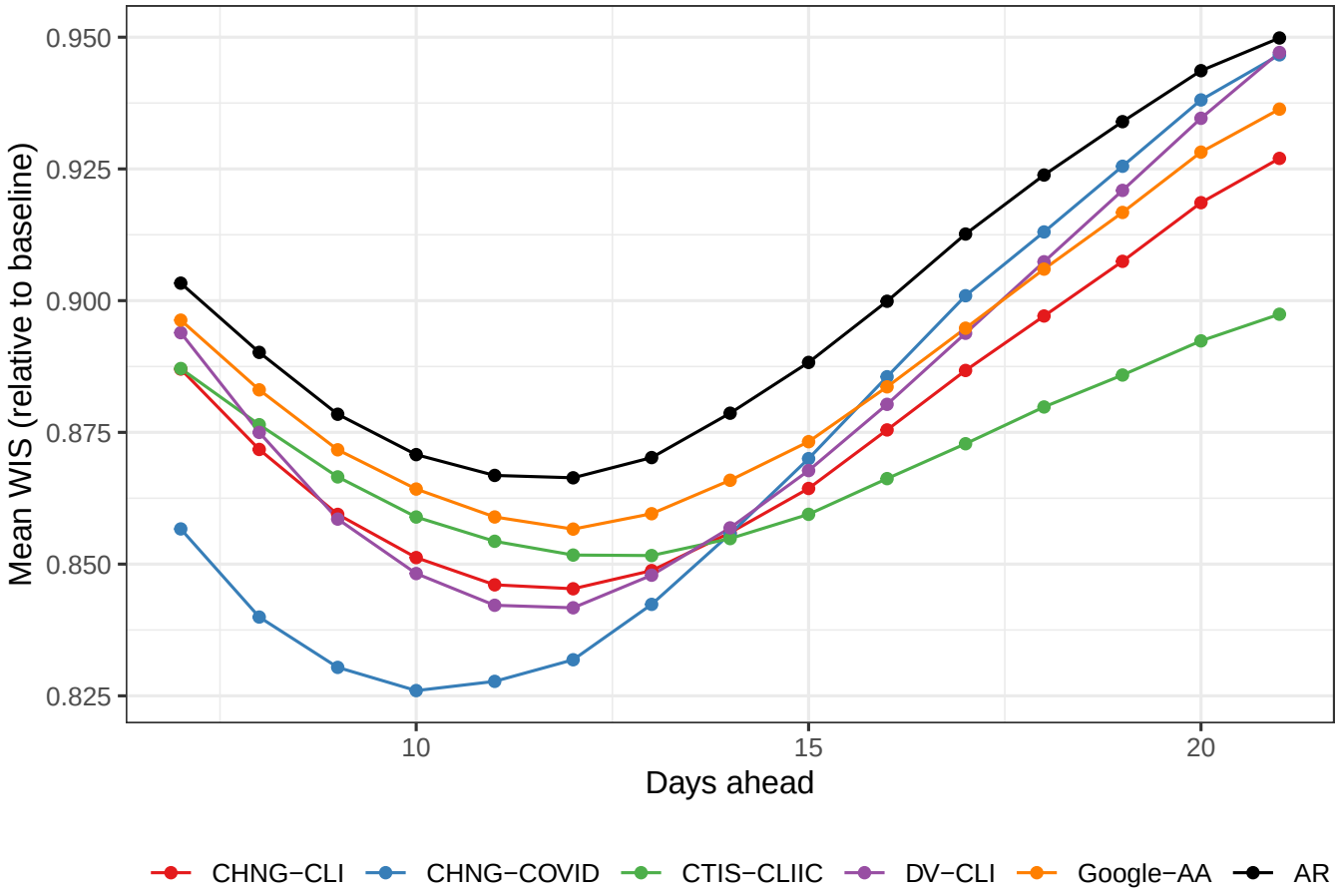

Fig. S1. Forecasting performance using finalized data. Compare to Figure 3 in the manuscript.

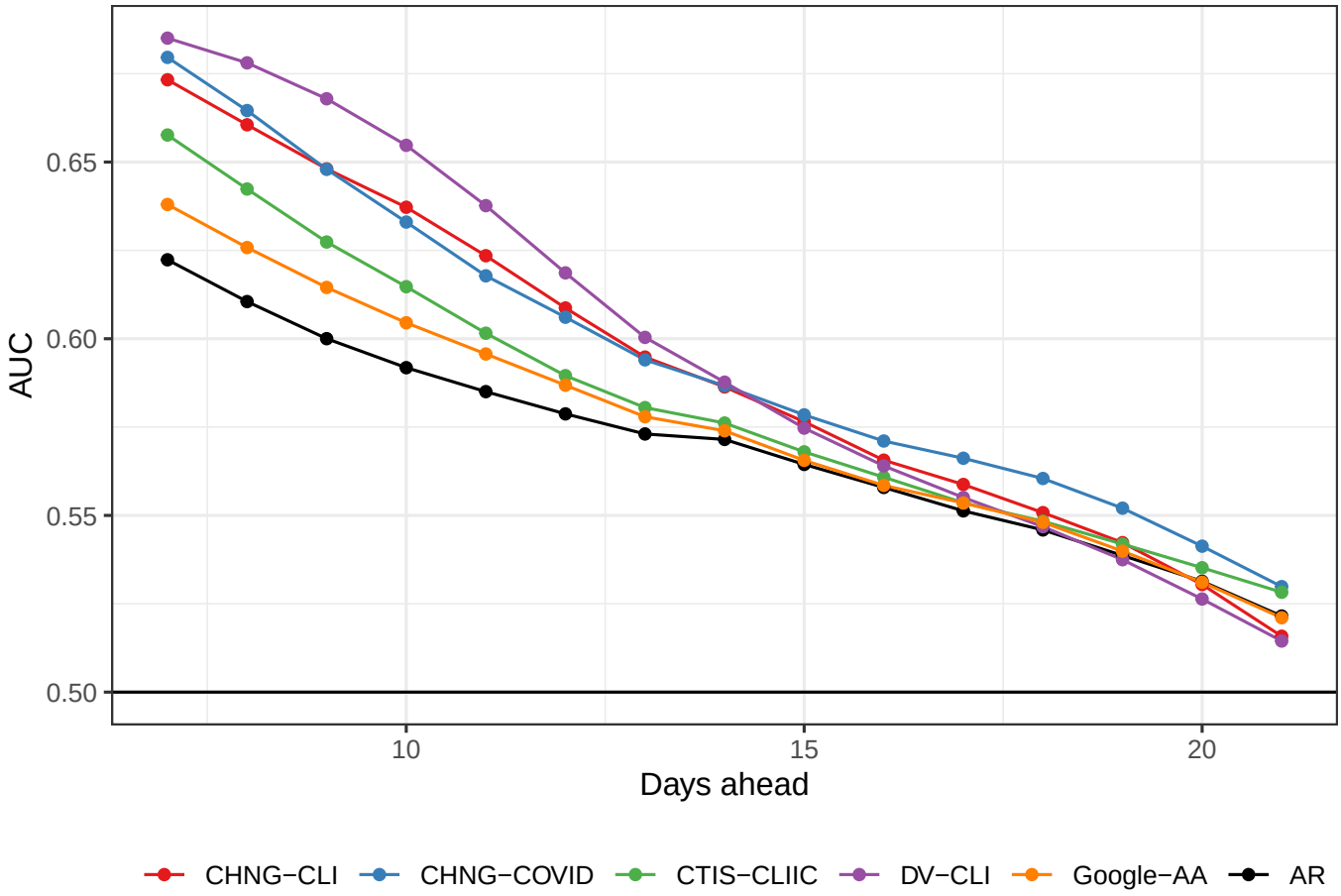

Fig. S2. Hotspot prediction performance using finalized data. Compare to Figure 4 in the manuscript.

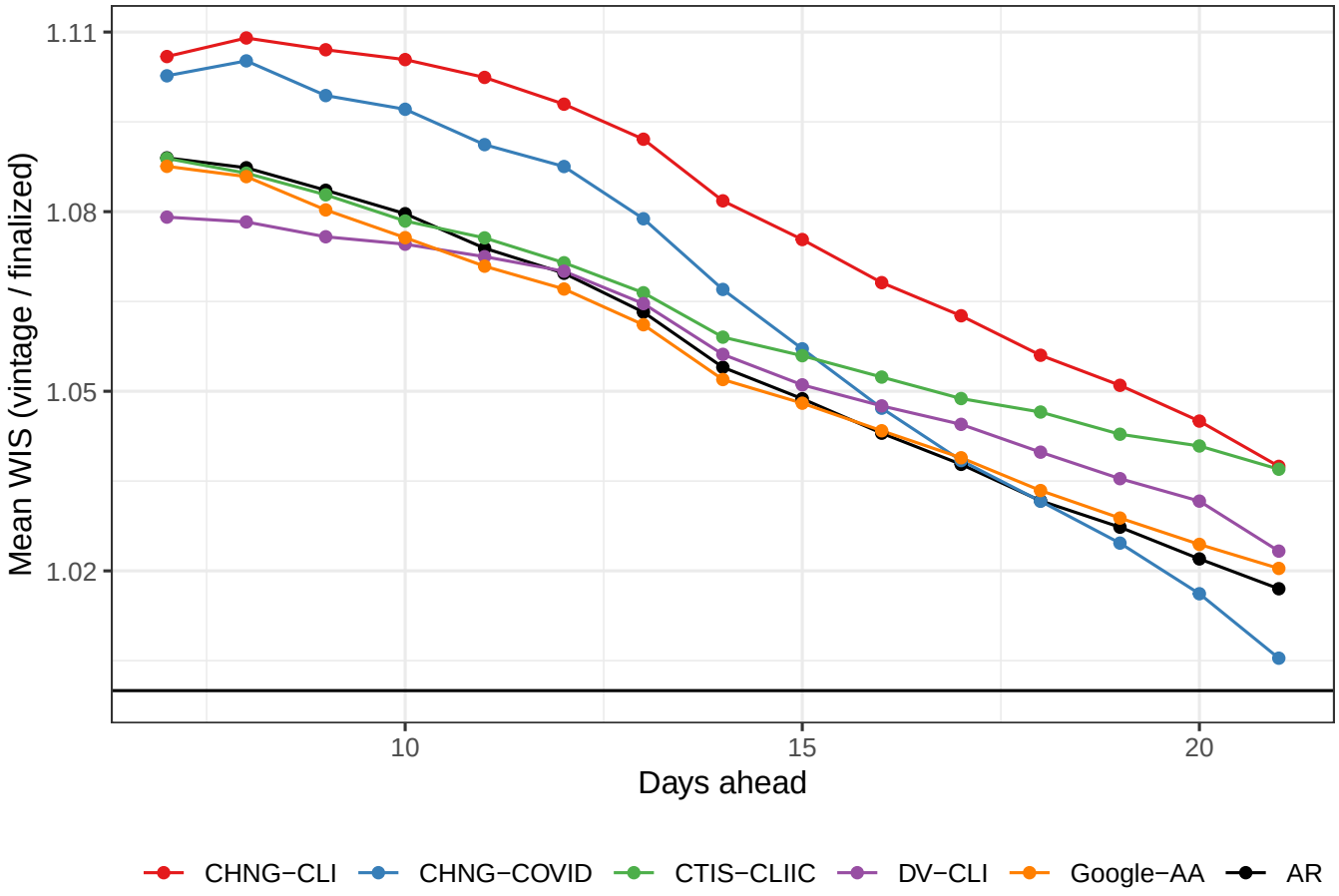

Fig. S3. Relative forecast WIS with vintage compared to finalized data. Using finalized data leads to overly optimistic performance.

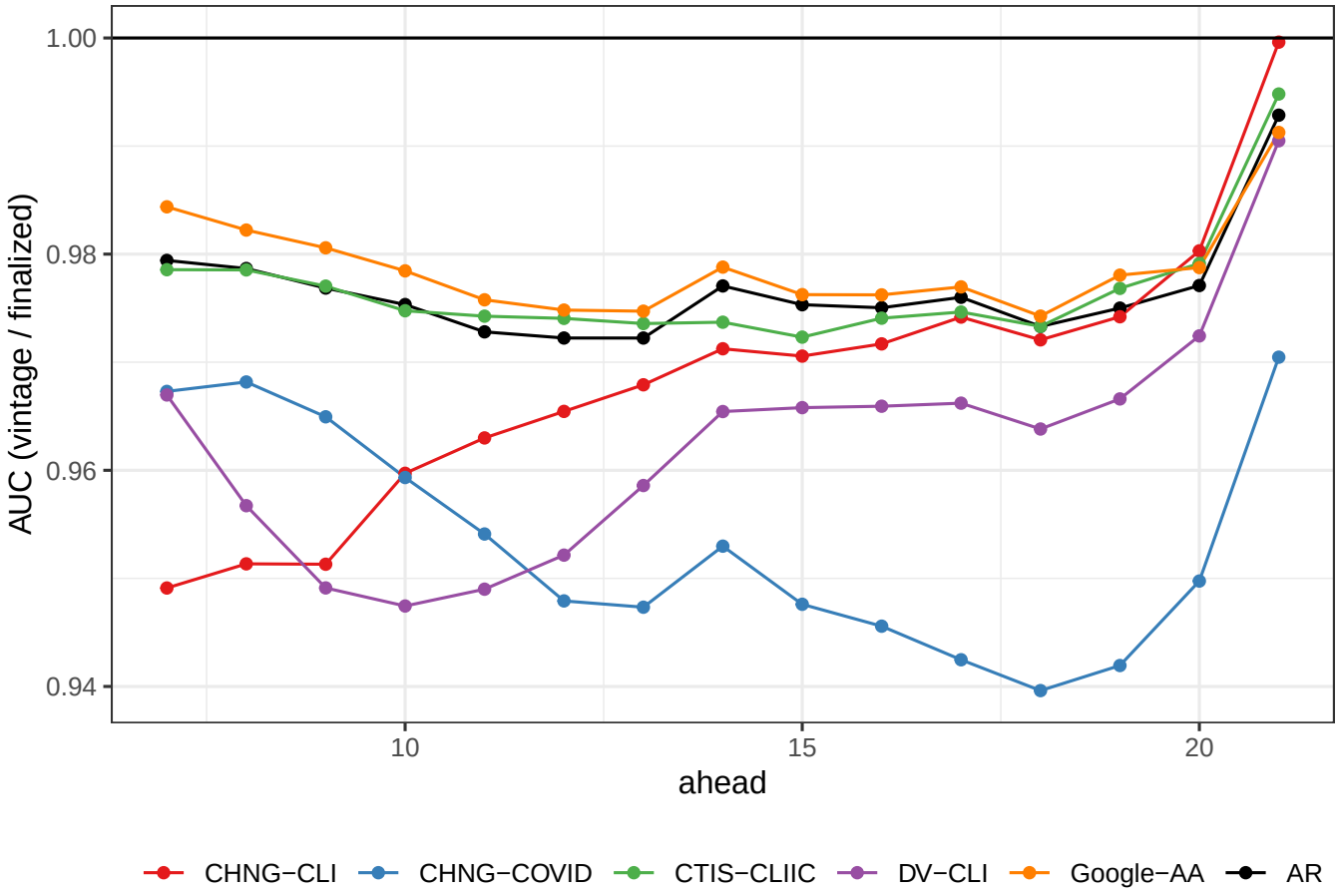

Fig. S4. Relative AUC with vintage compared to finalized data. Using finalized data leads to overly optimistic hotspot performance.

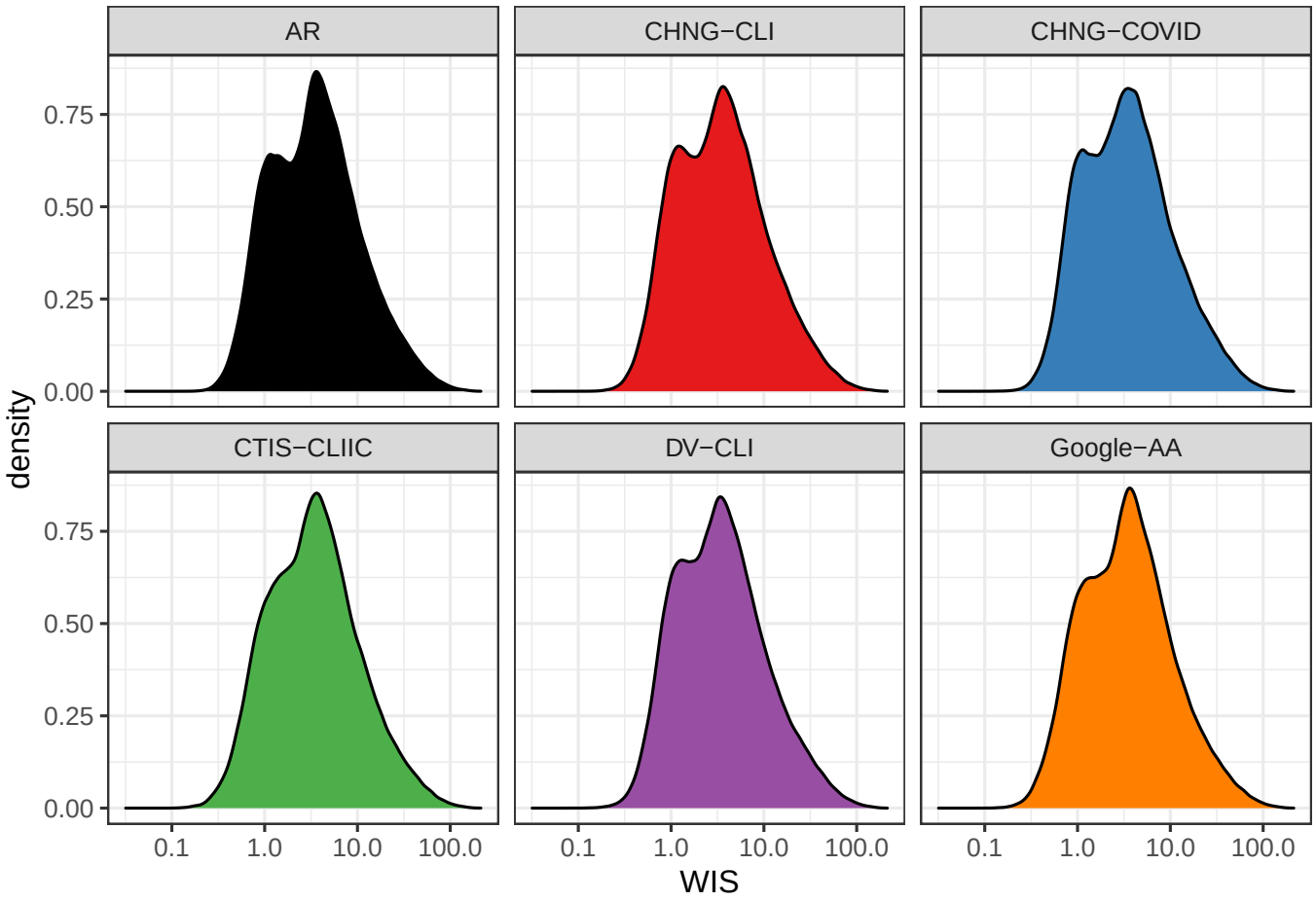

Fig. S5. Weighted interval score appears to more closely resemble a log-Gaussian distribution.

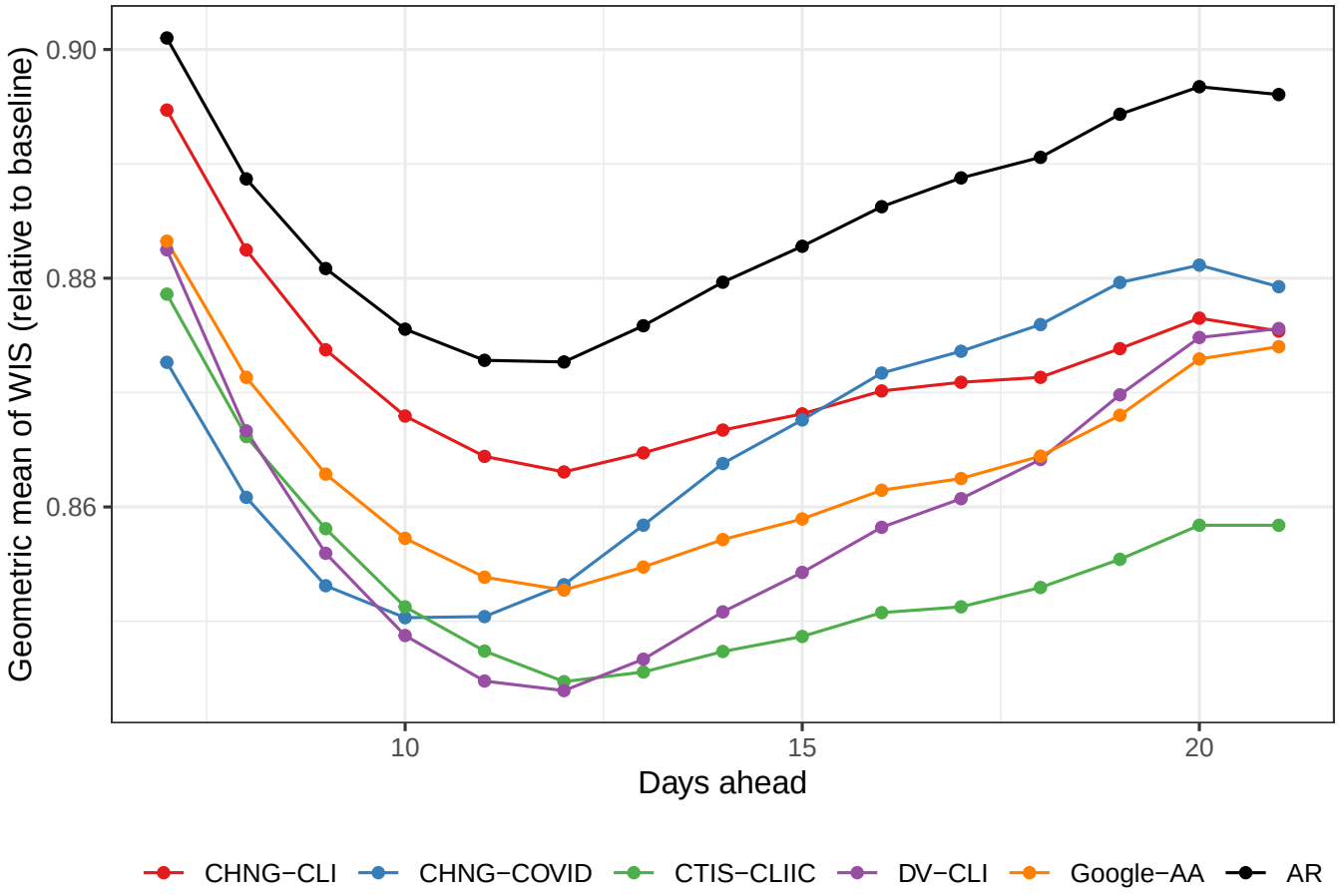

Fig. S6. Relative forecast performance using vintage data and summarizing with the more robust geometric mean.

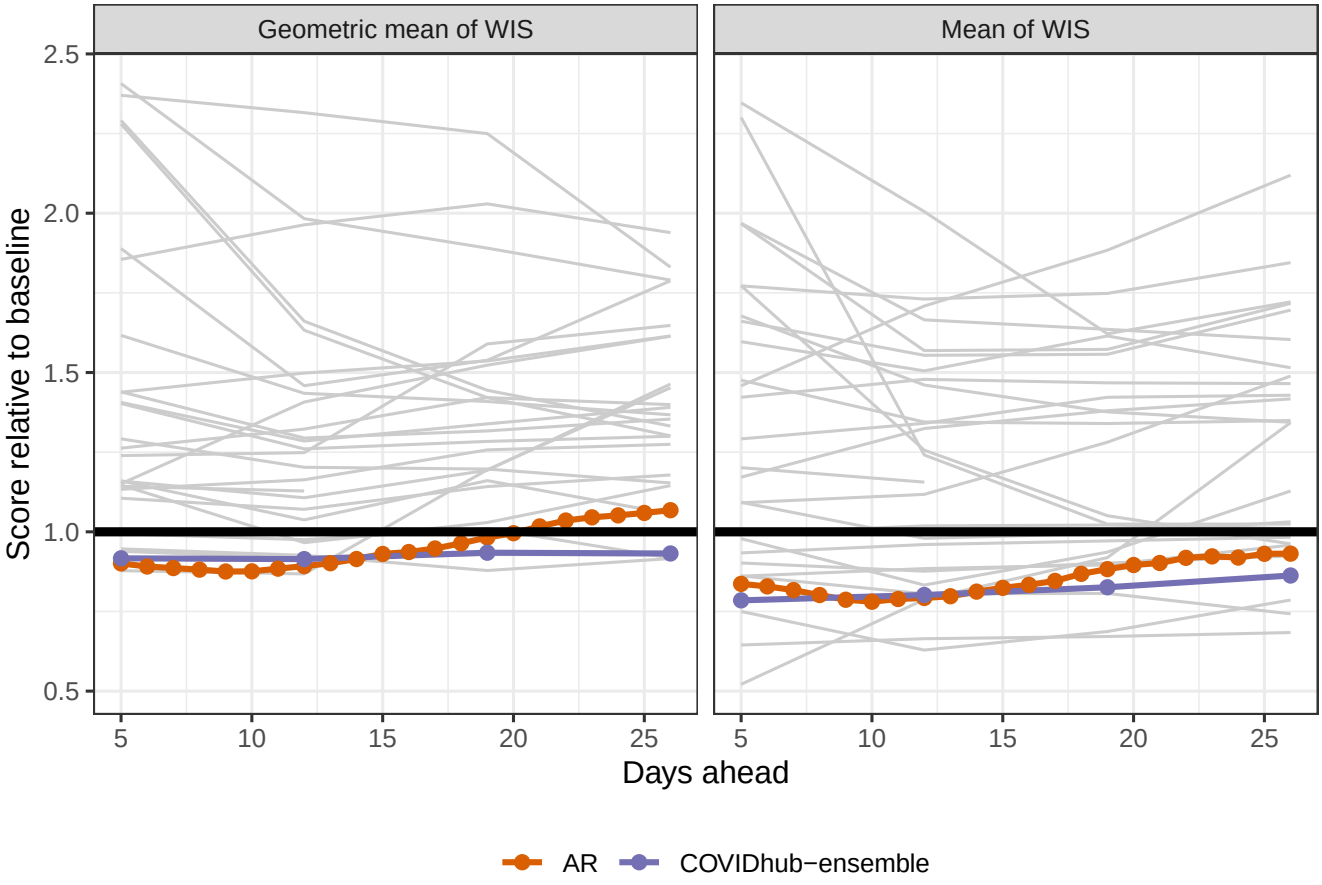

Fig. S7. Performance of the AR model re-estimated for state-level, epiweek case incidence. The thin grey lines are the forecasts of other teams.

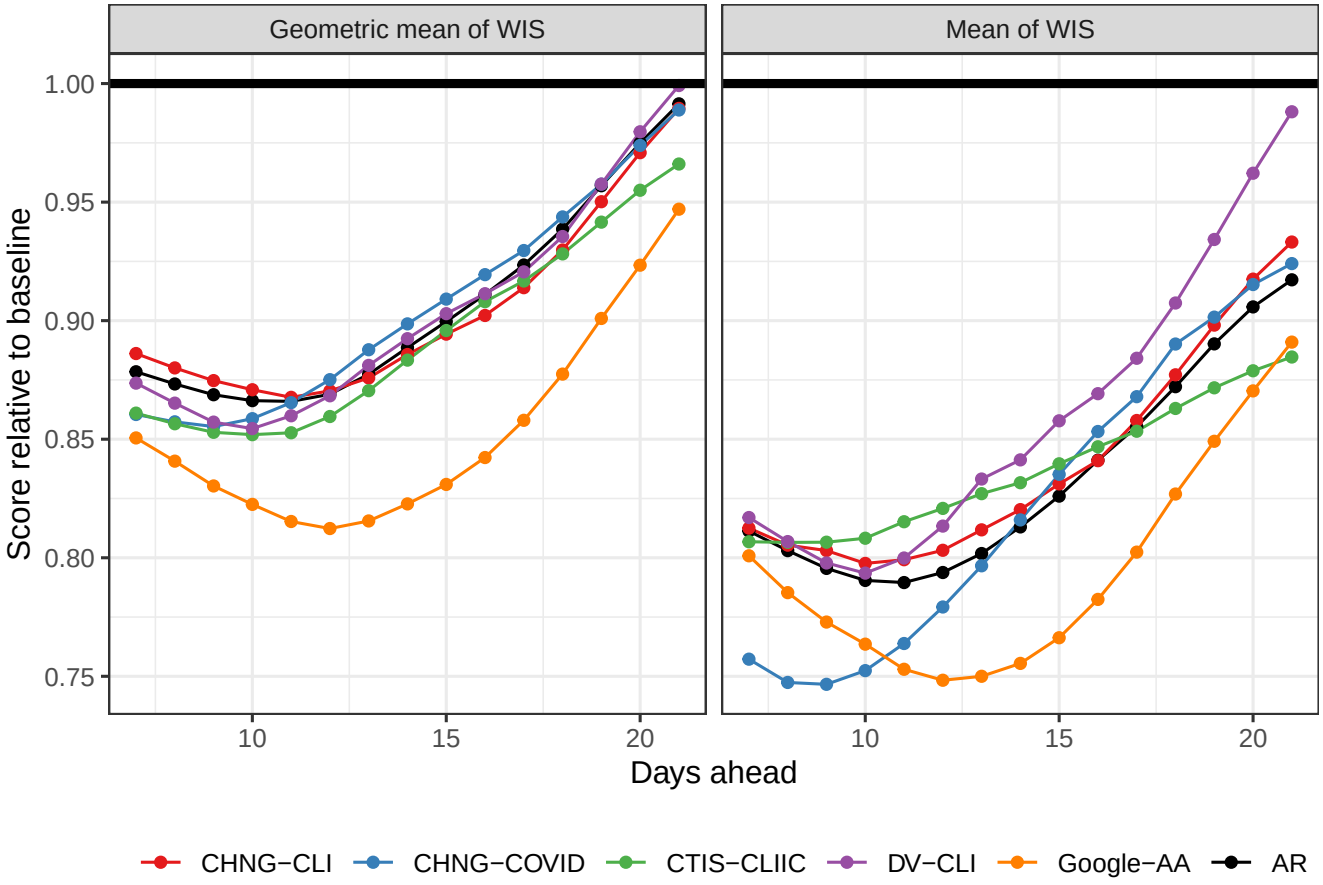

Fig. S8. Performance of our models re-estimated for state-level, epiweek case incidence.

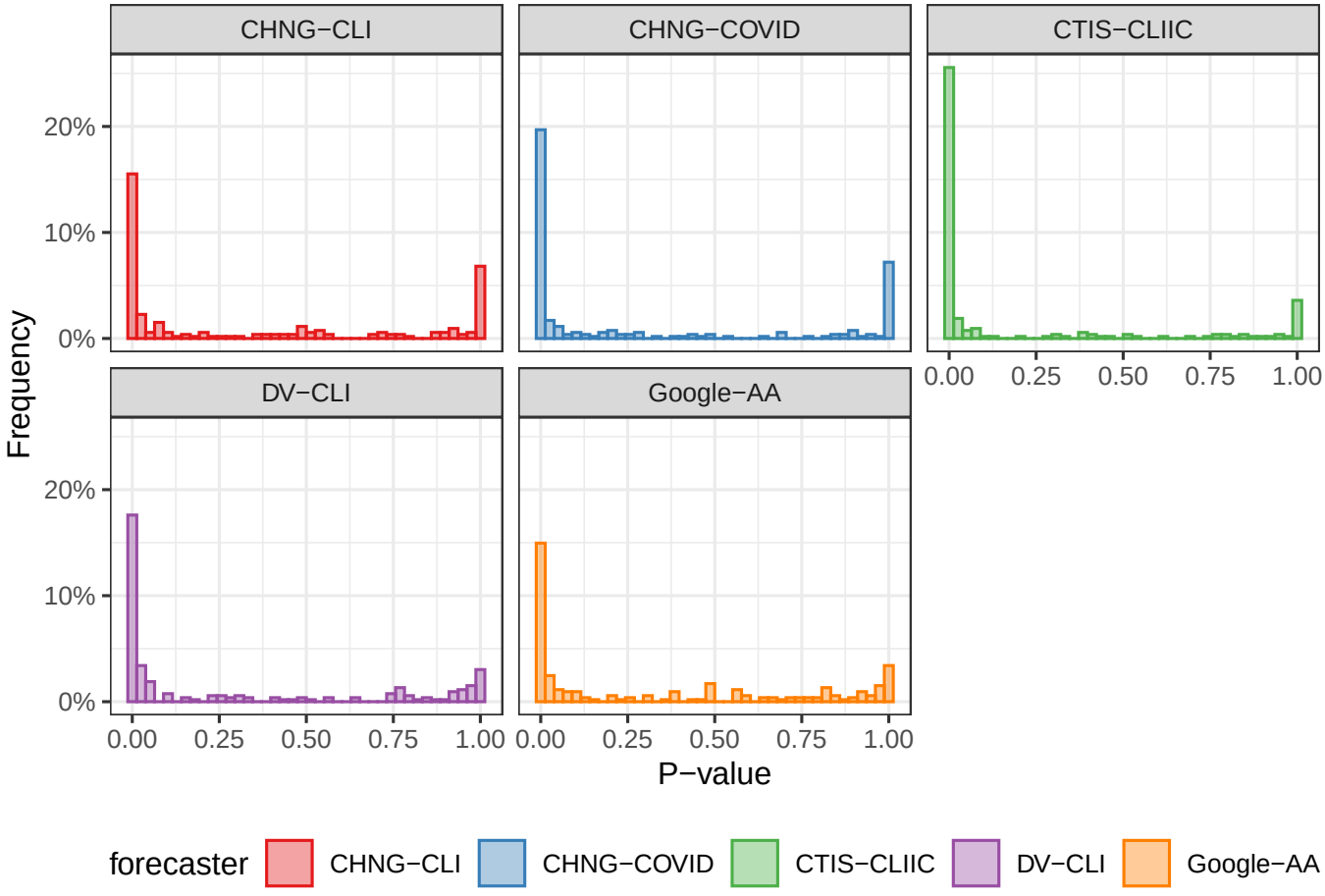

Fig. S9. P-values for a sign test that the WIS of the AR forecaster is smaller than that of the indicator-assisted model. Each P-value corresponds a particular forecast date.

Table S1. P**-values for a one-sided Diebold-Mariano test for improvement in forecast error. The test is for the null hypothesis of equal performance against the alternative that the indicator assisted model is better.**

metric	CHNG-CLI	CHNG-COVID	CTIS-CLIIC	DV-CLI	Google-AA
Geometric Mean Relative WIS	0.116	0.067	0.000	0.059	0.008
Mean Relative WIS	0.159	0.111	0.002	0.074	0.064

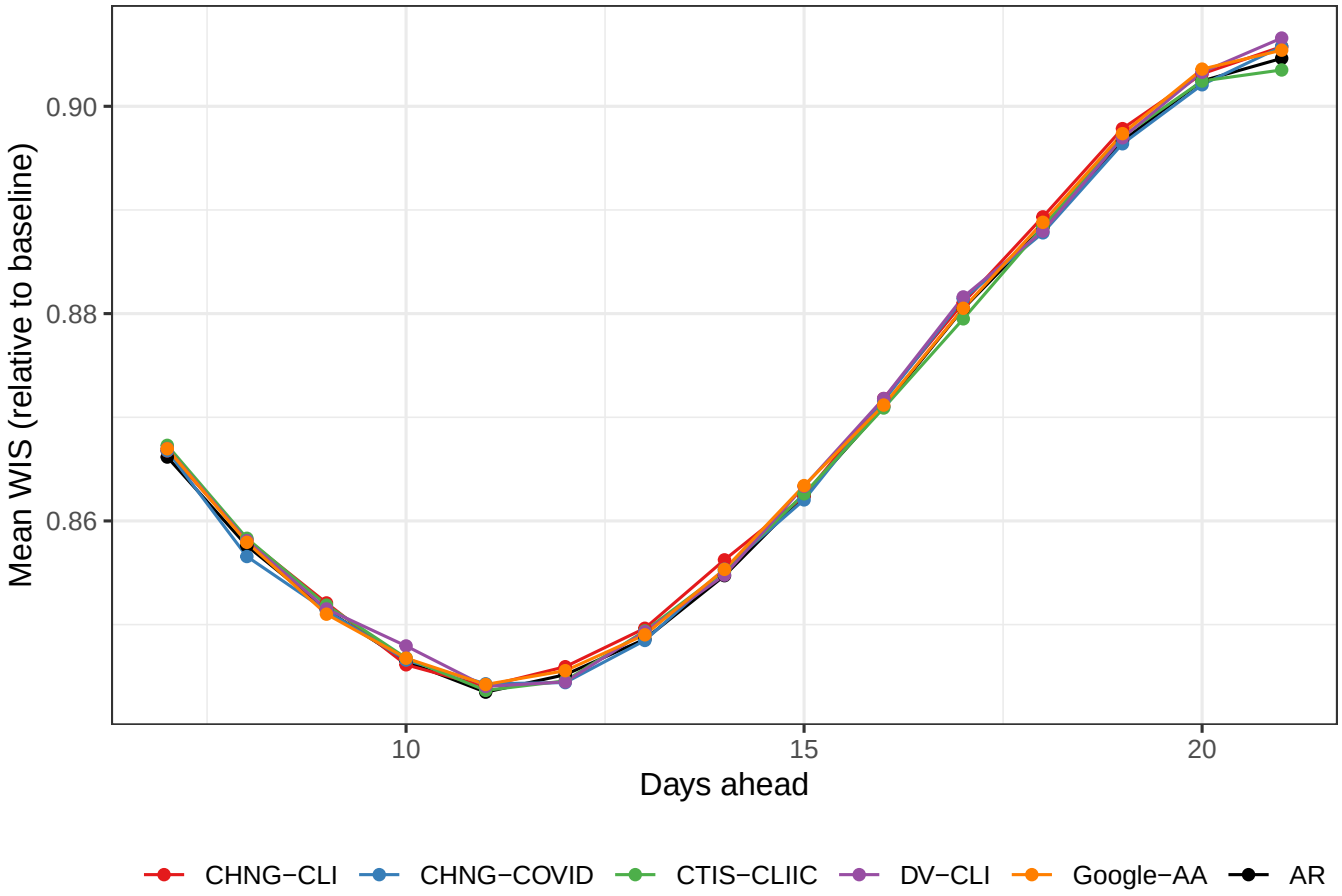

Fig. S10. Forecast performance when indicators are replaced with samples from their empirical distribution. Performance is largely similar to the AR model.

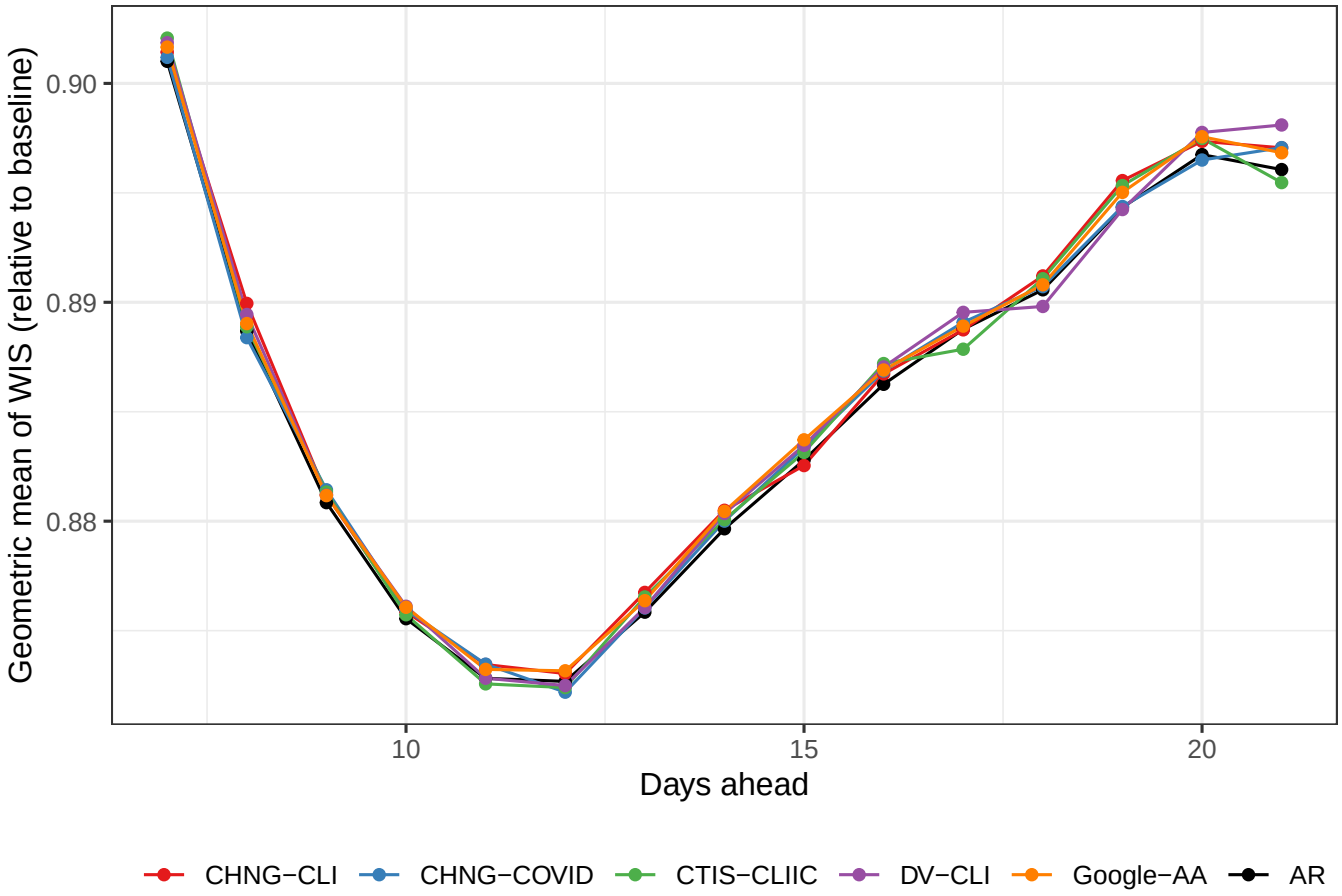

Fig. S11. Forecast performance as measured with the geometric mean when indicators are replaced with samples from their empirical distribution. Performance is largely similar to the AR model.

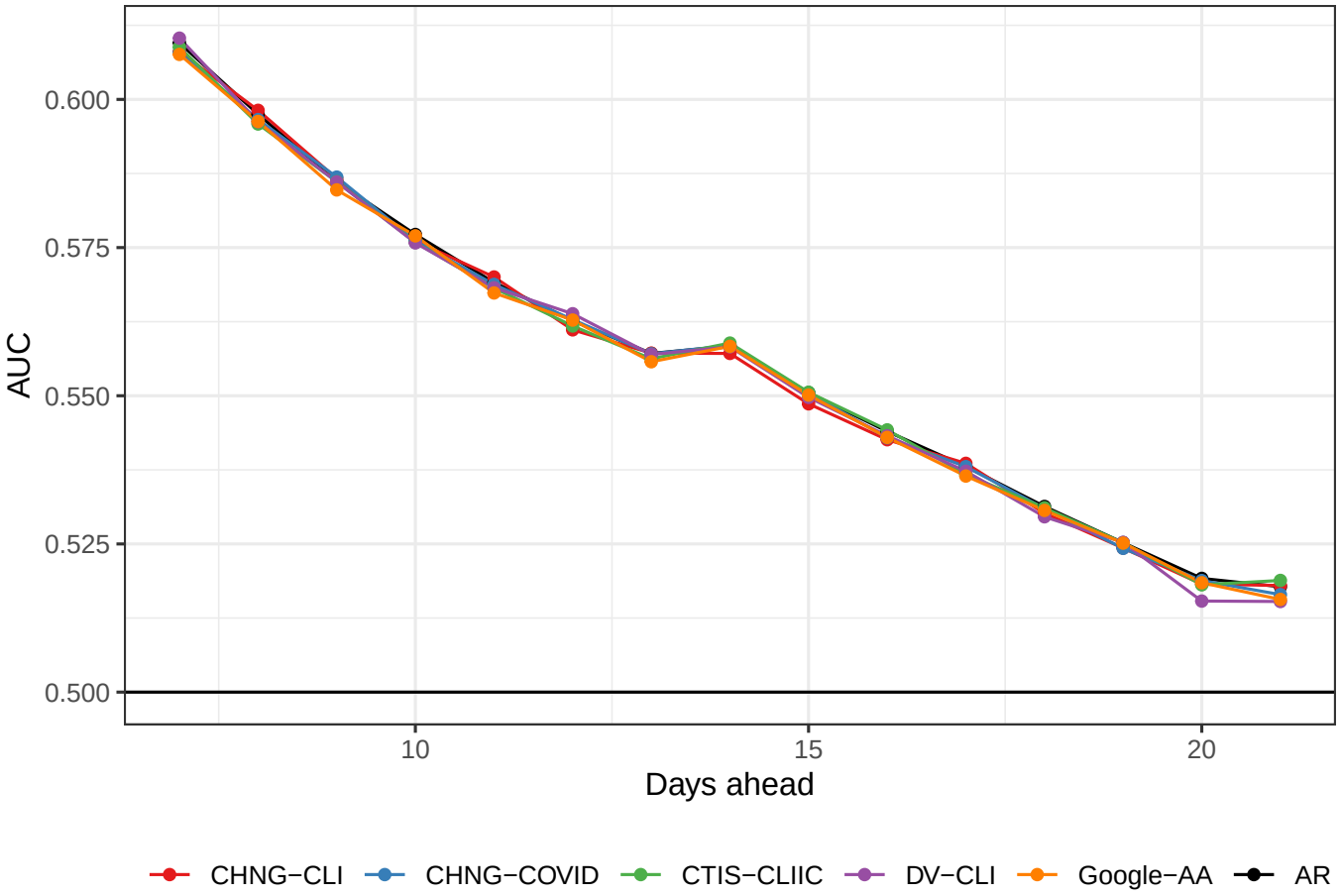

Fig. S12. Hotspot prediction performance when indicators are replaced with samples from their empirical distribution. Performance is largely similar to the AR model.

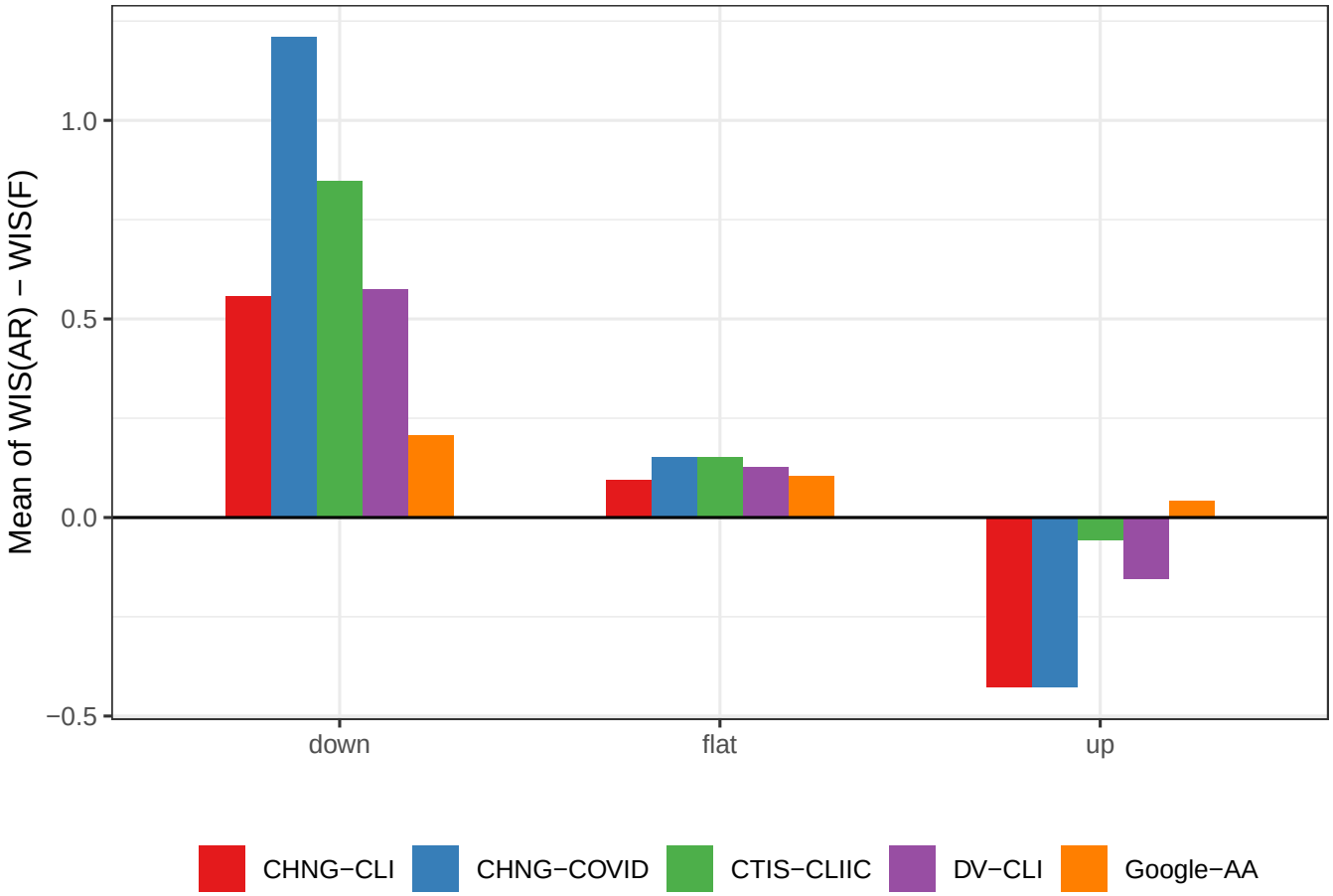

Fig. S13. Average difference between the WIS of the AR model and the WIS of the other forecasters. The indicator-assisted forecasters do best during down and flat periods.

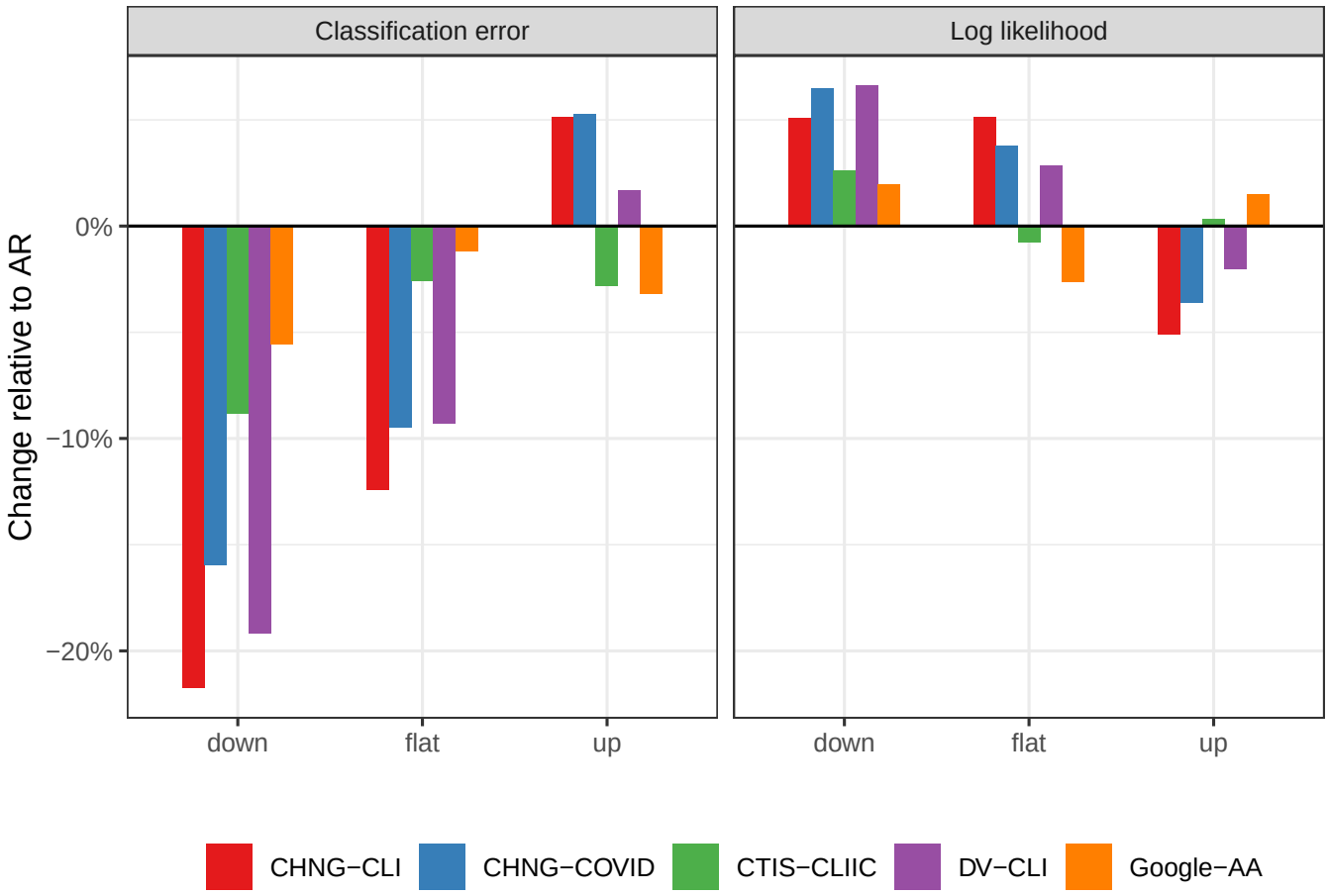

Fig. S14. Classification and loglikelihood separated into periods of upswing, downswing, and flat cases. Like the analysis of the forecasting task in the main paper (see Figure 7), relative performance is better during down and flat periods.

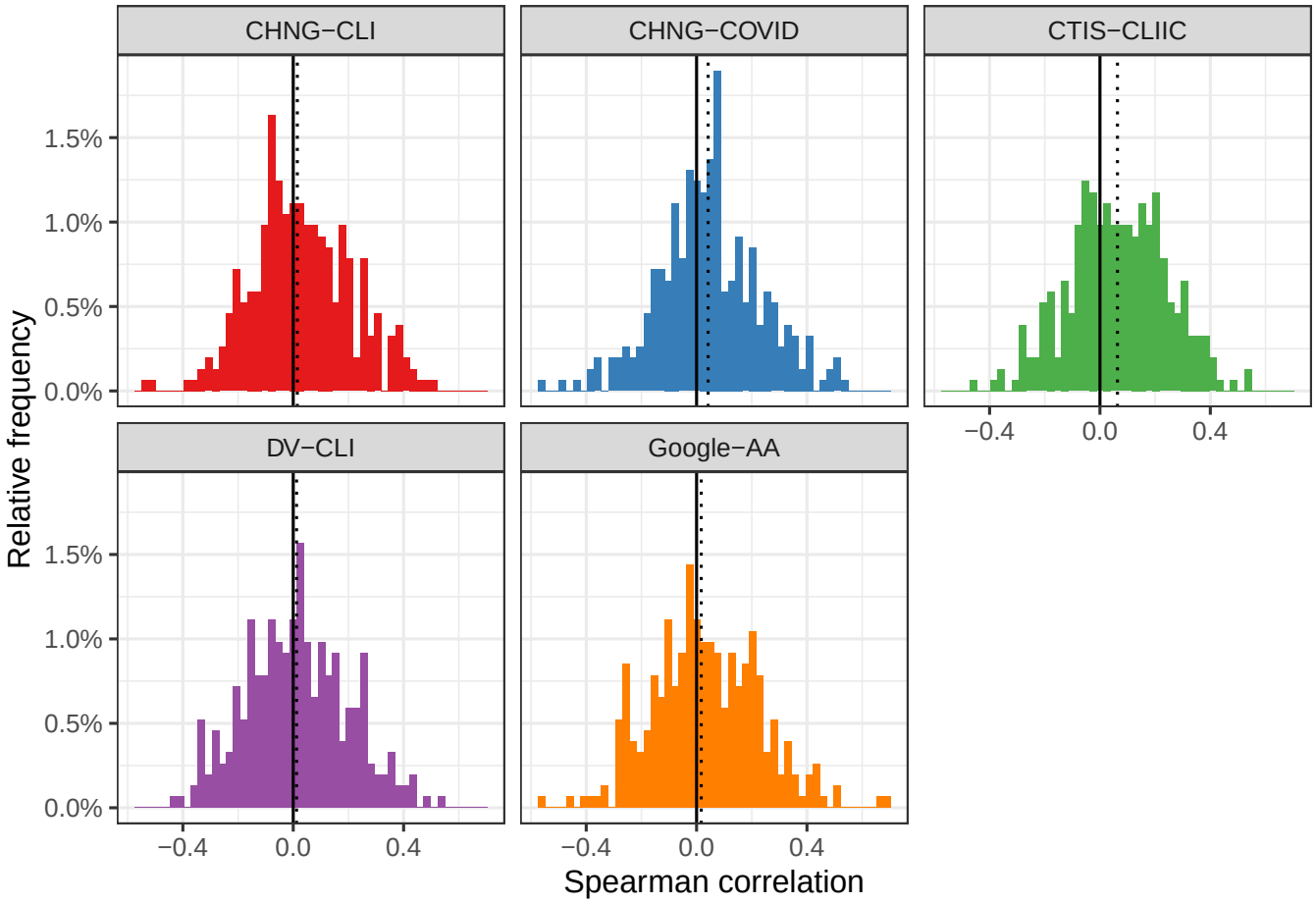

Fig. S15. Histograms of the Spearman correlation between the ratio of AR to AR WIS with the percent change in smoothed case rates relative to 7 days earlier.

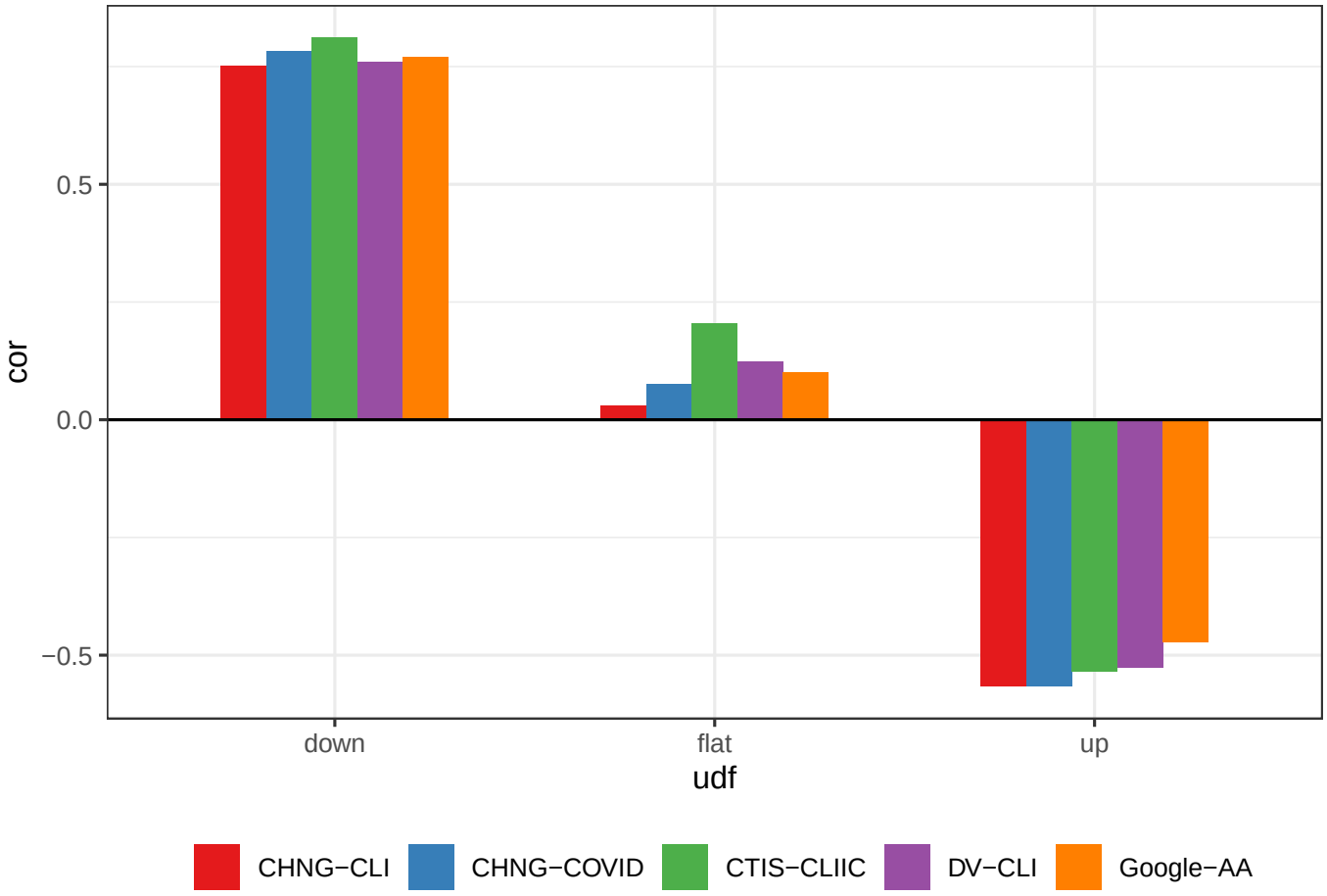

Fig. S16. Correlation of the difference in WIS with the difference in median predictions for the AR model relative to the indicator-assisted forecaster. In down periods, improvements in forecast risk are highlycorrelated with lower median predictions. The opposite is true in up periods. This suggests, as one might expect that improved performance of the indicator-assisted model is attributable to being closer to the truth then the AR model. This conclusion is stronger in down periods then in up periods.

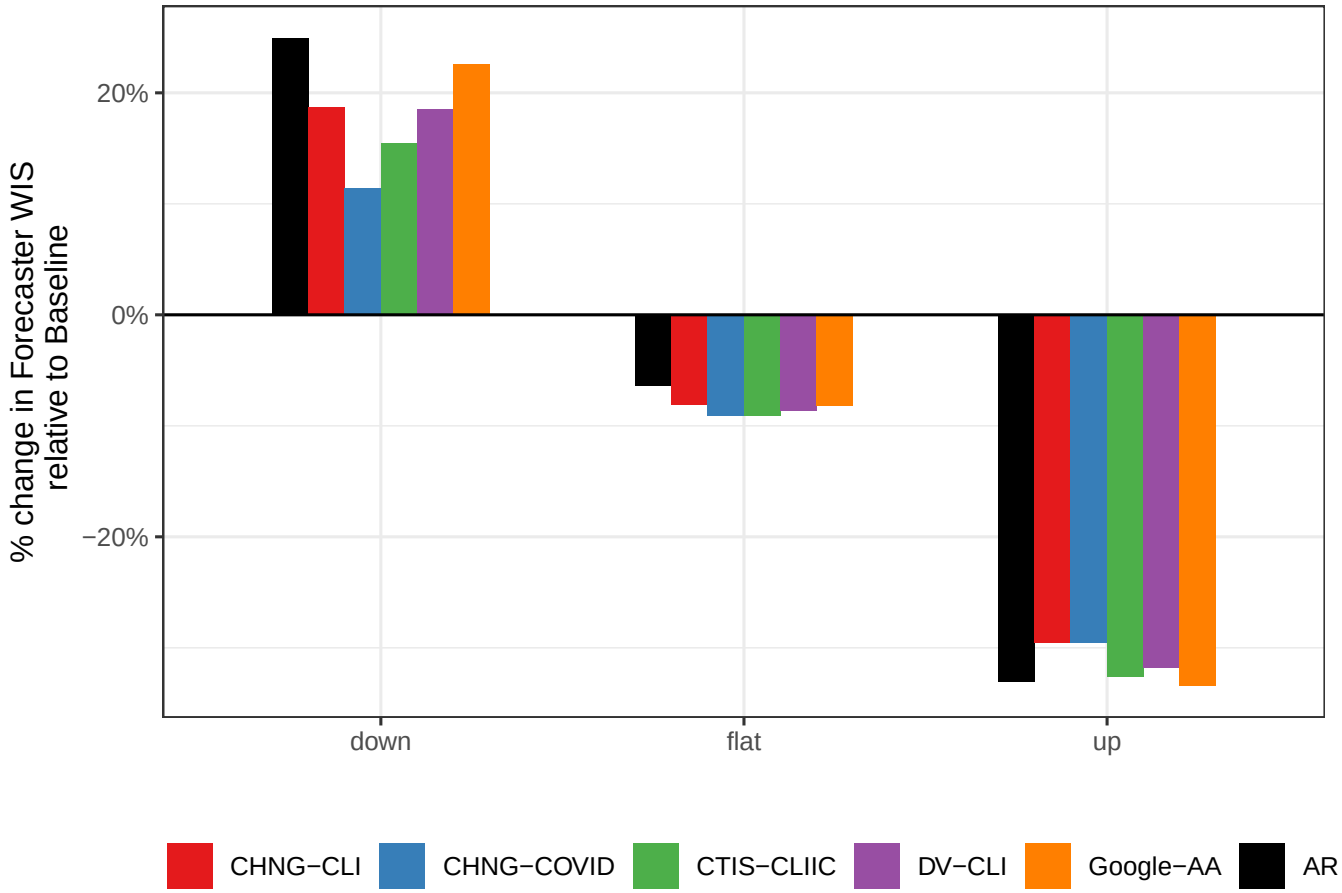

Fig. S17. Percent change in average WIS of the forecaster (AR or indicator assisted) relative to the baseline. All models perform poorly durning down periods, but the indicators help. During flat periods, the indicators improve slightly over the AR. During up periods, all forecasters do much better than the baseline, but only some do as well as AR.

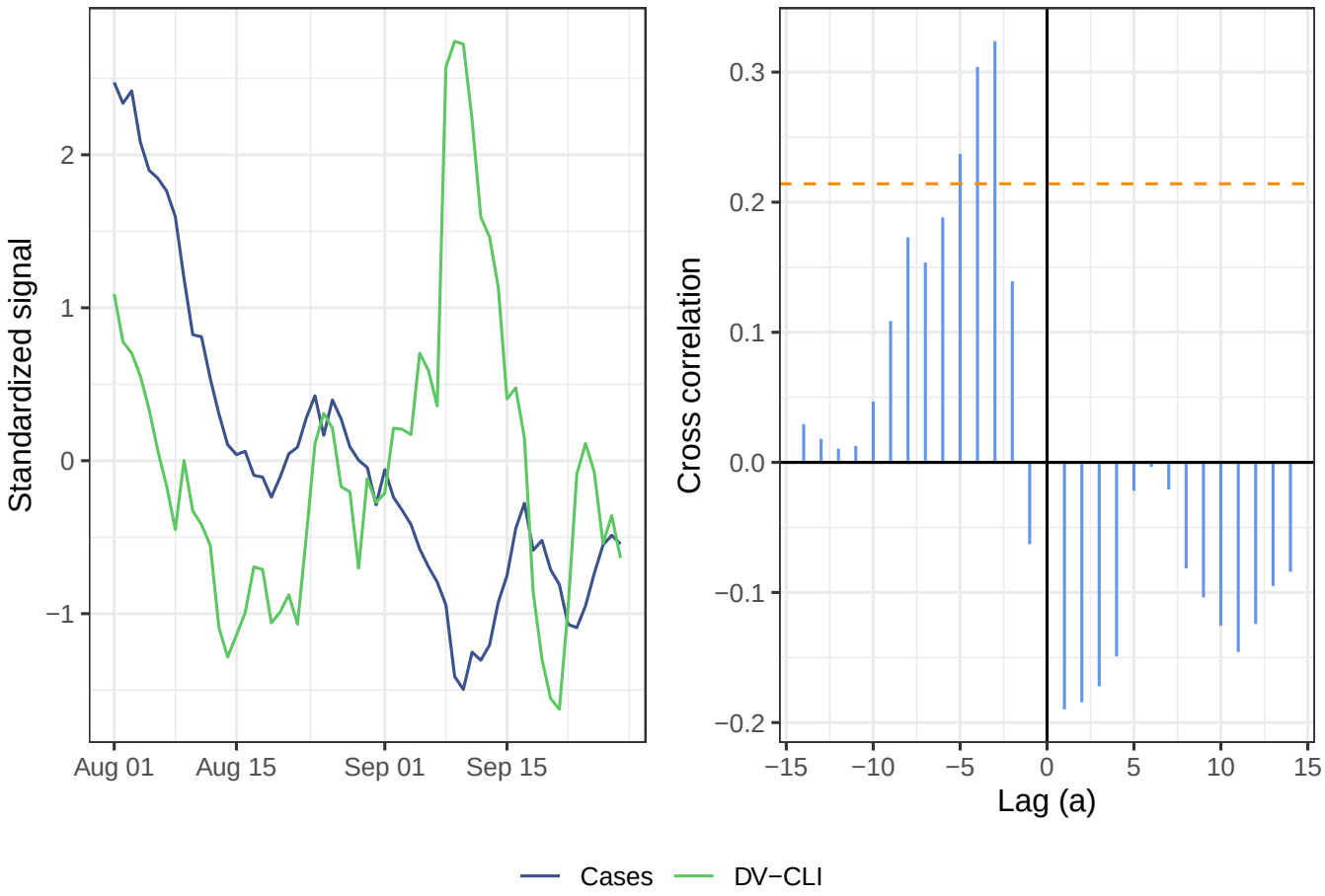

Fig. S18. Illustration of the cross-correlation function between DV-CLI and cases. The left panel shows the standardized signals over the period from August 1 to September 28 (as of May 15, 2021). The right panel shows $\mathrm{CCF}_\ell(a)$ for different values of a as vertical blue bars. The orange dashed lines indicate the 95% significance threshold. By our leadingness/laggingness metric, DV-CLI is leading (but notlagging) cases over this period.

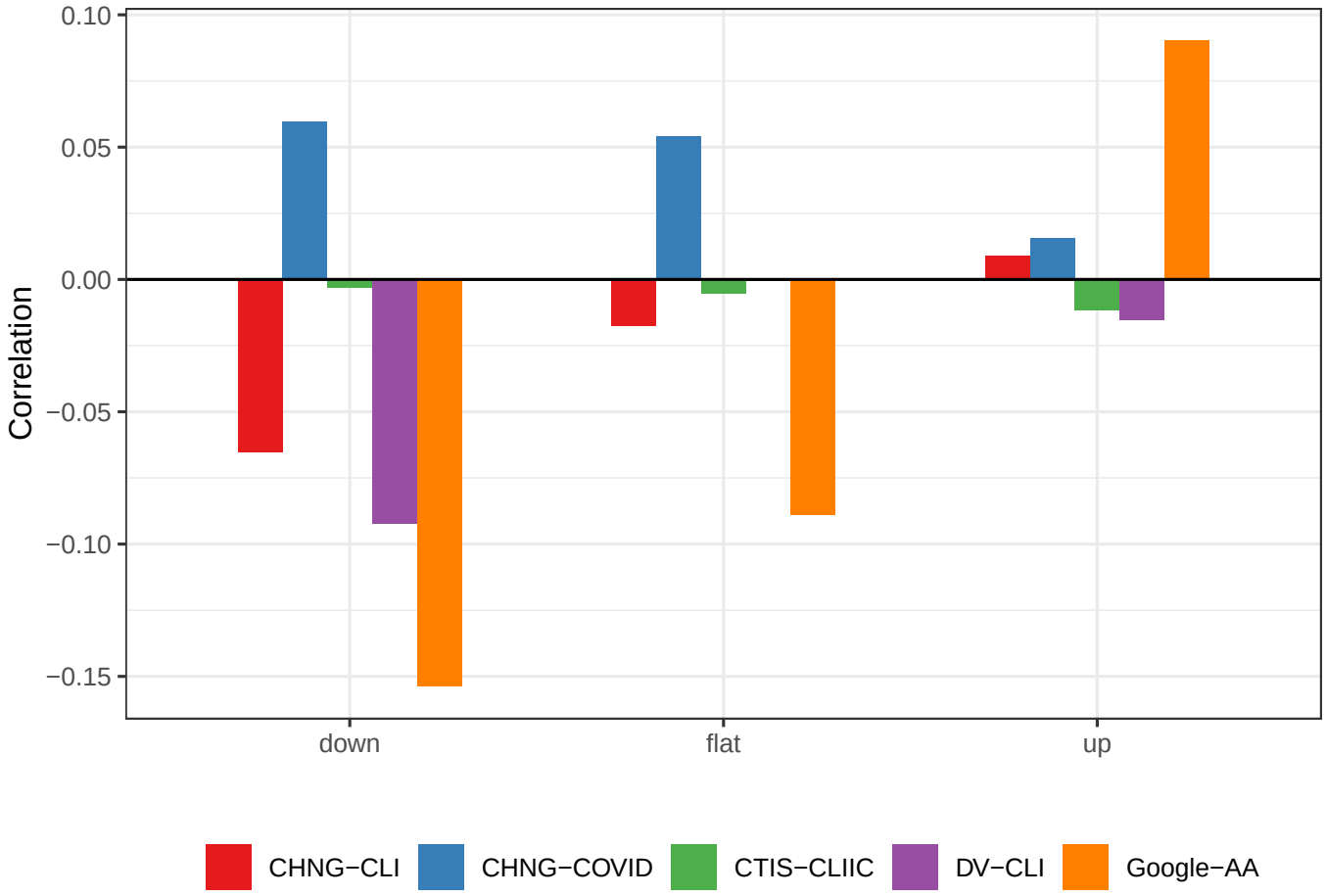

Fig. S19. Correlation of the difference in WIS with the laggingness of the indicator at the target date, stratified by up, down, or flat period. Compare to Figure 5 in the manuscript.

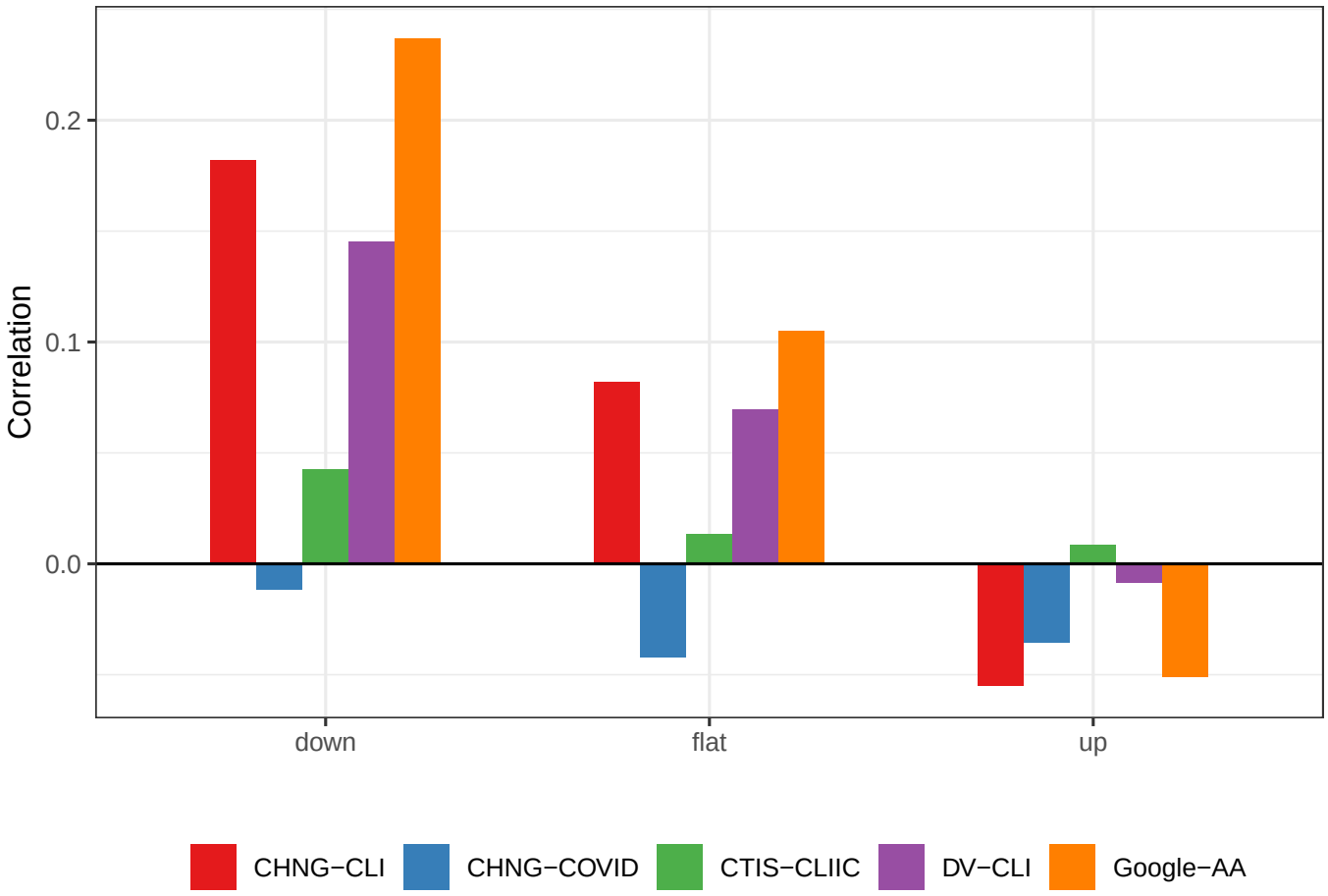

Fig. S20. Correlation of the difference between leadingness and laggingness with the difference in WIS. The relationship is essentially the same as described in the manuscript and shown in Figure 5.

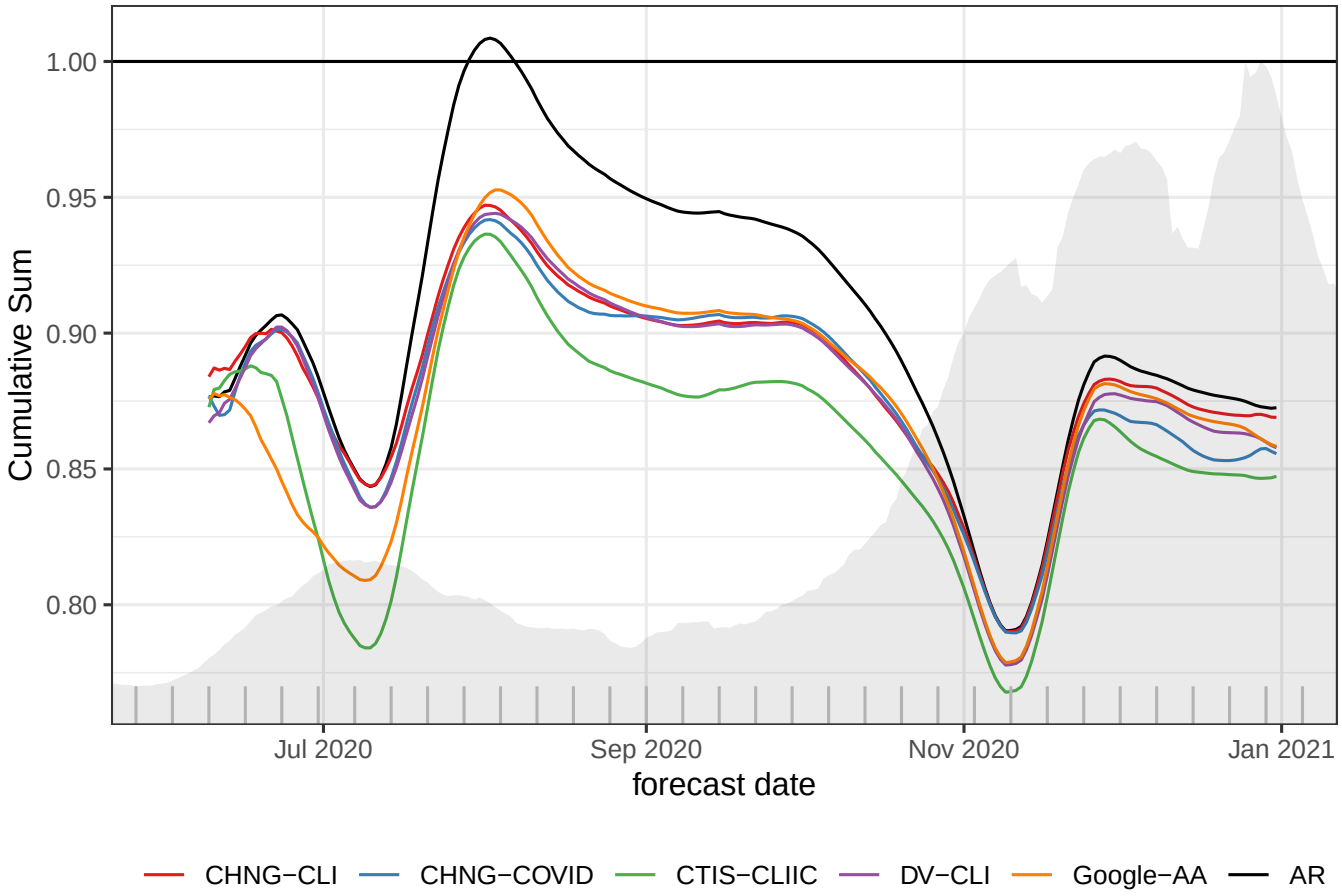

Fig. S21. This figure shows the cumulative sum of WIS for each forecaster divided by the cumulative sum of WIS for the baseline model. The shaded background shows total U.S. cases as reported by JHU-CSSE for the 14-day ahead target. Hashes along the x-axis denote weeks.

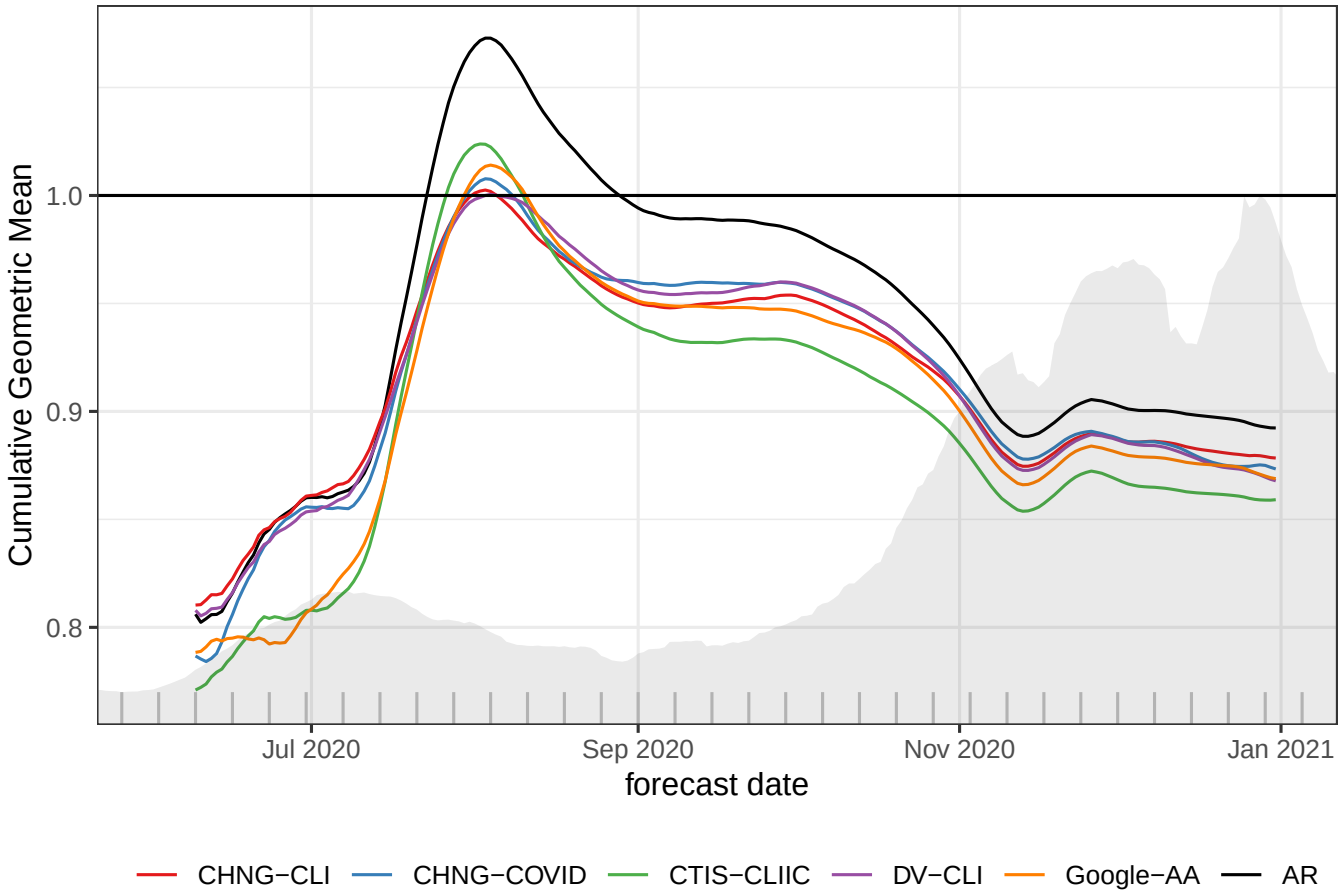

Fig. S22. This figure shows the cumulative geometric mean of WIS for each forecaster divided by the WIS for the baseline model. The shaded background shows total U.S. cases as reported by JHU-CSSE for the 14-day ahead target. Hashes along the x-axis denote weeks.

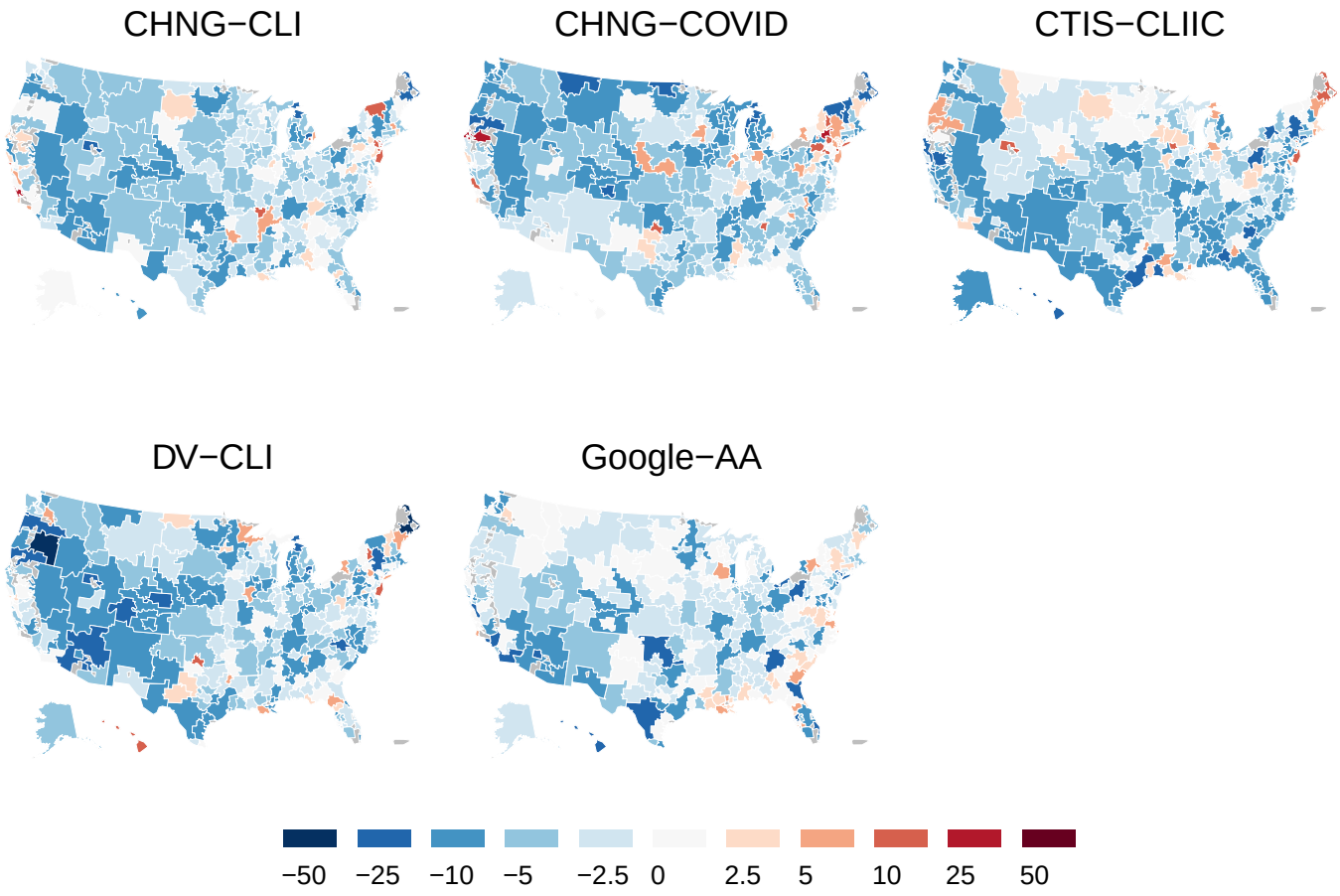

Fig. S23. Percent improvement in WIS relative to the AR forecaster by region (negative numbers indicate improved performance, positives indicate worsening).

SI Dataset S1 (predictions.zip)

Archived `.RDS` (R objects) files containing all predictions for forecasting and hotspots using vintage and finalized data. Persistent DOI to be added at publication.

SI Dataset S2 (evaluations.zip)

Archived `.RDS` (R objects) files containing all evaluations for forecasting and hotspots using vintage and finalized data. Persistent DOI to be added at publication.

SI Dataset S3 (analysis.zip)

Archived `.RDS` (R objects) files containing additional data used to produce graphics and conclusions in the manuscript. Persistent DOI to be added at publication.

SI Dataset S4 (code.zip)

R script files containing all code used to reproduce the analysis described in the manuscript. Persistent DOI to be added at publication.

References

1 Reich Lab, The COVID-19 Forecast Hub (https://covid19forecasthub.org) (2020).

2 Delphi Research Group, covidcast R package (https://cmu-delphi.github.io/covidcast/covidcastR) (2020).

3 FX Diebold, RS Mariano, Comparing predictive accuracy. *J. Bus. & Econ. Stat.* **20**, 134–144 (2002).

4 FX Diebold, Comparing predictive accuracy, twenty years later: A personal perspective on the use and abuse of diebold–mariano tests. *J. Bus. & Econ. Stat.* **33**, 1–24 (2015).

5 D Harvey, S Leybourne, P Newbold, Testing the equality of prediction mean squared errors. *Int. J. forecasting* **13**, 281–291 (1997).