Combining Probabilistic Forecasts of COVID-19 Mortality in the United States

James W. Taylor[a], Kathryn S. Taylor[b],

European Journal of Operational Research, 2023, 304(1), 25-41.

[a] Saïd Business School, University of Oxford, Park End Street, Oxford, OX1 1HP, UK
james.taylor@sbs.ox.ac.uk

[b] Nuffield Department of Primary Care Health Sciences, University of Oxford, Radcliffe Primary Care Building, Radcliffe Observatory Quarter, Woodstock Rd, Oxford OX2 6GG, UK
kathryn.taylor@phc.ox.ac.uk

Abstract

The COVID-19 pandemic has placed forecasting models at the forefront of health policy making. Predictions of mortality, cases and hospitalisations help governments meet planning and resource allocation challenges. In this paper, we consider the weekly forecasting of the cumulative mortality due to COVID-19 at the national and state level in the U.S. Optimal decision-making requires a forecast of a probability distribution, rather than just a single point forecast. Interval forecasts are also important, as they can support decision making and provide situational awareness. We consider the case where probabilistic forecasts have been provided by multiple forecasting teams, and we combine the forecasts to extract the wisdom of the crowd. We use a dataset that has been made publicly available from the COVID-19 Forecast Hub. A notable feature of the dataset is that the availability of forecasts from participating teams varies greatly across the 40 weeks in our study. We evaluate the accuracy of combining methods that have been previously proposed for interval forecasts and predictions of probability distributions. These include the use of the simple average, the median, and trimming methods. In addition, we propose several new weighted combining methods. Our results show that, although the median was very useful for the early weeks of the pandemic, the simple average was preferable thereafter, and that, as a history of forecast accuracy accumulates, the best results can be produced by a weighted combining method that uses weights that are inversely proportional to the historical accuracy of the individual forecasting teams.

Keywords: OR in health services; COVID-19; forecast combining; distributional forecasts; interval forecasts.

1. Introduction

The coronavirus 2019 (COVID-19) was declared a pandemic by the World Health Organisation on 11 March 2020 (WHO, 2020). By the end of January 2021, over 102 million individuals had been infected, and COVID-19 had caused over 2.2 million deaths worldwide (WHO, 2021). The COVID-19 pandemic has created enormous planning and resource allocation challenges. Governments are relying upon predictions of the numbers of COVID-19 cases, people hospitalised and deaths to help decide what actions to take (Adam, 2020; Nikolopoulos et al., 2021; Phelan et al., 2020). In this paper, we consider the short-term forecasting of reported deaths from COVID-19.

Forecasting methods are well established in providing predictions of uncertain events to decision makers across a variety of settings, ranging from energy providers and individuals relying on the weather outlook, to investors eager to gain insight into future economic conditions. Epidemiological forecasting models have been applied to both vector-borne diseases, including Dengue disease (Shi et al., 2016) and the Zika virus (Kobres et al., 2019), and contagious infectious diseases. These include the Severe Acute Respiratory Syndrome (SARS) (Ng et al., 2003), Ebola (Viboud et al., 2017) and the Middle East respiratory syndrome (MERS) (Da'ar et al., 2018). Numerous COVID-19 models have emerged (Adam, 2020; COVID-19 Forecast Hub, 2020; Nikolopoulos et al., 2021). These models are based on different assumptions and therefore answer different questions (Holmdahl and Buckee, 2020). Due to the lack of data, assumptions have to be made about several factors including the extent of immunity, transmission among people who are asymptomatic and how the public will react to new government restrictions. Paucity of data is a common challenge in forecasting infectious diseases (Lauer et al., 2020). Policy makers need to be aware of the limitations of the models, and need to be conscious of the uncertainty in predictions from these models (Sridhar and Majumder, 2020).

Gneiting and Katzfuss (2014) describe how optimal decision-making relies on the availability of a forecast of a probability distribution, rather than a single point forecast (see, for example, Gianfreda and Bunn, 2018). We refer to such a probabilistic forecast as a *distributional forecast*. Probabilistic predictions can also take the form of *interval forecasts*. A variety of definitions exist for interval forecasts (Brehmer and Gneiting, 2020). In this paper we use the most common, which is that a (1-α) interval forecast is an interval predicted to contain the true outcome with probability (1-α), and equal probability of being above or below the interval. Interval forecasts are valuable, as they can support real-time decision making and provide situational awareness (see, for example, Grushka-Cockayne and Jose, 2020; Bracher et al., 2021). In this paper, we consider the case where probabilistic forecasts of COVID-19 deaths are provided by multiple forecasters. We combine the forecasts to extract the wisdom of the crowd. Combining provides a pragmatic approach to synthesising the information underlying different forecasting methods. It also enables diversification of the risk inherent in selecting a single forecaster who may turn out to be poor, and it offsets statistical bias associated with individual forecasters who tend to be under- or overconfident. Harnessing the wisdom of a crowd has been found to be useful in many forecasting applications, ranging from sports betting to economics to weather and

climate modelling (see Brown and Reade, 2019; Budescu and Chen, 2015; Mote et al., 2016).

Since the early work by Bates and Granger (1969) on combining point forecasts, a variety of simplistic and sophisticated combining methods have been proposed. Recent work has involved combining probabilistic forecasts. Winkler et al. (2019) predicts that probabilistic forecast combining will become more common due to developments in the field, the rising popularity of forecasting competitions, and raised awareness by increased reporting of probabilistic predictions in the media.

In this paper, we evaluate combining methods applied to multiple probabilistic forecasts of cumulative COVID-19 mortality at the state and national level in the U.S, using data made publicly available (see COVID-19 Forecast Hub, 2020). A notable characteristic of this dataset is that the availability of forecasts from the participating forecasting teams varies greatly across the duration of the dataset. Without a comparable record of past accuracy, it is not clear how best to implement a weighted combining method. This situation has led researchers to focus on combining methods that do not rely on the historical accuracy of the individual forecasters. An example is the simple average. Its success for combining point forecasts has motivated its use for probabilistic forecasts. The median and trimmed means have also been proposed for forecasts of interval and distributional forecasts, as they provide simple, robust alternatives to the mean (Hora et al., 2013; Gaba et al., 2017). In this paper, we evaluate these combining methods for the COVID-19 dataset. We also introduce several weighted combining methods. These address whether it is beneficial to allocate weights using the historical accuracy of each forecaster when these accuracies are not directly comparable.

In Section 2, we describe the dataset and the rise in mortality due to COVID-19 in the U.S. We consider interval forecast combining in Section 3, and the combination of distributional forecasts in Section 4. We separate our consideration of interval and distributional forecasting because the combining methods differ for these two forms of probabilistic forecasts. Section 5 provides a summary and concluding comments.

2. The COVID-19 Mortality Dataset

In this section, we first summarise the progression in mortality due to COVID-19 across the U.S. We then describe the forecasts in the dataset, and the criteria that we applied to select forecasts from this dataset for inclusion in our analysis.

The COVID-19 Forecast Hub is curated by a group led by Nicholas Reich (COVID-19 Forecast Hub, 2020). The Hub provides open access to weekly observations and forecasts for the cumulative total of reported COVID-19 deaths, as well as observations and forecasts for the total deaths each week (incident deaths). These data are available for the U.S. both at the national and state levels. The forecasts are submitted by multiple forecasting teams from academia, industry and government affiliated groups.

2.1. Reported COVID-19 Mortality

The actual number of deaths from COVID-19 will be under-reported due to various factors

including reporting delays and the lack of testing. There will also be indirect deaths as hospitals, overwhelmed by COVID-19, have had to delay diagnosis and treatment for other conditions, such as cancer. Therefore, the impact of COVID-19 will be judged ultimately in terms of the excess mortality, that is, the number of deaths above the number that would have been expected (Weinberger, et al 2020). The actual cumulative deaths described in this paper are obtained from the COVID-19 Forecast Hub. These data are from the daily reports issued by the Centre for Systems Science and Engineering at Johns Hopkins University. Based on these data, the first reported death due to COVID-19 in the U.S. was in the State of Washington, in the week ending 29 February 2020. On 30 January 2021, the total number of COVID-19 deaths in the U.S. had risen to 439,530. Figure 1 shows the number of deaths across the U.S. on this date. Figure 2 shows the rise in the cumulative total number of COVID-19 deaths in the five states with highest cumulative total on this date.

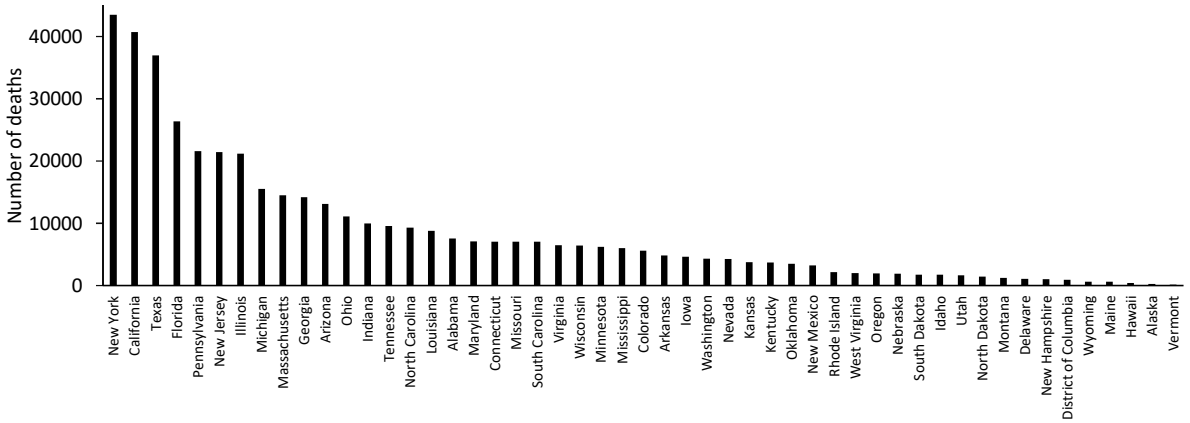

Figure 1. Number of reported COVID-19 deaths in the U.S. up to 30 January 2021.

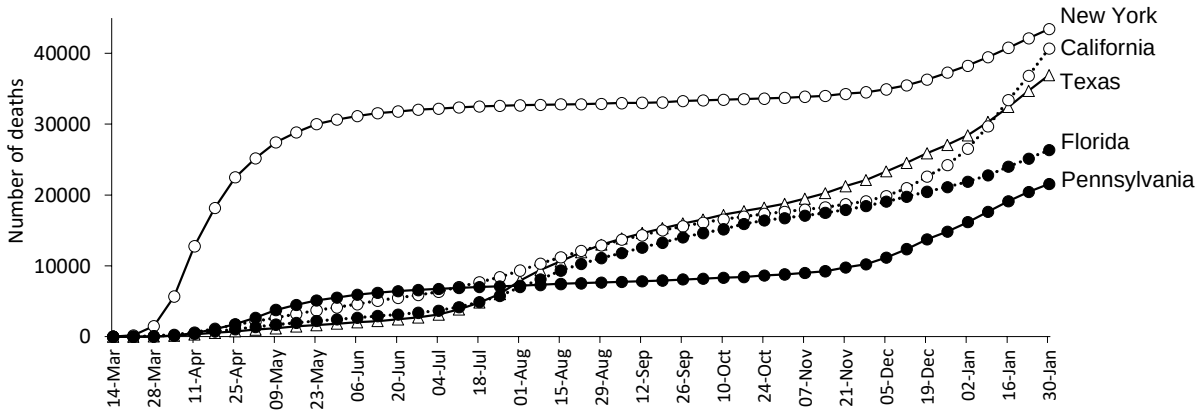

Figure 2. Rise in COVID-19 deaths in the five states with highest cumulative total on 30 January 2021.

2.2. Forecasts of COVID-19 Mortality

The curators of the COVID-19 Forecast Hub ask forecasting teams to submit forecasts for one-week periods ending at midnight on Saturday evenings. The weeks are numbered starting with Week 0 defined as the week ending on Saturday 21 December 2019. At the end of each week, the numbers of incident and cumulative deaths are published, and with that week as forecast origin, the teams submit forecasts for 1 to 4 weeks ahead. For each of these lead times, and for incident and cumulative deaths,

the teams provide a forecast of the probability distribution, which we refer to as the *distributional forecast*. It is provided in the form of forecasts of the quantiles corresponding to the following 23 probability levels: 1%, 2.5%, 5%, 10%, 15%, 20%, 25%, 30%, 35%, 40%, 45%, 50%, 55%, 60%, 65%, 70%, 75%, 80%, 85%, 90%, 95%, 97.5% and 99%. In addition, each team provides a point forecast of the central tendency of the distribution, which often coincides with their forecast of the median. The COVID-19 Forecast Hub provides data visualisations for the incident and cumulative numbers of deaths each week, with interactive plots so that the forecasts of different teams may be compared (see https://viz.covid19forecasthub.org/). Visualisations of the forecasts are also available from the website of the Centers for Disease Control and Prevention, and Nate Silver's FiveThirtyEight website.

As weekly cumulative and incident deaths are related, for simplicity, we focus only on cumulative deaths. Although the COVID-19 Forecast Hub provides forecasts for all U.S. states and territories, we followed the convention adopted for the Hub's visualisation by considering only the 50 states and the District of Columbia. For conciseness, we refer to these as 51 states. Given that we are also considering the national total, our dataset consists of 52 time series and associated forecasts.

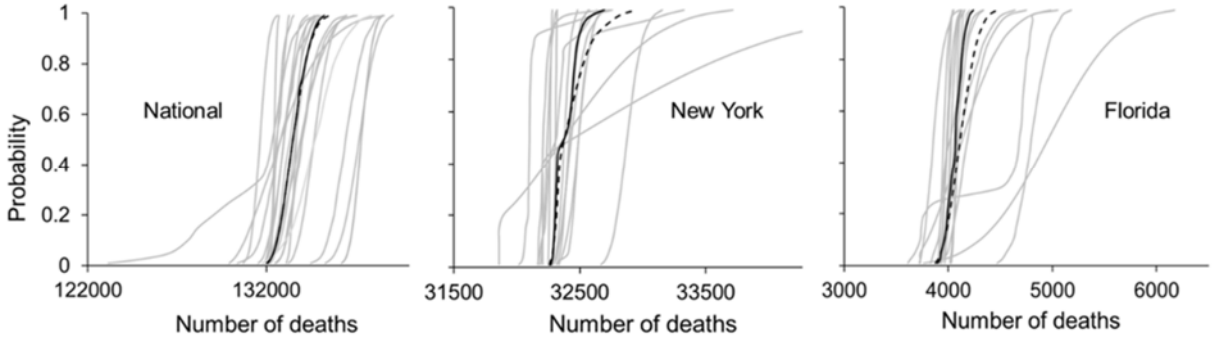

Figure 3. One week-ahead distributional forecasts produced with Week 28 as forecast origin.

In Figure 3, we give examples of distributional forecasts for the national level of cumulative mortality, and for the states of New York and Florida. In each plot, the distribution function shown as the dashed line is a forecast combination, proposed and made available by the curators of the Hub. They refer to it as the ensemble forecast, and explain that, up to Week 31, it was the simple average of forecasts provided by the individual participating teams, and after that date, it was the median of these forecasts. The median distributional forecast is highlighted in black in each plot. Figure 3 shows that there is considerable variation among the distributional forecasts in terms of their location, spread and shape. It is also interesting to note that, at least for New York and Florida, there are a number of outlying distributional forecasts, which motivates consideration of alternative combining methods, based on robust estimation, such as the median and trimming. We discuss this further in Sections 3 and 4.

The COVID-19 Forecast Hub provides information regarding the methods used by the various forecasting teams and licensing conditions for the use of each team's data. The supplementary material to this paper summarises this information. Approximately half of the teams use compartmental models. These involve the estimation of the rates at which the population, or sectors of the population, transfer

between the states of being susceptible, exposed, infected and recovered/removed. Hence, they are widely referred to as SEIR or SIR models. The other forecasting teams used a variety of approaches, including agent-based simulation, statistical models, and deep learning. The use of data-driven machine learning methods is consistent with the increasing use of such methods in healthcare (see Guha and Kumar, 2018). Surowiecki (2004) describes conditions under which wisdom can best be extracted from a crowd. These include independent contributors, diversity of opinions, and a trustworthy central convenor to collate the information provided. Forecasts from the Hub satisfy these conditions.

2.3. Inclusion Criteria for the Forecasts

In our analysis, we considered forecasts made with forecast origins as Week 18 up to Week 57, that is, from the week ending 25 April 2020 to the week ending 23 January 2021. In terms of observed values with which to compare the forecasts, we used weekly data up to Week 58. Week 18 seemed a reasonable starting point for our analysis because levels of mortality were relatively low prior to this, and it was the first forecast origin for which the ensemble forecast was included in the Hub's visualisation. From Week 20, the curators of the Hub have produced files listing the forecasts that they did not include in their ensemble, based on several data screening checks. We omitted these forecasts from our analysis, and also followed the curators by treating as ineligible any submission that did not provide forecasts for all 23 quantiles and all four lead times. Unless this criterion was not met, we included forecasts not recorded as being assessed for eligibility because we felt we had no clear justification for omitting them. We considered data screening checks for these forecasts, but concluded that setting our own thresholds for inclusion would be arbitrary. Consequently, we applied combining methods to a dataset of forecasts that included 11% more forecasts than were included in the ensemble combining method.

Figure 4 shows the total number of forecasting teams included in our study for each forecast origin from Weeks 18 to 57. For each week, the figure also shows the split between the number of teams that used compartmental models and the number using alternatives. Note that the number of teams shown for each week in Figure 4 is an upper bound for the number available for combining for any one time series. This is because some teams either did not provide forecasts or did not provide eligible forecasts for all time series.

For each of the 49 forecasting teams that submitted forecasts, Figure 5 shows, for each week, whether we were able to include that team in the combined forecasts for at least one series. A break in the horizontal line for any team indicates that, from the first week when the team submitted forecasts, it was not the case that forecasts from that team were available and eligible in all the following weeks. The circles in Figure 5 give an indication of the extent to which each forecasting team featured in our study. The figure shows that, even when a record of past historical accuracy becomes available for all teams, accuracy will not be available for the same past periods and same time series. This has potential implications for how to implement a weighted combining method based on the historical accuracy of

each method.

It would be interesting to compare the accuracy of the forecasts produced by the different teams. However, as Figure 5 shows, for most forecasting teams, forecasts were not available for many of the 52 series and 40 forecast origins in our study. Indeed, a full set of forecasts for all series and forecast origins were not available from any of the teams.

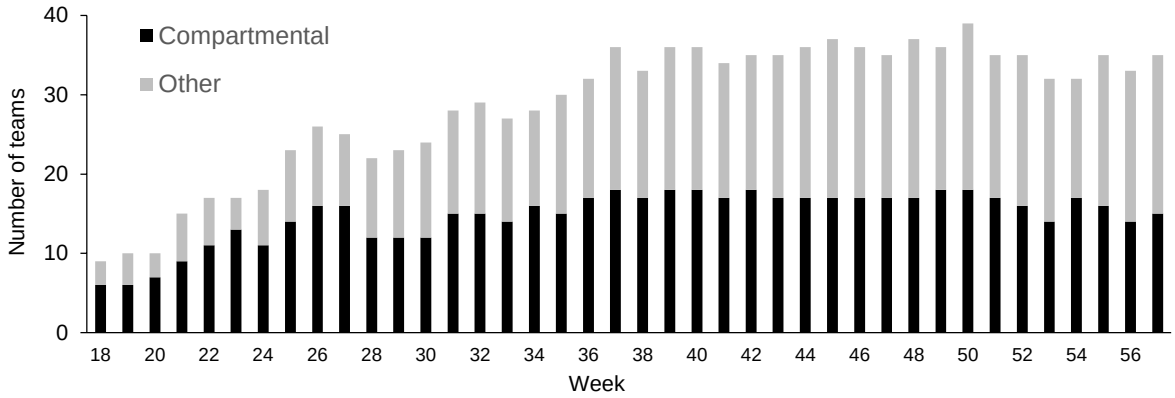

Figure 4. Number of forecasting teams included in our study for each forecast origin. The stacked bars indicate the split between teams using compartmental models, and alternatives.

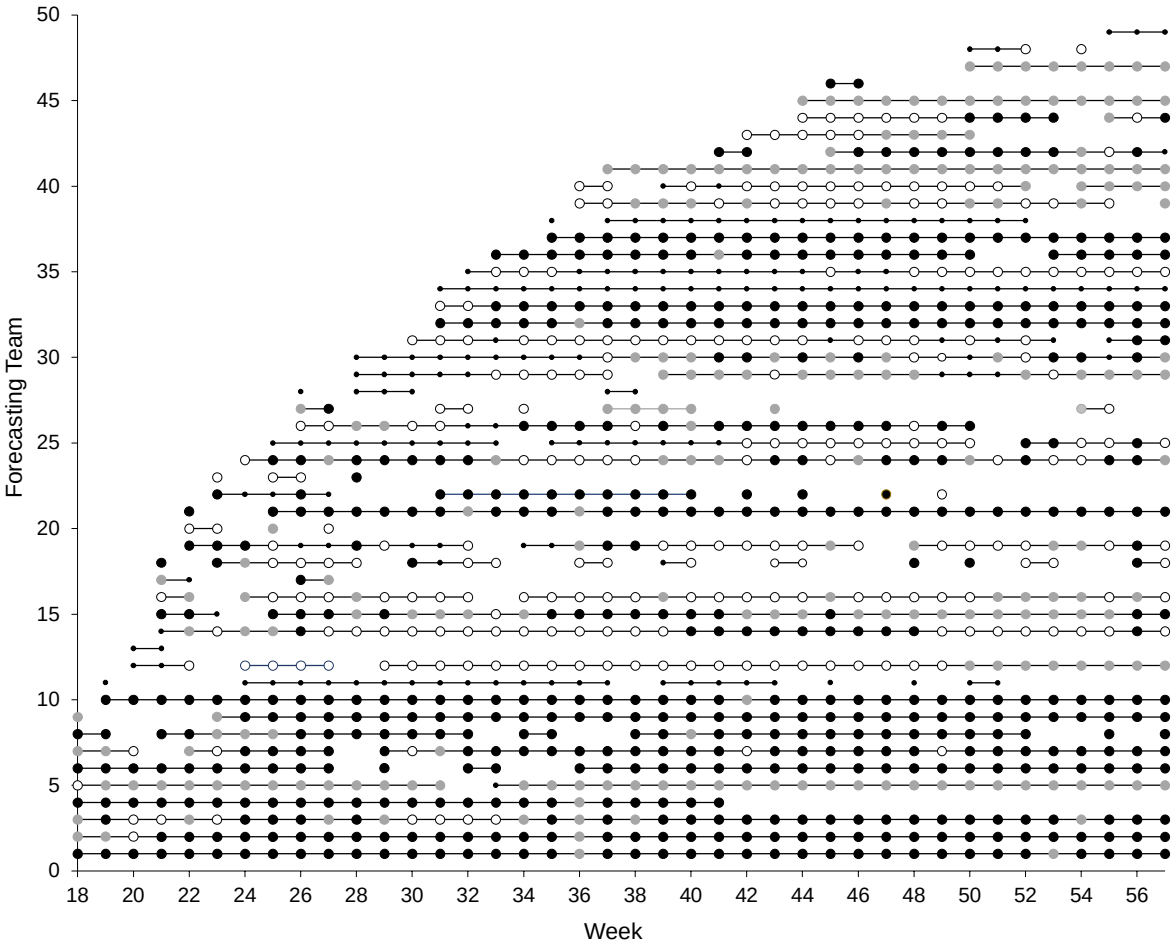

Figure 5. Timeline showing whether forecasts from each team were included in our study for each forecast origin. The circles indicate the number of the 52 series for which forecasts were available and eligible. Black, grey, white and small black circles indicate: all 52 time series, either 51 or 50, between 49 and 26, and 25 or fewer, respectively.

3. Combining Interval Forecasts of COVID-19 Mortality

In this section, we empirically compare a variety of interval forecast combining methods. After describing the structure of our empirical study, we discuss measures used to evaluate interval forecasts. We then review the literature on interval forecast combining methods, before introducing new weighted methods. We summarise the methods that we implemented, and then report the empirical results.

3.1. Structure of the Empirical Study

As we explained in Section 2.3, we used forecasts produced at Weeks 18 to 57, which amounted to 40 forecast origins. Some of the combining methods require parameter estimation, and for this we opted to use at least 10 weeks. Therefore, we evaluated out-of-sample forecasts for the final 30 weeks of our dataset. As we moved the forecast origin through the out-of-sample period, we re-estimated parameters using all weeks up to and including the forecast origin. As the parameter optimisation and several combinations involve forecast accuracy measures, we now discuss interval forecast evaluation.

3.2. Evaluation Measures for Interval Forecasts

In this paper, our interest is in (1-α) intervals that are bounded by the $\alpha/2$ and (1-$\alpha/2$) quantiles. We consider α=5% and 50%, which correspond to 95% and 50% intervals, respectively. Our choice of these intervals reflects those presented in the visualisation of the COVID-19 Forecast Hub.

The accuracy of a set of interval forecasts can be assessed by evaluating the forecasts of the quantiles bounding the interval. A simple measure of accuracy of forecasts of the θ quantile is to check the percentage of observations that fall below the forecast. We refer to this as the *hit percentage*. If this is equal to θ, the forecast is said to be *calibrated*. More precisely, we should refer to this as *unconditional calibration*, with *conditional calibration* being the property that the conditional expectation of the hit percentage is equal to θ (Nolde and Ziegel, 2017). Given the short length of the time series considered in our analysis, we assess only unconditional calibration.

In addition to calibration, a quantile forecast should be evaluated using a *score*. A score can be viewed as a measure of how closely the forecast varies over time with the actual quantile. The score is said to be *consistent* if it is minimised by the true quantile. The use of a consistent score ensures honest reporting by a forecaster (Gneiting and Raftery, 2007). For quantile forecasts, the most widely used consistent score is the quantile regression loss function (see Koenker and Machado, 1999; Taylor, 1999). We refer to it as the *quantile score*, and present it as follows:

$$S_\theta^q\left(q_t(\theta), y_t\right) = \left(\theta - I\left\{y_t \le q_t(\theta)\right\}\right)\left(y_t - q_t(\theta)\right) \tag{1}$$

where y_t is the observation in period t, $q_t(\theta)$ is the θ quantile, and $I\{\cdot\}$ is the indicator function. When θ=50%, the score reduces to the absolute error, showing that the widely used mean absolute error is an appropriate score for a point forecast defined as a prediction of the median. To summarise forecasting performance across a time series, the average of the score is computed. A consistent score for an interval forecast is produced by summing the quantile score for the quantiles bounding the interval (Gneiting

and Raftery, 2007). For an interval bounded by $q_t(\alpha/2)$ and $q_t(1\text{-}\alpha/2)$, if we sum the quantile scores and divide by $\alpha/2$, we get the following *interval score* (Winkler, 1972):

$$S_\alpha^{INT}\left(l_t, u_t, y_t\right) = \left(u_t - l_t\right) + \frac{2}{\alpha} I\left\{y_t \le l_t\right\}\left(l_t - y_t\right) + \frac{2}{\alpha} I\left\{y_t \ge u_t\right\}\left(y_t - u_t\right) \tag{2}$$

where l_t is the interval's lower bound $q_t(\alpha/2)$ and u_t is its upper bound $q_t(1\text{-}\alpha/2)$. The score has the intuitive interpretation that it rewards narrow intervals, with observations that fall outside the interval incurring a penalty, the magnitude of which depends on the value of α (Gneiting and Raftery, 2007). In an application to influenza forecasting, Bracher et al. (2021) use this interpretation to seek insight into why the interval score for one forecasting model is lower than another.

In reporting both interval and distributional forecasting results in this paper, we average results across the four lead times. We do this for three reasons. Firstly, the relative performances of the methods were similar across the four lead times. Secondly, with many results to report, we felt it impractical to provide results for each lead time. Thirdly, we have a relatively small out-of-sample period, which is of particular concern when evaluating forecasts of extreme quantiles, such as the 2.5% and 97.5% quantiles. To show that the results are consistent across the lead times, in Section 4.6, we provide empirical results for distributional forecasting for each lead time. This also enables us to report the results of statistical tests, which are not available for our results averaged across lead times.

3.3. A Review of Combining Methods for Interval Forecasts

The literature on combining interval forecasts is dominated by applications, such as ours, where there is a sizeable group of individual forecasters and a record of past accuracy is not available for the same past periods. The methods that have been proposed consider each bound of the interval separately. An obvious simplistic approach is to use the simple average of the forecasts. In the vast literature on combining point forecasts, it is well established that the simple average can be very competitive in a variety of applications. Interestingly, this is true regardless of whether a record of historical accuracy is available to enable far more sophisticated combining methods to be fitted (Larrick and Soll, 2006). An advantage of the simple average is its simplicity, and robustness to changes over time in the relative performance of the individual methods. However, consideration of robustness prompts the use of combining methods that are robust to outliers, with the obvious candidate being the median. The simple average and the median are both considered by Park and Budescu (2015) and Gaba et al. (2017).

Extending the idea of robustness to outliers, Park and Budescu (2015) propose that, for each bound, a chosen percentage of the highest and lowest forecasts are discarded, followed by averaging of the rest. We refer to this as *symmetric trimming*. The median is an extreme version of symmetric trimming, where all but one forecast is trimmed.

Rather than having robustness as motivation, Gaba et al. (2017) use trimming to address the situation where the individual forecasters tend to be either under- or overconfident. We refer to their methods as *asymmetric trimming*. Their *exterior trimming* involves removing a percentage of the

highest-valued upper bounds and lowest-valued lower bounds, with the combination computed by averaging the remaining upper and lower bounds, respectively. This approach is suitable when the forecasters have tended to be underconfident, with interval forecasts that are generally too wide. Gaba et al. (2017) also suggest *interior trimming*, which involves removing a percentage of the lowest-valued upper bounds and highest-valued lower bounds, followed by averaging. This approach to combining is suitable when the forecasters are overconfident. They also propose an extreme version of interior trimming, which involves discarding all but the highest upper bound and lowest lower bound. They refer to this as the *envelope* method. In addition, Gaba et al. (2017) describe a heuristic approach that views the bounds of each forecaster as having been produced based on a normal distribution.

The empirical studies of Park and Budescu (2015) involved 80% and 90% interval forecasts produced judgmentally by volunteers in experiments, where the interval forecasts related to general knowledge questions and estimates of financial and economic quantities. They found that the simple average and median were outperformed by trimming, and that symmetric was preferable to asymmetric trimming. Gaba et al. (2017) considered 90% interval forecasts produced for financial quantities by employees at a financial brokerage firm, and for macroeconomic variables by participants in the Federal Reserve Bank of Philadelphia's Survey of Professional Forecasters. These forecasts are produced using a variety of methods, including statistical models, expert judgment, and a mixture of the two. They found that exterior trimming performs very well for some of their data, but that the ranking of combining methods is dependent on the characteristics of the individual forecasts, such as under- or overconfidence. In a recent paper, Grushka-Cockayne and Jose (2020) compared combinations of 95% interval forecasts produced by time series methods for the 100,000 series from the M4-Competition. Overall, the best results were achieved with median combining and interior asymmetric trimming.

In our implementation of the combinations involving trimming, we optimised the percentage of forecasts to trim, represented by the parameter β, by finding the value that minimised the sum of the interval score of expression (2) for all lead times for all periods up to and including the forecast origin.

3.4. New Score-Based Weighted Combining Methods for Interval Forecasts

For point forecasting, many combining methods have been proposed that allocate weights according to the historical accuracy of the forecasters. Analogous methods have been proposed for probabilistic forecasts. However, we are not aware of approaches to weighting probabilistic forecasts when historical accuracy is not available for each forecaster for the same past periods, which is the case in our study. Capistrán and Timmermann (2009) address this situation for point forecasts from a survey of macroeconomic forecasters that had "frequent entry and exit" of forecasters. We draw on their ideas to propose weighted combinations for probabilistic forecasts. Each method is based on historical accuracy, and pragmatically disregards the fact that this accuracy has not been assessed using the same past periods for each forecaster. We do not consider the approach of Capistrán and Timmermann (2009) that involves imputation of the missing forecasts, as it is not clear how to adapt it for probabilistic

forecasts, and in any case, imputation is less appealing for short time series with many missing forecasts.

A simple suggestion of Capistrán and Timmermann (2009) is to select the forecaster with best historical accuracy. We implemented this *previous best* approach by selecting the forecasting team for which the in-sample interval score was lowest. With the shortest in-sample period being 10 weeks, we considered only forecasting teams for which we had forecasts for at least five past forecast origins. We imposed the same requirement on all the weighted combining methods. Clearly, five periods is a small number with which to assess interval forecast accuracy, but larger numbers led to the elimination of many forecasters for the early weeks in our out-of-sample period.

Capistrán and Timmermann (2009) implement the proposal of Bates and Granger (1969) to combine point forecasts using a convex combination with weights inversely proportional to the mean squared error. Stock and Watson (2001) explain that this simple method has the appeal of robustness when the estimation sample is small or there are many predictors, which are both issues in our study. Shan and Yang (2009) use the approach to combine quantile forecasts, with the weights computed by replacing squared error with the quantile score. A similar approach is adopted by Taylor (2020) in a study of value at risk and expected shortfall forecasting for financial data. We implemented *inverse score* combining using the interval score of expression (2), and the quantile score of expression (1).

The inverse score approach requires no parameter optimisation, which is appealing because our time series are quite short. However, with little past data, a simple average may be preferable. In view of this, we implemented a form of shrinkage, used for point forecasting by Stock and Watson (2004) and Capistrán and Timmermann (2009), which has the potential to reduce the combination to the simple average. The shrinkage forecast is a weighted average of the forecasts from the simple average and an inverse score method. We estimated the weight by minimising the interval score, after forecasts had become available for the inverse score method for a reasonable number of weeks. We set this to be the first 10 weeks of our 30-week out-of-sample period. This produced forecasts for the final 20 weeks of the out-of-sample period. For the first 10 weeks, we set the forecasts to be those of the inverse score method. We implemented *shrinkage* with the inverse interval score method, and separately with the inverse quantile score method.

Stock and Watson (2001), Shan and Yang (2009) and Taylor (2020) incorporate a *tuning* parameter $\lambda > 0$ in the inverse score approach to control the influence of the score on the combining weights. Expression (3) presents the resulting combining weight for forecaster i at forecast origin t:

$$w_{it} = \frac{\left(1/MS_{i,t}\right)^{\lambda}}{\sum_{j=1}^{J}\left(1/MS_{j,t}\right)^{\lambda}}, \tag{3}$$

where $MS_{i,t}$ is the historical mean of the score computed at forecast origin t from forecaster i, and J is the number of forecasting teams included in the combination. If λ is close to zero, the combination reduces to a simple average, whereas a large value for λ leads to the selection of the team with best historical accuracy. We followed the same approach that we used for the shrinkage method to estimate

λ and produce forecasts for the first 10 weeks of the out-of-sample period. We implemented the tuning method based on the interval score and another version of the method based on the quantile score.

The inverse quantile score methods, as well as asymmetric exterior trimming, sometimes delivered a lower bound above the upper bound. In this situation, we replaced the two bounds by their mean. Although the interval then has zero width, it is better than the upper bound being below the lower.

3.5. Interval Forecast Combining Methods Implemented in this Study

For each mortality series, forecast origin and lead time, we applied the following methods:

Ensemble: This is the combination produced by the COVID-19 Forecast Hub. For the first 13 of the 40 forecast origins in our study, the forecast for each bound was the simple average of the individual forecasts of that bound, and thereafter, it was the median. As noted in Section 2.3, the ensemble was computed from a subset of the forecasts used in the other combining methods that we consider.

Simple average: For each bound, we computed the arithmetic average of forecasts of this bound. We used our full set of forecasts for this and the other combining methods described below.

Geometric mean: For each bound, we computed the geometric mean of forecasts. This was the only combination using the geometric mean. It was motivated by the potentially exponential rise in mortality.

Median: For each bound, we computed the median of forecasts of this bound.

Symmetric trimming: For each bound, we averaged the forecasts remaining after the removal of the N lowest-valued and N highest-valued forecasts, where N is the largest integer less than or equal to the product of $\beta/2$ and the total number of forecasts, and β is the percentage of forecasts to trim.

Asymmetric exterior trimming: We first removed the N lowest-valued lower bound forecasts, as well as the N highest-valued upper bound forecasts, where N is the largest integer less than or equal to the product of β and the number of forecasts. For each bound, we averaged the remaining forecasts.

Asymmetric interior trimming: We removed the N highest-valued lower bound forecasts, as well as the N lowest-valued upper bounds, where N is defined as for asymmetric exterior trimming. For each bound, we averaged the remaining forecasts.

Envelope: This uses the lowest-valued lower bound forecast and highest-valued upper bound forecast.

Previous best: The interval forecast is provided by the forecasting team for which the interval score was the lowest when computed using the weeks up to and including the forecast origin.

Inverse interval score: This is a convex combination of forecasts, where the weights are inversely proportional to the interval score computed using the weeks up to and including the forecast origin.

Inverse interval score shrinkage: This is a weighted average of the simple average and inverse interval score combining methods.

Inverse interval score tuning: This applies a tuning parameter to the weights of the inverse interval score method, as shown in expression (3).

Inverse quantile score: This is a convex combination of forecasts, where the weights on the forecasts of each bound are inversely proportional to the quantile score computed for the forecasts of that bound

using the weeks up to and including the forecast origin.

Inverse quantile score shrinkage: The forecast for each bound is a weighted average of the forecasts of that bound from the simple average and inverse quantile score combining methods.

Inverse quantile score tuning: This weighted combining method applies a tuning parameter to the weights of the inverse quantile score method, as shown in expression (3).

3.6. Interval Forecasting Results

As we explained in Section 3.2, we averaged the results across the lead times. However, with the level of mortality varying greatly across the series, it is inevitable that averaging will lead to the interval score being dominated by its value for the high mortality series. In view of this, we report results for the following four categories of the series: all 52 series; the 17 series with the highest cumulative mortality at end of the final week of our dataset, which corresponded to 16 states and the national U.S. series; the 17 states with the next highest cumulative mortality at the end of the final week; and the 18 states with lowest cumulative mortality at the end of the final week. We refer to these categories as: *all*, *high*, *medium* and *low*.

For the four categories of series, the mean of the interval score for 95% interval forecasts is presented in the first four columns of values in Table 1. The unit of the score is deaths, and lower values of the score reflect greater accuracy. To summarise performance within each category of series, we calculated the geometric mean of the ratios of the (arithmetic) mean score for each method to the (arithmetic) mean score for the simple average method, then subtracted this from 1, and multiplied the result by 100. This can be viewed as an average skill score for each category, reflecting the percentage by which a method is more accurate than the simple average. We present this measure in the final four columns of Table 1. The first four rows of results correspond to simple benchmark combining methods. Of these, the simple average was the best for all four categories of the series. In the next four rows of results, we see that asymmetric interior trimming was the most successful trimming method. This method produced similar results to the simple average. The envelope method performed poorly. Turning to the score-based methods in the final seven rows, we find that the 'previous best' method was relatively poor, and the inverse interval score and inverse quantile score methods performed well. We note that incorporating shrinkage was only beneficial for the low mortality series, and tuning was not useful for any of the four categories of series. Overall, Table 1 shows that the best results were produced by the inverse score methods, with their improvements over the simple average being particularly good for the high and medium mortality series. As we said in Section 3.2, statistical tests are not available for the score averaged over lead times. (We return to the issue of statistical testing in Section 4.6.)

As we noted in Section 2.3, for most forecasting teams, forecasts were not available for many of the 52 series and 40 forecast origins in our study. Indeed, for the out-of-sample 30-week period, a full set of forecasts for each series and forecast origin were available from only one team. The results for this team were very poor, and so we have not included them in Table 1. (We return to the issue of

evaluating the individual forecasting teams in Section 4.6.)

Table 1. Interval score for 95% interval forecasts, averaged over the 30-week out-of-sample period.

	Interval Score				Skill Score (%)			
	All	High	Medium	Low	All	High	Medium	Low
Simple benchmarks								
Ensemble	1402	3423	689	166	-10.9	-9.3	-14.8	-8.9
Simple average	1199	2942	577	**140**	0.0	0.0	0.0	**0.0**
Geometric mean	1308	3163	601	224	-16.3	-11.9	-2.6	-35.8
Median	1505	3730	697	165	-11.3	-9.7	-17.1	-7.4
Trimming methods								
Sym trim	1380	3392	665	156	-7.9	-8.3	-12.9	-3.1
Asym ext trim	1507	3641	768	190	-27.2	-23.3	-35.3	-23.6
Asym int trim	1179	2874	582	**141**	-0.2	0.4	-0.1	-0.8
Envelope	3986	10054	1569	538	-240.1	-217.0	-210.3	-296.3
Score-based methods								
Previous best	1761	4062	1026	283	-70.1	-44.9	-84.6	-83.2
Inv interval score	**1087**	**2614**	558	145	**3.8**	**9.5**	2.2	-0.3
Inv interval score shrink	1114	2703	558	**138**	**4.3**	8.0	**2.4**	**2.4**
Inv interval score tuning	1246	3075	574	154	-0.4	6.9	-0.9	-7.3
Inv quantile score	**1070**	**2581**	**533**	149	3.6	**9.2**	**4.5**	-2.8
Inv quantile score shrink	**1086**	**2635**	**539**	**141**	**4.9**	**9.4**	**4.4**	**0.9**
Inv quantile score tuning	1101	2644	**553**	161	-0.9	5.8	1.7	-10.5

Note: The unit of the score is deaths. Lower values of the score and higher values of the skill score are better. Bold indicates the best three methods in each column.

Table 2. Interval score for 50% interval forecasts, averaged over the 30-week out-of-sample period.

	Interval Score				Skill Score (%)			
	All	High	Medium	Low	All	High	Medium	Low
Simple benchmarks methods								
Ensemble	559	1374	260	72	-1.3	-1.2	-3.1	0.2
Simple average	555	1371	252	72	0.0	0.0	0.0	0.0
Geometric mean	571	1406	255	80	-6.0	-3.9	-1.2	-12.9
Median	579	1429	263	73	-2.3	-2.8	-4.5	0.1
Trimming methods								
Sym trim	575	1424	258	72	-1.5	-2.9	-2.6	0.7
Asym ext trim	592	1463	267	76	-4.9	-4.2	-6.8	-3.8
Asym int trim	565	1399	254	72	-1.5	-2.8	-0.9	-0.9
Envelope	1540	3743	723	231	-228.8	-240.3	-198.7	-248.5
Score-based methods								
Previous best	578	1312	355	95	-29.0	-18.6	-34.6	-34.0
Inv interval score	520	1265	251	**71**	**2.2**	**4.9**	**0.5**	**1.2**
Inv interval score shrink	522	1272	251	**71**	1.9	4.6	**0.5**	0.5
Inv interval score tuning	**477**	**1134**	**250**	**71**	**2.5**	**6.1**	**0.6**	**0.8**
Inv quantile score	**512**	**1241**	**250**	**71**	**2.2**	**5.0**	**0.5**	**1.1**
Inv quantile score shrink	515	1249	**250**	**71**	2.0	4.7	**0.6**	0.6
Inv quantile score tuning	**490**	**1172**	252	72	1.7	4.7	0.2	0.2

Note: The unit of the score is deaths. Lower values of the score and higher values of the skill score are better. Bold indicates the best three methods in each column.

Table 2 presents the interval score results for the 50% interval forecasts. The results are broadly consistent with those for the 95% interval. One difference is that, for the 50% intervals, the tuning forms of the inverse score methods were the most accurate for the all and high categories of series. Further investigation revealed that this was due to very good accuracy for the national U.S. mortality series.

Figure 6 summarises calibration using Q-Q plots to report the hit percentages for the bounds of the 50% and 95% interval forecasts, which are forecasts of the quantiles with probability levels θ=2.5%, 25%, 75% and 97.5%. To ensure readability, we present the results for just four methods. We chose three of the simple benchmark methods and the inverse interval score method. This was the simplest of the inverse score methods, all of which performed well in terms of the interval score. Similar calibration results were produced by the other inverse score methods. In each plot, the dashed line indicates the ideal performance. The plots show that the ensemble and median are slightly outperformed by the simple average and inverse interval score, which have particularly good calibration for the 2.5% and 97.5% quantiles. For the 25% and 75% quantiles, the hit percentages for all four methods were too low, indicating that the forecasts of these quantiles tended to be too low.

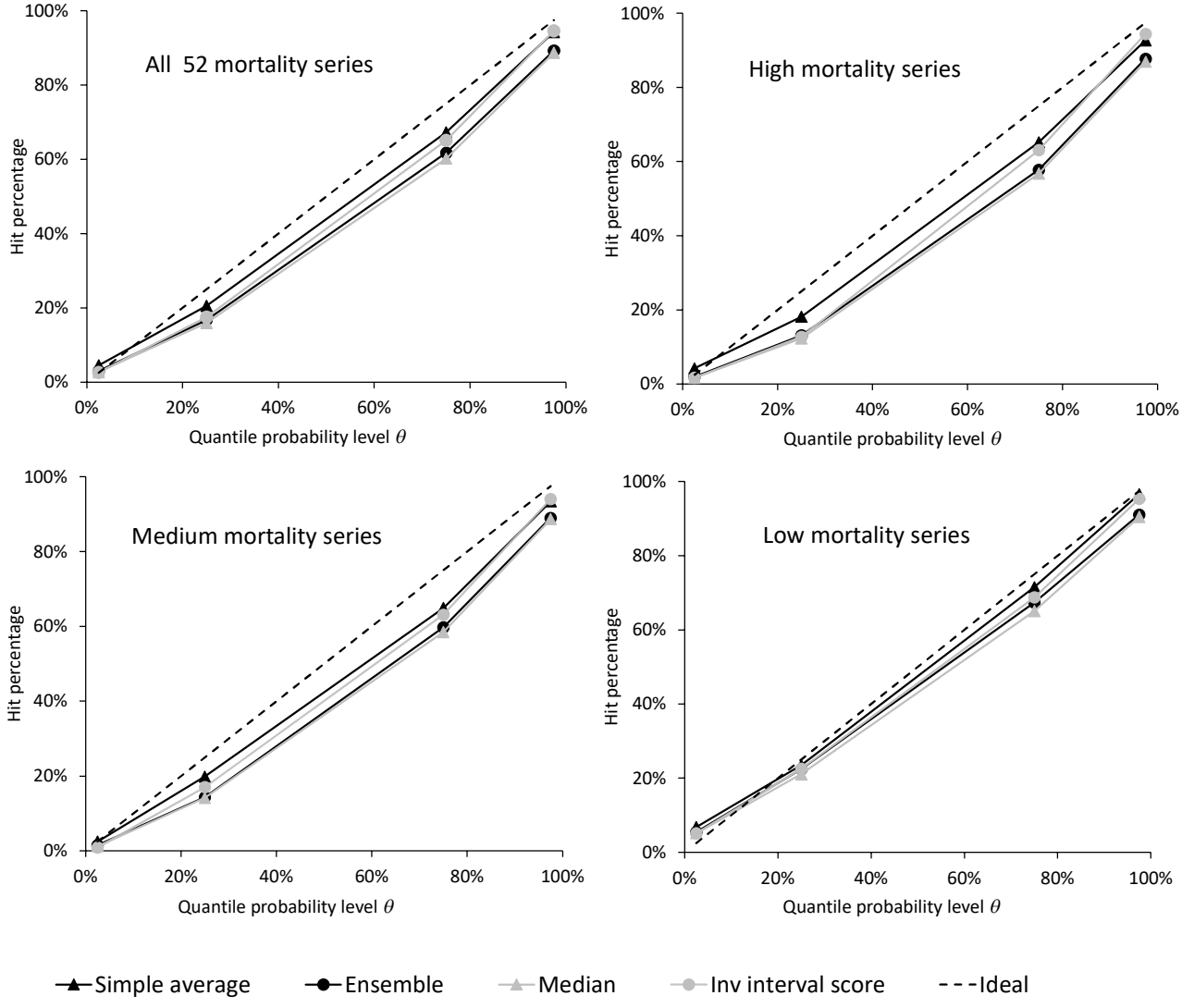

Figure 6. Calibration hit percentages for bounds on 50% and 95% intervals, computed using the 30-week out-of-sample period.

4. Combining Distributional Forecasts of COVID-19 Mortality

This section empirically compares combining methods for distributional forecasts. We first briefly discuss the structure of our empirical analysis, and describe measures used to evaluate distributional forecasts. We then review combining methods that have been proposed in the literature for data of the type that we consider, and present new weighted methods. We then summarise the combining methods that we implemented, and present our empirical results.

4.1. Structure of the Empirical Study

Our empirical analysis of distributional forecasts followed the same structure as our study of interval forecasts. The first 10 weeks of data were used as the initial estimation period, and out-of-sample forecasts were evaluated for the final 30 weeks. For each forecast origin, we re-estimated parameters using all weeks up to and including the forecast origin. To help us describe parameter optimisation and several combining methods, we next discuss distributional forecast evaluation.

4.2. Evaluation Measures for Distributional Forecasts

Gneiting et al. (2007) describe how the aim of distributional forecasting is to maximise *sharpness* subject to *calibration*. Sharpness concerns the concentration of the distributional forecast, and calibration assesses its statistical consistency with the data. Randomly sampled values from a calibrated distributional forecast are indistinguishable from the observations (Gneiting and Katzfuss, 2014). To evaluate calibration for distributional forecasts in our study, we computed the hit percentage for each of the 23 quantiles.

A score summarises calibration and sharpness, and is said to be *proper* if it is minimised by the true distribution. As with consistent scoring functions for quantiles, proper distributional scores are recommended to ensure forecasters report honest predictions (Gneiting and Raftery, 2007). A widely used proper score for distributions of continuous random variables is the continuous ranked probability score (CRPS) (see Gneiting and Raftery, 2007). It can be viewed in several different ways, including the integral of the quantile score of expression (1), with respect to the probability level θ. For our application, where we have quantile forecasts for just 23 different values of θ, we use the linear quantile score (LQS) (see Grushka-Cockayne et al., 2017) in expression (4), which is the sum of the quantile scores for the 23 quantile forecasts:

$$S_{\theta_i}^{DIST}\left(q_t(\theta_i), y_t\right) = \sum_{i=1}^{23}\left(\theta_i - I\left\{y_t \le q_t(\theta_i)\right\}\right)\left(y_t - q_t(\theta_i)\right) \tag{4}$$

This is a proper score, and this can be seen by viewing it as a quantile-weighted version of the CRPS (see Gneiting and Ranjan, 2011). We note that Bracher et al. (2021) present it as a weighted sum of the interval score of expression (2) and the quantile score of expression (1) for the median.

Although our main interest in this paper is probabilistic forecasting, we also consider point predictions of the median, as this conveys the accuracy of the centre of location of the distributional

forecasts, and of course such point forecasts are often the main focus of attention. We evaluate these point forecasts using the mean absolute error (MAE).

4.3. A Review of Combining Methods for Distributional Forecasts

In this brief review of distributional forecast combining methods, we focus on the literature that has considered applications, like ours, where there is a large group of individual forecasters, and there is not a sizeable record of past accuracy available for each forecaster for the same past periods.

The simple average is a well-established approach for combining distributional forecasts (see, for example, Stone, 1961). It has typically taken the form of the linear opinion pool, which, for any chosen value of the random variable, is the average of the corresponding cumulative probabilities obtained from the distributional forecasts. However, this form of averaging has been criticised because, even when the quantiles of the individual distributions are calibrated, the linear opinion pool will not itself provide perfect calibration (Hora, 2004; Ranjan and Gneiting, 2010). It has been noted that, when there is diversity among the means of the individual forecasts, this will tend to lead to an exaggerated variance in the forecast of the linear opinion pool (see, for example, Dawid, 1995). To address these problems, Lichtendahl et al. (2013) propose that, instead of averaging the cumulative probabilities, the distributional forecasts should be averaged by taking the mean of their corresponding quantile forecasts. In other words, for a chosen probability level θ, they suggest that combining is performed by averaging forecasts of the quantile $q_t(\theta)$ provided by each individual distributional forecast. They show how this leads to more attractive theoretical properties than the linear opinion pool. For our application in this paper, it provides a more convenient approach to averaging because each forecasting team submits the distributional forecast in terms of quantile forecasts for the same 23 values of the probability θ.

To provide robust combining, Hora et al. (2013) propose that the median is used. For any chosen value of the random variable, their approach finds the median of the corresponding values of the cumulative probability forecasts. In fact, they show that this delivers the same combined distributional forecast as an approach that uses the median of forecasts of the quantile $q_t(\theta)$ for a chosen θ.

As with interval forecast combining, trimming has also been proposed for distributional forecasts. Jose et al. (2014) propose *interior* and *exterior* trimming approaches, which involve trimming the innermost and outermost distributions, respectively, from a set of distributional forecasts. To enable the trimming, the distributional forecasts must essentially be ordered in some way, and for this, they propose two alternative approaches: the *CDF approach* (CA) and *mean approach* (MA). MA orders the distributional forecasts according to their means, and involves trimming entire distributional forecasts. After a proportion have been trimmed, the authors use a linear opinion pool to average the rest. CA orders the distributional forecasts separately for each of a set of values of the random variable. After the trimming is performed, the combined forecast is computed as the average of the cumulative probabilities given by the remaining distributional forecasts. Jose et al. (2014) note that CA could be adapted so that the trimming and averaging is performed on forecasts of the quantile $q_t(\theta)$ for any chosen

value of the probability θ, which would be more consistent with the advice of Lichtendahl et al. (2013) to average quantiles, rather than probabilities. For our application, this is a more convenient way to implement CA, as our distributional forecasts have each been provided in the form of a set of quantile forecasts. Following similar reasoning, we avoided the linear opinion pool with MA, so that following trimming, quantile forecasts are averaged.

For interval forecast combining, symmetric trimming is motivated by robustness, and asymmetric trimming enables the impact to be reduced of a tendency among the individual forecasters to be either under- or overconfident. It is worth noting that analogous asymmetric methods are not straightforward for distributional forecasts because of the need to ensure that the resulting distribution function is monotonically increasing. It is also interesting to note that, although the trimming methods proposed by Jose et al. (2014) are all symmetric, and hence their exterior trimming will enable the removal of outliers, their main motivation for trimming is to address under- or overconfidence among the individual forecasters. For example, in an application to forecasts from a survey of professional economic forecasters, they show that, in comparison with the linear opinion pool, exterior trimming enables the impact to be reduced of underconfident forecasts of inflation, while interior trimming allows the combined distributional forecast to reduce the impact of overconfident forecasts of growth. Grushka-Cockayne et al. (2017) investigate the theoretical properties of CA with exterior trimming, and compare it with the linear opinion pool. They show that exterior trimming can overcome the tendency for the linear opinion pool to produce an underconfident forecast when the individual distributional forecasts have diverse means. They demonstrate this with an ensemble of distributional forecasts from a quantile regression forest. They show that the tendency for this machine learning method to overfit leads to diverse means among the ensemble, which can be addressed by CA with exterior trimming.

In our study, for the distributional forecast combining methods that involve trimming, we optimised the trimming percentage β by finding the value that minimised the sum of the LQS of expression (4) for all four lead times using all periods up to and including the forecast origin.

4.4. New Score-Based Weighted Combining Methods for Distributional Forecasts

For distributional forecasting, we implemented a set of score-based combining methods analogous to those described in Section 3.4 for interval forecasting. For the 'previous best' and inverse interval score weighted combining methods, we replaced the use of the interval score by the LQS. The other three combining methods in Section 3.4 produced a forecast for each interval bound, based on the inverse of the quantile score. For distributional forecasting, we implemented these methods for each of the 23 quantiles that underlie each distributional forecast. Four of the inverse score combining methods in Section 3.4 involved a parameter, which we optimised using the interval score. For the analogous methods for distributional forecasting, we used the same optimisation procedure with the interval score replaced by the LQS. The inverse quantile score methods sometimes gave forecasts for the 23 quantiles that were not monotonically increasing with θ, the probability level. When this occurred for the quantile

forecasts corresponding to two adjacent values of θ, we replaced both quantile forecasts by their mean.

4.5. Distributional Forecast Combining Methods Implemented in this Study

For each mortality series, forecast origin and lead time, we implemented the following methods:

Ensemble: This is the combination produced by the COVID-19 Forecast Hub. For the first 13 of the 40 forecast origins in our study, the ensemble forecast of each of the 23 quantiles was the simple average of the corresponding quantile forecasts, and thereafter, it was the median. As noted in Section 2.3, the ensemble used a subset of the forecasts included in all other combining methods that we considered.

Simple average: For each of the 23 quantiles, we used the arithmetic mean of the corresponding quantile forecasts. We used our full set of forecasts for this and the other combining methods described below.

Geometric mean: For each of the 23 quantiles, we used the geometric mean of the corresponding quantile forecasts. This was the only combining method involving the geometric mean.

Median: For each of the 23 quantiles, we found the median of the quantile forecasts.

CA exterior trimming: For each of the 23 quantiles, we averaged the quantile forecasts remaining after we had removed the N lowest-valued and N highest-valued quantile forecasts, where N is the largest integer less than or equal to the product of $\beta/2$ and the total number of forecasts, and β is the percentage of forecasts to trim. For each bound, we averaged the remaining forecasts.

CA interior trimming: With this method, for each of the 23 quantiles, the innermost quantile forecasts were trimmed. The combination was computed as the average of the quantile forecasts that were either among the N lowest-valued or N highest-valued quantile forecasts, where N is the largest integer less than or equal to the product of $(1-\beta)/2$ and the total number of forecasts.

MA exterior trimming: This method involved trimming entire distributional forecasts. The trimming was based on the mean of each distributional forecast, which we estimated using the average of the 23 quantile forecasts. The forecast combination was computed by averaging the distributional forecasts that remain after the removal of the N distributional forecasts with lowest-valued mean and the N distributional forecasts with highest-valued mean, where N is the largest integer less than or equal to the product of $\beta/2$ and the total number of forecasts.

MA interior trimming: This was similar to MA exterior trimming, except the innermost distributional forecasts were trimmed. The combination was computed as the average of the distributional forecasts that were among the N distributional forecasts with lowest-valued mean and the N distributional forecasts with highest-valued mean, where N is the largest integer less than or equal to the product of $(1-\beta)/2$ and the total number of forecasts.

Previous best: The distributional forecast is provided by the forecasting team for which the LQS was the lowest when computed for the weeks up to and including the forecast origin.

Inverse LQS: This is a convex combination of forecasts, where the weights are inversely proportional to the LQS computed for the weeks up to and including the forecast origin.

Inverse LQS shrinkage: This is a weighted average of the simple average and inverse LQS methods.

Inverse LQS tuning: This applies a tuning parameter to the weights of the inverse LQS method.

Inverse quantile score: This is a convex combination of forecasts, where the weights on the forecasts of each of the 23 quantiles are inversely proportional to the quantile score computed for each quantile forecast using the weeks up to and including the forecast origin.

Inverse quantile score shrinkage: The forecast for each quantile is a weighted average of the forecasts of that quantile from the simple average and inverse quantile score combining methods.

Inverse quantile score tuning: This applies a tuning parameter to the weights of the inverse quantile score method.

4.6. Distributional Forecasting Results

Table 3 presents the mean of the LQS for the four categories of series, along with skill scores. The unit of the LQS is the number of deaths, and lower values of the score reflect greater accuracy. There are similarities between the results of Table 3 and the interval score results of Tables 1 and 2, with the simple average being the best of the four simple benchmarks; the trimming methods performing reasonably but unremarkably; the previous best method doing poorly; and the best results produced by the inverse score methods. For the medium and low mortality series, Table 3 shows that the inverse score methods performed similarly, while for the high mortality series, incorporating tuning was beneficial. Closer inspection revealed that this finding for the high mortality series was due to tuning improving accuracy for the national U.S. series. The incorporation of shrinkage was not beneficial.

Table 3. LQS for the distributional forecasts, averaged over the 30-week out-of-sample period.

	LQS				Skill Score (%)			
	All	High	Medium	Low	All	High	Medium	Low
Simple benchmarks methods								
Ensemble	1296	3182	606	168	-2.0	-1.7	-3.4	-0.9
Simple average	1274	3142	581	164	0.0	0.0	0.0	0.0
Geometric mean	1312	3227	587	188	-6.8	-4.4	-0.9	-15.3
Median	1344	3320	614	168	-2.7	-3.1	-4.8	-0.5
Trimming methods								
CDF ext trim	1317	3253	600	165	-1.7	-3.3	-2.8	0.8
CDF int trim	1330	3311	582	166	-2.7	-7.1	-0.2	-1.0
Mean ext trim	1308	3230	596	166	-0.9	-1.6	-1.9	0.6
Mean int trim	1298	3212	584	166	-1.2	-2.5	-0.4	-0.9
Score-based methods								
Previous best	1495	3496	833	229	-33.3	-24.2	-36.6	-39.4
Inv LQS	1194	2901	**579**	**163**	**2.3**	**5.3**	**0.6**	**0.9**
Inv LQS shrink	1200	2920	**579**	**163**	2.1	4.9	0.5	**0.9**
Inv LQS tuning	**1115**	**2658**	**579**	**163**	2.0	**5.8**	0.3	0.0
Inv quantile score	**1179**	**2859**	**573**	**163**	**2.5**	**5.5**	**1.3**	0.8
Inv quantile score shrink	1197	2911	**578**	**163**	**2.2**	5.1	**0.7**	**1.0**
Inv quantile score tuning	**1125**	**2690**	580	**163**	2.0	4.6	**0.6**	**0.9**

Note: The unit of the score is deaths. Lower values of the score and higher values of the skill score are better. Bold indicates the best three methods in each column.

As we explained in Section 3.6, for the out-of-sample 30-week period, a full set of forecasts for each series and forecast origin were available from only one of the individual forecasting teams. That team performed very poorly in terms of interval forecasting, and also for distributional forecasting. To gain some insight into the relative accuracy of the individual teams, we computed the skill scores, as in Table 3, for each team using the periods for which forecasts were available from that team. This led to a wide range of skill scores, due partly to the instability of this measure when forecasts were available from a team for only a small number of periods. Nevertheless, it was interesting to find that the best skill scores computed for the individual teams for the all, high, medium and low categories of series were -1.3%, -1.3%, -8.6% and -6.1%, respectively. These negative values indicate that the best individual team for each category was less accurate than the simple average combining method.

To look in more detail at the LQS results, Figure 7 reports the LQS for each of the 52 series for the simple average and the inverse LQS method, which is a simple form of inverse score method, involving no parameter estimation. The inverse LQS method can be seen to be slightly better than the simple average for most of the series, with this being most evident for the national U.S. series.

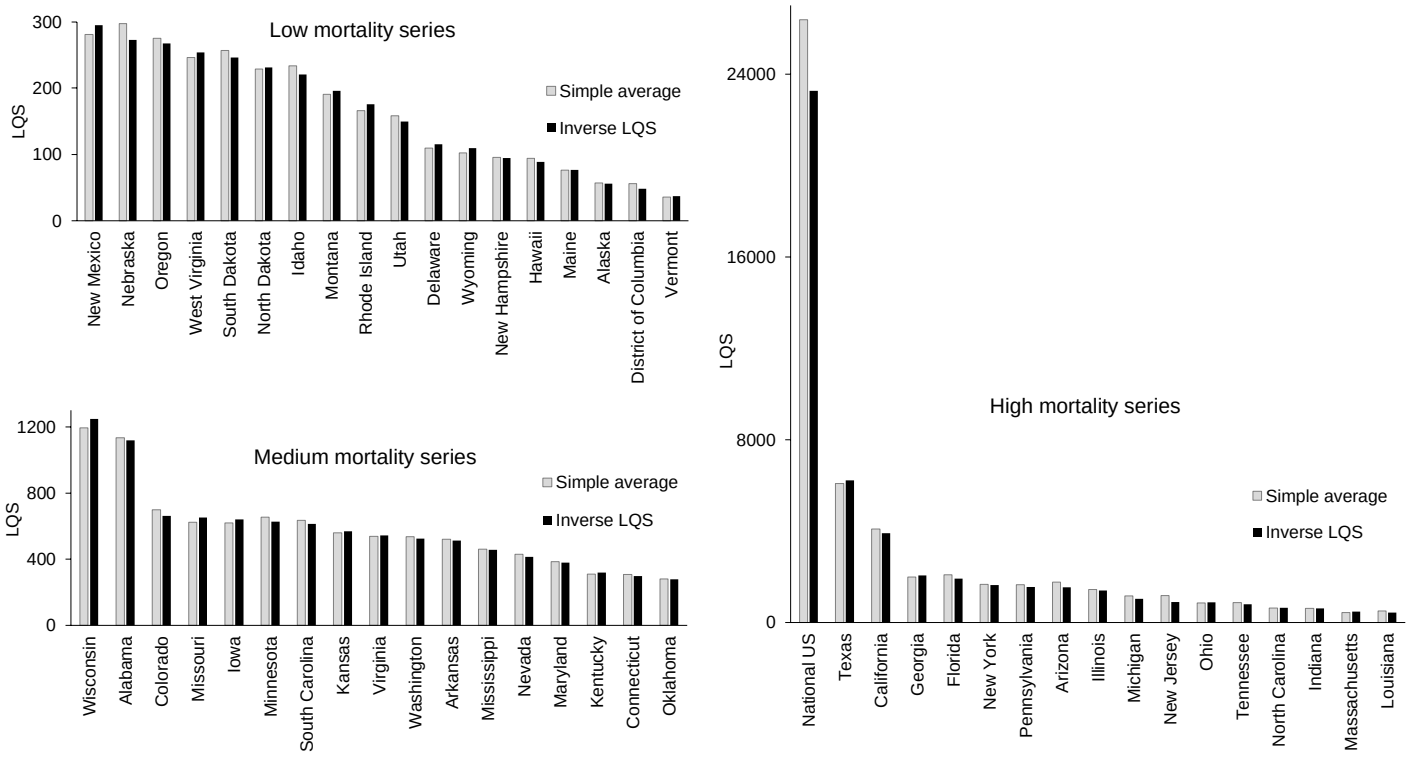

Figure 7. Comparison of the LQS for the simple average and inverse LQS combining methods for each of the 52 mortality series. LQS averaged over the 30-week out-of-sample period.

We also evaluated the distributional forecast combining methods of Table 3 in terms of their accuracy for producing 95% and 50% interval forecasts. The interval forecasts from the simple benchmarks were the same as those produced from these methods in our interval forecasting study of Section 3. However, this was not the case for the other methods in Table 3. We found that the best of the trimming methods in Table 3 was outperformed by the best of the interval forecast trimming methods in Section 3. For the inverse score methods, the best interval forecast accuracy achieved by the

methods of Table 3 was similar to the accuracy of the best of the inverse score methods in Section 3.

For the four categories of series, Figure 8 presents Q-Q plots to summarise the calibration hit percentages for each of the 23 quantiles. To ensure legibility, we include just four methods in each plot: three simple benchmarks and the inverse LQS method, which was one of the most competitive methods in terms of the LQS. The other inverse score methods delivered similar calibration results. All the Q-Q plots show good calibration for all four methods for the extreme quantiles, with forecasts for the other quantiles being, on average, too low. The figure shows that the methods performed similarly for the medium and low mortality series, while for the high mortality series and all 52 series considered together, there was better calibration from the simple average and inverse LQS method, with the simple average being the better calibrated for the lower quantiles.

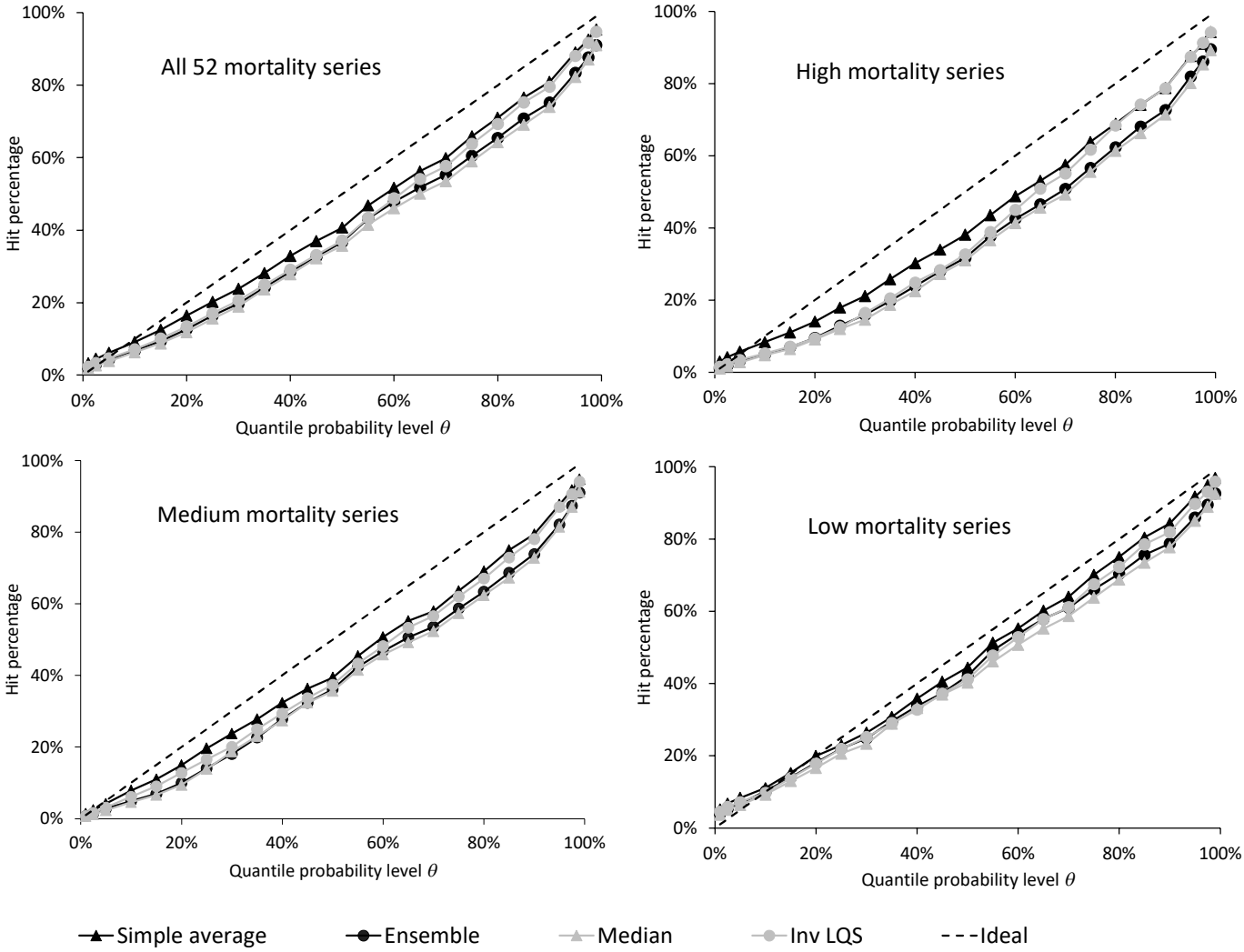

Figure 8. Calibration of distributional forecasts assessed using hit percentages for the 23 quantile probability levels θ. Hit percentages computed using the 30-week out-of-sample period.

In Table 4, we evaluate point forecast accuracy for forecasts of the median using the MAE averaged across the four lead times. The unit of the MAE is the number of deaths. The relative performances of the methods in Table 4 are similar to those in Table 3 for the LQS. This is quite common when there is sizeable variation over time in the location of the probability distribution,

because inaccuracy in the prediction of the distribution's location will affect the accuracy of all the quantiles, and hence the whole distribution.

Table 4. MAE for the median point forecasts derived from the distributional forecasts. MAE averaged over the 30-week out-of-sample period.

	MAE				Skill Score (%)			
	All	High	Medium	Low	All	High	Medium	Low
Simple benchmarks methods								
Ensemble	171	420	79	23	-0.6	-0.5	-1.2	-0.3
Simple average	170	420	**77**	**22**	0.0	0.0	0.0	0.0
Geometric mean	174	428	78	24	-4.9	-3.0	-1.0	-10.7
Median	175	431	79	**22**	-1.0	-1.5	-2.2	0.5
Trimming methods								
CDF ext trim	174	430	79	**22**	-0.9	-2.8	-1.4	1.4
CDF int trim	177	440	**77**	**22**	-2.2	-6.1	-0.2	-0.6
Mean ext trim	173	427	79	**22**	-0.2	-1.1	-1.3	**1.5**
Mean int trim	173	427	78	**22**	-1.0	-1.9	-0.5	-0.5
Score-based methods								
Previous best	184	431	102	29	-24.7	-16.6	-26.1	-31.6
Inv LQS	**160**	**389**	**77**	**22**	**2.0**	**4.5**	0.3	1.2
Inv LQS shrink	161	391	**77**	**22**	1.9	4.2	0.3	1.1
Inv LQS tuning	**147**	**349**	**77**	**22**	1.9	5.3	**0.5**	-0.1
Inv quantile score	161	393	**77**	**22**	**2.1**	**4.3**	**0.5**	**1.6**
Inv quantile score shrink	161	392	**77**	**22**	1.9	**4.3**	0.3	1.1
Inv quantile score tuning	**150**	**357**	**77**	**22**	**2.5**	4.2	**0.4**	**2.9**

Note: The unit of the score is deaths. Lower values of the score and higher values of the skill score are better. Bold indicates the best three methods in each column.

In the rest of this section, we extend our empirical analysis to consider the following issues: statistical testing, consistency of the results across lead times and across the 40 week-period, model diversity, use of different numbers of forecasts in the combination, and evaluation of average ranks.

In this paper, we have presented results averaged across lead times. As far as we are aware, statistical tests are not available to compare results averaged in this way from different methods. The work of Quaedvlieg (2021) considers multi-horizon comparisons, but it is not applicable to our study where we expand the length of the estimation sample for each new forecast origin, which we feel would be done in practice because our time series are short. To consider statistical testing, we focus on each lead time separately. This also enables us to compare accuracy across lead times. Table 5 presents LQS results for each of the four lead times for a subgroup of the combining methods. The symbols * and † indicate methods with LQS that is significantly less than the simple average and median combinations, respectively, using a 5% significance level.[1] For all the series considered together and the high mortality

[1] We based our test on the Diebold-Mariano test, for which the test statistic is the difference between the mean of an accuracy measure for two methods for a single time series (Diebold and Mariano, 2002). We needed to compare the difference averaged across multiple series. For this test statistic, we computed the variance of the sampling distribution by first summing each variance of the sampling distribution from the Diebold-Mariano test applied to each series, and then dividing this by the square of the number of series.

series, the table shows that the inverse score methods were generally significantly more accurate than the simple average and the median. For the medium mortality series, the inverse score methods were significantly more accurate than the median, but generally not significantly more accurate than the simple average. There are no cases of significance for the low mortality series. Finally, we note the table shows similar rankings of the methods for each lead time.

Table 5. LQS for the distributional forecasts for each of the four lead times, averaged over the 30-week out-of-sample period.

	LQS				Skill Score (%)			
	All	High	Medium	Low	All	High	Medium	Low
1 week-ahead								
Ensemble	592	1450	284	74	2.7	8.0	-0.9	1.1
Simple average	617	1532	276	74	0.0	0.0	0.0	0.0
Geometric mean	641	1591	280	84	-5.4	-3.6	-1.7	-10.9
Median	597	1466	284	73	3.1	8.0	-1.1	2.3
Inv LQS	570*	1395*	272†	**72**	4.8	8.7	2.6	**3.0**
Inv LQS tuning	538*†	1297*	272†	**72**	4.9	**10.5**	2.5	1.7
Inv quantile score	563*	1378*	**270**†	**72**	4.8	9.1	**2.7**	2.6
Inv quantile score tuning	**534***†	**1284***†	273†	**72**	**5.0**	10.4	1.9	2.6
2 weeks-ahead								
Ensemble	1003†	2458†	472	132	-1.2	1.3	-4.2	-0.6
Simple average	993	2449	450†	130	0.0	0.0	0.0	0.0
Geometric mean	1022	2514	456†	149	-4.9	-3.2	-1.2	-10.2
Median	1033	2544	476	131	-1.5	0.6	-5.3	0.0
Inv LQS	935*†	2272*†	449*†	130	2.2	5.8	0.5	0.3
Inv LQS tuning	**862***†	**2050***†	450†	**129**	**2.8**	**8.6**	0.1	-0.3
Inv quantile score	925*†	2245*†	**446**†	130	2.4	6.1	**0.8**	0.3
Inv quantile score tuning	872*†	2078*†	451†	**129**	**2.8**	7.7	0.3	**0.5**
3 weeks-ahead								
Ensemble	1492†	3664†	698†	192	-2.8	-2.9	-4.0	-1.6
Simple average	1453	3580	666†	**188**	0.0	0.0	0.0	**0.0**
Geometric mean	1496	3670	673†	220	-7.2	-2.5	-1.1	-18.2
Median	1551	3833	708	193	-3.9	-4.5	-5.7	-1.7
Inv LQS	1371*†	3328*†	666†	**188**	1.3	4.1	0.1	-0.3
Inv LQS tuning	**1273**†	**3025**†	668†	**188**	0.9	**4.7**	-0.4	-1.7
Inv quantile score	1352*†	3277*†	**660**†	**188**	**1.5**	4.3	**0.5**	-0.3
Inv quantile score tuning	1292*†	3085*†	668†	**188**	0.9	3.4	0.1	-0.7
4 weeks-ahead								
Ensemble	2097†	5156†	969†	274	-3.0	-5.1	-3.3	-0.9
Simple average	2034†	5007	933†	266	0.0	0.0	0.0	0.0
Geometric mean	2089	5133	940†	298	-6.3	-5.0	-0.5	-13.3
Median	2194	5435	985	274	-4.1	-7.1	-5.0	-0.5
Inv LQS	1902*†	4608*†	931†	263	2.3	**4.9**	0.5	1.4
Inv LQS tuning	**1787**†	**4261**†	927†	261	1.7	4.0	0.2	0.8
Inv quantile score	1874*†	4536*†	**916**†	264	**2.6**	**4.9**	**1.8**	1.3
Inv quantile score tuning	1804*†	4312*†	928†	**262**	1.7	2.6	0.8	**1.7**

Note: The unit of LQS is deaths. Lower LQS values are better. For each lead time, bold indicates the best method for each of the four categories of series: all series, high, medium and low. * and † indicate a score significantly lower than the simple average and median, respectively, at the 5% significance level.

Table 6. LQS for the distributional forecasts, averaged over each 10-week period.

	LQS				Skill Score (%)			
	All	High	Medium	Low	All	High	Medium	Low
Weeks 1-10								
Ensemble	1349	3528	476	114	-0.8	-3.1	-4.6	4.9
Simple average	1254	3256	455	117	0.0	0.0	0.0	0.0
Geometric mean	**1212**	**3210**	397	95	13.0	7.9	14.1	16.7
Median	1220	3270	**369**	**86**	**20.3**	**10.9**	**20.1**	**28.5**
Inv LQS	NA	NA	NA	NA	NA	NA	NA	NA
Inv LQS tuning	NA	NA	NA	NA	NA	NA	NA	NA
Inv quantile score	NA	NA	NA	NA	NA	NA	NA	NA
Inv quantile score tuning	NA	NA	NA	NA	NA	NA	NA	NA
Weeks 11-20								
Ensemble	832	2166	**311**	**65**	**13.6**	**17.6**	**8.4**	**14.4**
Simple average	855	2208	329	75	0.0	0.0	0.0	0.0
Geometric mean	930	2366	334	138	-16.7	-8.7	-1.8	-42.0
Median	865	2254	320	67	12.8	17.5	6.3	14.3
Inv LQS	786	2007	321	71	8.2	14.5	3.5	6.5
Inv LQS tuning	786	2007	321	71	8.2	14.5	3.5	6.5
Inv quantile score	**777**	**1982**	320	70	8.9	14.7	4.2	**7.5**
Inv quantile score tuning	**777**	**1982**	320	70	8.9	14.7	4.2	**7.5**
Weeks 21-30								
Ensemble	1120	2623	611	182	-9.5	-9.8	-7.9	-10.8
Simple average	1041	2460	548	**166**	0.0	0.0	0.0	**0.0**
Geometric mean	1048	2466	557	173	-0.8	0.2	0.0	-2.5
Median	1138	2673	614	182	-9.2	-9.0	-8.6	-10.0
Inv LQS	971	2233	558	170	-0.6	3.4	-1.0	-4.2
Inv LQS tuning	**849**	**1856**	560	172	-0.2	**6.6**	-1.5	-5.9
Inv quantile score	954	2195	**543**	170	**0.1**	3.5	**1.3**	-4.5
Inv quantile score tuning	878	1945	555	174	-1.2	4.3	-0.9	-7.0
Weeks 31-40								
Ensemble	2107	5187	966	276	-2.4	-4.3	-3.3	0.2
Simple average	2099	5192	939	273	0.0	0.0	0.0	0.0
Geometric mean	2130	5291	943	266	0.7	-1.2	-0.5	3.5
Median	2212	5500	978	272	-3.4	-7.2	-5.1	1.5
Inv LQS	1991	4879	929	267	**1.6**	**1.8**	0.4	2.5
Inv LQS tuning	**1868**	**4510**	**927**	262	0.9	0.3	**0.8**	1.6
Inv quantile score	1970	4811	928	270	1.3	1.5	0.2	2.1
Inv quantile score tuning	1881	4542	937	**261**	1.1	-0.5	0.0	**3.7**

Note: The unit of LQS is deaths. Lower LQS values are better. NA indicates not available. For each 10-week period, bold indicates the best method for each of the four categories of series: all series, high, medium and low.

To address whether the ranking of the methods varies over the 40 weeks of our dataset, Table 6 presents the LQS separately for the four successive 10-week periods. We have averaged the LQS across lead times, and we consider the same subgroup of methods as in Table 5. Note that the trimming and inverse score methods were not available for the first 10-week period, as this was the first in-sample estimation period for these methods. Although insight is limited from such short periods, especially for probabilistic forecasts, it is interesting to see that the inverse score methods were reasonably consistent

in performing well across the 10-week periods, and the median was very competitive for the first two 10-week periods.

In Section 2, we discussed the different methods used by the individual forecasting teams. We described how compartmental models were used by approximately half the teams, and we illustrated this in Figure 4. We were curious to see how the combining methods would perform if they were applied to only the teams using compartmental models. Given their widespread use by epidemiologists, one might surmise that combining only these models would be adequate, and that a combination using only the other types of models would deliver poor results. We investigate these issues in Table 7, where we report LQS results produced by applying each of the combining methods to the following three different sets of the individual forecasting teams: all the teams, the teams using compartmental models, and the teams not using compartmental models. For the low mortality series, Table 7 shows that combining only compartmental models was most accurate. For the medium mortality series, combining only compartmental models was more accurate than combining only non-compartmental models, and there was only a small benefit in including the latter in a combination with the former. For the high mortality series, the best results were produced by combining both types of model.

Table 7. LQS for the distributional forecasts, averaged over the 30-week out-of-sample period, for combining methods applied to three different sets of individual forecasts: all, compartmental models only, and non-compartmental models.

	All series			High			Medium			Low		
	All	Comp	Non-Comp	All	Comp	Non-Comp	All	Comp	Non-Comp	All	Comp	Non-Comp
Simple benchmarks methods												
Ensemble	1296	NA	NA	3182	NA	NA	606	NA	NA	168	NA	NA
Simple average	1274	1327	1381	3142	3294	3394	581	594	634	164	**161**	187
Geometric mean	1312	1351	1401	3227	3360	3432	587	601	637	188	163	206
Median	1344	1390	1431	3320	3445	3536	614	632	650	168	165	180
Trimming methods												
CDF ext trim	1317	1369	1374	3253	3403	3380	600	614	631	165	**162**	182
CDF int trim	1330	1353	1464	3311	3367	3600	582	598	668	166	163	198
Mean ext trim	1308	1368	1380	3230	3403	3393	596	609	636	166	163	181
Mean int trim	1298	1356	1441	3212	3373	3549	584	603	650	166	163	196
Score-based methods												
Previous best	1495	1474	1520	3496	3463	3510	833	803	887	229	229	238
Inv LQS	1194	1277	1239	2901	3153	2972	579	582	630	163	**162**	178
Inv LQS shrink	1200	1279	1244	2920	3151	2991	579	588	627	163	163	177
Inv LQS tuning	**1115**	**1163**	1195	**2658**	**2799**	2833	579	582	636	163	166	177
Inv quantile score	1179	1253	1230	2859	3087	2946	**573**	**576**	626	163	**161**	179
Inv quantile score shrink	1197	1277	1243	2911	3145	2988	**578**	587	628	163	163	177
Inv quantile score tuning	**1125**	1212	1211	**2690**	2947	2876	580	582	641	163	167	178

Note: The unit of LQS is deaths. Lower LQS values are better. NA indicates not available. Bold indicates the best three methods for each of the four categories of series: all series, high, medium and low.

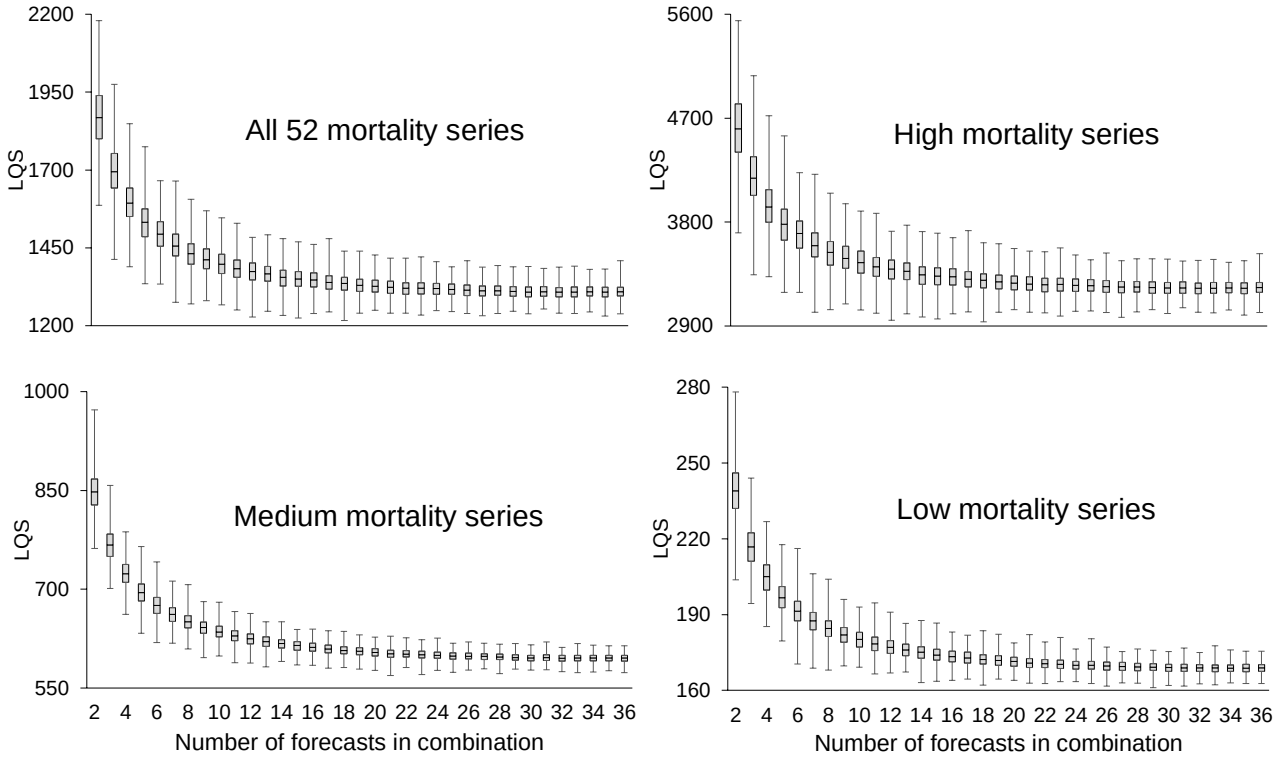

Figure 9. LQS for distributional forecasts from simple average combinations of different numbers of forecasts. LQS averaged over the 30-week out-of-sample period.

We investigated the impact on distributional forecast accuracy of using different numbers of forecasts in the combinations. We represent this number by *K*. In Section 2.3, we described how the availability of the forecasting teams varied across the series and forecast origins. The number of teams available varied between 6 and 36, with a median of 27. To investigate different values of *K*, for each location and forecast origin, we sampled *K* forecasts, with replacement, from the available forecasts, and evaluated combinations of the *K* forecasts. We did this 1,000 times for *K*=2 to 36. For each value of *K*, and simple average combining, Figure 9 uses a Box plot to summarise the resulting 1,000 LQS values. Each panel shows a noticeable improvement in the LQS as *K* increases to about 20. In fact, the LQS for each of the four categories of series continued to improve slightly up to about *K*=30, indicating the benefit of using a large pool of forecasts.

Our final set of results consists of ranks of the methods, averaged over the series within each of the four categories. Table 8 reports the average ranks for the LQS and MAE. (Prior to computing the ranks, each score was averaged across the four lead times.) Lower ranks are better. The average ranks provide a similar message to the average scores in Tables 3 and 4, with the inverse score methods performing the best. We implemented the statistical test for average ranks proposed by Koning et al. (Section 2.2, 2005) to enable multiple comparisons with the best method in each column of Table 8. In each column, the best average rank is highlighted in bold, and * indicates a method that has average rank significantly worse than the best method in that column, using a 5% significance level. Significance can be seen in all cases for the 'previous best' method, and in most cases for the median combining method.

Table 8. Average ranks of the LQS and MAE computed for the 30-week out-of-sample period.

	LQS				MAE			
	All	High	Medium	Low	All	High	Medium	Low
Simple benchmarks methods								
Ensemble	9.4*	9.7	9.3	9.2	8.8	9.2	7.8	9.4
Simple average	6.7	7.2	6.5	**6.4**	6.9	7.1	**5.9**	7.6
Geometric mean	8.9*	8.8	8.9	9.1	9.2*	9.7	8.7	9.2
Median	10.2*	11.5*	11.3*	7.9	9.5*	10.6*	9.8	8.2
Trimming methods								
CDF ext trim	9.6*	11.2*	9.9	7.6	9.1	11.3*	8.9	7.2
CDF int trim	7.7	8.3	6.9	7.9	7.8	8.9	6.5	8.1
Mean ext trim	8.4	9.0	8.6	7.6	8.3	9.2	8.7	7.1
Mean int trim	8.0	8.4	7.9	7.7	7.8	8.8	7.2	7.3
Score-based methods								
Previous best	13.7*	12.6*	14.5*	14.0*	12.9*	11.6*	12.8*	14.1*
Inv LQS	**5.8**	**4.5**	5.8	7.0	6.8	**4.9**	7.8	7.7
Inv LQS shrink	6.3	5.0	6.8	6.9	6.9	5.6	7.4	7.6
Inv LQS tuning	6.9	6.4	6.4	7.8	6.9	5.6	7.6	7.4
Inv quantile score	5.9	5.3	**5.2**	7.0	**6.2**	5.4	6.6	6.5
Inv quantile score shrink	5.9	5.0	6.1	6.6	6.7	5.7	7.2	7.2
Inv quantile score tuning	6.6	6.8	5.7	7.3	6.2	6.2	7.1	**5.4**

Note: Lower values are better. In each column, bold indicates the best method, and * indicates a value significantly worse than the best method, at the 5% significance level.

5. Summary and Concluding Comments

We have provided an empirical comparison of combining methods for interval and distributional forecasts of cumulative mortality due to COVID-19. The forecasts were produced by teams using a variety of approaches, including compartmental and statistical models. Combining provides a pragmatic way to synthesise the diverse information underlying these models. For combining probabilistic forecasting, methods proposed in the literature include the simple average, which is a natural benchmark, as well as the median and trimming methods, which enable robust estimation and adjustment for the case where forecasters tend to be under- or overconfident. For applications, such as ours, where there is frequent entry and exit of participating forecasting teams, it is not clear how best to form weighted forecast combinations, as a comparable history of accuracy is not available for the teams. We took a pragmatic approach and implemented weighted combinations based on the inverse of appropriate scoring functions, computed using whatever historical forecasts were available for each method. We are not aware of previous studies that have used this approach for probabilistic forecasting with data that has frequent entry and exit of forecasters. For our out-of-sample period of the most recent 30 weeks, these weighted combinations outperformed all other methods for high mortality series, while for the other series, accuracy of these methods matched the best of the other combining methods, which was the simple average. For the first 10 weeks of our dataset, insufficient historical accuracy was available with which to construct the weighted combinations. For these early weeks, the median was overall the most accurate method.

Acknowledgements

We are very grateful to all the forecasting groups who generously made their forecasts available on the Reich Lab COVID-19 Forecast Hub, and to Nicholas Reich and his team, for acting as curators for the Hub, and for providing such convenient access to the data, along with useful supplementary information. We would like to thank Nia Roberts for clarifying our understanding of the licence terms for the forecast data. We are also very grateful to four referees for providing very useful comments on the paper.

References

Adam, D. (2020). Special report: The simulations driving the world's response to COVID-19. *Nature*, 580. 10.1038/d41586-020-01003-6

Bates, J.M., & Granger, C.W.J. (1969). The combination of forecasts. *Journal of the Operational Research Society*, 20(4), 451-468.

Bracher, J., Ray, E. L., Gneiting, T., & Reich, N. G. (2021). Evaluating epidemic forecasts in an interval format. *PLoS Computational Biology*, 17(2), e1008618.

Brehmer, J., & Gneiting, T. (2020). Scoring interval forecasts: Equal-tailed, shortest, and modal interval. arXiv preprint arXiv: 2007.05709.

Brown, A., & Reade, J. J. (2019). The wisdom of amateur crowds: Evidence from an online community of sports tipsters. *European Journal of Operational Research*, 272(3), 1073-1081.

Budescu, D. V., & Chen, E. (2015). Identifying expertise to extract the wisdom of crowds. *Management Science*, 61(2), 267-280.

Capistrán, C., & Timmermann, A. (2009). Forecast combination with entry and exit of experts. *Journal of Business & Economic Statistics*, 27(4), 428-440.

COVID-19 Forecast Hub (2020). https://github.com/reichlab/covid19-forecast-hub [accessed 19 July 2020]

Da'ar, O.B., & Ahmed, A.E. (2018). Underlying trend, seasonality, prediction, forecasting and the contribution of risk factors: an analysis of globally reported cases of Middle East Respiratory Syndrome Coronavirus. *Epidemiology and Infection*, 146(11), 1343-1349.

Dawid, A.P., DeGroot, M.H., & Mortera, J. (1995). Coherent combination of experts' opinions. *Test*, 4(2), 263-313.

Diebold, F. X., & Mariano, R. S. (2002). Comparing predictive accuracy. *Journal of Business & Economic Statistics*, 20(1), 134-144.

Gaba, A., Tsetlin, I., & Winkler, R.L. (2017). Combining interval forecasts. *Decision Analysis*, 14(1), 1-20.

Gianfreda, A., & Bunn, D.W. (2018). A stochastic latent moment model for electricity price formation. *Operations Research*, 66(5), 1189-1203.

Gneiting, T., Balabdaoui, F., & Raftery, A.E. (2007). Probabilistic forecasts, calibration and sharpness. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 69(2), 243-268.

Gneiting, T., & Katzfuss, M. (2014). Probabilistic forecasting. *Annual Review of Statistics and Its Application*, 1, 125-151.

Gneiting, T., & Ranjan, R. (2011). Comparing density forecasts using threshold- and quantile-weighted scoring rules. *Journal of Business & Economic Statistics*, 29(3), 411-422.

Gneiting, T., & Raftery, A. (2007). Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, 102(477), 359-378.

Grushka-Cockayne, Y., & Jose, V.R.R. (2020). Combining prediction intervals in the M4 competition. *International Journal of Forecasting*, 36(1), 178-185.

Grushka-Cockayne, Y., Jose, V.R.R., & Lichtendahl KC Jr (2017). Ensembles of overfit and overconfident forecasts. *Management Science*, 63(4), 1110-1130.

Guha, S., & Kumar, S. (2018). Emergence of big data research in operations management, information systems, and healthcare: Past contributions and future roadmap. *Production and Operations Management*, 27(9), 1724-1735.

Holmdahl, I., & Buckee, C. (2020). Wrong but Useful - What Covid-19 Epidemiologic Models Can and Cannot Tell Us. *New England Journal of Medicine*, 10.1056

Hora, S.C. (2004). Probability judgments for continuous quantities: linear combinations and calibration. *Management Science*, 50(5), 597-604.

Hora, S.C., Fransen B.R., Hawkins, N., & Susel, I. (2013). Median aggregation of distribution functions. *Decision Analysis*, 10(4), 279-291.

Jose, V.R.R., Grushka-Cockayne, Y., & Lichtendahl, K.C. Jr. (2014). Trimmed opinion pools and the crowd's calibration problem. *Management Science*, 60(2), 463-475.

Kobres, P.Y., Chretien, J.P., Johansson, M.A., et al. (2019). A systematic review and evaluation of Zika virus forecasting and prediction research during a public health emergency of international concern. *PLoS Neglected Tropical Diseases*, 13.10

Koenker, R., & Machado, J.A. (1999). Goodness of fit and related inference processes for quantile regression. *Journal of the American Statistical Association*, 94(448), 1296-1310.

Koning, A. J., Franses, P. H., Hibon, M., & Stekler, H. O. (2005). The M3 competition: Statistical tests of the results. *International Journal of Forecasting*, 21(3), 397-409.

Larrick, R.P., & Soll, J.B. (2006). Intuitions about combining opinions: Misappreciation of the averaging principle. *Management Science*, 52(1), 111-127.

Lauer, S.A., Brown, A.C., & Riech, N.G. (2020). Infectious disease forecasting for public health. In Drake JM, Bonsall, Strand M (editors). *Population biology of vector-borne diseases*. Oxford University Press.

Lichtendahl, K.C. Jr., Grushka-Cockayne, Y., & Winkler, R.L. (2013). Is it better to average probabilities or quantiles?. *Management Science*, 59(7), 1594-1611

Mote, P.W., Allen, M.R., Jones, R.G., Li, S., Mera, R., Rupp, D.E., Salahuddin, A. & Vickers, D. (2016). Superensemble regional climate modeling for the western United States. *Bulletin of the American Meteorological Society*, 97(2), 203-215.

Ng, T.W., Turinici, G., & Danchin, A. (2003). A double epidemic model for the SARS propagation. *BMC Infectious Diseases*, 3:19.

Nikolopoulos, K., Punia, S., Schäfers, A., Tsinopoulos, C., & Vasilakis, C. (2021). Forecasting and planning during a pandemic: COVID-19 growth rates, supply chain disruptions, and governmental decisions. *European Journal of Operational Research*, 290(1), 99-115.

Nolde, N., & Ziegel, J.F. (2017). Elicitability and backtesting: Perspectives for banking regulation. *Annals of Applied Statistics*, 11(4), 1833-1874.

Park, S., & Budescu, D.V. (2015). Aggregating multiple probability intervals to improve calibration. *Judgment and Decision Making*, 10(2), 130-143

Phelan, A.L., Katz, R., & Gostin, L.O. (2020). The novel coronavirus originating in Wuhan, China: Challenges for global health governance. *Journal of the American Medical Association*, 323(8), 709–710.

Ranjan, R., & Gneiting, T. (2010). Combining probability forecasts. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 72(1), 71-91.

Quaedvlieg, R. (2021). Multi-horizon forecast comparison. *Journal of Business & Economic Statistics*, 39(1), 40-53.

Shan, K., & Yang, Y. (2009). Combining regression quantile estimators. *Statistica Sinica*, 1171-1191.

Shi, Y., Liu, X., Kok, S.Y., et al. (2016). Three-Month Real-Time Dengue Forecast Models: An Early Warning System for Outbreak Alerts and Policy Decision Support in Singapore. *Environmental Health Perspectives*, 124(9), 1369-1375.

Sridhar, D., & Majumder, M.S. (2020). Modelling the pandemic. *BMJ*;369: m1567.

Stock, J.H., & Watson, M. (2001). A comparison of linear and nonlinear univariate models for forecasting macroeconomic time series. In R.F. Engle, & White, H. (eds.), *Cointegration, Causality and Forecasting: A Festschrift in Honour of Clive Granger*, pp. 1-44. Cambridge, Cambridge University Press.

Stock, J. H., & Watson, M. W. (2004). Combination forecasts of output growth in a seven-country data set. *Journal of Forecasting*, 23(6), 405-430.

Stone, M. (1961). The opinion pool. *The Annals of Mathematical Statistics*, 32(4), 1339-1342.

Surowiecki, J. (2004). *The wisdom of crowds: why the many are smarter than the few*. New York: Doubleday

Taylor, J.W. (1999). Evaluating volatility and interval forecasts. *Journal of Forecasting*, 18(2), 111-128.

Taylor, J. W. (2020). Forecast combinations for value at risk and expected shortfall. *International Journal of Forecasting*, 36(2), 428-441.

Viboud, C., Sun, K., Gaffey, R. et al. (2017). The RAPIDD Ebola forecasting challenge: Synthesis and lessons learnt. *Epidemics*, 22, 13-21.

Weinberger, D.M., Chen, J., Cohen, T., et al. (2020). Estimation of excess deaths associated with the COVID-19 pandemic in the United States, March to May 2020. *JAMA Internal Medicine*, 180(10), 1336-1344.

Winkler, R.L. (1972). A decision-theoretic approach to interval estimation. *Journal of the American Statistical Association*, 67(337), 187-191.

Winkler, R.L., Grushka-Cockayne, Y., Lichtendahl, K.C. Jr., & Jose, V.R.R. (2019). Probability forecasts and their combination: a research perspective. *Decision Analysis*, 16(4), 239-260.

World Health Organization (2020). Coronavirus disease (COVID-19). Situation Report – 209. https://www.who.int/emergencies/diseases/novel-coronavirus-2019/situation-reports [accessed 28 February 2021]

World Health Organization (2021). WHO Coronavirus Disease (COVID-19) Dashboard. https://covid19.who.int/ [accessed 28 February 2021]

Supplementary Information

The terms and conditions of the forecasts that were analysed are recorded in the forecasting groups' supplementary files on the Reich Lab COVID-19 Forecast GitHub website: https://github.com/reichlab/COVID19-forecast-hub/tree/master/data-processed.
A number of the forecasting teams released their data under one of the following licences:
https://creativecommons.org/licenses/by/4.0/,
https://creativecommons.org/licenses/by-nc/4.0/,
https://creativecommons.org/licenses/by-nc-sa/4.0/

Details of forecast models, ordered by model name

Contributors and citations	**Model (short name)**	**Model description***	**Access information**
Wattanachit N, Ray EL, Reich N https://www.medrxiv.org/content/10.1101/2020.08.19.20177493v1 https://www.medrxiv.org/content/10.1101/2021.02.03.21250974v1	COVID hub-ensemble	An ensemble, or model average, of submitted forecasts to the COVID-19 Forecast Hub.	https://github.com/reichlab/covid19-forecast-hub/tree/master/data-processed/COVIDhub-ensemble
COMPARTMENTAL MODELS			
Tomar V, Jain C	Auquan-SEIR	Modified SEIR model with compartments for reported and unreported infections. Non-linear mixed effects curve-fitting.	https://github.com/reichlab/covid19-forecast-hub/tree/master/data-processed/Auquan-SEIR
Panano B. https://bobpagano.com/covid-19-modeling/	BPangano-RtDriven	Projects infections and deaths for 223 locations using an SIR model.	https://github.com/reichlab/covid19-forecast-hub/tree/master/data-processed/BPagano-RtDriven
Carlson E, Henderson M, Kelly C, Kofman I, Zhang X	CovidActNow-SEIR_CAN	SEIR model forecasts of cumulative deaths, incident deaths, incident hospitalizations by fitting predicted cases, deaths, and hospitalizations to the observations.	https://github.com/reichlab/covid19-forecast-hub/tree/master/data-processed/CovidActNow-SEIR_CAN

Contributors and citations	Model (short name)	Model description*	Access information
Li ML, Bouardi HT, Lami OS, Trikalinos TA, Trichakis NK, Bertsimas D https://www.covidanalytics.io/DELPHI_documentation_pdf	CovidAnalytics-DELPHI	SEIR model augmented with underdetection and interventions. Projections account for reopening and assume interventions would be re-enacted if cases continue to climb.	https://github.com/reichlab/covid19-forecast-hub/tree/master/data-processed/CovidAnalytics-DELPHI
Chhatwal J, Ayer T, Linas B, Dalgic O, Mueller P, Adee M, Ladd MA, Xiao J (Mass General Hospital, Harvard Medical School, Georgia Tech and Boston Medical Centre)	Covid19Sim-Simulator	An interactive tool that uses a validated SEIR compartment model.	https://github.com/reichlab/covid19-forecast-hub/tree/master/data-processed/Covid19Sim-Simulator
Pei S, Yamana T, Kandula S, Yang W, Galanti M, Shaman J https://doi.org/10.1101/2020.03.21.20040303	CU-select	Metapopulation county-level SEIR model for projecting future COVID-19 incidence and deaths. This forecast is the scenario we believe to be most plausible given the current setting.	https://github.com/reichlab/covid19-forecast-hub/tree/master/data-processed/CU-select
Pei S, Yamana T, Kandula S, Yang W, Galanti M, Shaman J https://doi.org/10.1101/2020.03.21.20040303	CU-nochange	This metapopulation county-level SEIR model assumes that current contact rates will remain unchanged in the future.	https://github.com/reichlab/covid19-forecast-hub/tree/master/data-processed/CU-nochange
Max A, Epshteyn A, Kang B, Li C-L, Sava D, Parish D, Miller D, Kanal E, Liu H, Nakhost H, Jones I, Lai J, Repenning J, Yoon J, Ramasamy K, Zhang L, Le L, Nikoltchev M, Siegler M, Dusenberry M, Yoder N, Rozenfeld O, Rangaswamy P, Sinha R, Xie R, Arik S, Singh S, Tsai T, Pfister T, Menon V, Karande V, Y, Li Y https://arxiv.org/abs/2008.00646	Google-Harvard-CPF	Our model improves upon standard compartmental models by using temporally and spatially rich data, and integrating covariate encodings into compartment transitions via end-to-end learning.	https://github.com/reichlab/covid19-forecast-hub/tree/master/data-processed/Google_Harvard-CPF
Lemaitre JC, Bi Q , Hulse JD, Grabowski MK, Grantz KH, Kaminsky J, Lauer SA, Lee EC, Meredith HR, Perez-Saez J, Truelove SA,	JHU_IDD-CovidSP	County-level metapopulation model with commuting and stochastic SEIR	https://github.com/reichlab/covid19-forecast-

Contributors and citations	Model (short name)	Model description*	Access information
Keegan LT, Kaminsky K, Shah S, Wills J, Aquilanti P-Y, Raman K, Subramaniyan A, Thursam G, Tran A. https://www.medrxiv.org/content/10.1101/2020.06.11.20127894v1		disease dynamics with social-distancing indicators.	hub/tree/master/data-processed/JHU_IDD-CovidSP
Kinsey M, Tallaksen K, Obrecht RF, Asher L, Costello C, Kelbaugh M, Wilson S	JHUAPL_Bucky	Metapopulation model using public mobility data. Local parameters (case reporting rates, doubling times, etc) are estimated using data from CSSE and CDC scenario 5. Primary output is case incidence.	https://github.com/reichlab/covid19-forecast-hub/tree/master/data-processed/JHUAPL-Bucky
Baek J, Farias V, Georgescu A, Levi R, Sinha D, Wilde J, Zheng A https://arxiv.org/abs/2006.06373	MITCovAlliance-SIR	SIR model trained on public heath regions. SIR parameters are functions of static demographic and time-varying mobility features. Two-stage approach that first learns magnitude of peak infections.	https://github.com/reichlab/covid19-forecast-hub/tree/master/data-processed/MITCovAlliance-SIR
Vespignani A, Chinazzi M, Davis JT, Mu K, Pastore y Piontti A, Samay N, Xiong X, Halloran ME, Longini IM, Dean NE, Viboud C, Sun K, Litvinova M, Gioannini C, Rossi L, Ajelli M https://uploads-ssl.webflow.com/58e6558acc00ee8e4536c1f5/5e8bab44f5baae4c1c2a75d2_GLEAM_web.pdf	MOBS-GLEAM_COVID	Metapopulation, age structured SLIR model. Superimposed on the worldwide population and mobility layers is an agent-based epidemic model that defines the infection and population dynamics. Makes predictions about the future that are dependent on the assumption that current interventions continue.	https://github.com/reichlab/covid19-forecast-hub/tree/master/data-processed/MOBS-GLEAM_COVID
Gao Z, Li C, Zheng S, Bian J, Xie X, LiuT-Y	MSRA-DeepST	A deep spatio-temporal network with knowledge based SEIR as a regularier under the assumption of spatio-temporal process in pandemic of different regions.	https://github.com/reichlab/covid19-forecast-hub/tree/master/data-processed/MSRA-DeepST
Espana G, Oidtman R,	NotreDame-Mobility	Ensemble of nine models that are identical except that they are driven by different mobility indices from	https://github.com/reichlab/covid19-forecast-hub/tree/master/data-

Contributors and citations	Model (short name)	Model description*	Access information
Cavany S, Costello A, Wieler A, Lerch A, Barbera C, Poterek M, Tran Q, Moore S, Perkins A		Apple and Google. The model underlying each is a deterministic, SEIR-like model.	processed/NotreDame-mobility
Koyluoglu U, Milliken J	OliverWyman-Navigator	Forecasts and scenario analysis for Detected and Undetected cases and death counts following a compartmental formulation with non-stationary transition rates.	https://github.com/reichlab/covid19-forecast-hub/tree/master/data-processed/OliverWyman-Navigator
Turtle J, Ben-Nun M, Riley P	PSI-DRAFT	A stochastic/deterministic, single-population SEIRX model that stratifies by both age distribution and disease severity and includes generic intervention fitting.	https://github.com/reichlab/covid19-forecast-hub/tree/master/data-processed/PSI-DRAFT
Shi Y, Shah T, Ban X https://www.medrxiv.org/content/10.1101/2020.07.25.20162016v1	RPI-UW-Mob_Collision	A mobility-informed simplified SIR model motivated by collision theory.	https://github.com/reichlab/covid19-forecast-hub/tree/master/data-processed/RPI-UW-Mob-Collision
Snyder TL, Wilson DD	SWC-TerminusCM	Mechanistic compartmental model using disease parameter estimates from literature. It uses Bayesian inference to predict the most likely model parameters.	https://github.com/reichlab/covid19-forecast-hub/tree/master/data-processed/SWC-TerminusCM
Cobey S, Arevalo P, Baskerville E, Carran S, Gostic K, McGough L, Ranjeva S, Wen F	UChicago-COVIDIL	Compartmental, age-structured SEIR model that infers past SARS-CoV-2 transmission rates and forecasts mortality under current and hypothetical public health interventions.	https://github.com/reichlab/covid19-forecast-hub/tree/master/data-processed/UChicago-CovidIL
Gu Q, Xu P, Chen J, Wang L, Zou D, Zhang W https://www.medrxiv.org/content/10.1101/2020.05.24.20111989v1	UCLA-SuEIR	Variant of the SEIR model considering both untested and unreported cases. The model considers reopening and assumes susceptible population will increase after the reopen.	https://github.com/reichlab/covid19-forecast-hub/tree/master/data-processed/UCLA-SuEIR
Chen YQ, Zhao Y, Guo L	UCM-MESALab-FoGSEIR	FoGSEIR model is a modification of integer order SEIR model considering	https://github.com/reichlab/covid19-forecast-

Contributors and citations	Model (short name)	Model description*	Access information
		fractional integrals. The model considers the age structure and reopening intervention to minimize infections and deaths.	hub/tree/master/data-processed/UCM_MESALab-FoGSEIR
Sheldon D, Gibson G, Reich N	UMass-MechBayes	Bayesian compartmental model with observations on cumulative case counts and cumulative deaths. Model is fit independently to each state. Model includes observation noise and a case detection rate.	https://github.com/reichlab/covid19-forecast-hub/tree/master/data-processed/UMass-MechBayes
Mayo ML, Rowland MA, Parno MD, Detwiller ID, Farthing MW, England WP George GE	USACE-ERDC_SEIR	The ERDC SEIR model makes predictions of several variables (e.g., reported new/cumulative cases per day, etc.). Model parameters are estimated using historical data using Bayesian inference.	https://github.com/reichlab/covid19-forecast-hub/tree/master/data-processed/USACE-ERDC_SEIR
Jain S, Tiwari A, Deva A, Kulkarni M, Shingi S, Bannur N, White J, Merugu S, Raval A	Wadhwani_AI-BayesOpt	A novel model-agnostic Bayesian optimization ("BayesOpt") approach for learning the parameters of the SEIR model from observed data.	https://github.com/reichlab/covid19-forecast-hub/tree/master/data-processed/Wadhwani_AI-BayesOpt
Gu Y https://covid19-projections.com/about/	YYG-ParamSearch	Based on the SEIR model with hyperparameter optimization to make daily projections regarding COVID-19 infections and deaths in 50 US states. The model accounts for state reopenings and its effects on infections and deaths.	https://github.com/reichlab/covid19-forecast-hub/tree/master/data-processed/YYG-ParamSearch
OTHER MODELS			
O'Dea E	CEID-Walk	A random walk model with drift. A least squares line is fitted to the tail observations of a target time series to estimate the drift and step variance of a random walk model.	https://github.com/reichlab/covid19-forecast-hub/blob/master/data-processed/CEID-Walk/metadata-CEID-Walk.txt

Contributors and citations	Model (short name)	Model description*	Access information
Wang Y, Zeng D, Wang Q, Xie S https://www.frontiersin.org/article/10.3389/fpubh.2020.00325	Columbia_UNC-SurvCon	Survival-convolution model with piece-wise transmission rates that incorporates latent incubation period and provides time-varying effective reproductive number.	https://github.com/reichlab/covid19-forecast-hub/tree/master/data-processed/Columbia_UNC-SurvCon
Ray EL, Tibshirani R	COVIDhub-baseline	Baseline prediction model.	https://github.com/reichlab/covid19-forecast-hub/tree/master/data-processed/COVIDhub-baseline
Kalantari R, Zhou M. https://dds-covid19.github.io/	DDS-NBDS	Jointly modeling daily deaths and cases using a negative binomial distribution based nonparametric Bayesian generalized linear dynamical system.	https://github.com/reichlab/covid19-forecast-hub/tree/master/data-processed/DDS-NBDS
Sherratt K, Bosse N, Abbott S, Hellewell J, Meakin S, Munday J, Funk S https://doi.org/10.12688/wellcomeopenres.16006.1	epiforecasts-ensemble1	A deaths forecast using the renewal equation and time-series forecasts of the time-varying reproduction number.	https://github.com/reichlab/covid19-forecast-hub/tree/master/data-processed/epiforecasts-ensemble1
Keskinocak P, Aglar BEO, Baxter A, Asplund J, Serban N	GT_CHHS-COVID19	Agent-based simulation model to project COVID19 infection spread.	https://github.com/reichlab/covid19-forecast-hub/tree/master/data-processed/GT_CHHS-COVID19
Prakash BA, Rodriguez A, Cui J, Tabassum A, Adhikari B, Sun J, Xiao D, Qiang C	GT-DeepCOVID	Data-driven approach based on deep learning for forecasting mortality and hospitalizations using syndromic, clinical, demographic, mobility and point-of-care data.	https://github.com/reichlab/covid19-forecast-hub/tree/master/data-processed/GT-DeepCOVID
Murry C and the IHME-CurveFitTeam https://www.medrxiv.org/content/10.1101/2020.03.27.20043752v1	IHME-CurveFit	Non-linear mixed effects curve-fitting. This model makes predictions about the future that are dependent on the assumption	https://github.com/reichlab/covid19-forecast-hub/tree/master/data-processed/IHME-CurveFit

Contributors and citations	Model (short name)	Model description*	Access information
		that current interventions continue.	
Wang L, Wang G, Gao L, Li X, Yu S, Kim M, Wang Y, Gu Z. https://arxiv.org/abs/2004.14103	IowaStateLW-STEM	A nonparametric space-time disease transmission model. The projections assume that the data used is reliable, the future will continue to follow the current pattern, and current interventions will remain the same till the end of forecasting period.	https://github.com/reichlab/covid19-forecast-hub/tree/master/data-processed/IowaStateLW-STEM
Marshall M, Gardner L, Drew C, Burman E, Nixon K	JHU_CSSE-DECOM	County-level, empirical machine learning model driven by epidemiological, mobility, demographic, and behavioral data.	https://github.com/reichlab/covid19-forecast-hub/blob/master/data-processed/JHU_CSSE-DECOM
Karlem D. https://arxiv.org/abs/2007.07156	Karlen-pypm	python Population Modeller	https://github.com/reichlab/covid19-forecast-hub/tree/master/data-processed/Karlen-pypm
Osthus D, Del Valle S, Manore C, Weaver B, Castro L, Shelley S, Smith M, Spencer J, Fairchild G, Travis Pitts T, Gerts D, Dauelsberg L, Daughton A, Gorris M, Hornbein B, Israel D, Parikh N, Shutt D, Ziemann A	LANL-GrowthRate	Statistical dynamical growth model accounting for population susceptibility. Makes predictions about the future, unconditional on particular intervention strategies.	https://github.com/reichlab/covid19-forecast-hub/tree/master/data-processed/LANL-GrowthRate
Gao Z, Li C, Cao W, Zheng S, Bian J, Xie X, Liu T-Y, Zhang S, Lavista Ferres J	Microsoft-DeepSTIA	A deep spatio-temporal network with intervention and hosptial gate under the assumption of spatio-temporal process in pandemic of different regions.	https://github.com/reichlab/covid19-forecast-hub/tree/master/data-processed/Microsoft-DeepSTIA
Espana G, Oidtman R, Cavany S, Costello A, Wieler A, Lerch A, Barbera C, Poterek M, Tran Q, Moore S, Perkins A	NotreDame-FRED	Agent-based model developed for influenza with parameters modified to represent the natural history of COVID-19	https://github.com/reichlab/covid19-forecast-hub/tree/master/data-processed/NotreDame-FRED
Walraven R	RobertWalraven-ESG	Multiple skewed gaussian distribution peaks fitted to raw data.	https://github.com/reichlab/covid19-forecast-

Contributors and citations	Model (short name)	Model description*	Access information
			hub/tree/master/data-processed/RobertWalraven-ESG
Nagraj VP, Turner SD, Hulme-Lowe C	SigSci-TS	Time series forecasting using ARIMA for case forecasts and lagged cases for death forecasts.	https://github.com/reichlab/covid19-forecast-hub/tree/master/data-processed/SigSci-TS
McConnell S, Donaldson B https://stevemcconnell.com/covid	SteveMcConnell_COVID Complete	National level and state level, near-term (1-4 week) fatality forecasts.	https://github.com/reichlab/covid19-forecast-hub/tree/master/data-processed/SteveMcConnell-CovidComplete
Bieggel H, Lega J	UA-EpiCovDA†	SIR mechanistic model with data assimilation. EpiCovDA is an extension of the EpiGro model. Model parameters are fit to Covid-19 data using a variational data assimilation method. A prior distribution of the parameters is estimated by fitting an SIR Incidence-Cumulative Cases curve to data from states that had at least 1000 cases by 04/01/2020.	https://github.com/reichlab/covid19-forecast-hub/tree/master/data-processed/UA-EpiCovDA
Wu D, Gao L, M Yian, Yu R, Vespignani A, Chinazzi M, Davis JT, Mu K, Pastore y Piontti A, Xiong X	UCSD-NEU_DeepGLEAM	Combines the signal of a discrete stochastic epidemic computational model GLEAM with a deep learning spatiotemporal forecasting framework to further improve predictions.'	https://github.com/reichlab/covid19-forecast-hub/tree/master/data-processed/UCSD_NEU-DeepGLEAM
Corsetti S, Schwarz T	UMich-RidgeTfReg	Nation-level model of confirmed cases and deaths based on ridge regression. No assumptions made about social distancing.	https://github.com/reichlab/covid19-forecast-hub/tree/master/data-processed/UMich-RidgeTfReg

Contributors and citations	Model (short name)	Model description*	Access information
Jin X, Wang Y-X, Yan X	UCSB-ACTS	Data-driven machine learning model makes predictions by referring to other regions with similar growth patterns and assuming the similar development will take place in the current region.	https://github.com/reichlab/covid19-forecast-hub/tree/master/data-processed/UCSB-ACTS
Srivastava A, Prasanna VK, Tianjian Xu F. https://arxiv.org/abs/2007.05180	USC-SI_kJalpha_RF	A heterogeneous infection rate model with human mobility for epidemic modeling. The model adapts to changing trends and provide predictions of confirmed cases and deaths.	https://github.com/reichlab/covid19-forecast-hub/tree/master/data-processed/USC-SI_kJalpha_RF
Woody S, et al. at the University of Texas	UT-Mobility	This model makes predictions assuming that social distancing patterns, as measured by anonymized mobile-phone GPS traces, remain constant in the future. Only models *first-wave deaths*.	https://github.com/reichlab/covid19-forecast-hub/tree/master/data-processed/UT-Mobility
Mehrotra P, Ivan JI, and the Walmart Labs COVID-19 Team	WalmartLabsML_LogForecasting	A logistic growth prophet forecasting model fit using case counts and deaths as features.The Model is built by Propeht model with logistic growths to forecast the US cumulative deaths. By sampling from uniform distribution to get the quantiles.	https://github.com/reichlab/covid19-forecast-hub/tree/master/data-processed/WalmartLabsML-LogForecasting

* Based on information recorded on the GIT Hub
† Classed as other model as it is an extension to a non-linear growth model with a prior distribution SIR curve fitted.