Ensemble forecasts of COVID-19 cases and deaths in the United States

Evan L. Ray[a,1,*], Logan C. Brooks[b], Yijin Wang[a], Aaron Gerding[a], Estee Cramer[a], Martha Zorn[a], Jacob Bien[c], Johannes Bracher[d,e], Aaron Rumack[b], Matthew Biggerstaff[f], Michael A. Johansson[f], Ryan J. Tibshirani[b], Nicholas G. Reich[a]

[a] *School of Public Health and Health Sciences, University of Massachusetts Amherst*
[b] *Machine Learning Department, Carnegie Mellon University*
[c] *Department of Data Sciences and Operations, University of Southern California*
[d] *Chair of Econometrics and Statistics, Karlsruhe Institute of Technology*
[e] *Computational Statistics Group, Heidelberg Institute for Theoretical Studies*
[f] *COVID-19 Response, U.S. Centers for Disease Control and Prevention*

Abstract

Abstract goes here.

Keywords: COVID-19, forecast, ensemble, keyword 4

2010 MSC: 00-01, 99-00

1. Introduction

- General overview
  - Forecasts of short-term trajectory can help inform public health response.
  - Ensemble forecasts are generally performant
  - We consider how to effectively construct ensemble forecasts of the short-term trajectory of COVID-19 to support public health end users.
- Review literature on ensemble forecasting in general and for infectious disease in particular

*Corresponding author
Email address: elray@umass.edu (Evan L. Ray)

- Review literature on ensemble methodology
  - quantile averaging and density averaging
  - exponentially weighted averaging (weights are sigmoid of a measure of forecast skill)
- Summary of our motivations and contributions
  - Goal is to produce ensemble forecasts of short-term trajectory of COVID-19 that have good average performance and stable performance across time and locations.
  - Real-time forecasting introduces challenges that we need to be able to handle; most importantly,
    * Outlying component forecasts due to software errors or difficulties handling data reporting anomalies
    * Missing component forecasts from teams that join at different times or submit forecasts for a subset of locations
    * mention other things?
  - We explore and compare variations on ensemble methods designed to address these challenges by using combination mechanisms that are robust to outliers and may allow for giving more weight to better component forecasters.

2. Context and Analysis Set Up: forecasting COVID-19 burden in the United States

- Describe hub and parameters of our analysis
  - Dates active, geographic scales, forecast targets, representation in terms of quantiles
  - In this analysis, we focus on state level forecasts of incident deaths between July 27, 2020 and May 31, 2021 and forecasts of incident cases between ... and May 31, 2021.

- Exhibit features of component forecasts that motivate our methods
  - Occasional outlying forecasts motivate robust methods
  - Some component forecasters are consistently better than others; motivates trained methods
- Figure with data and component/ensemble forecasts?

3. Methods

- Approaches to evaluation
  - WIS/pinball loss
  - relative WIS to handle missingness
- Two-by-two table with methods we primarily consider:
  - equal-weighted mean
  - equal-weighted median
  - weighted mean
  - weighted median, based on relative WIS
- Details about trained methods that we will consider in the primary analysis
  - Training set window size: 4, 8, 12, or "full history" weeks
  - Number of component forecasters included: all eligible, top 10, or top 5
  - Handle missing forecasts by setting their weight to 0 and "renormalizing"
- Other details for which careful exploration is deferred to the supplement
  - Introduce extra parameters for each forecast horizon? (or, turn this around and use per-horizon weights in the main analysis and simplified version in supplement?)

- Introduce extra parameters for each quantile level (or for 3 groups of quantile levels)?

4. Results

- Figures 2 and 3 display summaries of overall performance across all locations, forecast dates, and forecast horizons. The main take aways are:
  - Robust methods are helpful:
    * Comparison of equally-weighted approaches: For both incident cases and incident deaths, the equally weighted median had better mean and worst-case weighted interval scores than the equally weighted mean. However, the equally-weighted median ensemble did have lower coverage rates in the upper tail than the equally-weighted mean ensemble.
    * Comparison of weighted approaches: The relative WIS weighted median was generally at least as good as the weighted mean; differences in mean WIS between these methods were more pronounced for cases than for deaths. Even for cases, there were more outliers in values of WIS for the weighted mean approach, where skill was substantially worse than the equally-weighted median approach. For trained approaches, the weighted median had almost strictly better probabilistic calibration than the trained mean. For the weighted mean approach, subsetting to top-performing models led to improved calibration, but there were not corresponding gains from subsetting to fewer forecasters when using the weighted median ensemble. This may be because the weighted median approach has fewer parameters to estimate, and so is less likely to overfit the training data.
  - Trained methods can have improved mean performance, with caveats
    * In our evaluations for incident cases and deaths, averaging across all forecasts, the weighted median was better than equally-weighted

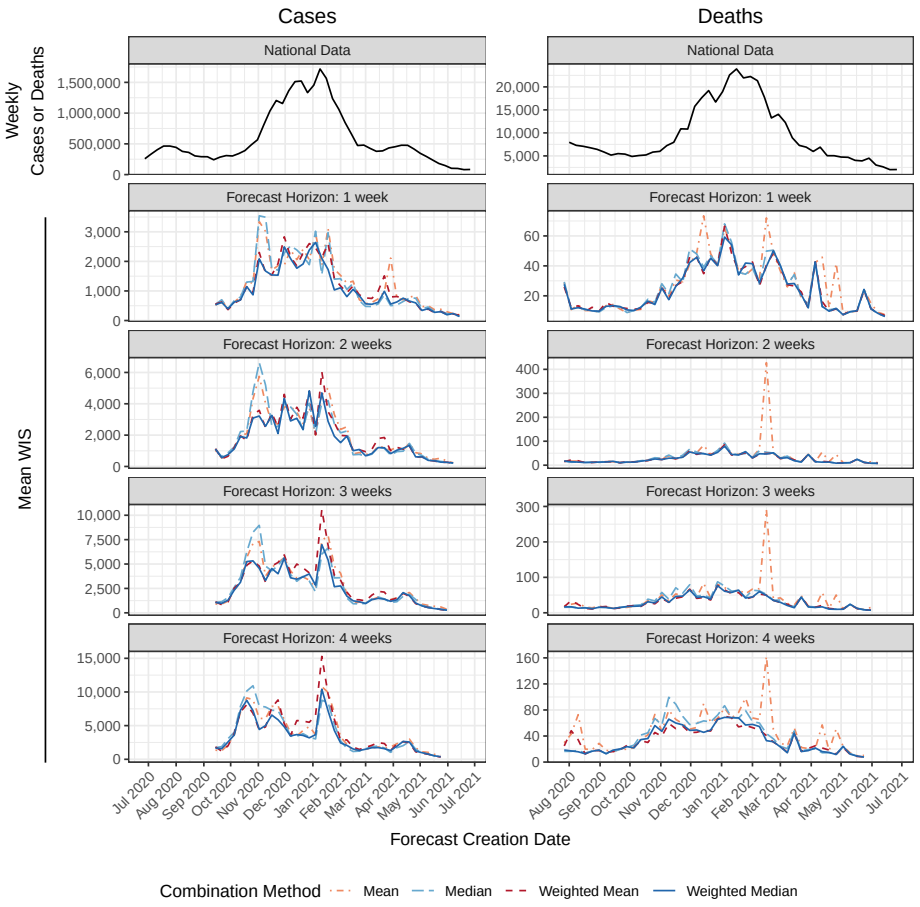

Figure 1: Weekly reported cases and deaths at the national level and mean weighted interval scores (WIS) for state-level forecasts over time for four ensembles: 1) an equally weighted mean ensemble, 2) an equally weighted median ensemble, 3) a weighted mean ensemble, and 4) a weighted median ensemble. 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 each week are updated each week based on performance during the training window. Means are calculated separately for each combination forecast horizon and forecast creation date, averaging across all states and territories. Lower scores indicate better forecast performance.

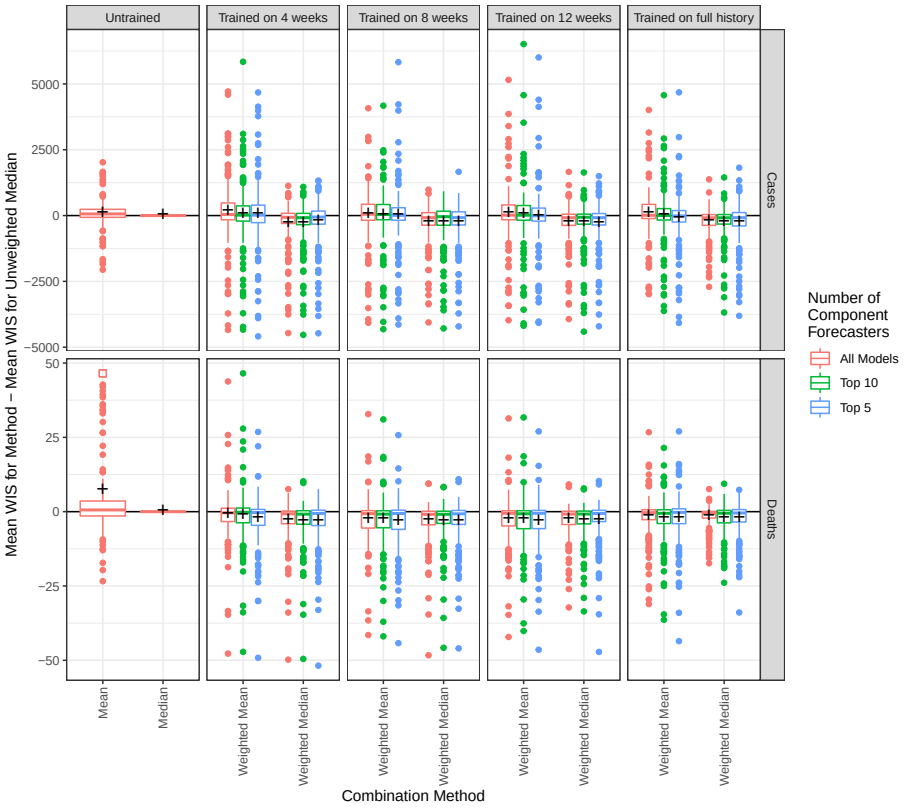

Figure 2: Boxplots summarizing forecast skill for forecasts of weekly cases (top row) and deaths (bottom row). The vertical axis is the difference in mean skill for the given method and the equally-weighted median; the boxplots summarize the distribution of these differences for each combination of forecast date and horizon, averaging across all locations. A cross is displayed at the difference in mean scores for the specified combination method and the equally weighted median. A negative value indicates that the given method outperformed the equally weighted median. Columns indicate the size of the training set, and colors indicate the number of component forecasters included in the ensemble. For readability of the plot, four large outliers for the equally weighted mean ensemble forecasts of incident deaths are truncated, indicated with a hollow square point.

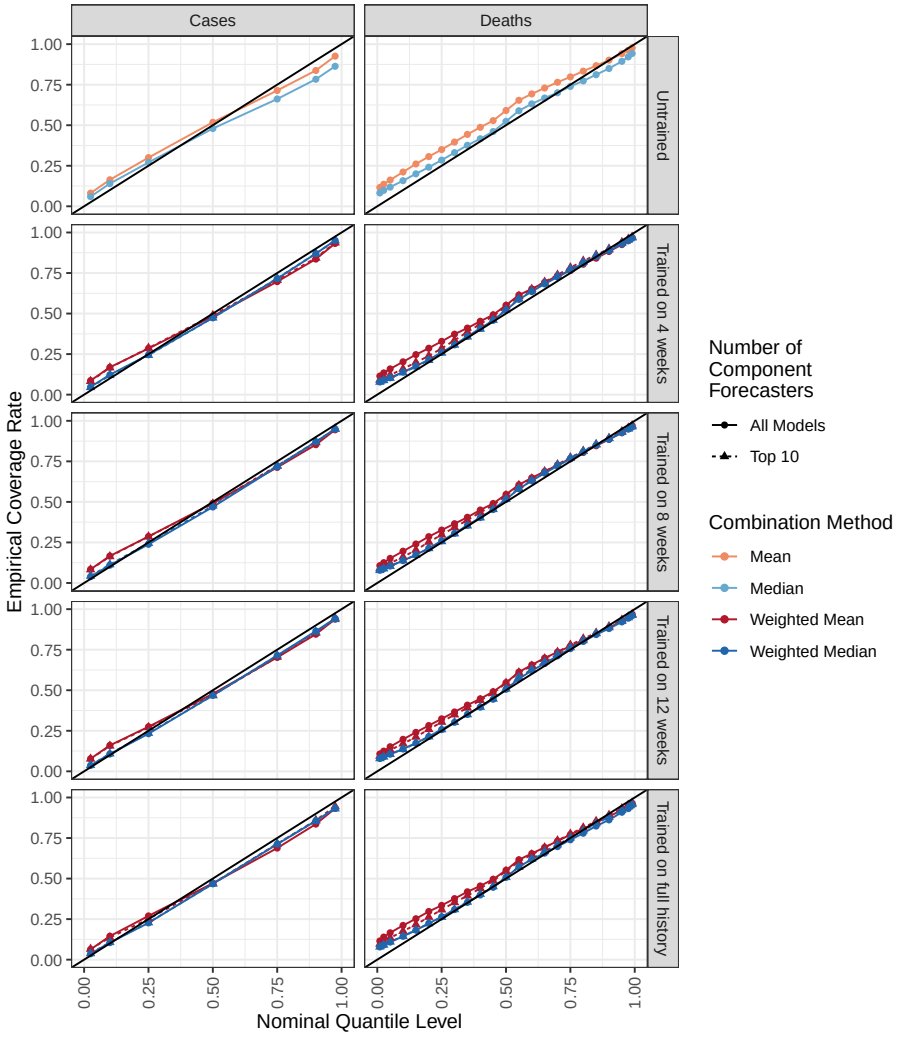

Figure 3: Quantile coverage rates for ensemble forecasts of weekly reported cases and deaths. The horizontal axis gives the nominal quantile level of the prediction; there are seven quantile levels for forecasts of cases and 23 quantile levels for forecasts of deaths. The vertical axis gives the empirical coverage rate of the forecasts at each quantile level, calculated as the proportion of eventually observed values that were less than or equal to the predictive quantile. A well calibrated model will have empirical coverage rates approximately equal to the nominal quantile level.

median and weighted mean was better than equally-weighted mean. These improvements were especially pronounced during times of stable trends that many component forecasters did not capture.

    * But there are some specific times when the weighted approaches were not as good as an equally-weighted median. Specifically, trained methods have more of a tendency to "miss" at turning points by predicting a continuation of recent trends.

- Although the magnitude of scores was larger at larger forecast horizons, these general trends in the relative performance of different ensemble specifications were stable across different forecast horizons (Supplemental Figures 1 and 2).

- For both trained ensembles, using separate model weights at each forecast horizon led to some small improvents in forecast skill at short term forecast horizons of 1 and 2 weeks ahead, but generally worse forecast skill at longer forecast horizons of 3 and 4 weeks ahead (Supplemental Figures 3 and 4). These differences in forecast skill were more pronounced for the weighted mean ensemble than for the weighted median ensemble.

- Performance for the trained median was not very sensitive to other details of the ensemble specification.

    - improvements in mean skill of the weighted median were consistent across target variables (incident cases and deaths) and choices for training set window size and number of component forecasters included (Figure 2).

    - Allowing for separate parameters per quantile level had limited impact for the weighted mean approach, but was unhelpful for the weighted median (Supplemental Figures ...). Note that reductions in coverage concentrated in a few times. Need to make some plots to try to understand what was happening.

– We also considered other possible formulations of a weighted median, including fitting an unweighted median to a subset of top-performing models and calculating the weighted median using the weights that were obtained from the weighted mean. As measured by WIS, the best versions of these other variations on weighted medians had similar performance to the best versions of the relative WIS weighted median considered in the primary analysis. However, the method using weights transferred from a weighted mean ensemble was more sensitive to settings like the number of component forecasters included (Supplemental Figure ...).

5. Discussion

- Robust methods are helpful
- Trained methods can have improved mean performance, with caveats
- Important note: our analysis uses forecasts and ground truth data as they were available in real time, but it is not a prospective analysis. We examined a large number of methods and selected a few to discuss here. Our results should be taken as a statement of how these methods would have done over the past year, and we do not necessarily claim that these findings generalize to the future.
- Summary: robust methods are appealing for public health end users.

6. Supplemental Materials (to be moved to a separate file once we're confident about what goes where)

6.1. Scores broken down by forecast horizon

6.2. Separate weights at different forecast horizons

6.3. Separate weights at different quantile levels

6.4. Impact of reporting anomalies

Nothing here yet. Preliminary examinations indicate "not much".

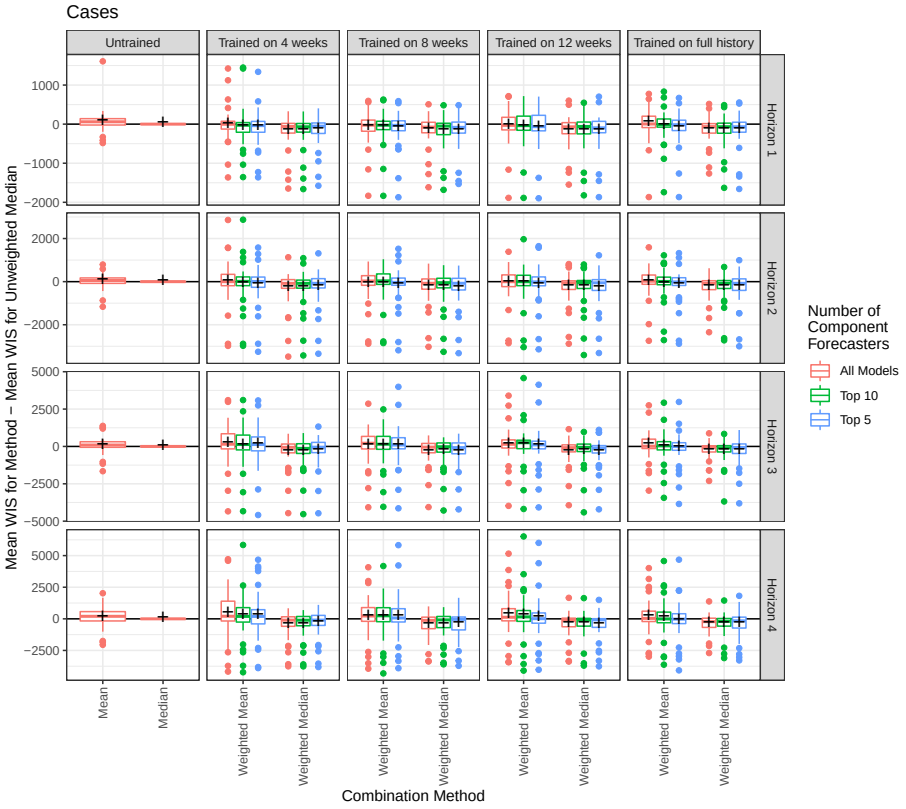

Figure 4: Boxplots summarizing forecast skill for forecasts of weekly cases, broken down by forecast horizon (in rows). The vertical axis is the difference in mean skill for the given method and the equally-weighted median; the boxplots summarize the distribution of these differences for each combination of forecast date and horizon, averaging across all locations. A cross is displayed at the difference in mean scores for the specified combination method and the equally weighted median. A negative value indicates that the given method outperformed the equally weighted median. Columns indicate the size of the training set, and colors indicate the number of component forecasters included in the ensemble. For readability of the plot, four large outliers for the equally weighted mean ensemble forecasts of incident deaths are truncated, indicated with a hollow square point.

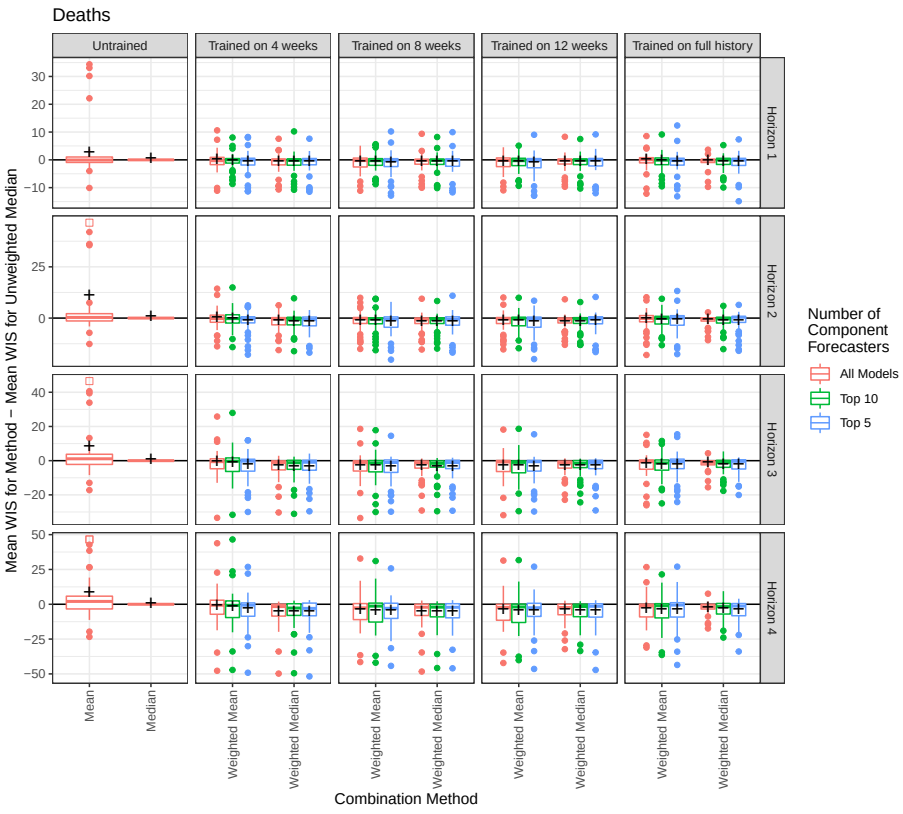

Figure 5: Boxplots summarizing forecast skill for forecasts of weekly deaths, broken down by forecast horizon (in rows). The vertical axis is the difference in mean skill for the given method and the equally-weighted median; the boxplots summarize the distribution of these differences for each combination of forecast date and horizon, averaging across all locations. A cross is displayed at the difference in mean scores for the specified combination method and the equally weighted median. A negative value indicates that the given method outperformed the equally weighted median. Columns indicate the size of the training set, and colors indicate the number of component forecasters included in the ensemble. For readability of the plot, four large outliers for the equally weighted mean ensemble forecasts of incident deaths are truncated, indicated with a hollow square point.

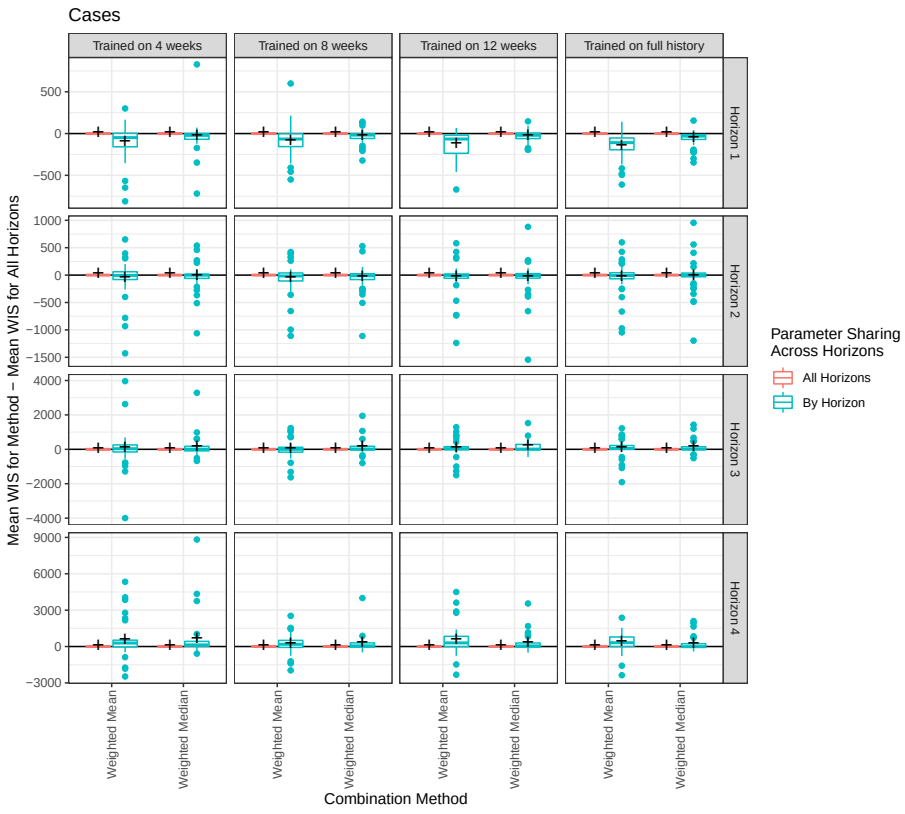

Figure 6: Boxplots summarizing forecast skill for forecasts of weekly cases, varying whether model weights are shared across all forecast horizons or are estimated separately for each forecast horizon. The vertical axis is the difference in mean skill for the given ensemble specification when component 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 trained ensembles combining the ten best individual component forecasters are presented. Columns indicate the size of the training set.

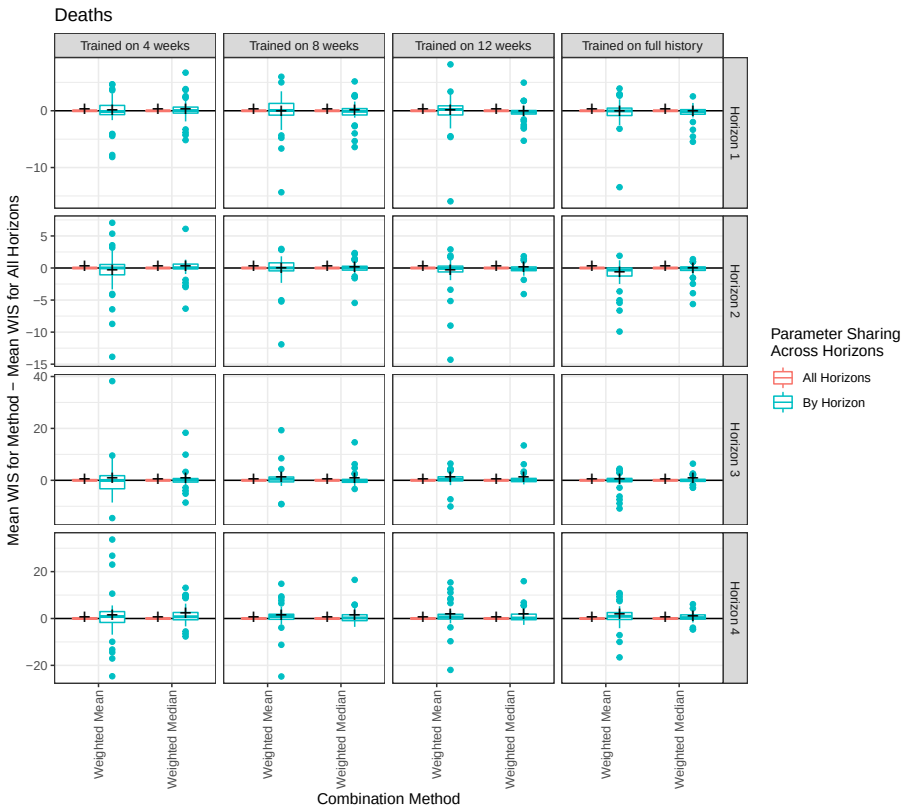

Figure 7: Boxplots summarizing forecast skill for forecasts of weekly deaths, varying whether model weights are shared across all forecast horizons or are estimated separately for each forecast horizon. The vertical axis is the difference in mean skill for the given ensemble specification when component 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 trained ensembles combining the ten best individual component forecasters are presented. Columns indicate the size of the training set.

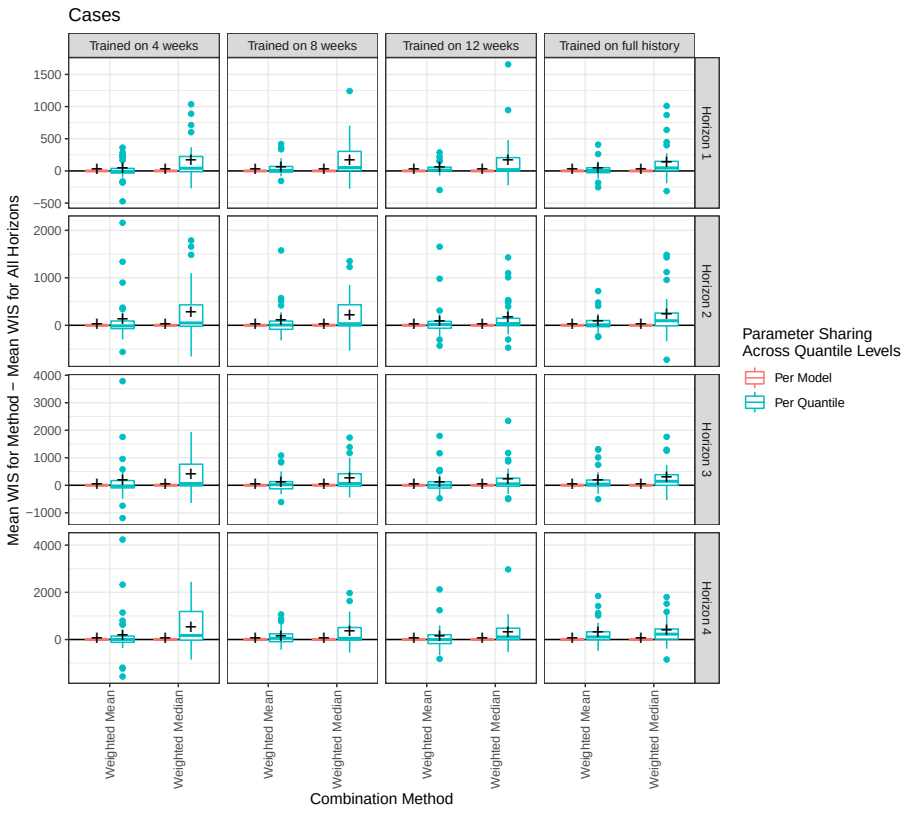

Figure 8: Boxplots summarizing forecast skill for forecasts of weekly cases, varying whether model weights are shared across all quantile levels or are estimated separately for each quantile level. The vertical axis is the difference in mean skill for the given ensemble specification when component weights 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. For this analysis, only results for trained ensembles combining the ten best individual component forecasters are presented. Columns indicate the size of the training set.

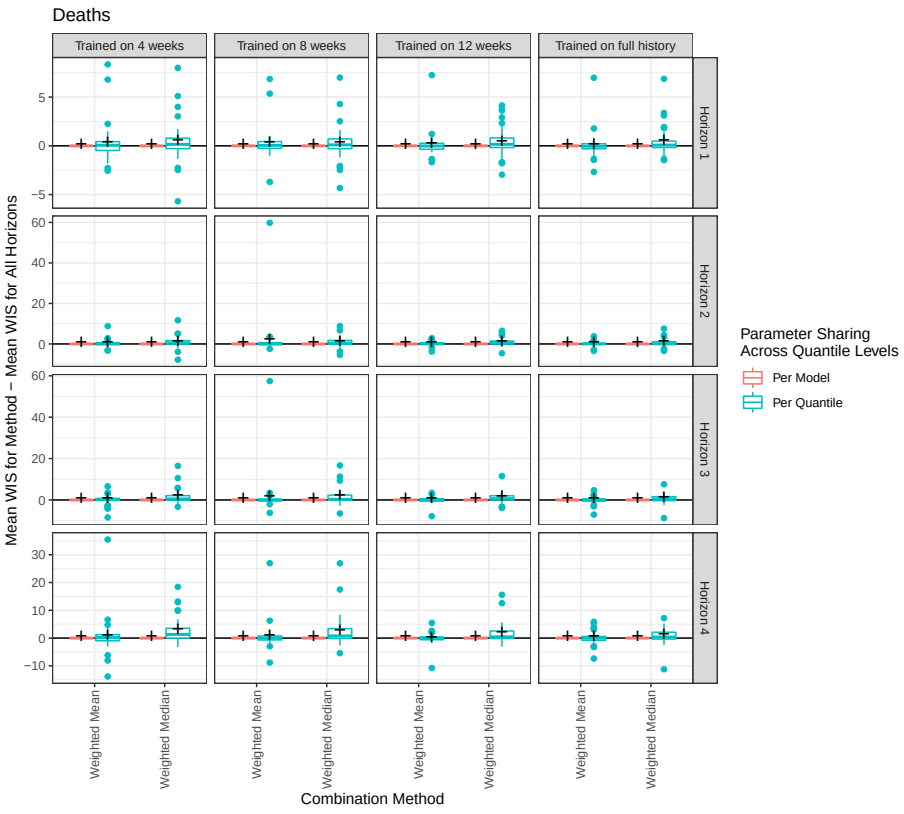

Figure 9: Boxplots summarizing forecast skill for forecasts of weekly deaths, varying whether model weights are shared across all quantile levels or are estimated separately for each quantile level. The vertical axis is the difference in mean skill for the given ensemble specification when component 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. For this analysis, only results for trained ensembles combining the ten best individual component forecasters are presented. Columns indicate the size of the training set.

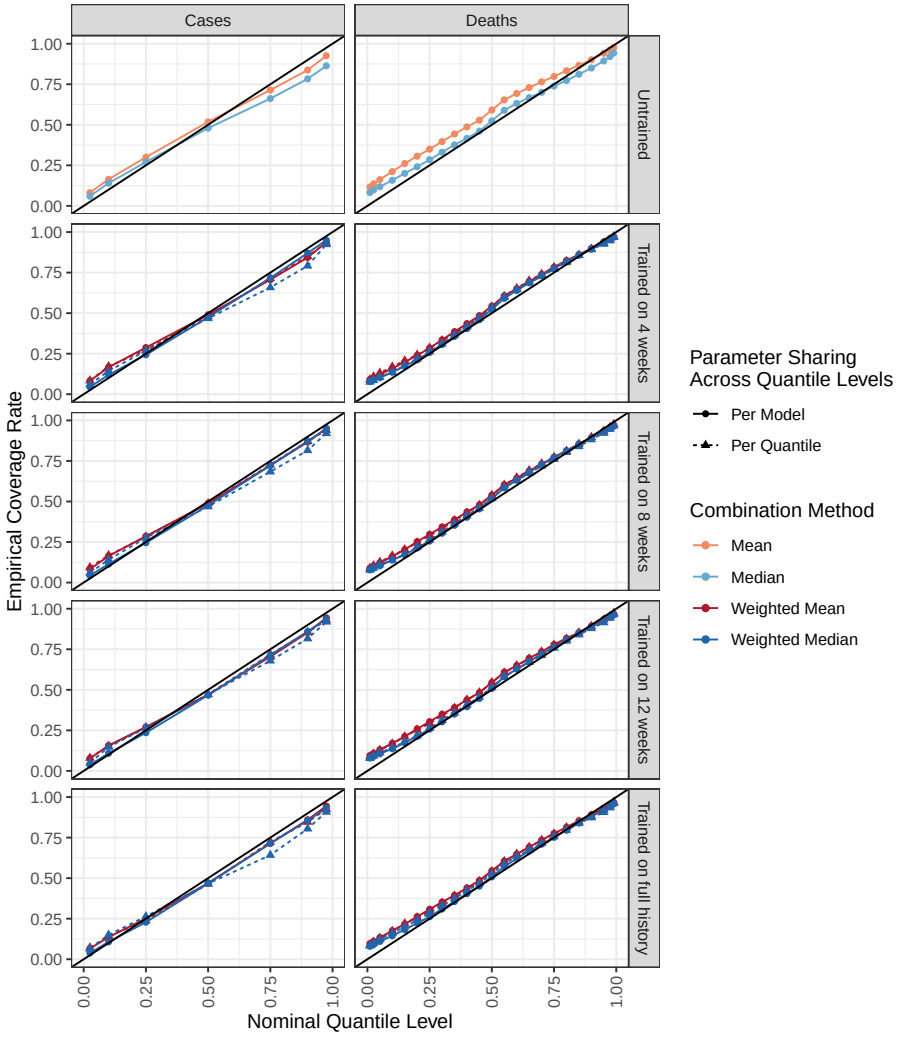

Figure 10: Quantile coverage rates.

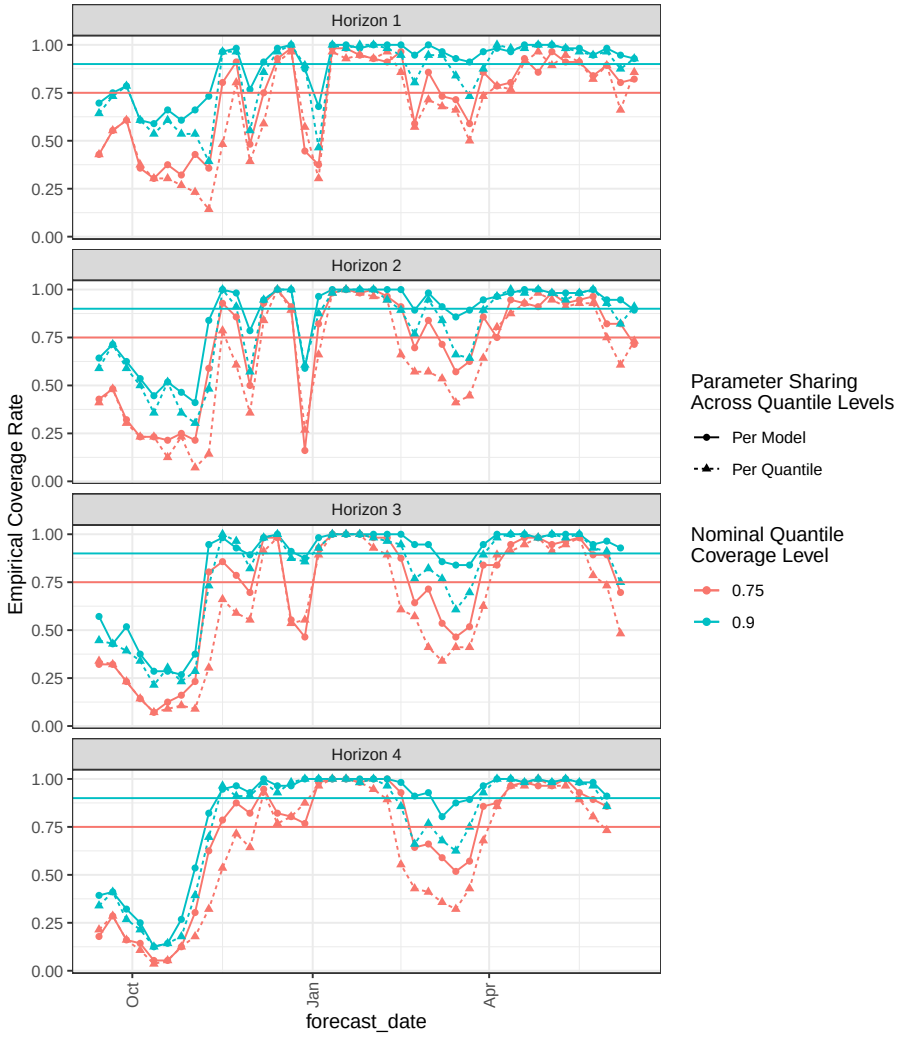

Figure 11: Quantile coverage rates over time.

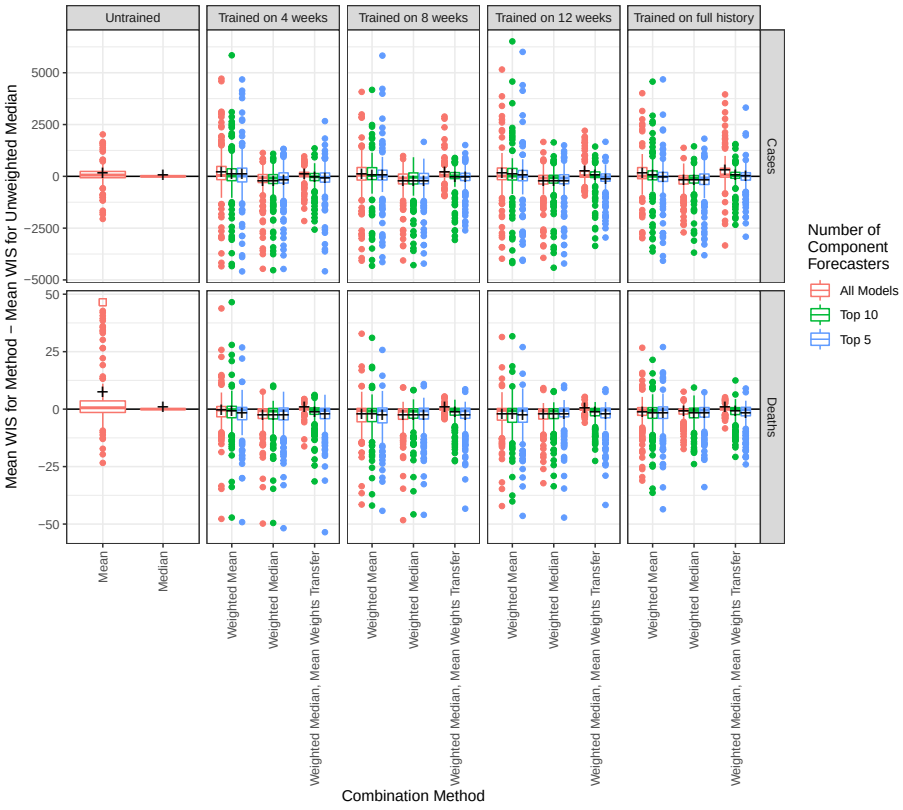

Figure 12: WIS boxplots including additional variations on weighted medians. Need to add equally-weighted median of individual top-performing models back into the mix here.

6.5. Other approaches to weighted medians

- Approaches we have tried – generally similar to the relative WIS weighted approach presented in the main text, occasionally a little worse.
  - Equally-weighted median of best component forecasters
  - Transfer weights from weighted mean to weighted median – similar to the relative WIS weighted approach
- Estimate weighted median weights directly – challenging computationally, Ryan has offered to write this up.

References