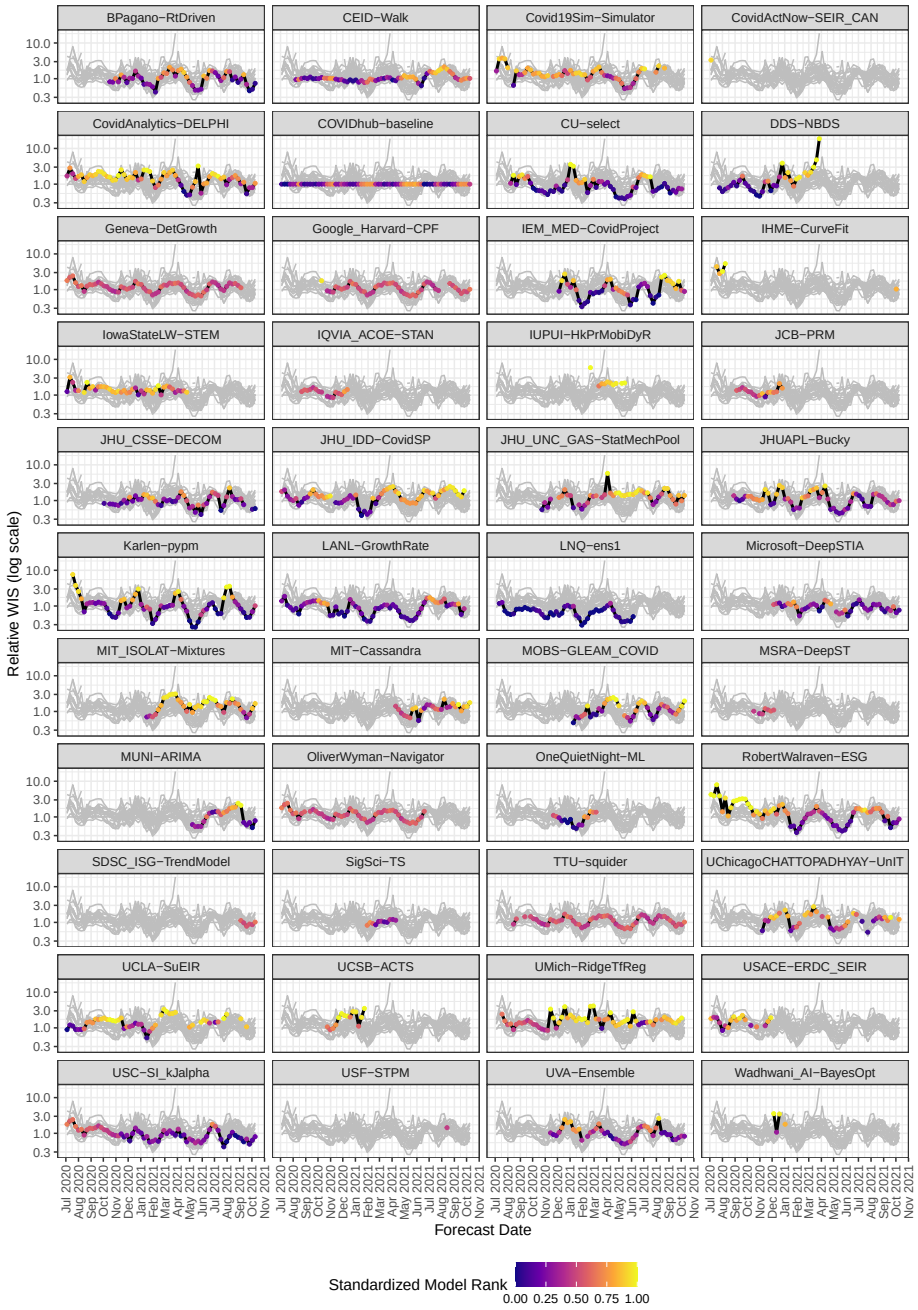
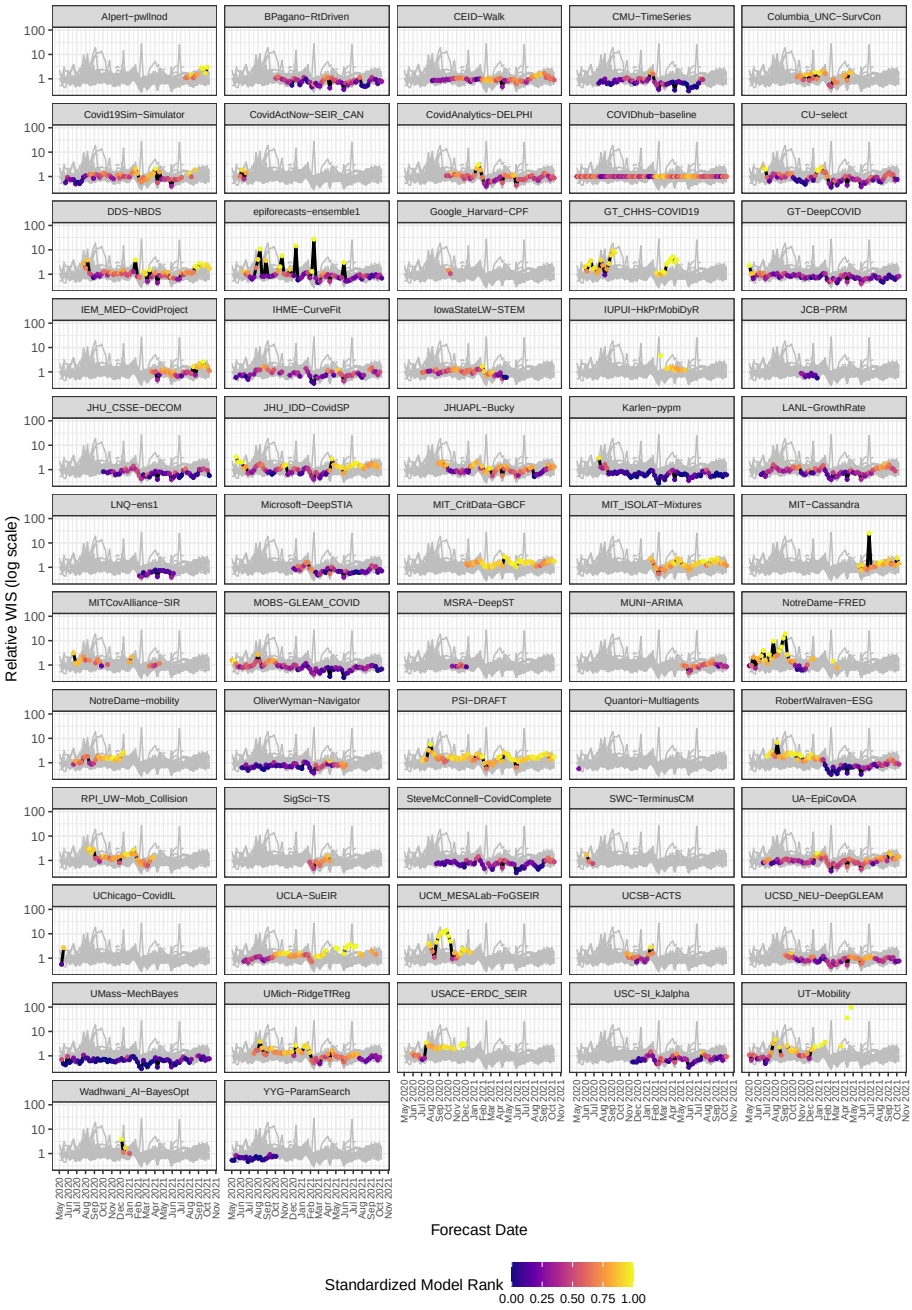
Supplemental Materials for Comparing trained and untrained probabilistic ensemble forecasts of COVID-19 cases and deaths in the United States

1 Relative WIS of component forecasters

Supplemental Figures 1 and 2 show the relative WIS of each component forecaster as a function of the forecast week for the weekly cases and weekly deaths targets respectively. For each week, we calculated a standardized rank for each model based on where that model's relative WIS fell relative to all other models that had submissions that week. In these rankings, 0 indicates the model with the best performance, and 1 indicates the model with the worst performance as measured by relative WIS. We see that nonstationarity of relative performance is very common, with many models alternating between weeks with top-ranking performance and bottom-ranking performance.



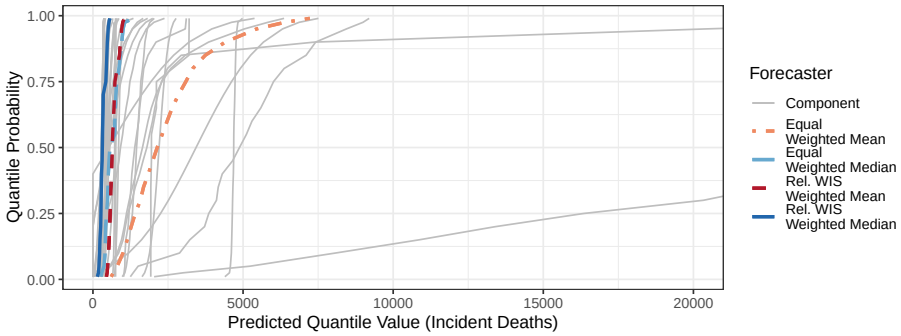
Supplemental Figure 1: Weekly component ranks according to relative WIS for forecasts of cases in the U.S. A rank of 0 indicates that the model has the best performance in a given week, and a rank of 1 indicates that it has the worst performance.



Supplemental Figure 2: Weekly component ranks according to relative WIS for forecasts of deaths in the U.S. A rank of 0 indicates that the model has the best performance in a given week, and a rank of 1 indicates that it has the worst performance.

2 Quantile ensembles as horizontal combinations of predictive cdfs

Supplemental Figure 3 illustrates the ensemble methods considered in this manuscript as horizontal combinations of the cumulative distribution functions of predictive distributions from component forecasters, computing a weighted or unweighted mean or median at each quantile probability level along the vertical axis.



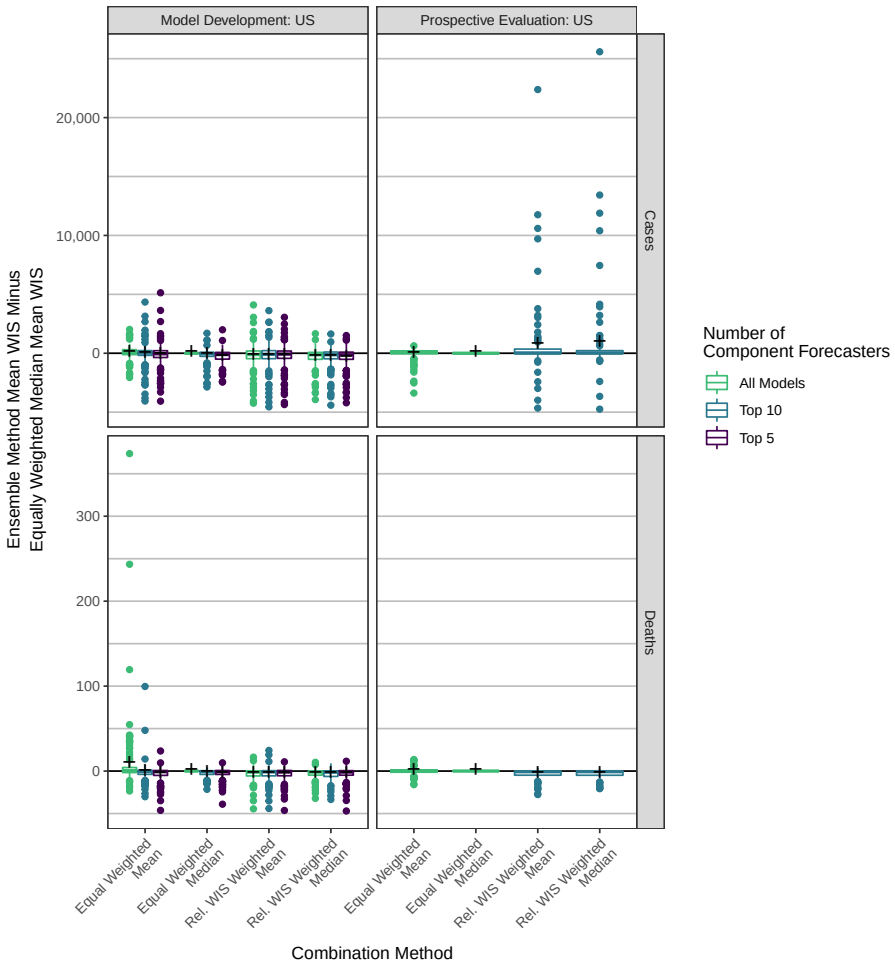
Supplemental Figure 3: Illustration of four ensemble methods for forecasting incident deaths in Ohio at a forecast horizon of 1 week from February 15, 2021. Each line corresponds to the forecast distribution from one component model or ensemble, and is obtained by interpolating between the 23 predictive quantiles; the resulting curves approximate the predictive CDFs associated with these forecasts. The curves are cut off for two component forecasters with extremely wide predictive distributions. At each quantile level along the vertical axis, the ensemble forecasts are obtained as a combination of the component model forecasts at that quantile level.

3 Expanded results from primary analysis

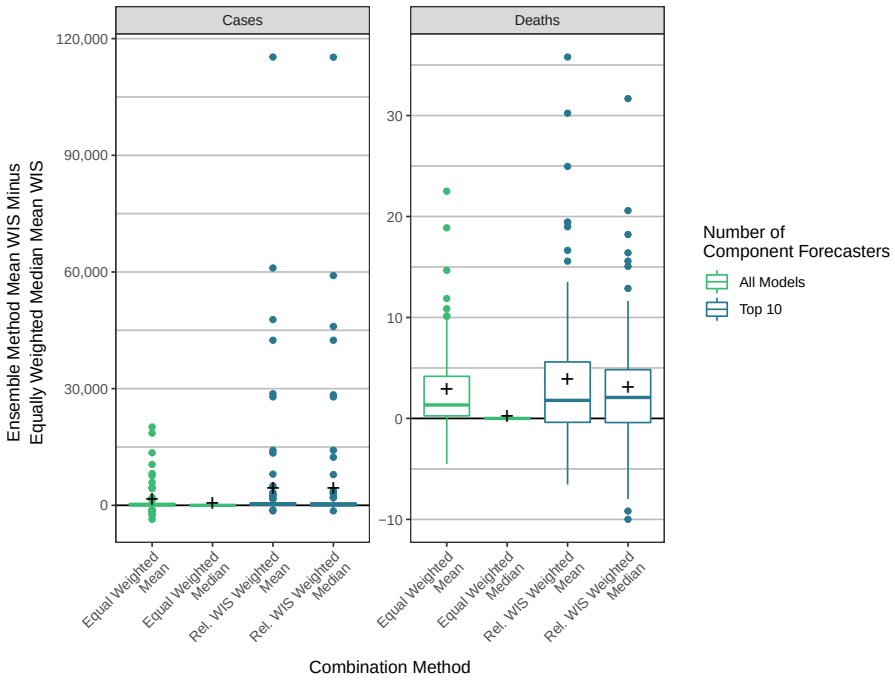
This section includes figures giving additional views of the primary results from Figures 3 and 4 in the article.

3.1 Full distributions of Weighted Interval Score differences

For legibility, Figures 3 and 4 in the main text displayed the central tendency (central quantiles and means) of differences in weighted interval scores between the different methods, but suppressed outliers corresponding to individual combinations of forecast dates and horizons with large differences between the equal weighted median ensemble and another ensemble method. Supplemental Figures 4 and 5 display full box plots including outliers that were suppressed in the main text.



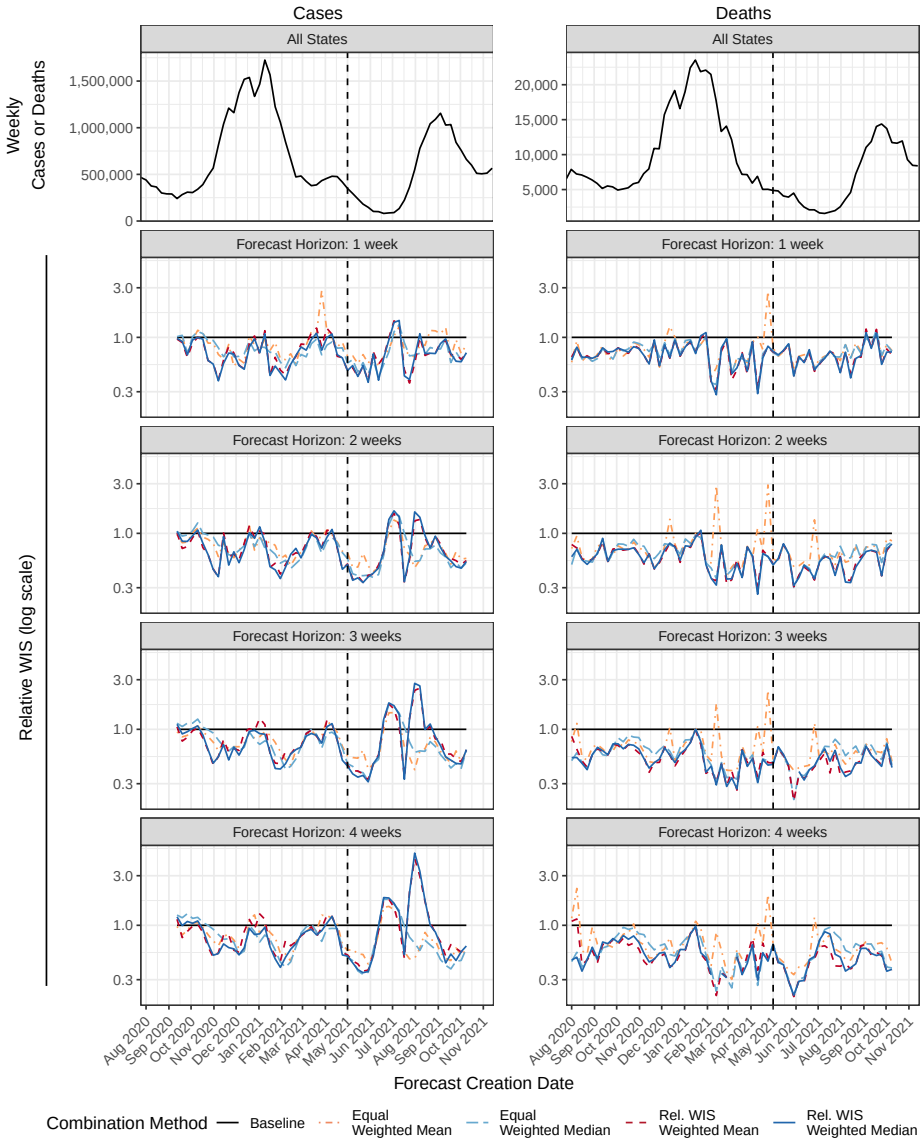
Supplemental Figure 4: Performance measures for ensemble forecasts of weekly cases and deaths in the U.S. The vertical axis is the difference in mean WIS for the given ensemble method and the equally weighted median ensemble. Boxplots summarize the distribution of these differences in means, averaging across all locations for each combination of forecast date and horizon. A cross is displayed at the difference in overall mean scores for the specified combination method and the equally weighted median averaging across all locations, forecast dates, and horizons. A negative value indicates that the given method outperformed the equally weighted median.



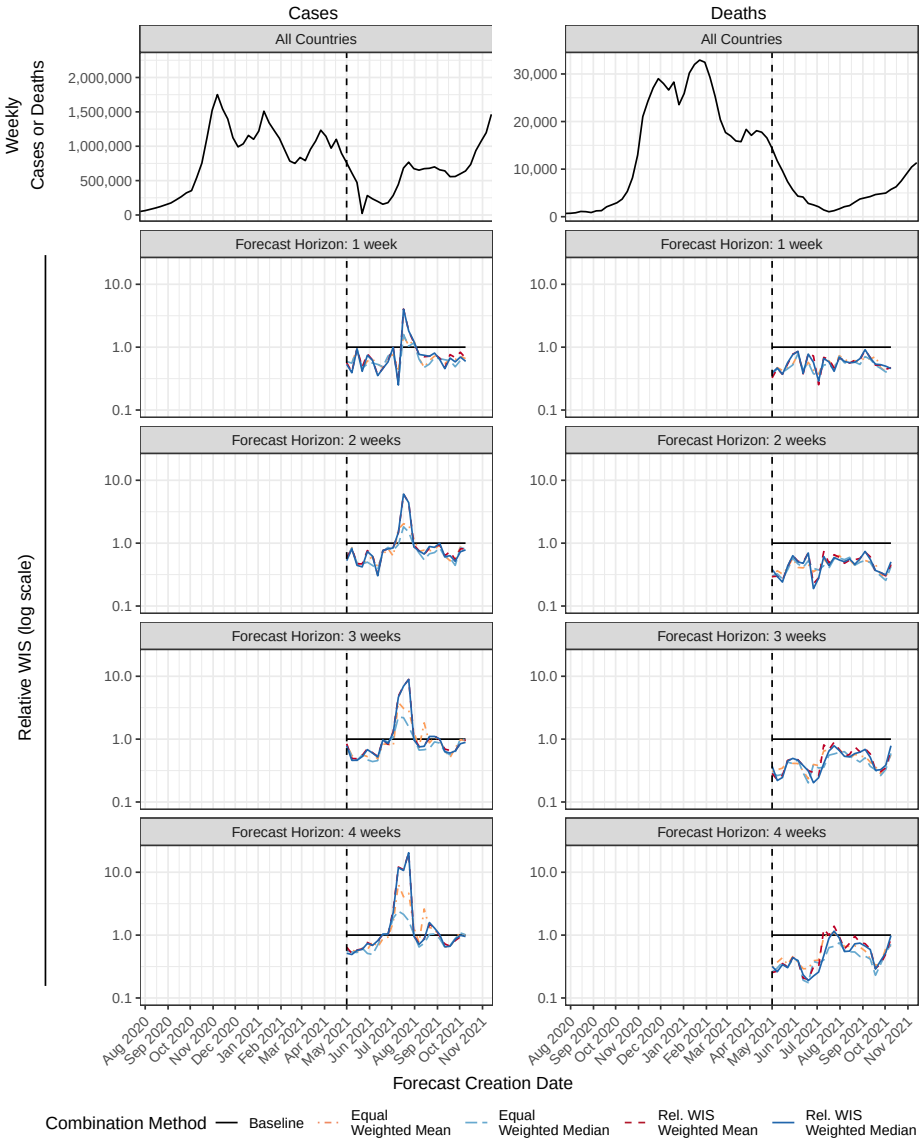
Supplemental Figure 5: Performance measures for ensemble forecasts of weekly cases and deaths in Europe. The vertical axis is the difference in mean WIS for the given ensemble method and the equally weighted median ensemble. Boxplots summarize the distribution of these differences in means, averaging across all locations for each combination of forecast date and horizon. A cross is displayed at the difference in overall mean scores for the specified combination method and the equally weighted median averaging across all locations, forecast dates, and horizons. A negative value indicates that the given method outperformed the equally weighted median.

3.2 Scores by forecast creation date

Supplemental Figures 6 and 7 show relative WIS for the ensemble methods over time for forecasts in the US and in Europe respectively. The included ensemble methods are 1) an equally weighted mean ensemble, 2) an equally weighted median ensemble, 3) a weighted mean ensemble, and 4) a weighted median ensemble. Both of the weighted ensembles combine the ten component forecasters with best individual performance as measured by the relative WIS, and are trained on a sliding 12-week window. The component forecasters included in the trained ensembles are updated each week based on performance during the training window.



Supplemental Figure 6: Weekly reported cases and deaths at the national level in the United States and mean weighted interval scores (WIS) relative to the baseline for state-level forecasts over time for four ensembles. Mean WIS calculated separately for each combination forecast horizon and forecast creation date, averaging across all states and territories, and then normalized relative to the mean WIS for the baseline model. Lower scores indicate better forecast performance. A vertical dashed line is shown at the start of the prospective evaluation phase.



Supplemental Figure 7: Weekly reported cases and deaths aggregated across all European countries included in the European Forecast Hub and mean weighted interval scores (WIS) relative to the baseline for state-level forecasts over time for four ensembles. Mean WIS calculated separately for each combination forecast horizon and forecast creation date, averaging across all states and territories, and then normalized relative to the mean WIS for the baseline model. Lower scores indicate better forecast performance. A vertical dashed line is shown at the start of the prospective evaluation phase.

3.3 Impact of reporting anomalies

We conducted a supplemental analysis in which we removed forecasts that were affected by reporting anomalies before calculating summaries of forecast performance. We catalogued two types of reporting anomalies, as illustrated in Supplemental Figure 8:

1. **Outliers** were identified manually by examining plots of the data. Negative weekly counts and other observations that did not appear to match local trends were recorded as outliers.

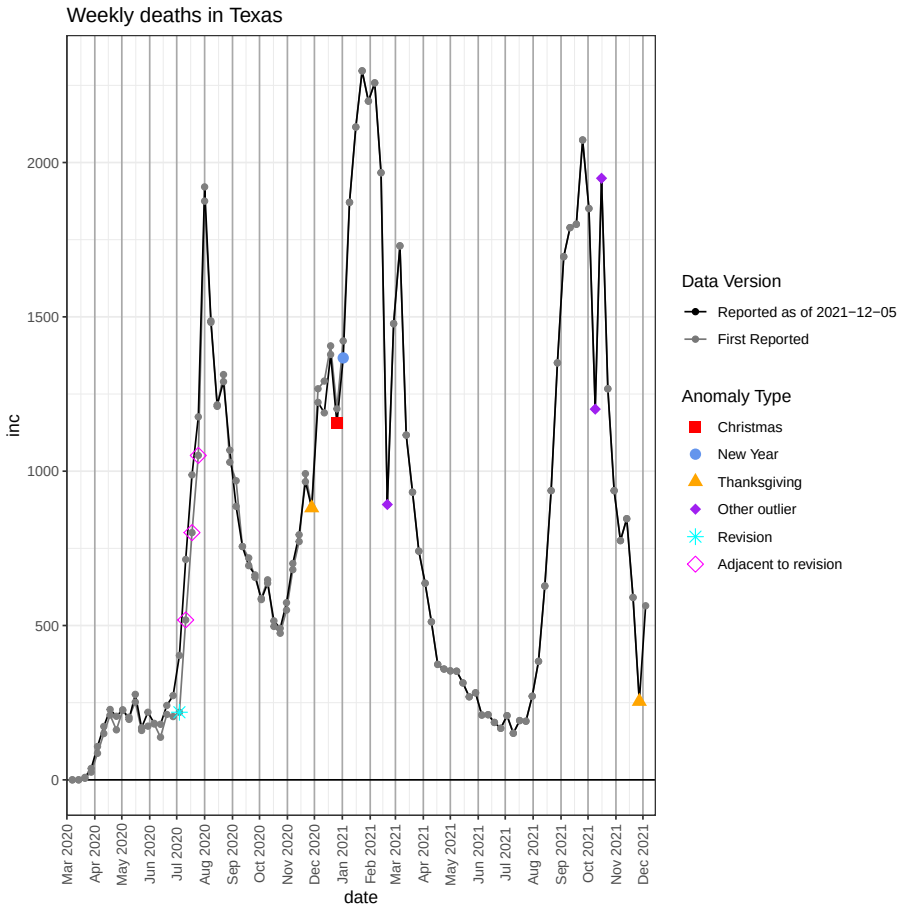
2. **Revisions** were identified automatically. The value for a particular week was identified as a large revision if the difference between the original reported value and the final reported value was at least 20, and that difference was at least 40% of the initial reported value or the final reported value.

For the purpose of this analysis, forecasts with a target end date coinciding with an observation that was identified as an outlier were excluded. This is because we would prefer forecasting methods to focus on capturing the epidemiological process rather than aspects of the reporting process that lead to outliers.

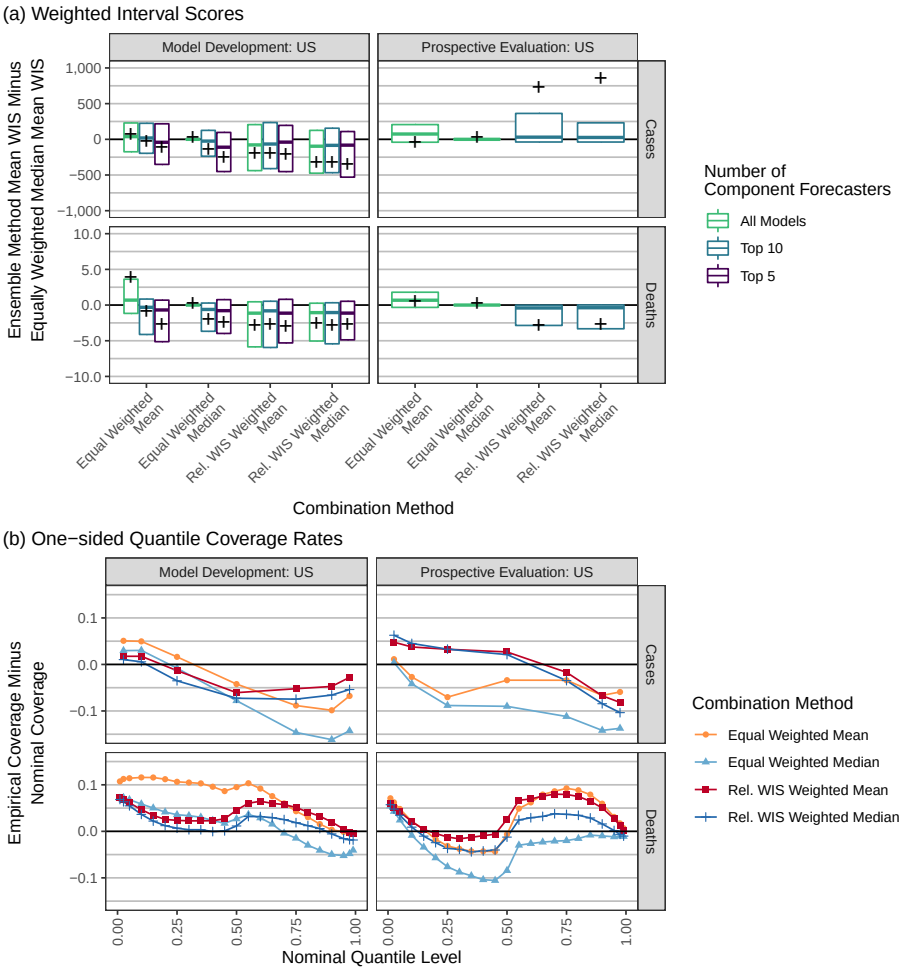
Forecasts with a forecast date on the date of a value that was later revised, or within the following 3 weeks unless the revision had already been made by the time of the forecast, were excluded. These forecasts were excluded because the input data used for the forecast were not a reliable indicator of the state of the epidemic at the time of the forecast. One might reasonably expect forecasters to account for the possibility of such data revisions, but this analysis represents a conservative examination of whether these revisions affected the main results in the article.

Together, these criteria led to removal of 294 combinations of location, forecast date, and forecast horizon out of 14,332 such combinations throughout the model development and prospective evaluation phases in the U.S.

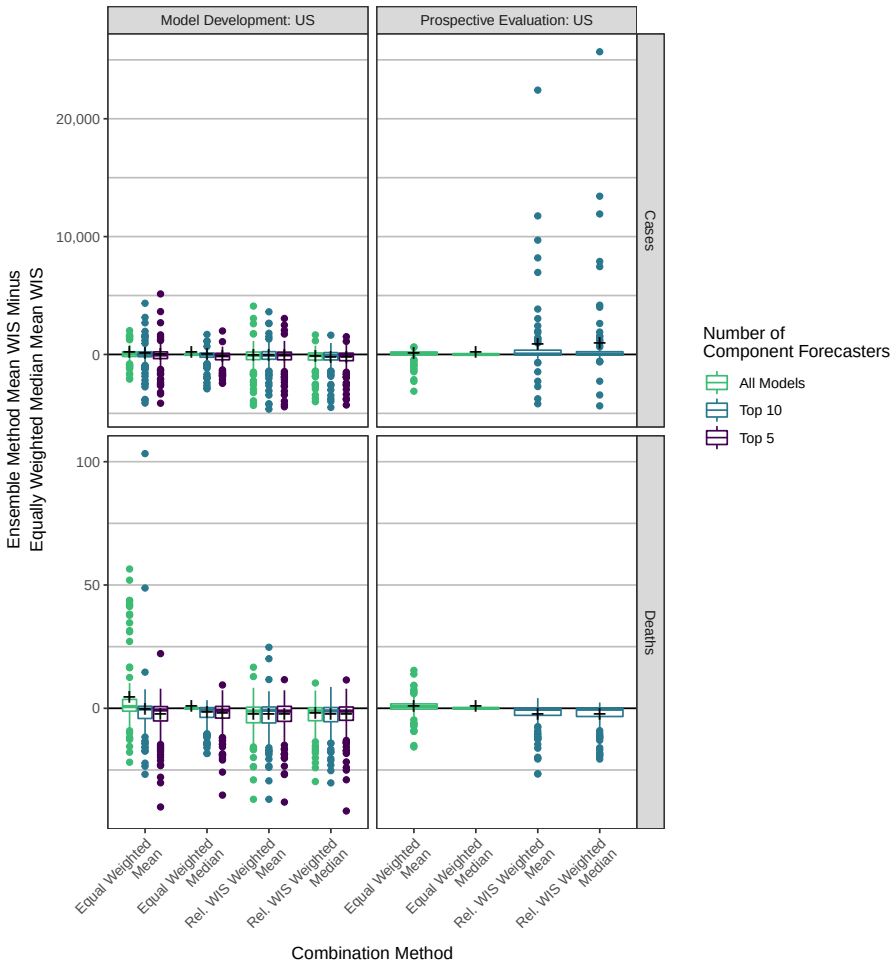
Supplemental Figures 9 and 10 mirror Figure 3 in the primary text and Supplemental Figure 4, summarizing forecast skill after removing scores affected by reporting anomalies. Although the forecasts affected by reporting anomalies generally had higher WIS values than other forecasts, they did not have unusually large *differences* in WIS between forecasting methods. The results about the relative performance of ensemble methods hold stable whether or not the forecasts affected by data anomalies are removed.



Supplemental Figure 8: An illustration of data anomalies identified for weekly deaths in the state of Texas. Identified outliers include suppressed reporting during holidays and a winter storm in February 2021, as well as a period of reduced reporting followed by catch-up reporting in October 2021.



Supplemental Figure 9: Summaries of forecast performance after removing forecasts affected by reporting anomalies. Panel (a) shows the 25th percentile, median, and 75th percentile of differences in mean WIS between specified ensemble methods and the equally weighted median ensemble, where the means average across locations for each combination of forecast date and forecast horizon. Crosses show the difference in overall mean WIS averaging across all locations, forecast dates, and forecast horizons. Panel (b) shows the calibration of predictive quantiles, with the difference between the empirical coverage rate and the nominal coverage rate on the vertical axis. A well calibrated model will have a difference between the empirical coverage rate and the nominal quantile level that is approximately zero. A method that generates conservative two-sided intervals would have a difference that is negative for nominal quantile levels less than 0.5 and positive for nominal quantile levels greater than 0.5.



Supplemental Figure 10: Summaries of forecast performance after removing forecasts affected by reporting anomalies. Boxplots summarize the distribution of differences in mean WIS between specified ensemble methods and the equally weighted median ensemble, where the means average across locations for each combination of forecast date and forecast horizon. Crosses show the overall difference in mean WIS averaging across all locations, forecast dates, and forecast horizons.

4 Model variations considered during development phase

Here we present some results for model variations considered during the model development phase. All figures here represent only scores for forecast dates before May 3, 2021.

4.1 Additional combination methods and training window sizes

The manuscript gives results for equally weighted mean and median ensembles, and relative WIS weighted mean and median ensembles, using a fixed training set window size of the 12 weeks prior to the forecast date. Here we show results on the development set for forecasts using a range of training set window sizes including 4 weeks, 8 weeks, 12 weeks, and all available forecast history. We also consider two additional combination methods along with those presented in the main text.

The first of these new combination methods is a weighted mean. As a reminder, $q^m_{l,s,t,k}$ denotes the predictive quantile at probability level k from component model m at location l, forecast date s, and target end date t. With this notation, the weighted mean ensemble forecast quantiles are calculated as

$$q^{\text{ens}}_{l,s,t,k} = \sum_{m=1}^{M} w^m_s q^m_{l,s,t,k}.$$

The model weights w^m_s are constrained to be non-negative and sum to one; in case of missing forecasts, the weights for any missing models are set to zero and the remaining weights are rescaled to sum to 1. As indicated by the subscript s, the weights w^m_s are updated each week by optimizing the ensemble WIS over the training window of the specified number of weeks before the forecast date s.

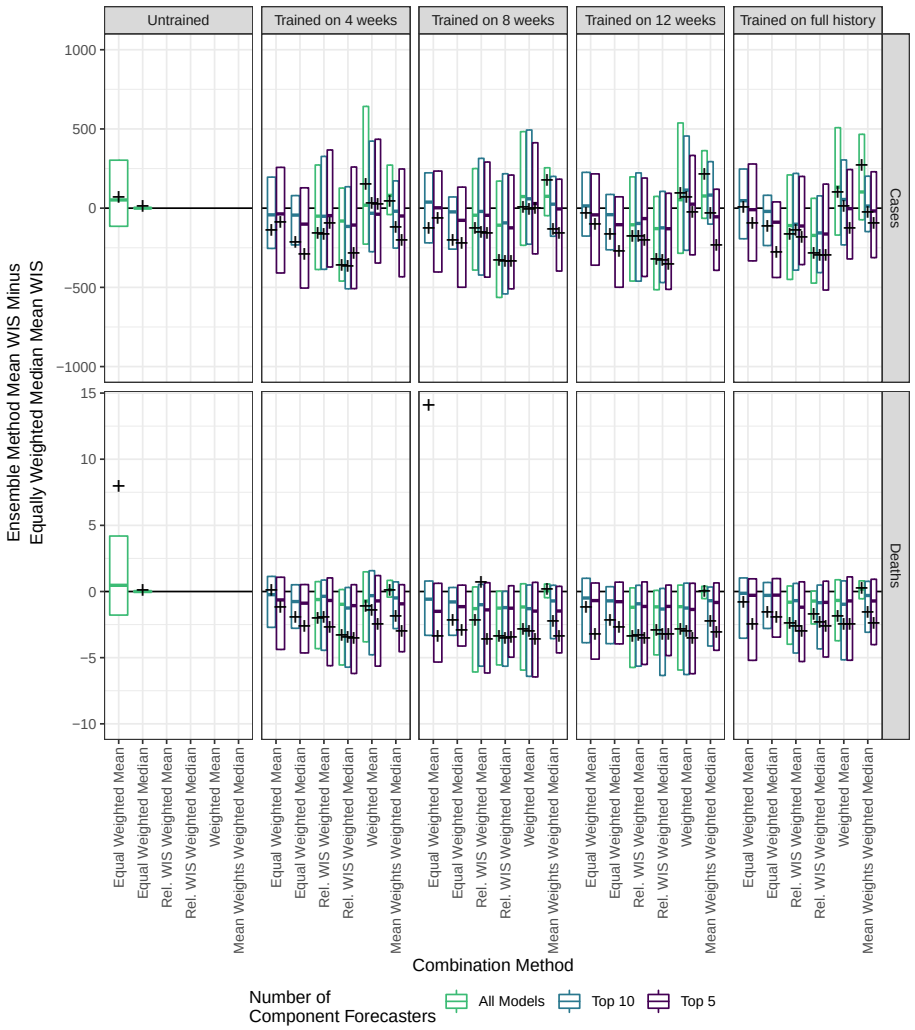
The second of the new combination methods is a weighted median ensemble that uses the weights estimated for the weighted mean ensemble. This offers comparable flexibility to the weighted mean ensemble, but has the disadvantage that the weights are not obtained by optimizing the forecast skill of the method that is actually used for forecast combination. Direct estimation of the weights for a weighted median by optimizing ensemble WIS is challenging because the objective function is not differentiable in the weights; the optimization problem is a mixed integer linear program, which is computationally demanding. In other experiments, we also considered a method for computing an approximate weighted median by the smoothing weighted distribution of predictive quantiles from component forecasters. However, this method's performance was not substantively different from the other methods considered here and we omit those results for brevity.

Supplemental Figures 11 and 12 display the results of this expanded comparison including all combinations of the training set window size, the six combina-

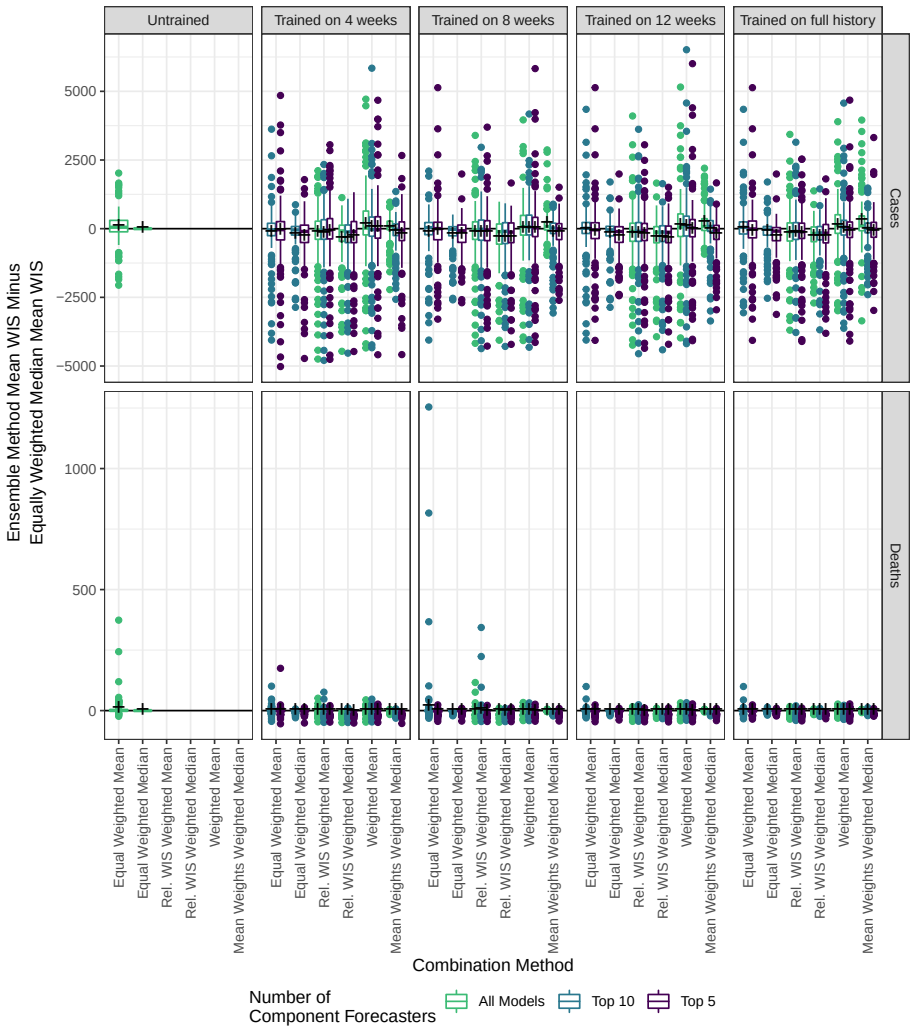
tion methods, and three variations on the number of top-performing component forecasters included in the ensemble. Across all combinations of training set window size, number of component forecasters included, and target variable (cases or deaths), the relative WIS weighted median ensemble had the most stable performance. For deaths, it had the best mean WIS for all training set window sizes, though it had similar performance to the equally weighted median of the top 5 models. For cases, it was more often matched by other methods, though the performance of the other methods was more inconsistent across different settings for other tuning parameters. Across most settings, using a relative WIS weighted mean or median offered an improvement in mean WIS over taking an equally weighted mean or median of top performing models.

There is perhaps a slight indication that an intermediate training set size of 8 to 12 weeks is better than training on 4 weeks or the full available history, but this signal is not strong. Some of the combination methods were better when fewer top models were included, but the relative WIS weighted median method was not sensitive to this setting.

We selected the relative WIS weighted mean and median ensembles for the prospective evaluation because they were consistently better than both more flexibly weighted methods and equally weighted combinations of top-performing components when forecasting cases, and were comparable to the best of the other approaches when forecasting deaths. We selected an intermediate training set window size of 12 weeks because both the relative WIS weighted mean and median methods did well with that training set size. We selected including the top ten component forecasters as an intermediate setting for that tuning parameter, though we did not see a strong reason to prefer it to the other possibilities we considered.



Supplemental Figure 11: The 25th percentile, median, and 75th percentile of differences in mean WIS between specified ensemble methods and the equally weighted median of all component forcasters, where the means average across locations for each combination of forecast date and forecast horizon. Crosses show the overall difference in mean WIS averaging across all locations, forecast dates, and forecast horizons. A negative value indicates that the method corresponding to a particular combination of training set size, number of component forecasters included, and combination method outperformed the ensemble calculated as an equally weighted median of all component forecasts.



Supplemental Figure 12: Boxplots summarizing the full distribution of differences in mean WIS between specified ensemble methods and the equally weighted median of all component forcasters, where the means average across locations for each combination of forecast date and forecast horizon. Crosses show the overall difference in mean WIS averaging across all locations, forecast dates, and forecast horizons. A negative value indicates that the method corresponding to a particular combination of training set size, number of component forecasters included, and combination method outperformed the ensemble calculated as an equally weighted median of all component forecasts.

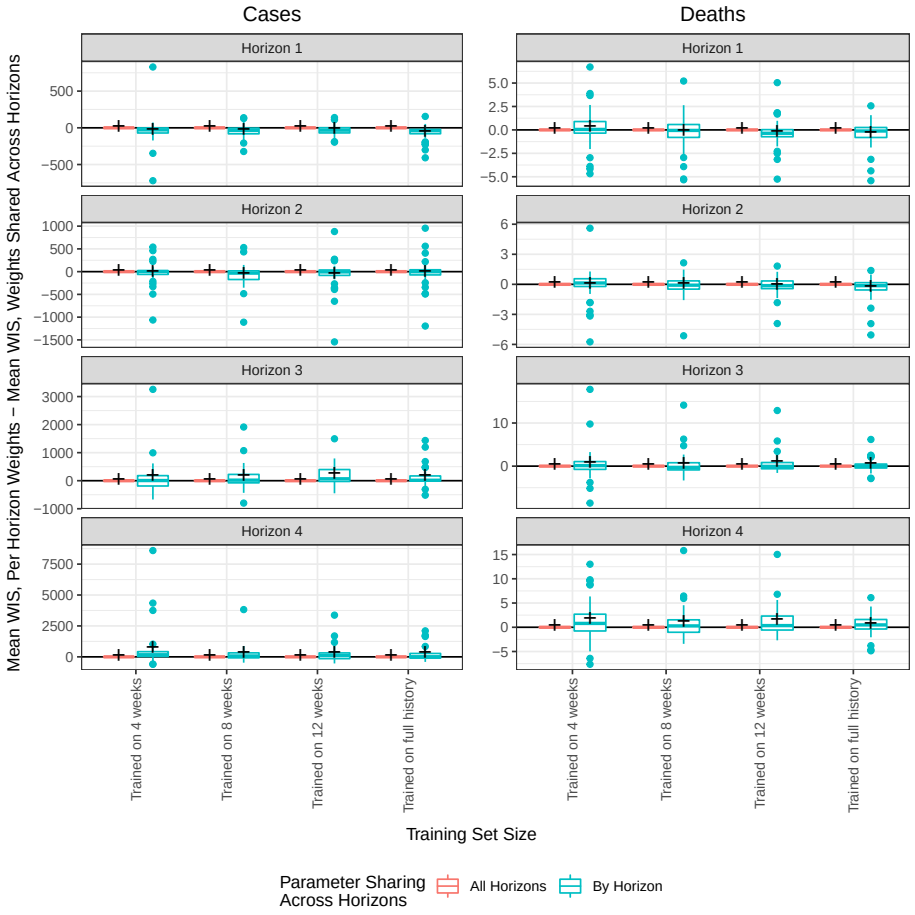
4.2 Separate weights at different forecast horizons

We considered a variation on the relative WIS weighted median ensemble that estimated separate weights for each of the one through four week ahead forecast horizons. In this approach, the relative WIS of component forecasters was calculated separately at each forecast horizon, and the estimation of the weighting parameter θ was performed separately to optimize the forecast skill of the ensemble at each horizon.

Supplemental Figure 13 shows the results of this per-horizon weighting scheme for the relative WIS weighted median ensemble combining the top 10 component forecasters. We found that using separate model weights at each forecast horizon led to small improvements in mean WIS at short-term forecast horizons of one to two weeks ahead, but slightly worse mean WIS at longer forecast horizons of three to four weeks ahead.

We see two possible contributing factors to these results. First, forecasts at long horizons have scores that are larger in magnitude than forecasts at short horizons, and so tend to dominate the overall score when averaging across horizons. This may result in weights that favor performance at longer term horizons when weights are shared across horizons, thereby harming performance of forecasts at short horizons. Second, sharing weights across horizons may be particularly helpful for longer term forecasts because of the gain in the training set sample size that comes with weight sharing. For example, with a training set size of four weeks and weights estimated separately by horizon, only one week's worth of forecasts are actually included in the training set for weight estimation at a horizon of four weeks, because the target data for four-week ahead forecasts made within the past three weeks have not yet been observed at the time of weight estimation. Sharing weights across horizons means that more information about model performance is available for weight estimation at these longer horizons. In support of this explanation, note that the magnitude of relative losses in forecast skill from estimating per-horizon weights at longer forecast horizons decreases as the training set size increases.

These results suggest the possibility of a blended strategy, where per-horizon estimation is used to obtain the weights for short horizons but the weights for longer horizons are estimated by sharing information across all horizons. In light of the small magnitude of the gains from using a per-horizon weighting at short horizons, we decided to pursue a unified approach of using shared weights across all horizons to reduce methodological and narrative complexity.



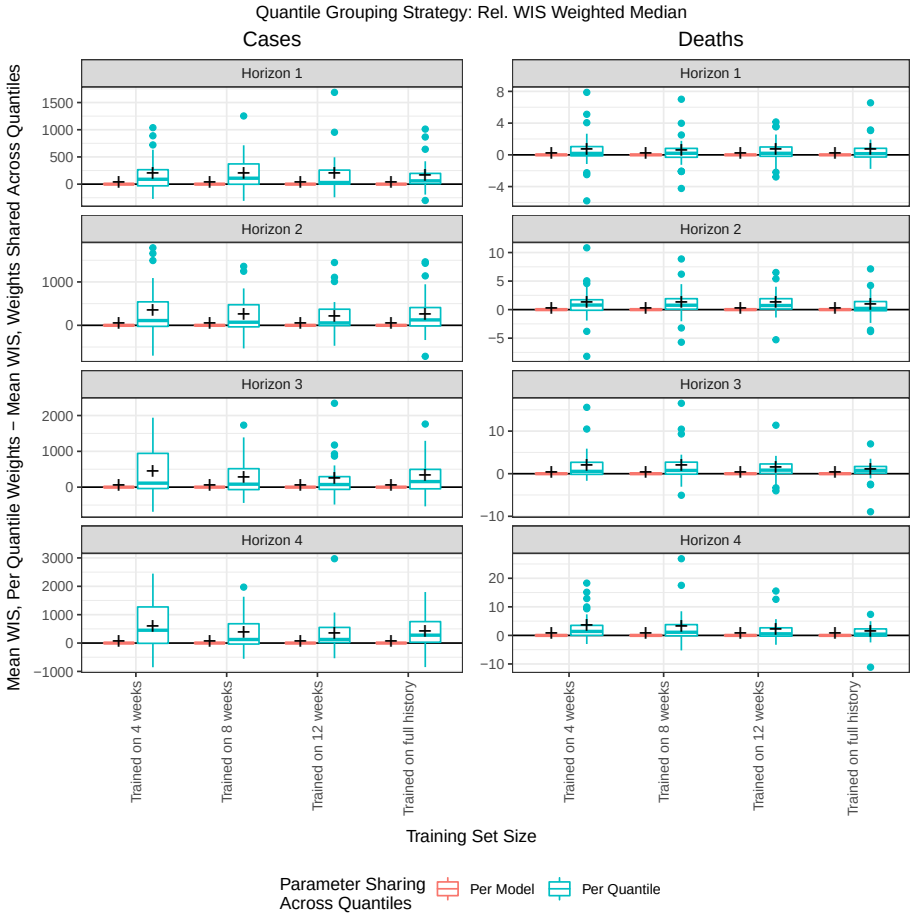
Supplemental Figure 13: Boxplots summarizing forecast skill for forecasts of weekly cases, varying whether model weights are shared across all forecast horizons or are estimated separately for each forecast horizon. The vertical axis is the difference in mean skill for the given ensemble specification when component weights are shared across all horizons and the same specification with separate component weights for each forecast horizon. The boxplots summarize the distribution of these differences for each combination of forecast date and horizon, averaging across all locations. A cross is displayed at the difference in overall mean scores. A negative value indicates that the method with separate component weights for each forecast horizon outperformed the corresponding specification with weights shared across forecast horizons. For this analysis, only results for relative WIS weighted ensembles combining the ten best individual component forecasters are presented.

4.3 Separate weights at different quantile levels

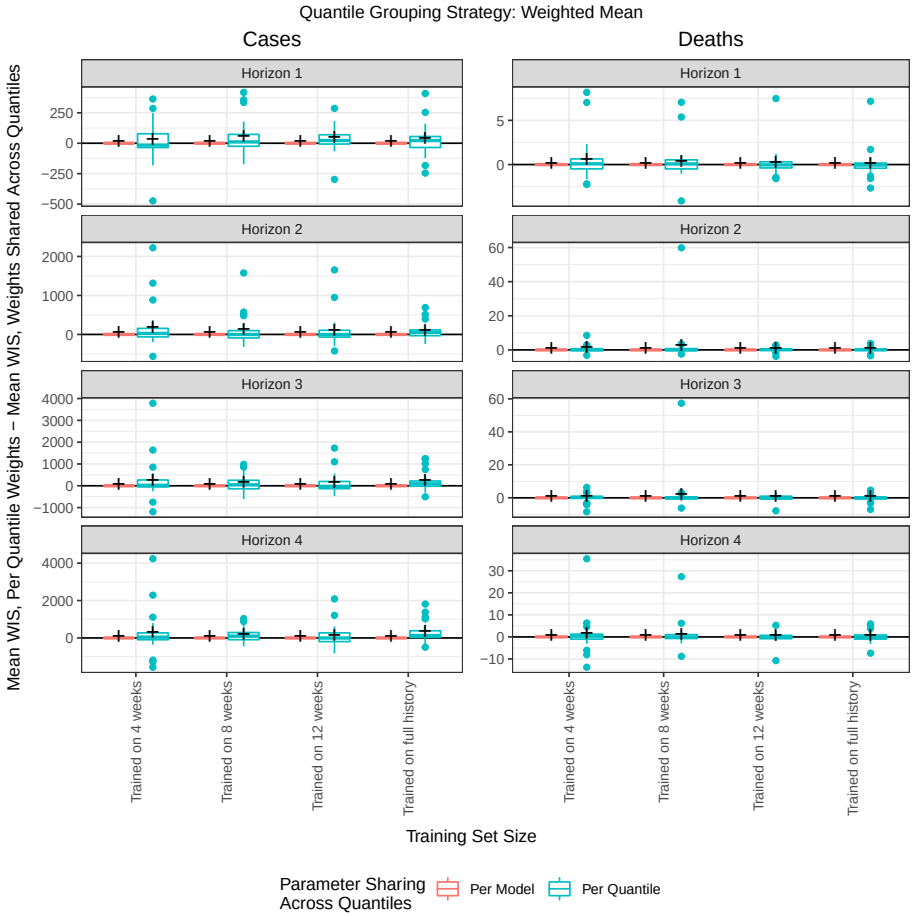
We considered strategies for estimation of separate model weights at each quantile level rather than sharing model weights across all quantile levels. We considered this possibility for both the relative WIS weighted median ensemble and the convex mean ensemble that directly optimizes the component weights rather than setting them to be a sigmoid function of the relative WIS. In both variations, weights were estimated by optimizing the contribution to the WIS from each quantile level separately, i.e., the pinball loss. Similarly, for the relative WIS weighted median, the relative WIS for component models that is used as an input for calculating model weights was obtained separately based on the contribution to WIS at each quantile level.

Supplemental Figures 14 and 15 summarize the WIS of these methods on the model development set, comparing these approaches to the corresponding methods with a single weight per model that is shared across all quantile levels. Supplemental Figure 16 displays the probabilistic calibration of these model variations in terms of one-sided coverage rates for predictive quantiles. In this experiment, all methods combine the top ten component forecasters; we consider varying training set window sizes.

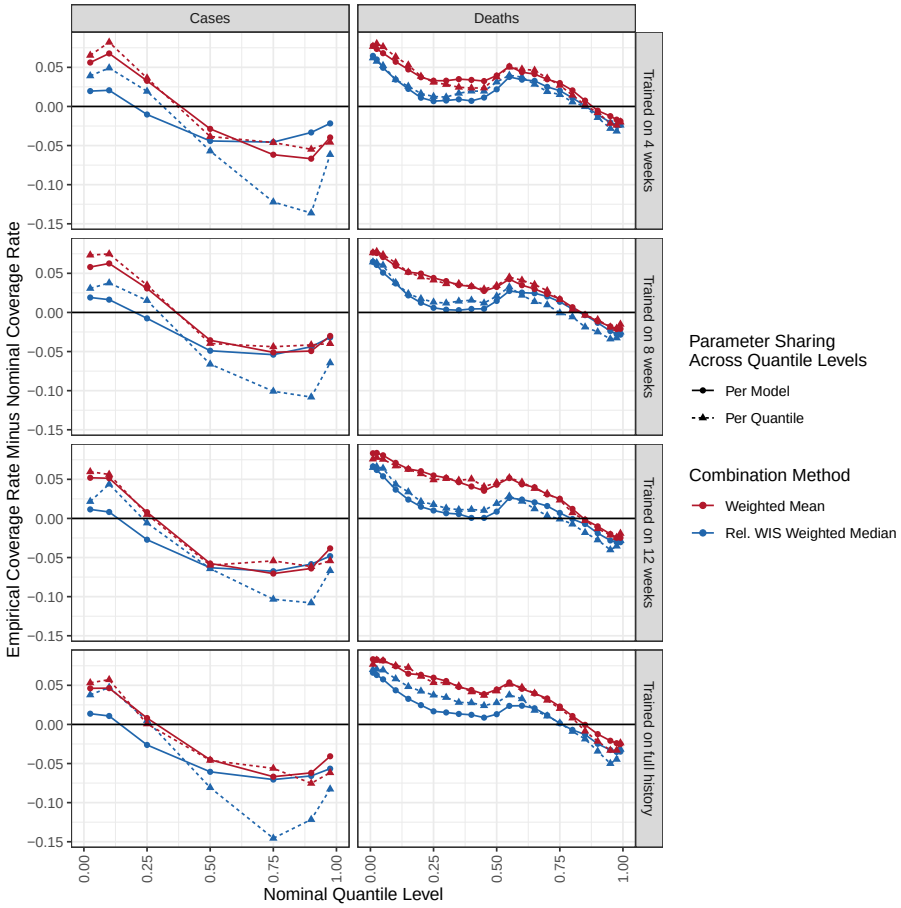
Allowing for separate parameters per quantile level led to worse mean WIS. Additionally, for both cases and deaths, the per-quantile weighting schemes led to generally narrower predictive distributions with worse calibration in the tails. This effect was much stronger for forecasts of cases than forecasts of deaths, and it was stronger for the relative WIS weighted median ensemble than the convex weighted mean ensemble.



Supplemental Figure 14: Boxplots summarizing forecast skill for forecasts of weekly cases from a relative WIS weighted median ensemble, varying whether model weights are shared across all quantile levels ("Per Model") or are estimated separately for each quantile level ("Per Quantile"). The vertical axis is the difference in mean skill for the given ensemble specification when component weights are shared across all quantile levels and the same specification with separate component weights for each quantile level. The boxplots summarize the distribution of these differences for each combination of forecast date and horizon, averaging across all locations. A cross is displayed at the difference in overall mean scores. A negative value indicates that the method with separate component weights for each quantile level outperformed the corresponding specification with weights shared across quantile levels.



Supplemental Figure 15: Boxplots summarizing forecast skill for forecasts of weekly cases from a convex weighted mean ensemble with directly estimated weights, varying whether model weights are shared across all quantile levels ("Per Model") or are estimated separately for each quantile level ("Per Quantile"). The vertical axis is the difference in mean skill for the given ensemble specification when component weights are shared across all quantile levels and the same specification with separate component weights for each quantile level. The boxplots summarize the distribution of these differences for each combination of forecast date and horizon, averaging across all locations. A cross is displayed at the difference in overall mean scores. A negative value indicates that the method with separate component weights for each quantile level outperformed the corresponding specification with weights shared across quantile levels.



Supplemental Figure 16: Quantile coverage rates for the convex weighted mean ensemble and the relative WIS weighted median ensemble, varying whether weights are estimated separately per quantile level ("Per Quantile") or shared across all quantile levels ("Per Model"). All methods combine the top 10 component forecasters in the training set window size specified in facet rows. The vertical axis is the difference between the empirical coverage rate and the nominal coverage rate. A well calibrated method would have a difference of 0 between the empirical and nominal coverage rates, and a method that generates conservative (wide) interval forecasts would have a negative difference for quantile levels less than 0.5 and a positive difference for quantile levels greater than 0.5.

5 Performance of trained methods near local peaks

We identified local peaks in state level weekly cases and deaths as weeks that had the maximum incidence within a centered rolling window of 11 weeks (i.e., weeks that had the largest reported weekly counts among the preceding five and following five weeks). By visual inspection, we made manual adjustments to the weeks identified using this rule to remove outliers and weeks that did not correspond to a visually distinct peak. The resulting weeks identified as local peaks are shown in Supplemental Figures 17 and 18. Within our evaluation time frame, there were 122 local peaks for weekly cases and 90 local peaks for weekly deaths.

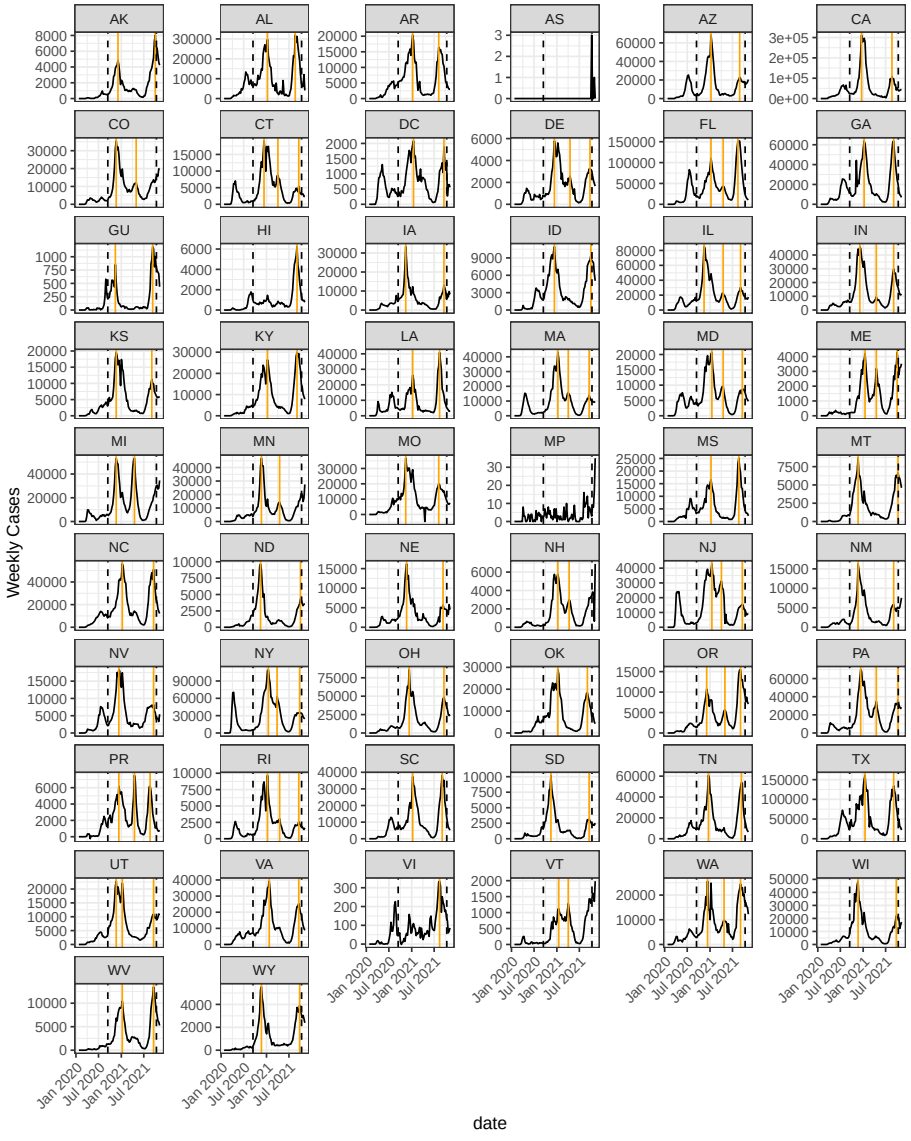
Supplemental Figure 19 summarizes the central tendency of the errors for predictive medians across all forecasts at the state level in the U.S. and for just those forecasts issued in the week before a local peak. The errors are calculated as the value of the predictive median minus the observed count of cases or deaths in the target week, so that errors close to zero are preferred. We show results for three ensemble specifications: the equally weighted median of all component forecasters, the equally weighted median of the top 10 forecasters, and the relative WIS weighted median of the top 10 forecasters. Note that the equally weighted median of all components is untrained, the relative WIS weighted median of the top 10 forecasters is trained, and the equally weighted median of the top 10 forecasters represents an intermediate strategy: it is also trained, but it is not as strongly adaptive to component performance as the weighted median. Both trained ensembles use a training set window size of 12 weeks. We summarize our observations about these errors as follows:

1. Across all forecasts of cases, the median error is similar for all three strategies, but the magnitude of the average error from the relative WIS weighted median is slightly larger than the magnitude of the average error from the equally weighted median of all components.

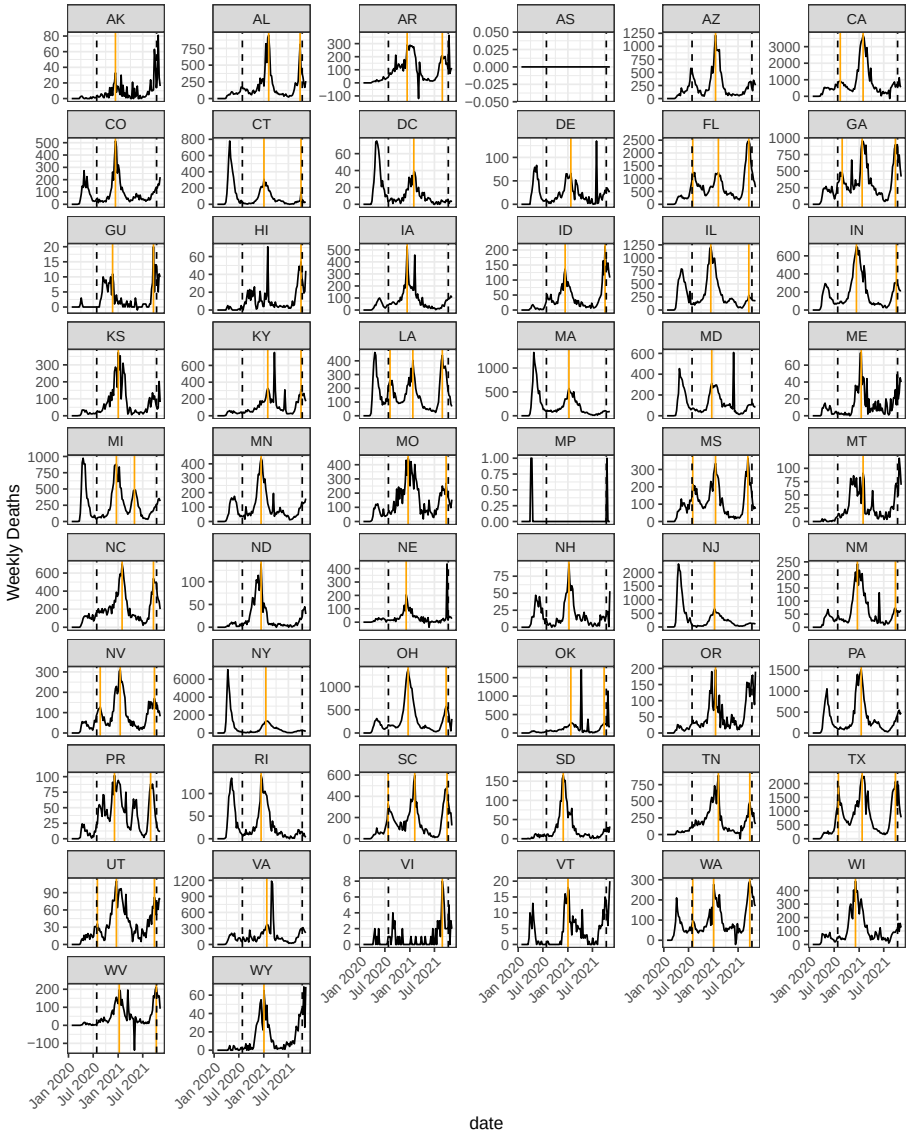
2. Across all forecasts of deaths, the median error is similar for all three strategies, but the magnitude of the average errors from the relative WIS weighted median is slightly smaller than the magnitude of the average error from the equally weighted median of all components.

3. For forecasts of cases near peaks, the trained methods were better than the untrained approach in the peak week, but have much larger errors at longer horizons. This indicates that forecasts from the trained methods "overshot" and missed the turning points by a larger margin than the untrained method.

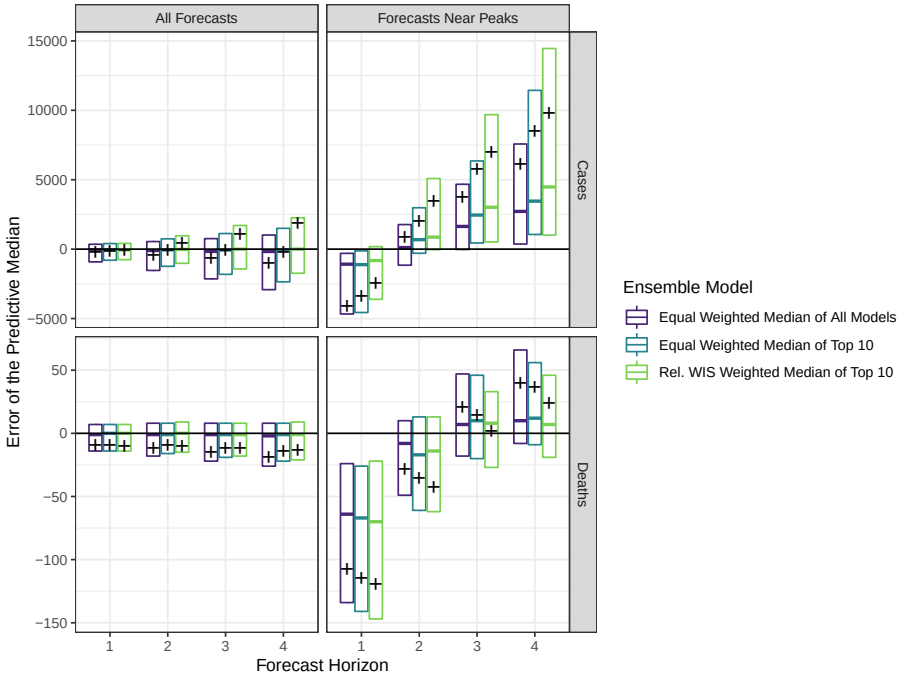
4. For forecasts of deaths near peaks, the trained methods are clearly better than the untrained method at longer forecast horizons, indicating that they did a better job of identifying the turnaround coming.



Supplemental Figure 17: Identified local peaks for weekly cases at the state level. Vertical dashed lines indicate the boundaries of the evaluation phase (the combined model development and prospective evaluation phases). Vertical orange lines indicate the locations of identified local peaks.

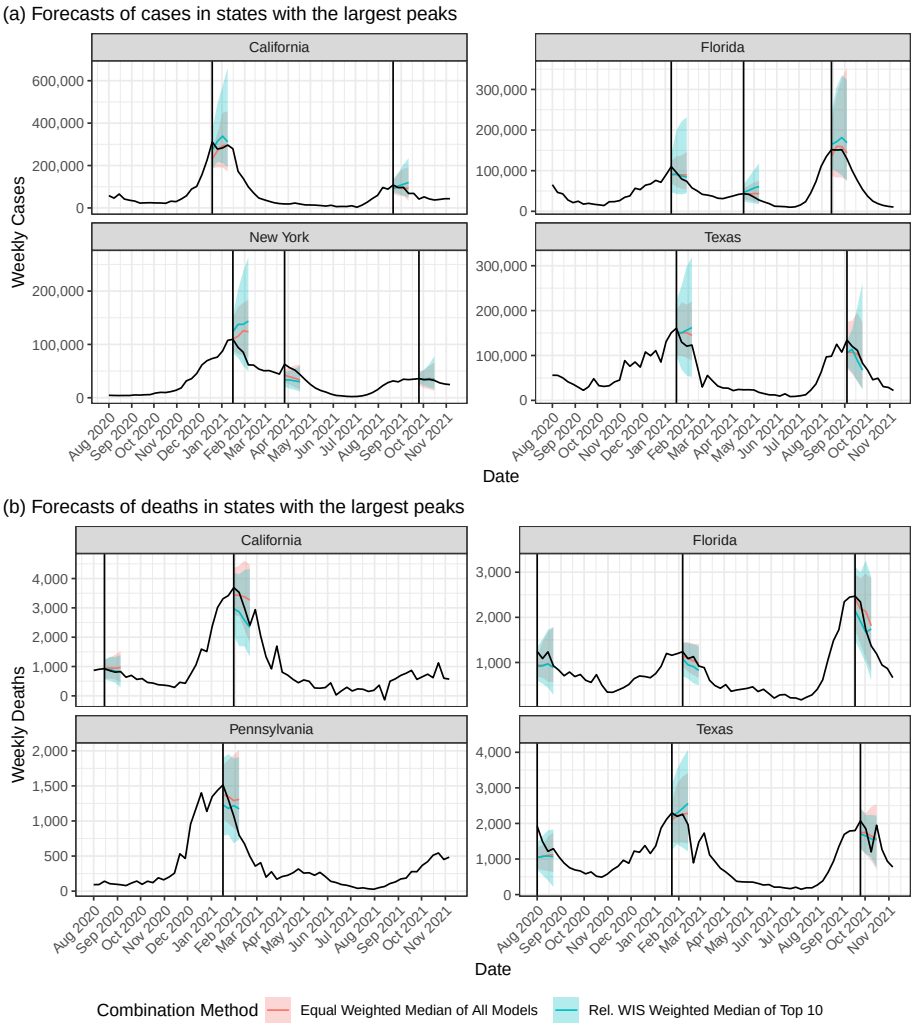


Supplemental Figure 18: Identified local peaks for weekly deaths at the state level. Vertical dashed lines indicate the boundaries of the evaluation phase (the combined model development and prospective evaluation phases). Vertical orange lines indicate the locations of identified local peaks.



Supplemental Figure 19: Errors of the predictive median across all forecasts, and for forecasts issued the week before a local peak in weekly cases or deaths. For forecasts issued before a peak, the one week ahead forecasts are for cases in the week of the local peak and forecasts at longer horizons are for cases in weeks after the local peak. The vertical axis is the difference between the predictive median and the observed value. A positive value indicates that the predictive median was larger than the eventually observed value, and a negative value indicates that the predictive median was less than the eventually observed value; a difference of zero is best. Boxes summarize the 25th percentile, median, and 75th percentile of these errors, and crosses show the mean error.

Supplemental Figure 20 shows illustrative examples of forecasts made the week before a local peak from the equally weighted median of all components and the relative WIS weighted median of the top 10 components. We can see a systematic tendency for the predictions from the trained method near local peaks to predict a continuation of rising trends for cases, whereas the forecasts of deaths more often capture the coming downturn.

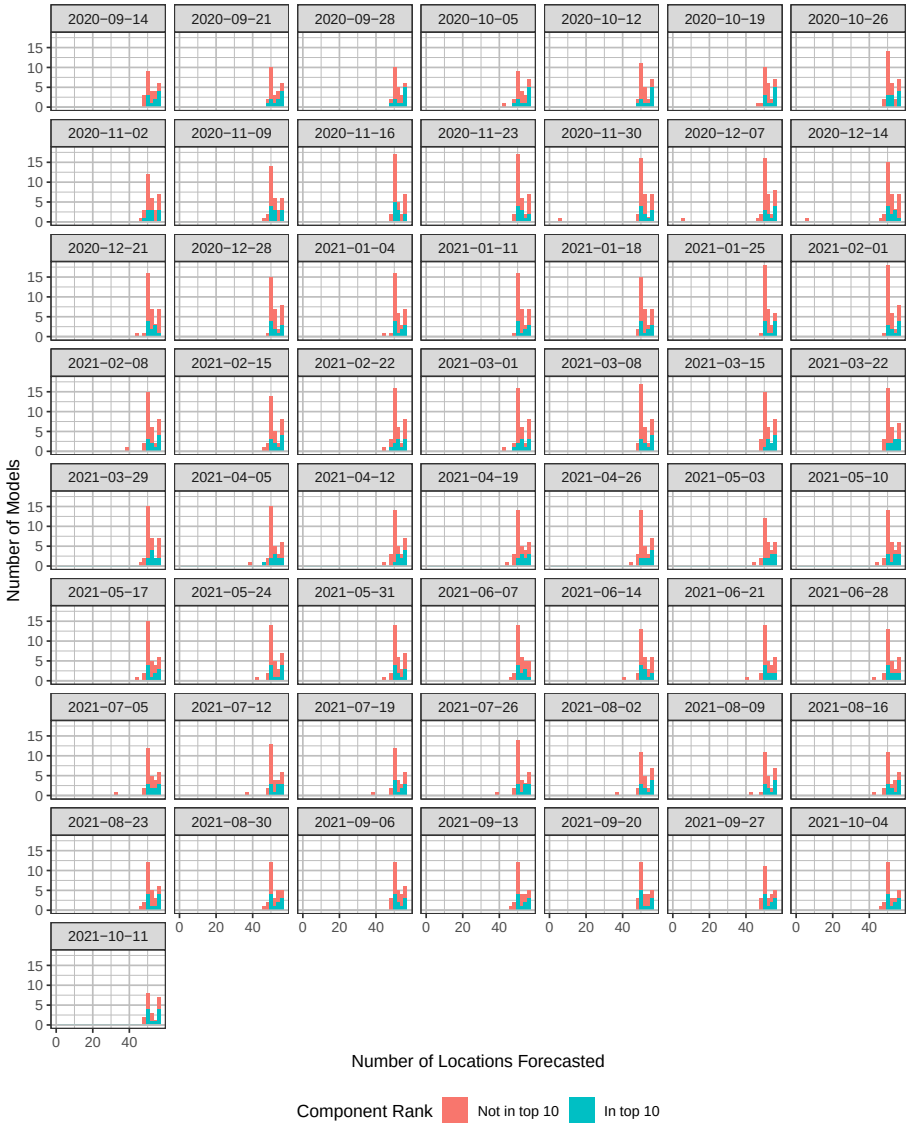


Supplemental Figure 20: Forecast distributions for weekly cases and deaths issued the week before a local peak. Forecasts are shown for all peaks during the evaluation period in the four states with the largest local peaks. Vertical lines indicate the time of the peak, corresponding to the date of a one-week-ahead forecast.

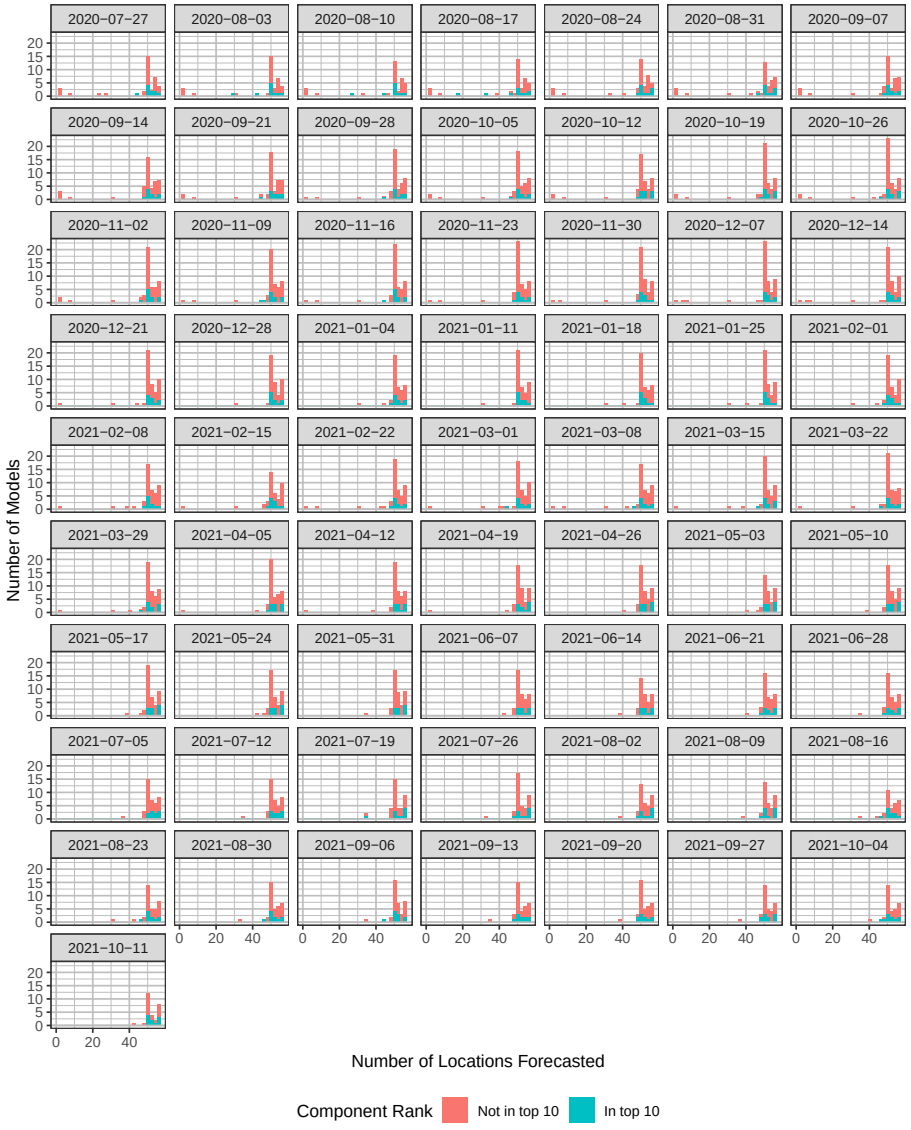
6 Differences in forecast missingness in the U.S. and Europe

Supplemental Figures 21 through 24 show histograms of the number of locations forecasted by the component models contributing to the U.S. COVID-19 Forecast Hub and the European COVID-19 Forecast Hub. Patterns of submission are starkly different for the U.S. and the EU. In the U.S., nearly all models submit forecasts for at least the 50 US States, and many additionally submit forecasts for the District of Columbia and US territories. In the EU, roughly half of forecasters provide forecasts for all or most European countries, while the other half provide forecasts for only a few countries.

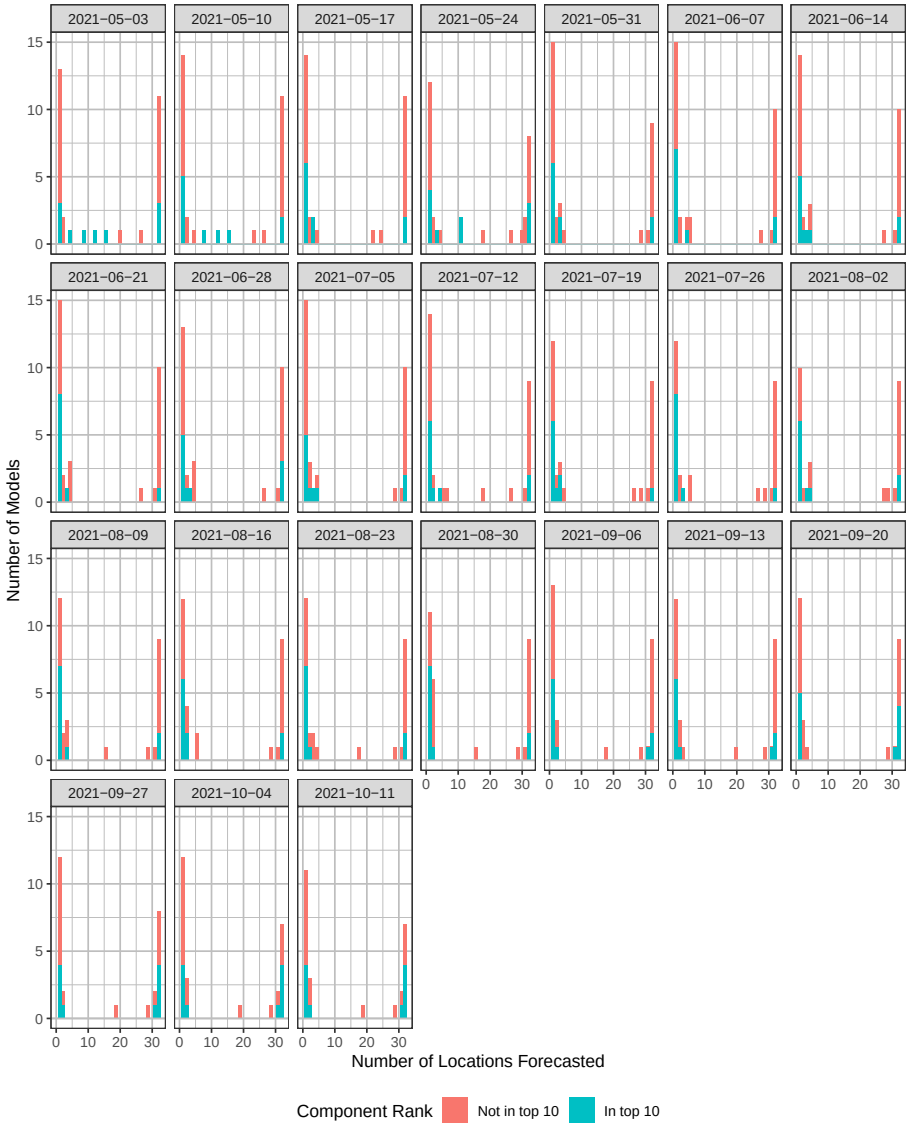
Supplemental Figures 25 through 28 show the effective weights used in each location after accounting for forecast missingness by rescaling the weights assigned to available models so that they sum to one. In the U.S., the effective weights closely match the nominal estimated weights in nearly all states, differing only slightly in the territories. In the EU, missingness is more prevalent and it is common for only a few of the selected top 10 component forecasters to provide forecasts for many countries.



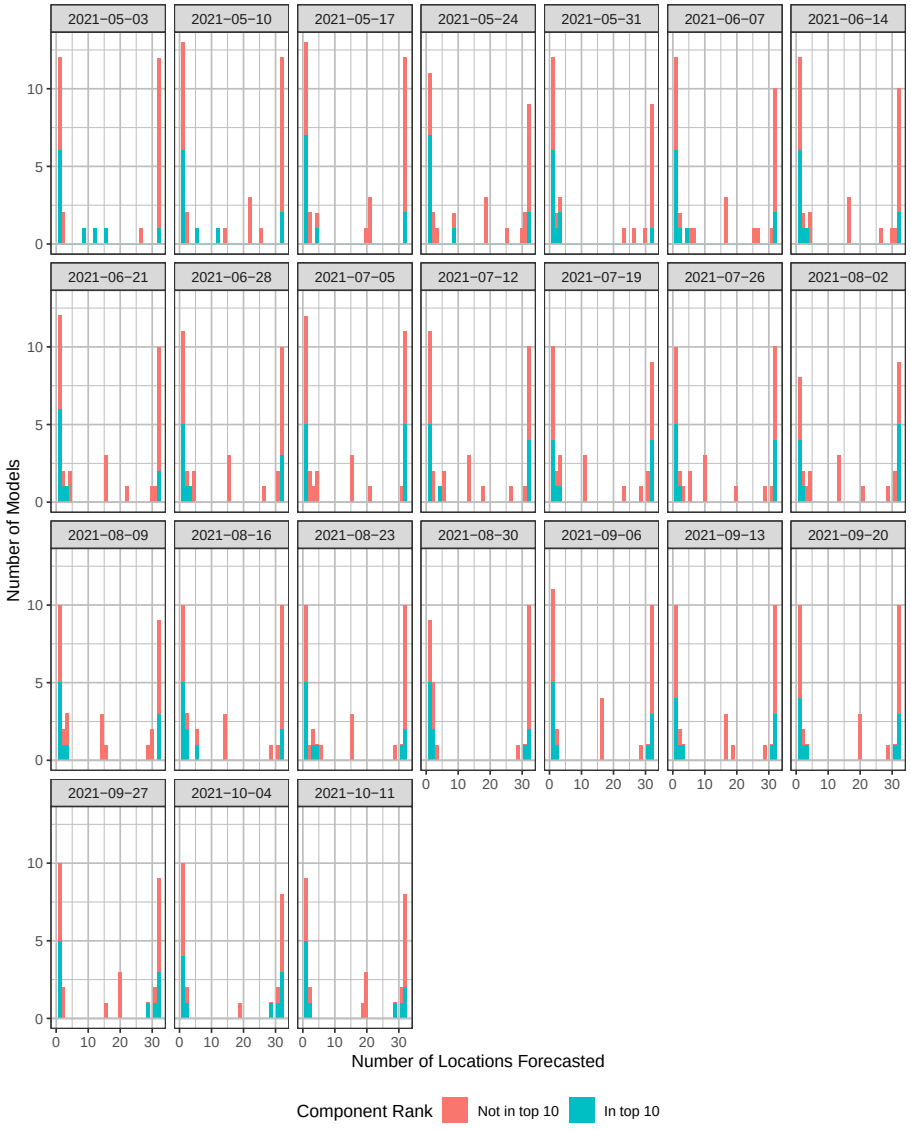
Supplemental Figure 21: Histograms of the number of locations forecasted by each contributing forecaster for weekly cases in the U.S. The top 10 forecasters, indicated with blue shading, were selected for inclusion in the weighted ensembles used for prospective evaluation



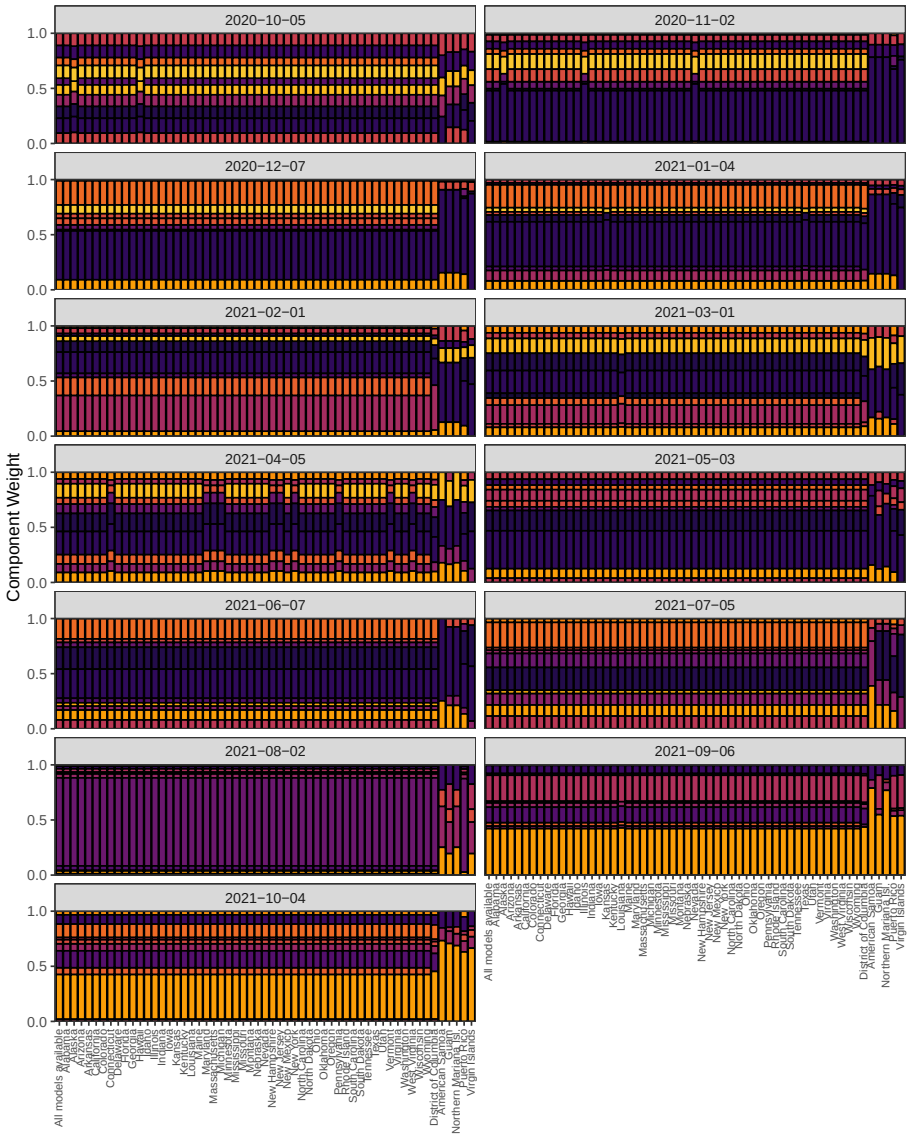
Supplemental Figure 22: Histograms of the number of locations forecasted by each contributing forecaster for weekly deaths in the US. The top 10 forecasters, indicated with blue shading, were selected for inclusion in the weighted ensembles used for prospective evaluation



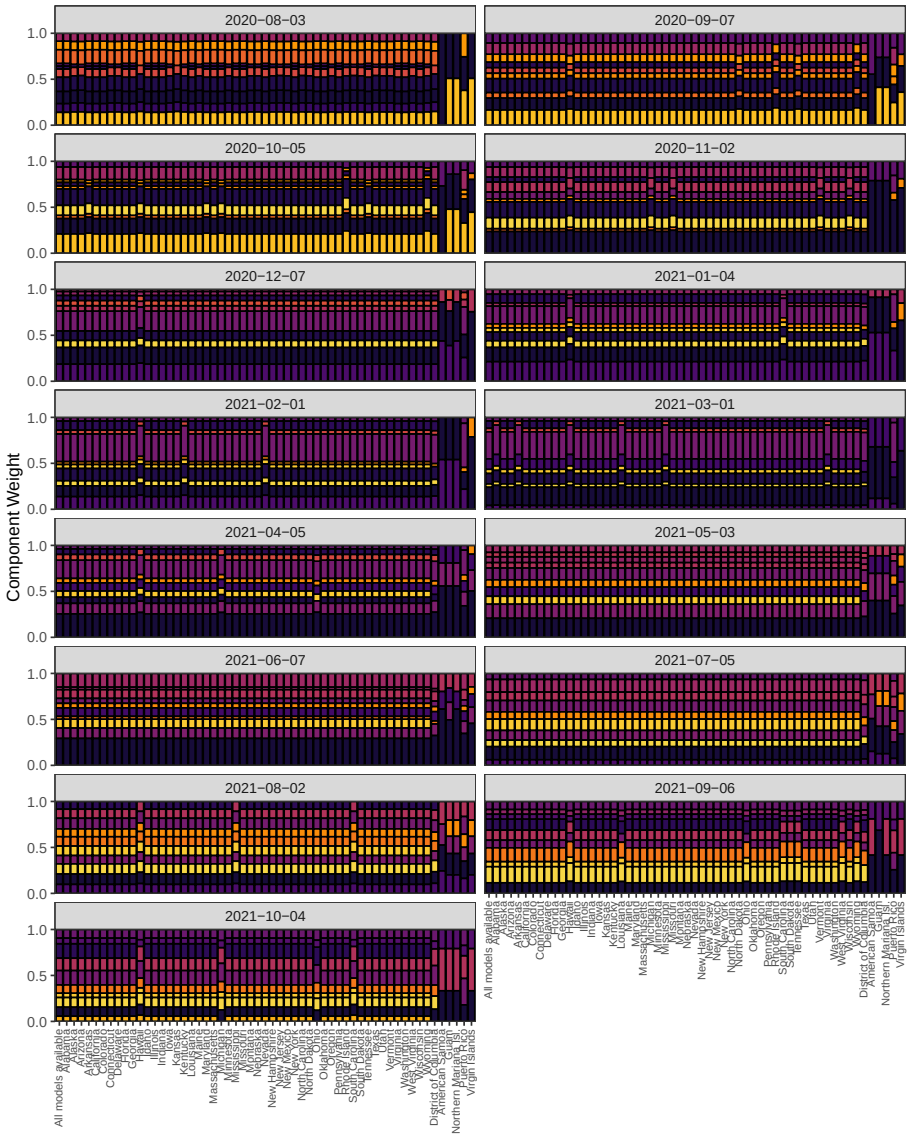
Supplemental Figure 23: Histograms of the number of locations forecasted by each contributing forecaster for weekly cases in Europe. The top 10 forecasters, indicated with blue shading, were selected for inclusion in the weighted ensembles used for prospective evaluation



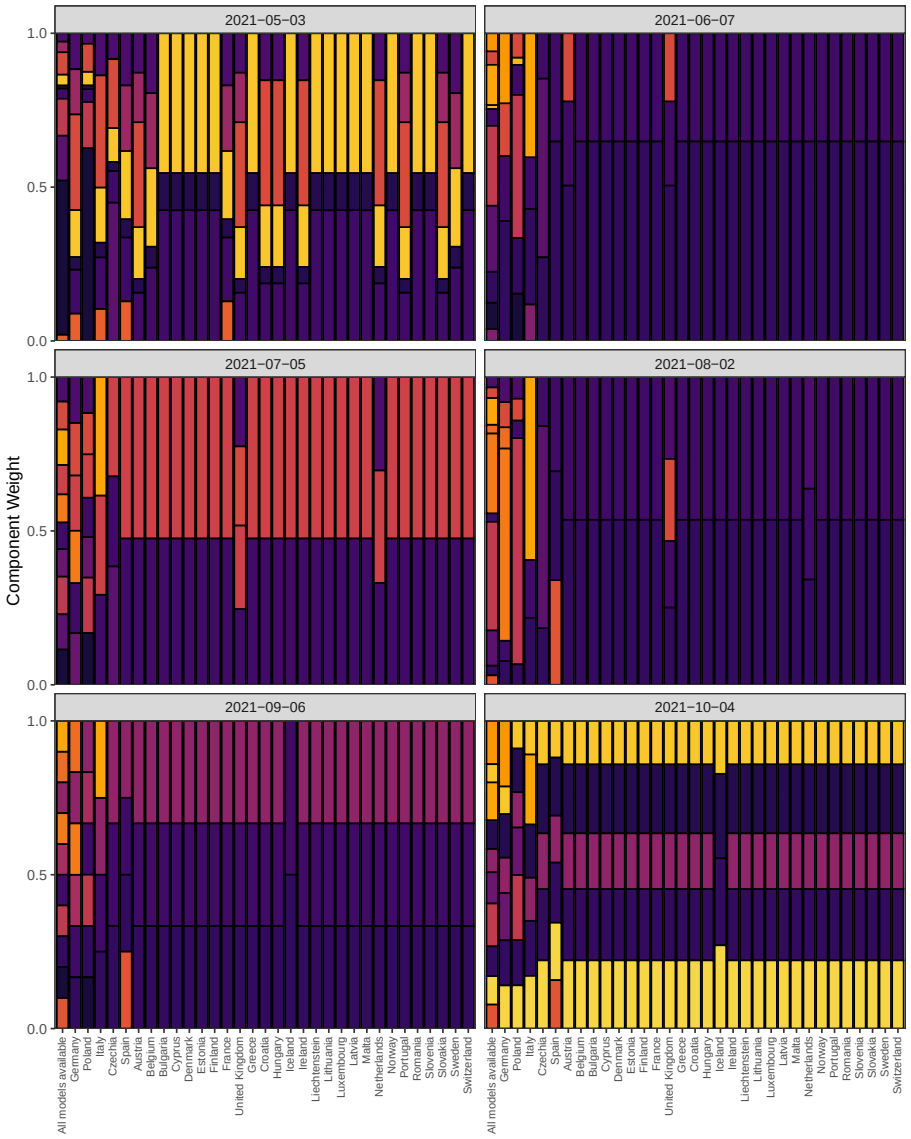
Supplemental Figure 24: Histograms of the number of locations forecasted by each contributing forecaster for weekly deaths in Europe. The top 10 forecasters, indicated with blue shading, were selected for inclusion in the weighted ensembles used for prospective evaluation



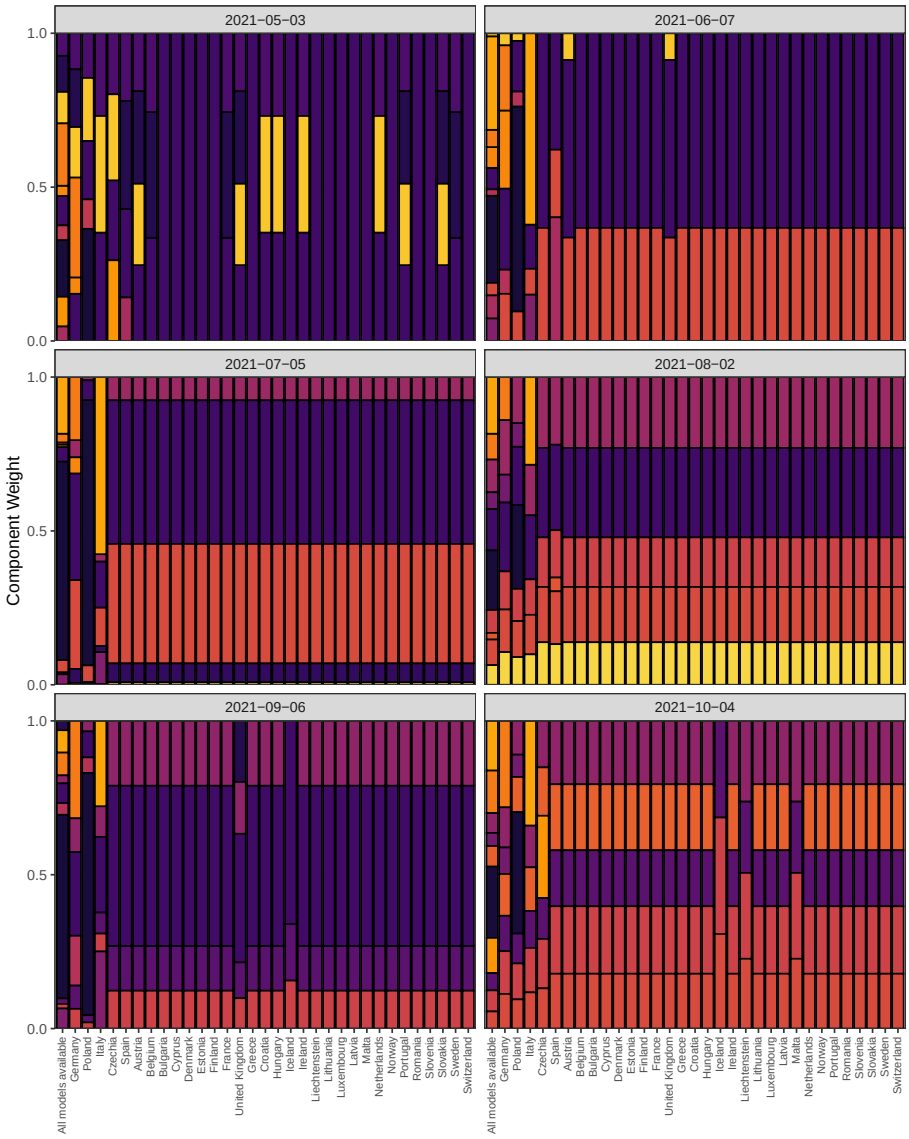
Supplemental Figure 25: Component weights for forecasts of weekly cases in the U.S., facetted by forecast date. Only the weights for the first week in each month are shown due to space constraints. The estimated weights that would be used if all models were available for a particular location are shown at left within each facet. The weights actually used for each location are obtained by setting the weight for components that are missing forecasts for that location to 0 and rescaling the others proportionally so that they sum to 1.



Supplemental Figure 26: Component weights for forecasts of weekly deaths in the U.S., facetted by forecast date. Only the weights for the first week in each month are shown due to space constraints. The estimated weights that would be used if all models were available for a particular location are shown at left within each facet. The weights actually used for each location are obtained by setting the weight for components that are missing forecasts for that location to 0 and rescaling the others proportionally so that they sum to 1.



Supplemental Figure 27: Component weights for forecasts of weekly cases in Europe, facetted by forecast date. Only the weights for the first week in each month are shown due to space constraints. The estimated weights that would be used if all models were available for a particular location are shown at left within each facet. The weights actually used for each location are obtained by setting the weight for components that are missing forecasts for that location to 0 and rescaling the others proportionally so that they sum to 1.



Supplemental Figure 28: Component weights for forecasts of weekly deaths in Europe, facetted by forecast date. Only the weights for the first week in each month are shown due to space constraints. The estimated weights that would be used if all models were available for a particular location are shown at left within each facet. The weights actually used for each location are obtained by setting the weight for components that are missing forecasts for that location to 0 and rescaling the others proportionally so that they sum to 1.

7 Adherence to EPIFORGE Guidelines

Section	Item	Item Description	Pages
Title/Abstract	1	Study described as a forecast or prediction research in at least the title or abstract	1
Introduction	2	Purpose of study and forecasting targets defined	4
Methods	3	Methods fully documented	7-13
Methods	4	Identify whether the forecast was performed prospectively, in real-time, and/or retrospectively	8
Methods	5	Origin of input source data explicitly described with reference	7
Methods	6	Source data made available, or reasons why this was not possible documented	13
Methods	7	Input data processing procedures described in detail	8
Methods	8	Statement and description of model type, with model assumptions documented with references	11-13
Methods	9	Model code made available, or reasons why this was not possible documented	14
Methods	10	Description of model validation, with justification of approach.	7-11
Methods	11	Description of forecast accuracy evaluation method, with justification	10-11
Methods	12	Where possible, compare model results to a benchmark or other comparator model, with justification of comparator choice	10-11
Methods	13	Description of forecast horizon, and justification of its length	7
Results	14	Uncertainty of forecasting results presented and explained	16, 22
Results	15	Results briefly summarized in lay terms, including a lay interpretation of forecast uncertainty	-

Section	Item	Item Description	Pages
Results	16	If results are published as a data object, encourage a time-stamped version number	14
Discussion	17	Limitations of forecast described, including limitations specific to data quality and methods	21-24
Discussion	18	If the research is applicable to a specific epidemic, comment on its potential implications and impact for public health action and decision making	26
Discussion	19	If the research is applicable to a specific epidemic, comment on how generalizable it may be across populations	24