Australian National University

Crawford School of Public Policy

CAMA

Centre for Applied Macroeconomic Analysis

Testing the predictive accuracy of COVID-19 forecasts

CAMA Working Paper 52/2021
July 2021

Laura Coroneo
University of York

Fabrizio Iacone
Universit`a degli Studi di Milano
University of York

Alessia Paccagnini
University College Dublin
Centre for Applied Macroeconomic Analysis, ANU

Paulo Santos Monteiro
University of York

Abstract

We test the predictive accuracy of forecasts of the number of COVID-19 fatalities produced by several forecasting teams and collected by the United States Centers for Disease Control and Prevention during the first and second waves of the epidemic in the United States. We find three main results. First, at the short horizon (1-week ahead) no forecasting team outperforms a simple time-series benchmark. Second, at longer horizons (3- and 4-week ahead) forecasters are more successful and sometimes outperform the benchmark, in particular during the first wave of the epidemic. Third, one of the best performing forecasts is the Ensemble forecast, that combines all available predictions using uniform weights. In view of these results, collecting a wide range of forecasts and combining them in an ensemble forecast may be a superior approach for health authorities, rather than relying on a small number of forecasts.

Keywords

Forecast evaluation, Forecasting tests, Epidemic

JEL Classification

C12, C53, I18

Address for correspondence:

(E) cama.admin@anu.edu.au

ISSN 2206-0332

Testing the predictive accuracy of COVID-19 forecasts*

Laura Coroneo[†1], Fabrizio Iacone[2,1], Alessia Paccagnini[3], and Paulo Santos Monteiro[1]

[1]University of York
[2]Università degli Studi di Milano
[3]University College Dublin & CAMA

24th June 2021

Abstract

We test the predictive accuracy of forecasts of the number of COVID-19 fatalities produced by several forecasting teams and collected by the United States Centers for Disease Control and Prevention during the first and second waves of the epidemic in the United States. We find three main results. First, at the short horizon (1-week ahead) no forecasting team outperforms a simple time-series benchmark. Second, at longer horizons (3- and 4-week ahead) forecasters are more successful and sometimes outperform the benchmark, in particular during the first wave of the epidemic. Third, one of the best performing forecasts is the Ensemble forecast, that combines all available predictions using uniform weights. In view of these results, collecting a wide range of forecasts and combining them in an ensemble forecast may be a superior approach for health authorities, rather than relying on a small number of forecasts.

JEL classification codes: C12; C53; I18.

Keywords: Forecast evaluation, Forecasting tests, Epidemic.

*We thank Valentina Corradi, Paul Levine, Massimiliamo Marcellino, Elmar Mertens, Barbara Rossi and participants to the seminars at the University of Surrey, the University of York, University College Dublin, Bank of Greece, the 2nd Vienna Workshop on Economic Forecasting 2020 (IHS), the COVid-19 Empirical Research Workshop (University of Milan). Andrea Ierardi, Fabio Caironi and Marzio De Corato provided excellent research assistance.

[†]Corresponding author: Department of Economics and Related Studies, University of York, YO22 1DL, York, United Kingdom. Email: laura.coroneo@york.ac.uk.

1 Introduction

Forecasting the evolution of an epidemic is of utmost importance for policymakers and health care providers. Timely and reliable forecasts are necessary to help health authorities and the community at large coping with a surge of infections and to inform public health interventions, for example, to enforce (or ease) a lockdown at the local or national level. Accordingly, in recent months there has been a rapidly growing number of research teams developing forecasts for the evolution of the current COVID-19 pandemic caused by the new coronavirus, SARS-CoV-2.

In the United States, the Centers for Disease Control and Prevention (CDC) collects weekly forecasts of the evolution of the COVID-19 pandemic produced by different institutions and research teams. These forecasts are aimed at informing public health decision-making by projecting the probable impact of the COVID-19 pandemic at horizons up to four weeks. The forecasting teams that submit their forecasts to the CDC include data scientists, epidemiologists, and statisticians, and use different models and methods (e.g. SEIR, Bayesian, and Deep Learning models), combining a variety of data sources and assumptions about the impact of non-pharmaceutical interventions on the spread of the epidemic (such as social distancing and the use of face coverings). This wealth of forecasts can be extremely valuable for decision-makers, but it also poses a problem: how to act when confronted with heterogeneous forecasts and, in particular, how to select the most reliable projections. Decision-makers are thus faced with the task of comparing the predictive accuracy of different forecasts. Indeed, selecting models and comparing their predictive accuracy are different tasks, and in this paper we focus on the latter.

As the Diebold and Mariano (DM) test of equal predictive accuracy (see Diebold and Mariano 1995, Giacomini and White 2006) adopts a model-free perspective to compare competing forecasts, imposing assumptions only on the forecast errors loss differential, we use it to compare competing forecasts for the number of COVID-19 fatalities collected by

the CDC. The application of the DM test is particularly challenging when only a few out-of-sample observations are available, as the standard test is unreliable, especially for multi-step forecasts (Clark and McCracken 2013). To overcome this small-sample problem, we apply fixed-smoothing asymptotics, as recently proposed for this test by Coroneo and Iacone (2020).

With fixed-smoothing asymptotics, the limit distribution of the DM statistic is derived under alternative assumptions. In particular, when the long-run variance in the test statistic is estimated as the weighted autocovariances estimate, the asymptotic distribution is derived assuming that the bandwidth-to-sample ratio (denoted as b) is constant, as recommended by Kiefer and Vogelsang (2005). With this alternative asymptotics, usually known as fixed-b, the test of equal predictive accuracy has a nonstandard limit distribution that depends on b and on the kernel used to estimate the long-run variance. The second alternative asymptotics that Coroneo and Iacone (2020) consider is the fixed-m approach, as in Sun (2013) and Hualde and Iacone (2017). In this case, the estimate of the long-run variance is based on a weighted periodogram estimate, the asymptotic distribution is derived assuming that the truncation parameter m is constant, and the test of equal predictive accuracy has a t distribution with degrees of freedom that depend on the truncation parameter m. Both approaches have been shown to deliver correctly sized predictive accuracy tests, even when only a small number of out-of-sample observations is available (see Coroneo and Iacone 2020, Harvey, Leybourne and Whitehouse 2017).

We evaluate forecasts for the cumulative number of COVID-19 fatalities produced at the national level for the United States by the eight forecasting teams that submitted their forecasts to the CDC without interruptions during the period June 20, 2020 to March 20, 2021. Although the evaluation period includes only 40 observations, we document an increase in the volatility of the forecasting errors in November 2020. Accordingly, we perform our forecast evaluation separately on two sub-samples: the first wave (from June 20, 2020 to October 31, 2020) and the second wave (from November 7, 2020 to March 20, 2021). This

implies that for each wave we can base our inference only on 20 observations, making the use of fixed-smoothing asymptotics crucial for obtaining reliable results.

We compare the predictive accuracy of the forecasts of each team relative to the forecasts of a simple benchmark model, obtained by fitting a second-order polynomial using a rolling window of the last five available observations. We also consider two ensemble forecasts that combine the forecasts from several models using equal weights: one published by the CDC and another one (the core ensemble) computed by us combining only the forecasts included in our evaluation exercise.

A feature that makes forecast evaluation important in its own right, especially when dealing with predicting the spread of COVID-19, is that the cost of under-predicting the spread of the disease can be greater than the cost of over-predicting it. In the midst of a public health crisis, the precautionary principle implies that erring in the side of caution is less costly than predicting the tapering off of the disease too soon. Scale effects may also be important in the evaluation of forecasts of an epidemic outbreak, since the same forecast error may be considered differently when the realized level of fatalities is small, and when there is a large number of fatalities. These effects may be taken into account in the forecast evaluation exercise by a judicious choice of the loss function. Therefore, we evaluate the predictive accuracy of each forecasting team using several loss functions, that include the widely used quadratic and absolute value loss, the absolute percentage loss (that takes into account the scale of the number of fatalities), and a linear exponential loss function (that penalizes under-prediction more than over-prediction).

Our findings can be summarized as follows. First, the simple polynomial benchmark outperforms the forecasters at the short horizon (1-week ahead), often significantly so. Second, at longer horizons (3- to 4-week ahead), the forecasters become more competitive and some statistically outperform the simple benchmark, especially in the first wave. This suggests that forecasters can successfully help inform forward looking policy decisions. Third, the ensemble

forecasts are among the best performing forecasts. This is particularly true in the first wave, but even in the second wave the ensemble forecast combinations outperform the benchmark, although in this sub-sample the DM test statistics are not statistically significant. The reliability of ensemble forecasts underlines the virtues of model averaging when uncertainty prevails, and supports the view in Manski (2020) that data and modelling uncertainties limit our ability to predict the impact of alternative policies using a tight set of models. Overall, our findings hold for all the loss functions considered and caution health authorities not to rely on a single forecasting team (or a small set) to predict the evolution of the pandemic. A better strategy appears to be to collect as many forecasts as possible and to use an ensemble forecast.

The remainder of the paper is organized as follows. Section 2 lays out the methodology to implement the test of equal predictive accuracy. Section 3 describes the data and the models. Results are documented and discussed in Section 4 and Section 5 concludes.

2 Forecast Evaluation

We consider the time series of cumulative daily deaths $\{y_1, ..., y_T\}$, for which we want to compare two h-week ahead forecasts $\widehat{y}^{(1)}_{t|t-h}\left(\widehat{\theta}^{(1)}_{w_1}\right)$ and $\widehat{y}^{(2)}_{t|t-h}\left(\widehat{\theta}^{(2)}_{w_2}\right)$, where $\widehat{\theta}^{(i)}_{w_i}$ for $i = 1, 2$ denote the estimates obtained with a rolling window of size w_i used to construct forecast i, if known.

The forecast error for forecast i is $e^{(i)}_{t|t-h} = y_t - \widehat{y}^{(i)}_{t|t-h}\left(\widehat{\theta}^{(i)}_{w_i}\right)$ and the associated loss is $L^{(i)}_{t|t-h} \equiv L\left(e^{(i)}_{t|t-h}\right)$, for example, a quadratic loss would be $L\left(e^{(i)}_{t|t-h}\right) = \left(e^{(i)}_{t|t-h}\right)^2$. The null hypothesis of equal predictive ability of the two forecasts is

$$H_0 : E\left[L\left(e^{(1)}_{t|t-h}\right) - L\left(e^{(2)}_{t|t-h}\right)\right] = 0. \tag{1}$$

Let

$$d_t \equiv L\left(e^{(1)}_{t|t-h}\right) - L\left(e^{(2)}_{t|t-h}\right),$$

denote the time-t loss differential between the two forecasts and let

$$\overline{d} = \frac{1}{T}\sum_{t=w+h}^{w+h+T-1} d_t,$$

denote the sample mean of the loss differential, where $w \equiv \max(w_1, w_2)$.

When a large sample is available, standard asymptotic theory may provide a valid guidance for the statistical evaluation of $\overline{d}$, see Diebold and Mariano (1995) and Giacomini and White (2006). However, the same inference may be severely biased when the sample has only a moderate size, as it is indeed the case when comparing forecast accuracy of predictions of the number of fatalities of COVID-19. In this case, fixed-b and fixed-m asymptotics can be used to overcome the small-sample size bias, see Coroneo and Iacone (2020), Choi and Kiefer (2010) and Harvey et al. (2017).

As for the fixed-b asymptotics, following Kiefer and Vogelsang (2005), under the null in (1)

$$\sqrt{T}\frac{\overline{d}}{\widehat{\sigma}_{BART,M}} \to_d \Phi_{BART}(b), \text{ for } b = M/T \in (0,1], \tag{2}$$

where $\widehat{\sigma}^2_{BART,M}$ denotes the weighted autocovariance estimate of the long-run variance of d_t using the Bartlett kernel and truncation lag M. Kiefer and Vogelsang (2005) characterize the limit distribution $\Phi_{BART}(b)$ and provide formulas to compute quantiles. For example, for the Bartlett kernel with $b \leq 1$, these can be obtained using the formula

$$q(b) = \alpha_0 + \alpha_1 b + \alpha_2 b^2 + \alpha_3 b^3,$$

where

$$\alpha_0 = 1.2816,\ \alpha_1 = 1.3040,\ \alpha_2 = 0.5135,\ \alpha_3 = -0.3386 \text{ for } 0.900 \text{ quantile}$$

$$\alpha_0 = 1.6449,\ \alpha_1 = 2.1859,\ \alpha_2 = 0.3142,\ \alpha_3 = -0.3427 \text{ for } 0.950 \text{ quantile}$$

$$\alpha_0 = 1.9600,\ \alpha_1 = 2.9694,\ \alpha_2 = 0.4160,\ \alpha_3 = -0.5324 \text{ for } 0.975 \text{ quantile}$$

When testing assumptions about the sample mean, Kiefer and Vogelsang (2005) show in Monte Carlo simulations that the fixed-b asymptotics yields a remarkable improvement in size. However, while the empirical size improves (it gets closer to the theoretical size) as b is closer to 1, the power of the test worsens, implying that there is a size-power trade-off.

For fixed-m asymptotics, following Hualde and Iacone (2017), under the null in (1) we have

$$\sqrt{T}\frac{\overline{d}}{\widehat{\sigma}_{DAN,m}} \rightarrow_d t_{2m}, \tag{3}$$

where $\widehat{\sigma}^2_{DAN,m}$ is the weighted periodogram estimate of the long-run variance of d_t using the Daniell kernel and truncation m. Similar results, with a slightly different standardisation and therefore a slightly different limit, are in Sun (2013). Monte Carlo simulations in Hualde and Iacone (2017) and Lazarus, Lewis, Stock and Watson (2018) show that fixed-m asymptotics has the same size-power trade-off documented for fixed-b asymptotics: the smaller the value for m, the better the empirical size, but also the weaker the power.

Coroneo and Iacone (2020) analyze the size and power properties of the tests of equal predictive accuracy in (2) and (3) in an environment with asymptotically non-vanishing estimation uncertainty, as in Giacomini and White (2006). Results indicate that the tests in (2) and (3) deliver correctly sized predictive accuracy tests for correlated loss differentials even in small samples, and that the power of these tests mimics the size-adjusted power. Considering size control and power loss in a Monte Carlo study, they recommend the bandwidth $M = \lfloor T^{1/2} \rfloor$ for the weighted autocovariance estimate of the long-run variance using the Bartlett kernel

(where $\lfloor \cdot \rfloor$ denotes the integer part of a number) and $m = \lfloor T^{1/3} \rfloor$ for the weighted periodogram estimate of the long-run variance using the Daniell kernel.

3 Forecasting Teams and Benchmark

3.1 Data and forecasting teams

In our empirical investigation, we use forecasts for the cumulative number of deaths collected by the Centers for Disease Control and Prevention (CDC). The CDC is a federal agency in charge of protecting public health through the control and prevention of diseases. It is also the official source of statistics on the COVID-19 pandemic evolution in the US. In particular, in collaboration with independent teams of forecasters, the CDC has set up a repository of weekly forecasts for the numbers of deaths, hospitalizations, and cases. These forecasts are developed independently by each team and shared publicly.[1] We focus on forecasts of the number of deaths for three main reasons. First, the number of fatalities is more accurate than the number of cases and hospitalizations, since the latter ignores asymptomatic cases and other diseases that are undetected. Second, the number of deaths is reported with less spatial and temporal biases. Third, when faced with a pandemic, the number of fatalities is arguably the primary concern of the health authorities and of the public.

Our sample includes projections for national COVID-19 cumulative deaths made for the period between June 20, 2020 and March 20, 2021 by eight forecasting teams. The deadline for the teams to submit their weekly forecasts is on the Monday of each week and they are usually published online on Wednesdays. Weekly cumulative data is the cumulative data up to and including Saturday. This means that, for example, the forecasts submitted by June 22 had as targets the cumulative number of deaths as of June 27 (1-week ahead), July 2 (2-week ahead), July 7 (3-week ahead), and July 12 (4-week ahead). Realised values are also taken

[1] Background information on each forecasting teams, along with a summary explanation of their methods are available via the link https://www.cdc.gov/coronavirus/2019-ncov/covid-data/forecasting-us.html.

Table 1: Forecasting Teams, Methods, and Assumptions

Code	Team	Model	Method	Change
CO	COVID Analytics - MIT Sloan	DELPHI Model	Deep Learning model	no
UM	University of Massachusetts, Amherst	UMass - MB	Mechanistic Bayesian compartment model	no
UA	University of Arizona	UA - EpiCovDA	Modified SEIR model	yes
GT	Georgia Institute of Technology, Deep Outbreak Project	GT - Deep COVID	Deep Learning model	no
MO	Northeastern University, Laboratory for the Modeling of Biological and Socio-technical Systems	MOBS - GLEAM COVID	Metapopulation, age structured SEIR model	no
PS	Predictive Science, Inc.	PS - DRAFT	SEIR model	yes
LA	Los Alamos National Laboratory	LANL - Growth Rate	Statistical dynamical growth model	no
JH	Johns Hopkins University, Infectious Disease Dynamics Lab	JHU - IDD - CovidSP	Metapopulation SEIR model	yes

Notes: The column code describes the code given in the empirical analysis to each team. A forecasting team is included if it submitted its predictions for all the weeks in our sample. The table reports for each forecasting team the modelling methodology and whether the model considers a change in the assumptions about policy interventions. In the fourth column, "yes" means that the modelling team makes assumptions about how levels of social distancing will change in the future, while "no" means that it is assumed that the existing measures will continue through the projected 4-week time period.

from the CDC website. Notice that when COVID-19 is reported as a cause of mortality on the death certificate, it is coded and counted as a fatality due to COVID-19.

The eight forecasting teams selected are those that submitted their predictions with no interruptions for all the weeks in our sample. We list the selected teams in Table 1, and report the main features of the selected forecasts. They vary widely with regards to their modelling choice, information input (for example, how the information on infected people is used), and in their assumptions about the evolution and impact of non-pharmaceutical interventions (for example regarding social distancing).[2]

3.2 Ensemble forecasts

In our forecast evaluation exercise, we also consider two Ensemble forecasts: one published by the CDC (that combines the individual forecasts from several teams and that we label Ensemble - EN) and one computed by us (that combines the individual forecasts from the eight teams listed in Table 1 and that we label Core Ensemble - CE).

Combining forecasts is an effective procedure when there is uncertainty about the model and

[2]Additional summary information about the models is on the CDC repository page https://github.com/cdcepi/COVID-19-Forecasts/blob/master/COVID-19_Forecast_Model_Descriptions.md, where links to the modelling teams are also provided.

the data, as it is indeed the case here, where differences also include alternative assumptions on the responses of the public and of the health authorities. In this situation, combining forecast is useful as it helps to diversify risk and to pool information (see Bates and Granger 1969). In particular, forecast combination is most advantageous when there is pervasive uncertainty, as the ranking of best-performing forecasts may be very unstable and therefore forecast combination provides a robust alternative (see Stock and Watson 1998, Timmermann 2006). Optimal weights that give the best combination, in the sense of minimizing a given loss function, can actually be derived, but in many practical applications estimated optimal weights schemes result in a combined forecast that does not improve simple averaging (see Clemen 1989, Smith and Wallis 2009, Claeskens, Magnus, Vasnev and Wang 2016).

In epidemiology, forecast combination has proved its ability to improve on the performance of individual competing models. For example, Reich et al. (2019) found that ensemble forecasting for influenza performed better on average against the constituting models; similar results have been obtained by Chowell et al. (2020) in the Ebola Forecasting Challenge. Both these works had access to a sufficiently long history of data, making a data-driven selection of the weights assigned to the contributing models possible. Interestingly, Reich et al. (2019) considered also the equal weighting scheme in their exercise, and found that this naive ensemble performed quite well even against the one with data-driven weights, making it a reasonable choice for the current situation of a new epidemic, in which no previous outbreaks exist and no previous track record of past models is available.

The Ensemble forecast produced by the CDC is also naive, in the sense that it is based on an equal weighting of all the available forecasts. Specifically, it is obtained by averaging forecasts across all teams, as long as they publish forecasts up to four weeks ahead and these forecasts are at least equal to the level observed on the day in which the forecast is submitted. The weekly composition of the pool of models contributing to the Ensemble forecast changes, and it includes, in general, a larger number of teams than the one we consider in our evaluation

exercise. This loose criterion allows to include as many forecasts as possible, which may be desirable, but there is also the risk of including poorly performing teams. For this reason, we also consider the Core Ensemble constructed by us, which uses only the forecasts (equally weighted) by the eight teams that are included in our forecast evaluation exercise. The conjecture motivating this choice is that as these are the most long standing forecasting teams, they should also be the most experienced ones, and this experience may give them the edge to outperform the other teams. In addition, by comparing the performance of the individual forecasts with the Core Ensemble forecast, we can reliably assess the value added by the combination of the forecasts, as the Core Ensemble uses only forecasts that are included in our exercise.

3.3 Benchmark forecasts

The benchmark against which we compare the forecasts collected by the CDC is a polynomial function. That is, benchmark forecasts are obtained as projections from the model:

$$y_t = \beta_0 + \beta_1\, t + \beta_2\, t^2 + u_t, \tag{4}$$

where y_t is the cumulative number of fatalities, t is the time trend, and u_t is an unobserved error term. To accommodate the fact that the forecasted patterns may need changing even over a short span of time, we fit the quadratic polynomial model using Least Squares with a rolling window of the last five observations (using weekly data, this covers approximately a month). To ensure that the benchmark forecasts for the cumulative number of deaths are not decreasing, we compute the benchmark predictions as the maximum between the current value and the prediction from (4).

This very simple statistical model has been chosen because any continuous and differentiable function can be approximated locally by a polynomial, and we take the second degree polynomial as a local approximation. In recent works, the choice of a polynomial benchmark

has also been considered by Jiang, Zhao and Shao (2020) and Li and Linton (2020), among others, although with some small differences. In Jiang et al. (2020), the intrinsic instability of the forecasted patterns is accommodated by fitting occasional breaks; whereas Li and Linton (2020) fitted the model to the incidence of deaths, rather than to the cumulative deaths.

3.4 Preliminary Analysis

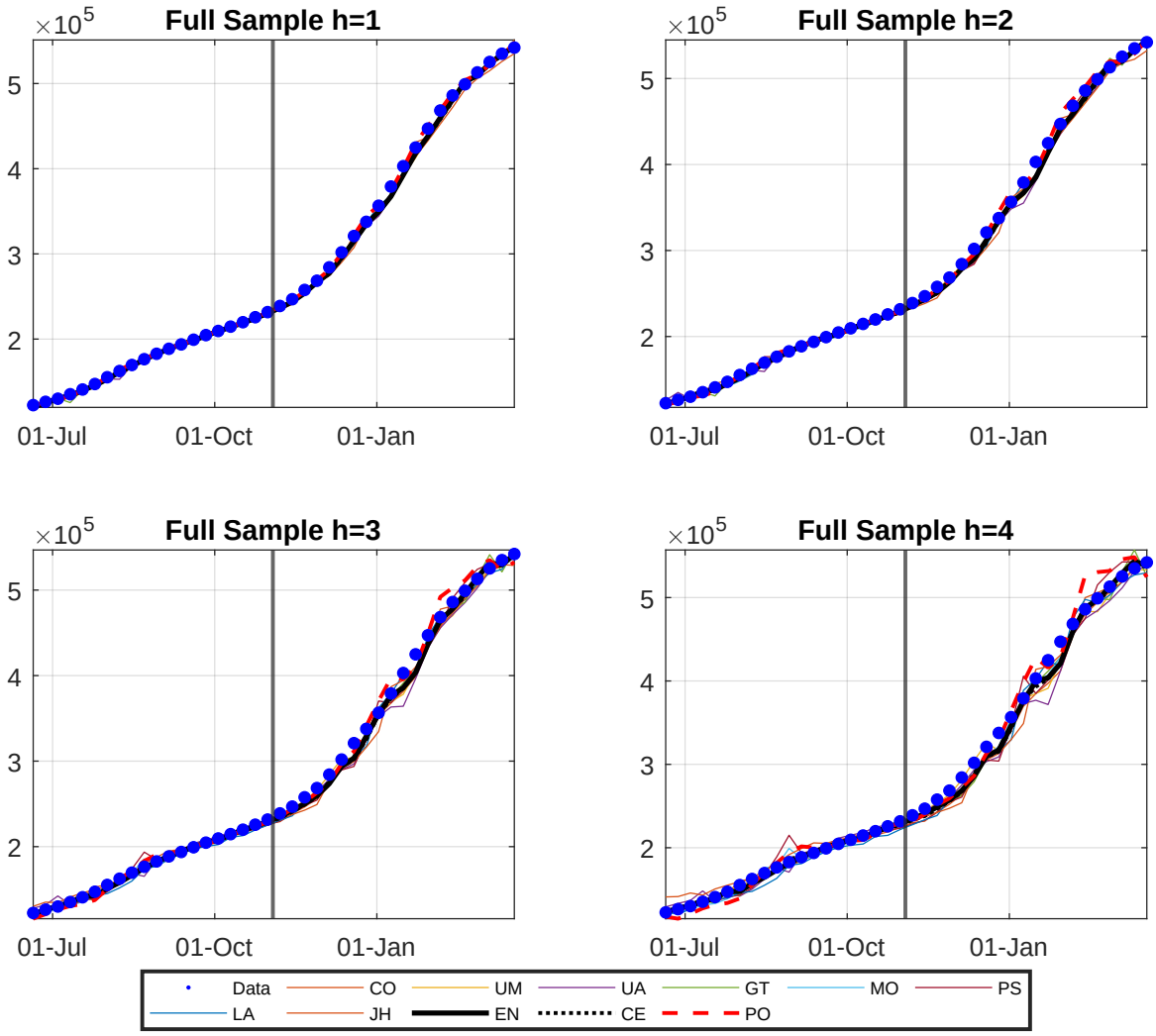
In this section, we present some preliminary analysis of the forecasts submitted by the forecasting teams in Table 1, the Ensemble (EN) forecast published by the CDC, the Core Ensemble (CE) constructed combining all the forecasts of the teams in Table 1, and the forecast of the polynomial benchmark (PO), described above.

Figure 1 plots all the forecasts considered for the 1 to 4-week ahead forecasting horizons, alongside the realised data. Comparing the graphs in Figure 1 at different horizons, it is apparent that the heterogeneity in forecasts grows with the forecasting horizon and, concurrently, that the forecasts are less precise as the forecast horizon increases. This simple observation may make the case for forecast combination at longer horizons more compelling.

Figure 2 plots the forecast errors for each model (computed as the difference between the realization and the point forecast). The figure indicates that most forecast teams seem to have systematically under-predicted the target, in particular in the second second part of the sample. This is, of course, relevant for policy makers if the costs of over-prediction and under-prediction are different.

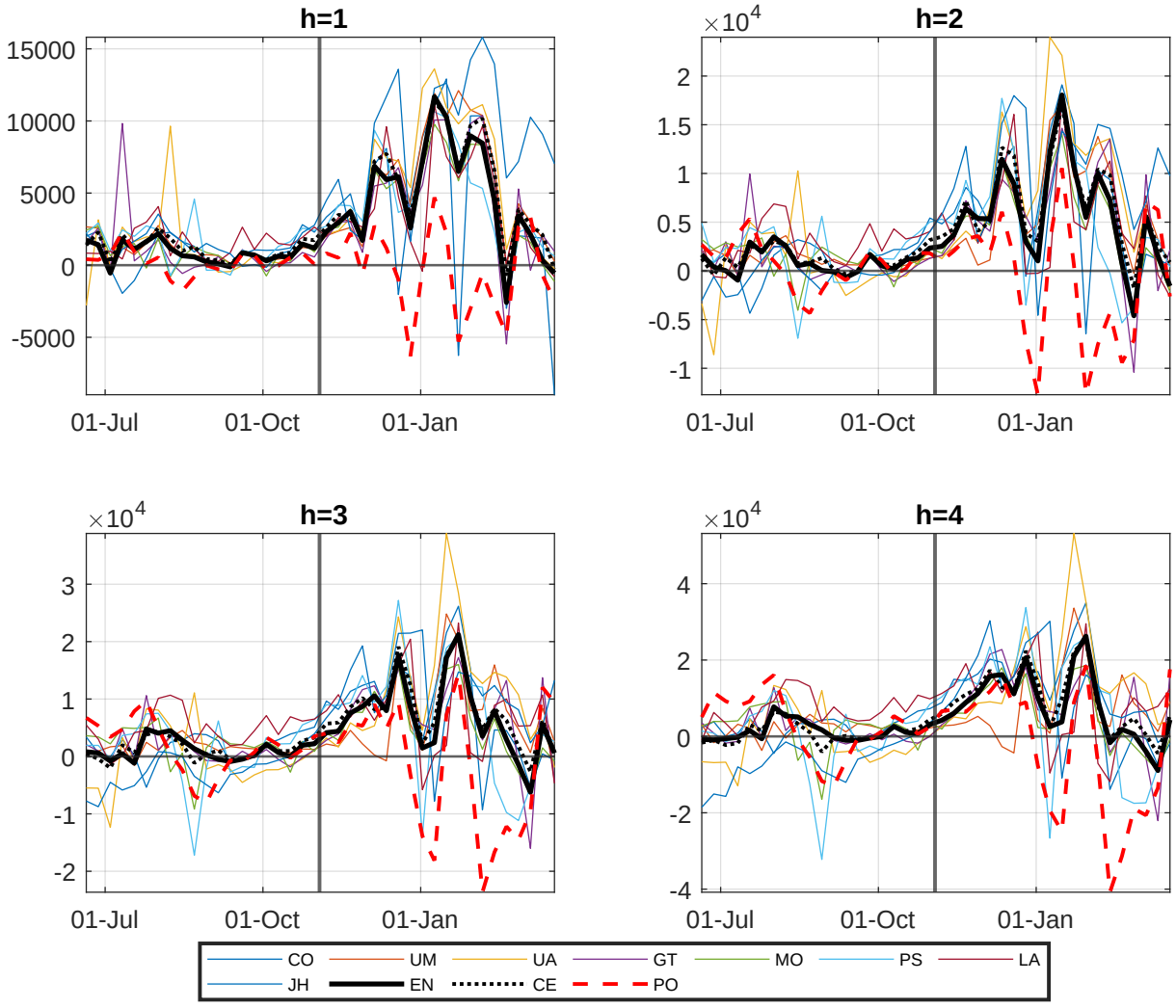
Figure 2 also reveals a clear break in the volatility of the forecast errors across the first and the second halves of the sample, with the change point at the beginning of November 2020 (as illustrated by the vertical line in each diagram of Figures 1-2). This is also confirmed in Tables A.1-A.2 in the Appendix, that report summary statistics for the forecast errors in the two sub-samples and, in particular, report the volatility of the forecast errors to be considerably higher in the second sub-sample. Such decline in the quality of the forecasts

Figure 1: Cumulative deaths in US, observed vs. forecasts



Note: Forecasts at forecasting horizons h from 1 to 4 weeks, along with realised cumulative fatalities. Weekly observations from June 20, 2020 to March 20, 2021. The vertical line indicates November 3, 2020 and delimits the two sub-samples. The names of the forecasting teams are as in Table 1; EN denotes the Ensemble published by the CDC, CE denotes the Core Ensemble constructed combining all the forecasts of the teams in Table 1, and PO denotes the polynomial benchmark.

Figure 2: Forecast errors



Note: Forecast errors at forecasting horizons h from 1 to 4 weeks. Weekly observations from June 20, 2020 to March 20, 2021. The vertical line indicates November 3, 2020 and delimits the two sub-samples. The names of the forecasting teams are as in Table 1; EN denotes the ensemble forecast, CE denotes the core ensemble, and PO the polynomial benchmark. Forecast errors are defined as the realised value minus the forecast.

in the most recent sub-sample may at first be puzzling: one would expect the forecasting teams to improve their performance as more information becomes progressively available. However, this structural break in the forecasting ability of all models could in part be related to the emergence of a new strain of the virus in the end of 2020, with specific mutations in the spike protein of SARS-CoV-2 resulting in increased transmissibility. Consistent with this explanation, research from the CDC reports that the B.1.1.7 virus strain (often referred to as the "Kent" variant) is estimated to have emerged in September 2020. This variant exhibited rapid growth in the US in early 2021, and was predicted to be the predominant variant by March 2021 (Galloway, Paul, MacCannell, Johansson, Brooks, MacNeil, Slayton, Tong, Silk and Armstrong 2021).

At any rate, the volatility of the forecasting errors increases markedly starting from the beginning of November 2020 and, thus, we perform our forecast evaluation separately on two sub-samples: the first wave (from June 20, 2020 to October 31, 2020), and the second wave (from November 7, 2020 to March 20, 2021). This means that for each wave we base our inference on just 20 observations. With such small sample sizes, fixed-smoothing asymptotics is crucial to obtain correctly sized tests for equal predictive accuracy.

Table 2 presents some summary statistics for the forecast errors. The table reports for each forecasting horizon and forecasting scheme (team, Ensemble, Core Ensemble or polynomial) the sample mean, median, standard deviation, skewness, and the sample autocorrelation coefficients up to order 4 (in the columns AC(1), AC(2), AC(3) and AC(4), respectively). With the exception of the polynomial model, the average of the forecast errors are positive for all forecasts, at each horizon, meaning that the forecasters tend to under-predict the number of fatalities.

At the 1-week horizon, the benchmark polynomial model appears to outperform all the competitors, with a much smaller average error and smaller dispersion. However, its performance deteriorates at longer horizons, with the volatility of the forecast errors increasing substan-

Table 2: Summary Statistics of Forecast Errors

1-week ahead	Mean	Median	Std	Skew	AC(1)	AC(2)	AC(3)	AC(4)
CO	2475.85	1920.00	4429.68	0.22	0.19	0.06	0.15	0.43
UM	3363.88	1705.50	3693.80	1.19	0.83	0.72	0.62	0.52
UA	3663.35	2062.00	4310.21	0.87	0.73	0.65	0.55	0.51
GT	2695.21	1334.09	3651.89	0.71	0.43	0.40	0.38	0.39
MO	2520.66	1549.87	2965.75	0.90	0.70	0.56	0.52	0.54
PS	3057.91	2454.00	3028.03	0.92	0.70	0.46	0.43	0.38
LA	2860.78	2108.79	2883.87	1.31	0.47	0.22	0.23	0.40
JH	5096.85	2333.36	4921.48	0.75	0.78	0.61	0.60	0.71
EN	2754.32	1678.50	3246.07	1.11	0.71	0.54	0.52	0.55
CE	3216.81	2015.61	3256.36	1.10	0.74	0.56	0.56	0.62
PO	−97.66	91.00	2210.91	−0.74	0.22	−0.44	−0.16	0.35
2-week ahead	Mean	Median	Std	Skew	AC(1)	AC(2)	AC(3)	AC(4)
CO	3737.88	2200.00	5428.87	0.87	0.42	0.11	0.25	0.44
UM	3830.68	1697.00	4541.00	1.52	0.70	0.43	0.52	0.59
UA	4586.73	2705.50	6911.65	0.98	0.71	0.52	0.51	0.52
GT	3330.98	2166.46	4945.81	0.25	0.31	0.20	0.36	0.35
MO	3004.34	2365.75	3780.38	0.81	0.56	0.26	0.33	0.44
PS	3759.05	3333.50	5381.69	0.60	0.56	0.15	0.14	0.19
LA	4128.15	3400.92	4235.16	1.37	0.16	-0.14	-0.06	0.35
JH	4986.19	2812.73	6491.23	0.49	0.68	0.44	0.47	0.61
EN	3322.70	1808.50	4459.98	1.27	0.61	0.25	0.36	0.47
CE	3920.50	2627.34	4299.01	1.28	0.66	0.33	0.47	0.60
PO	−89.38	1053.40	5021.96	−0.75	0.42	−0.26	−0.14	0.13
3-week ahead	Mean	Median	Std	Skew	AC(1)	AC(2)	AC(3)	AC(4)
CO	4971.90	3164.00	7114.77	0.88	0.60	0.27	0.25	0.36
UM	4410.38	2082.00	5849.59	2.00	0.52	0.22	0.35	0.58
UA	5644.75	3893.50	9930.94	1.20	0.70	0.48	0.43	0.51
GT	3988.10	2844.71	6407.18	−0.20	0.25	0.26	0.37	0.32
MO	3496.54	2592.72	5166.81	0.47	0.48	0.27	0.05	0.20
PS	4129.98	3890.75	9066.45	−0.03	0.52	0.09	0.06	0.06
LA	5992.51	5576.28	5790.07	0.72	0.05	−0.20	−0.18	0.28
JH	4343.70	2386.35	8874.59	0.43	0.71	0.55	0.53	0.60
EN	3951.13	2309.00	5562.79	1.32	0.60	0.28	0.21	0.35
CE	4622.23	3005.15	5569.31	1.32	0.66	0.35	0.39	0.54
PO	27.63	2232.54	8841.85	−0.87	0.58	0.01	−0.13	0.01
4-week ahead	Mean	Median	Std	Skew	AC(1)	AC(2)	AC(3)	AC(4)
CO	6594.93	4090.00	9744.80	0.70	0.73	0.36	0.26	0.26
UM	4963.63	3037.50	7521.18	2.13	0.43	0.01	0.18	0.44
UA	7021.95	4854.00	12912.43	1.37	0.72	0.44	0.41	0.48
GT	4875.40	3314.53	8662.06	-0.14	0.36	0.18	0.37	0.30
MO	3882.06	3566.27	7309.12	-0.11	0.54	0.30	-0.01	0.01
PS	4058.40	5026.75	13878.24	-0.50	0.50	0.11	0.05	0.01
LA	8273.18	7802.85	8444.13	0.06	0.16	-0.06	-0.08	0.37
JH	2769.82	16.31	12755.55	0.54	0.75	0.63	0.60	0.63
EN	4641.08	2610.50	7286.16	1.16	0.66	0.31	0.22	0.29
CE	5304.92	2759.80	7359.80	1.11	0.70	0.37	0.37	0.45
PO	-301.20	4249.03	13860.09	-1.03	0.64	0.18	0.02	0.07

Notes: The table reports summary sample statistics of forecast errors for the teams, the Ensemble (EN), the Core Ensemble (CE) and the polynomial (PO) forecasts. The table reports mean, median, standard deviation (std), skewness (skew), and autocorrelation coefficients up to order 4 (AC(1), AC(2), AC(3), AC(4)). Weekly observations from June 20, 2020 to March 20, 2021.

tially, and becoming larger than those of most forecasting teams. This is not surprising, as epidemiological models are designed to predict the evolution of a pandemic in the medium and the long run, and we observe here that even a very simple forecast does better when the horizon is very short. At longer horizons, however, epidemiological models should be expected to produce forecasts that are more stable than simple statistical benchmarks.

Finally, from Table 2 we can also observe that the forecast errors are autocorrelated, as documented in the columns AC(1), AC(2), AC(3) and AC(4). This happens even for one-step-ahead forecasts, where the first order sample autocorrelation may be as high as 0.83. This is interesting because, under mean squared error loss, optimal h-step ahead forecast errors should be at most MA($h-1$): so 1-step ahead forecast errors should be Martingale Differences, 2-step ahead errors should be at most MA(1), and so on. Indeed, this is the very argument given in Diebold and Mariano (1995) to justify the choice of the rectangular kernel to estimate the long-run variance. However, this condition is clearly violated by all modelling teams.

One explanation of this higher order autocorrelation in the forecast errors and the fact that the forecasting teams systematically underpredict the number of fatalities could be that the forecasting teams use alternative loss functions to produce their forecasts. Indeed, Patton and Timmermann (2007) show that, under asymmetric loss and nonlinear data generating processes, forecast errors can be biased and serially correlated of arbitrarily high order.

4 Forecast Evaluation Results

Our main results for the test of equal predictive ability of each forecasting team vis-à-vis the benchmark model (4) are reported in Table 3.

We conduct the analysis separately for the two sub-samples (the first and second waves of the epidemic) identified in Figure 2. This yields 20 observations for each sub-sample, underlying

the importance of alternative asymptotics in evaluating predictive ability. In the baseline analysis, we evaluate forecast errors relying on the quadratic loss function. We present the test statistics using both the weighted covariance estimator with Bartlett kernel (WCE) and the weighted periodogram estimator with Daniell kernel (WPE) of the long-run variance. A positive value for the test statistic indicates that the forecast in question is more accurate than the polynomial benchmark.

We report two-sided significance at the 5% and 10% levels, using fixed smoothing asymptotics (fixed-b for WCE and fixed-m for WPE) to establish the critical values. In particular, for $T = 20$ and the bandwidths recommendations in Coroneo and Iacone (2020), the critical values are ± 2.57 and ± 2.09 with fixed-b asymptotics and ± 2.78 and ± 2.13 with fixed-m asymptotics. For comparison, we also report significance based on bootstrap critical values, constructed using the overlapping stationary block-bootstrap of Politis and Romano (1994) using an average block length of $T^{1/4} \approx 2$ and a circular scheme, as described in Coroneo and Iacone (2020).

We consider first the upper panel of Table 3, which reports results for the first sub-sample from June 20, 2020 to October 31, 2020. No forecasting scheme predicts better than the benchmark at 1-week forecasting horizon. In fact, the benchmark often significantly outperforms the forecasting teams. Instead, differences at the 2-week horizon are almost never significant and in most cases the sign of the test statistic turns from negative to positive, signalling a smaller relative loss by the forecasting teams. The improvement in the relative performance of the forecasters rises further at longer horizons (3 and 4 weeks ahead), where we begin to observe statistically significant relative gains in performance. The best performing teams are the Georgia Institute of Technology (GT) and the University of Massachusetts (UM). Both the ensemble forecasts (the EN ensemble provided by the CDC and the CE core ensemble constructed combining all the forecasts of the teams in Table 1) are also among the best performing models and outperform the benchmark. This finding is consistent with the

Table 3: Tests for Equal Predictive Ability

First wave: Jun 20, 2020 – Oct 31, 2020

	1-week ahead		2-week ahead		3-week ahead		4-week ahead	
	WCE	WPE	WCE	WPE	WCE	WPE	WCE	WPE
CO	-2.656**	-1.950	-0.104	-0.096	1.896	1.719	2.407*	2.081
UM	-1.704	-1.985	1.878	2.246*	2.878**	2.823**	2.974**	2.676*
UA	-1.465	-1.358	-1.720	-1.982	-0.978	-2.210*	1.219	1.524
GT	-1.166	-1.014	-0.361	-0.306	2.964**	2.623*	3.476**	3.098**
MO	-1.534	-1.384	0.569	0.703	1.778	1.900	1.818	1.702
PS	-4.119**	-3.257**	-2.604**	-2.018	-0.685	-0.588	-0.461	-0.395
LA	-2.787**	-2.564*	-1.794	-1.494	-0.905	-0.742	-0.074	-0.060
JH	-2.338*	-2.145*	1.134	1.019	0.729	0.590	-0.328	-0.263
EN	-1.555	-1.700	2.005	2.237*	2.982**	2.914**	3.195**	2.874**
CE	-2.657**	-2.445*	1.656	1.750	2.877**	2.832**	3.127**	2.806**

Second wave: Nov 7, 2020 – Mar 20, 2021

	1-week ahead		2-week ahead		3-week ahead		4-week ahead	
	WCE	WPE	WCE	WPE	WCE	WPE	WCE	WPE
CO	-3.006**	-2.459*	-1.559	-1.322	-0.166	-0.147	0.371	0.321
UM	-2.293*	-1.835	-1.334	-1.111	0.773	0.706	1.398	1.206
UA	-2.586**	-2.076	-2.004	-1.688	-1.114	-0.972	-0.340	-0.298
GT	-2.535*	-2.035	-2.099*	-2.109	0.733	0.763	1.013	0.927
MO	-2.415*	-1.924	0.219	0.202	1.607	1.493	1.772	1.585
PS	-2.351*	-2.008	-1.241	-1.161	-0.588	-0.535	-0.084	-0.072
LA	-2.618**	-2.114	-1.051	-1.110	0.589	0.542	0.776	0.671
JH	-4.395**	-3.786**	-4.680**	-5.169**	-0.731	-0.628	0.305	0.258
EN	-2.365*	-1.901	-0.922	-0.849	0.996	0.956	1.383	1.250
CE	-2.729**	-2.180*	-1.354	-1.233	0.693	0.645	1.211	1.068

Note: test statistics for the test of equal predictive accuracy using the weighted covariance estimator (WCE) and the weighted periodogram estimator (WPE) of the long-run variance. The benchmark is a second degree polynomial fitted on a rolling window of 5 observations. The forecast errors are evaluated using the quadratic loss function, and a positive value of the test statistic indicates lower loss for the forecaster (i.e. better performance of the forecaster relative to the polynomial model). ** and * indicate, respectively, two-sided significance at the 5% and 10% level using fixed-b asymptotics for WCE and fixed-m asymptotics for WPE. ▇ and ▇ indicate, respectively, two-sided significance at the 5% and 10% level using the bootstrap. Bootstrap critical values are constructed using the overlapping stationary block-bootstrap of Politis and Romano (1994) using an average block length of $T^{1/4} \approx 2$ and a circular scheme, as described in Coroneo and Iacone (2020).

consensus in the literature about the advantages of forecast combination (see Stock and Watson 1998, Timmermann 2006).

Turning to the results in the bottom part of Table 3 that refer to the second sub-sample from November 7, 2020 to March 20, 2021, we can see that the benchmark still significantly outperforms all teams and the ensemble forecasts at short horizons. However, in this sub-sample the forecasting teams and the ensembles fail to significantly outperform the benchmark even at the long-term horizon, suggesting that the epidemic in the second wave was indeed more difficult to forecast. In both cases, however, at three to four weeks ahead, the two ensembles performed better than most of the individual teams.

Results are overall very similar regardless of the type of estimator of the long-run variance. We also notice that results from the bootstrap are largely the same, and confirm that fixed-smoothing asymptotics is a suitable and computationally much less time-consuming alternative to bootstrapping, as also found in Coroneo and Iacone (2020) and Gonçalves and Vogelsang (2011). Moreover, by using fixed-smoothing asymptotics we have known critical values for each test, given the sample size and choice of bandwidth.

Finally, notice that the performance of the Ensemble forecast published by the CDC is similar to the one of the Core Ensemble constructed combining all the forecasts of the teams in Table 1. However, the test statistics for the Ensemble are always strictly larger than the ones for the Core Ensemble, indicating that combining a larger set of forecasts than the ones considered in Table 1 can provide some benefits in terms of predictive ability, albeit a small one.

4.1 Alternative Loss Functions

The quadratic loss function that we use in the baseline forecast evaluation reported in Table 3 is a common choice in forecast evaluation. In particular, the null hypothesis is the equality of the mean square prediction error. However, in relation to predicting the spread of COVID-19

(and, more generally, predicting the spread of an epidemic), the cost of under-predicting the spread of the disease can be greater than the cost of over-predicting it. Similarly, scale effects are important, since the same forecast error may be more costly for public health policy interventions when the number of fatalities is small, compared to when it is large.

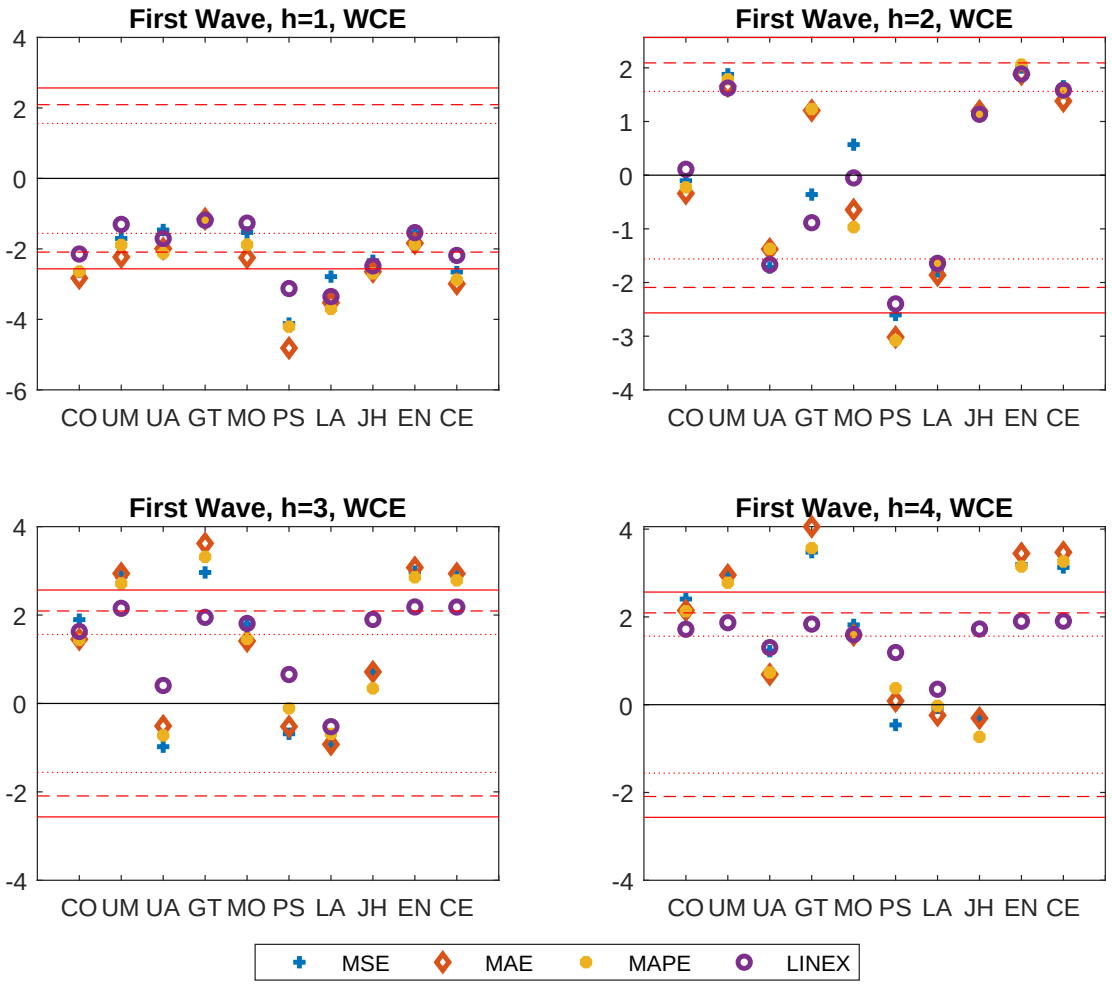
The DM test can be applied directly to alternative loss functions. Thus, we consider three alternative loss functions that provide alternative criteria for forecast comparison. Denoting e_t as the forecast error (thus abbreviating in this way $e^{(i)}_{t|t-h}$), the alternative loss functions considered are the following:

- Absolute: $L(e_t) = |e_t|$;
- Absolute Percentage: $L(e_t) = |e_t|/(y_t - y_{t-1})$;
- Linex: $L(e_t) = \exp\left(e_t/(y_t - y_{t-1})\right) - e_t/(y_t - y_{t-1}) - 1$.

The absolute loss function is an alternative measure that seems justified when forecast errors have the same importance: in this case, it seems natural to interpret it as giving all fatalities the same weight. This was also considered by Diebold and Mariano (1995) in their empirical application. The absolute percentage loss considers the scale of the number of fatalities occurring in the period, thus allowing to evaluate differently the same forecast error when only a few fatalities occur, as opposed to when there is a large number of fatalities. Finally, with the linear exponential (linex) loss function we impose asymmetric weights, with more penalty given to under-prediction than to over-prediction. This reflects the fact that the social cost of the two errors, under- and over-prediction, are different, as the cost of not responding to the pandemic and incurring in a large loss of lives in the future is often regarded to be much higher than the economic and social cost of responding too quickly, imposing a lockdown when it is not necessary (on the precautionary principle in public health see, for example, Goldstein 2001).

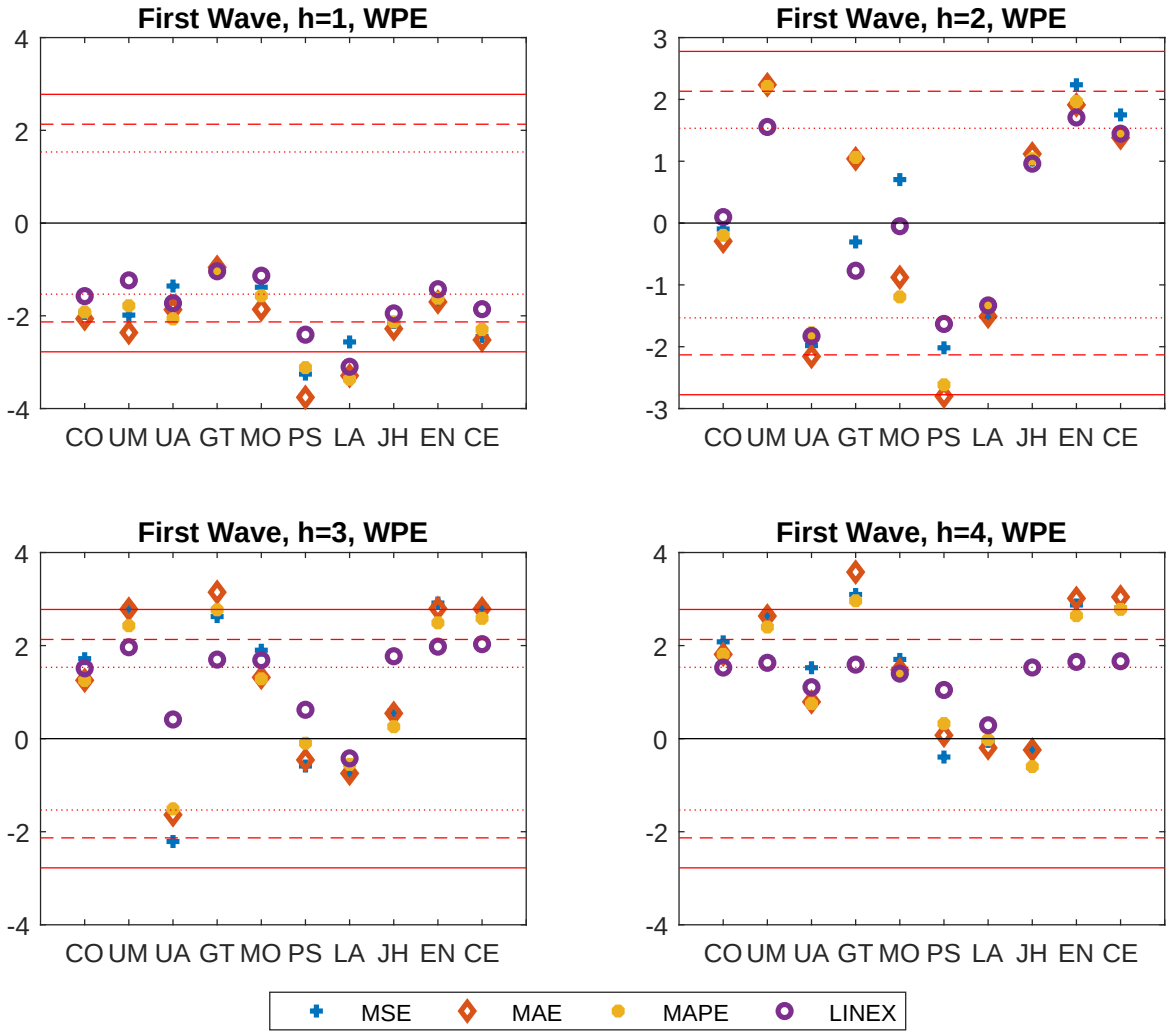
The findings are summarized in Figures 3 and 4 for the first wave, and in Figures 5 and 6 for

Figure 3: Forecast evaluation with WCE - First Wave



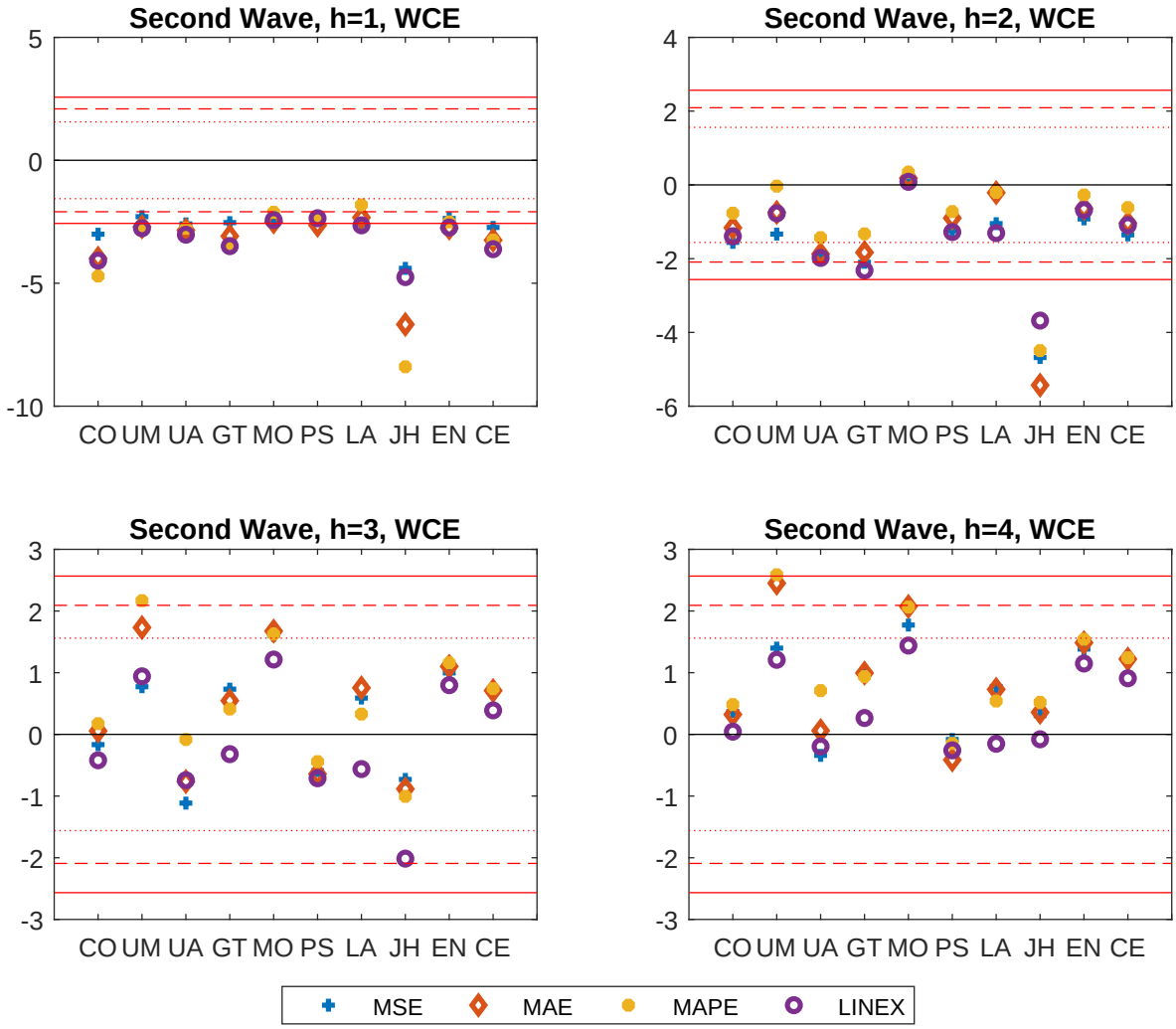
This figure reports the test statistic for the test of equal predictive accuracy using the weighted covariance estimator (WCE) of the long-run variance and fixed-b asymptotics. The benchmark is a second degree polynomial model fitted on a rolling window of 5 observations. A positive value of the test statistic indicates lower loss for the forecaster, i.e. better performance of the forecaster relative to the polynomial model. Different loss functions are reported with different markers: plus refers to a quadratic loss function, diamond to the absolute loss function, filled circle to the absolute percentage loss function and empty circle to the asymmetric loss function. The dotted, dashed and continuous red horizontal lines denote respectively the 20%, 10% and 5% significance levels. The forecast horizons h are 1, 2, 3 and 4 weeks ahead.

Figure 4: Forecast evaluation with WPE - First Wave



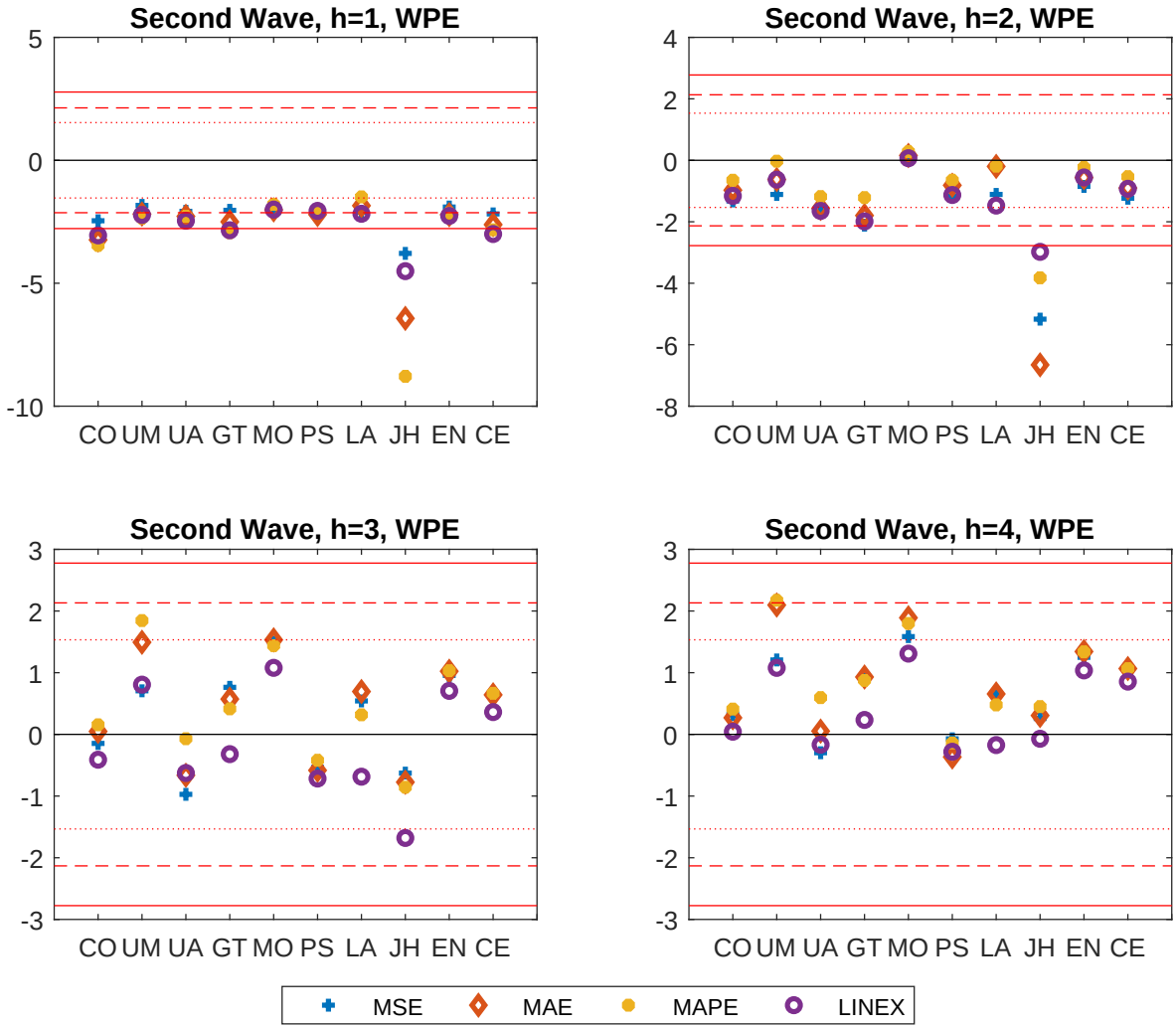
This figure reports the test statistic for the test of equal predictive accuracy using the weighted periodogram estimator (WPE) of the long-run variance and fixed-m asymptotics. The benchmark is a second degree polynomial model fitted on a rolling window of 5 observations. A positive value of the test statistic indicates lower loss for the forecaster, i.e. better performance of the forecaster relative to the polynomial model. Different loss functions are reported with different markers: plus refers to a quadratic loss function, diamond to the absolute loss function, filled circle to the absolute percentage loss function and empty circle to the asymmetric loss function. The dotted, dashed and continuous red horizontal lines denote respectively the 20%, 10% and 5% significance levels. The forecast horizons h are 1, 2, 3 and 4 weeks ahead.

Figure 5: Forecast evaluation WCE - Second Wave



This figure reports the test statistic for the test of equal predictive accuracy using the weighted covariance estimator (WCE) of the long-run variance and fixed-b asymptotics. The benchmark is a second degree polynomial model fitted on a rolling window of 5 observations. A positive value of the test statistic indicates lower loss for the forecaster, i.e. better performance of the forecaster relative to the polynomial model. Different loss functions are reported with different markers: plus refers to a quadratic loss function, diamond to the absolute loss function, filled circle to the absolute percentage loss function and empty circle to the asymmetric loss function. The dotted, dashed and continuous red horizontal lines denote respectively the 20%, 10% and 5% significance levels. The forecast horizons h are 1, 2, 3 and 4 weeks ahead.

Figure 6: Forecast evaluation WPE - Second Wave



This figure reports the test statistic for the test of equal predictive accuracy using the weighted periodogram estimator (WPE) of the long-run variance and fixed-m asymptotics. The benchmark is a second degree polynomial model fitted on a rolling window of 5 observations. A positive value of the test statistic indicates lower loss for the forecaster, i.e. better performance of the forecaster relative to the polynomial model. Different loss functions are reported with different markers: plus refers to a quadratic loss function, diamond to the absolute loss function, filled circle to the absolute percentage loss function and empty circle to the asymmetric loss function. The dotted, dashed and continuous red horizontal lines denote respectively the 20%, 10% and 5% significance levels. The forecast horizons h are 1, 2, 3 and 4 weeks ahead.

the second wave (the results for the quadratic loss function are also included, to facilitate the comparison). The dotted, dashed and continuous red horizontal lines denote respectively the 20%, 10% and 5% significance levels, which are, respectively, ±1.56, ±2.09 and ±2.57 for fixed-b asymptotics and ±1.53, ±2.13 and ±2.78 for fixed-m asymptotics.

First of all, we can observe how the results are not substantially different adopting a WCE or WPE estimator, as already documented in Table 3. Figures 3-6 also show that changing the loss function to absolute or absolute percentage does not have impact on the evaluation of predictive ability. On the other hand, the results are different if the linex function is used, as in this case for forecast horizons larger than one week ahead the null hypothesis is almost never rejected at the 5% significance level.

Comparing the forecasters predictive ability across the two sub-samples, it is clear that the forecasting teams performed relatively better during the first sub-sample compared to the second sub-sample. Whilst during the first sub-sample (Figures 3 and 4), many forecasting teams and the ensemble forecasts outperform the benchmark at the 3- and 4-week horizons, this is no longer true in the second sub-sample (Figures 5 and 6). As discussed earlier, it is possible that the increased transmissibility of the new strains of the virus that emerged in late 2020 and in 2021 were not immediately picked-up by the forecasting teams.

Considering the forecast horizon, it is clear that the simple polynomial benchmark outperforms all the teams, sometimes significantly so at the 1-week horizon, and often at the 2-week horizon. However, as the forecasting horizon is moved to three and four weeks, the teams improve their performance with respect to the polynomial benchmark. In the first wave, the Georgia Institute of Technology, Deep Out-break Project (GT) and the University of Massachusetts, Amhers (UM) teams, and also the Ensemble and the Core Ensemble outperform the benchmark at almost any level of statistical significance when the absolute and the absolute percentage loss functions are used. In the second wave, when we use the the absolute and the absolute percentage loss functions, we still document more accurate

predictions for several teams, for example for the University of Massachusetts, Amhers (UM), for the Northeastern University, Laboratory for the Modeling of Biological and Socio-technical Systems (MO) and for the ensembles, although these findings are seldom statistically significant. On the other hand, neither the forecasting teams nor the ensemble forecasts outperform the benchmark significantly, if the linex loss function is used. This seems to be mainly due to the fact that most forecasting teams (and the ensembles) under-predicted the fatalities, and this is more penalized with this loss function. Our empirical findings may also be evidence of the results discussed by Elliott and Timmermann (2004), who show how the equal weights ensemble is less appropriate in the presence of asymmetries in the loss function and in the distribution of the errors.

In general, we conclude that the ensemble forecasts deliver some of the best performing predictions. They often achieve statistically significant outperformance against the benchmark. This is the case for the 3- and 4-week ahead predictions during the first wave of the epidemic, when losses are evaluated using the quadratic, absolute or absolute percentage loss functions. Even when the outperfornance is not significant, ensemble predictions perform relatively well, in the sense of not underperforming the benchmark, even during the second wave when the individual models significantly underperform. Between the two ensemble forecasts, the wider ensemble obtained by the CDC performs slightly better against the benchmark compared to the Core Ensemble, illustrating once again the gains from combining a large number of predictions.

5 Conclusion

We evaluate the relative predictive accuracy of real-time forecasts for the number of COVID-19 fatalities in the US, produced by several competing forecasting teams in the first wave (20 June - 31 October 2020) and second wave (7 November 2020 - 20 March 2021) of the epidemic. Ensemble forecasts, that combine all available forecasts using an equal weights scheme, are

also included. Since sample sizes are small, we use fixed-smoothing asymptotics for the limit distribution of the scaled expected loss differential between two competing forecasts. We find that none of the forecasting teams outperform a simple statistical benchmark at the 1-week horizon; however, at longer forecasting horizons some teams show superior predictive ability at least during the first wave. The results for the second wave are less favorable for the forecasting teams, as none of the teams nor indeed the ensembles were able to perform statistically better than the benchmark even in the medium to long term horizons.

The ensemble forecasts deliver some of the most competitive predictions. Whilst they do not yield the best forecasts overall, they are competitive in the sense of delivering predictions that significantly outperform the benchmark at longer horizons during the first wave, and also never performing statistically worse than the benchmark even in during the second wave. In this sense, the Ensemble forecast may be seen as a robust choice. We also document that the broad based Ensemble published by the CDC is more accurate than the Core Ensemble, that only pools forecasts from the teams that we include in our exercise.

Overall, these results indicate that forecasts of the COVID-19 epidemic are valuable but need to be used with caution, and decision-makers should not rely on a single forecasting team (or a small set) to predict the evolution of the pandemic, but should hold a large and diverse portfolio of forecasts.

References

Bates, J. M., and C. W. J. Granger (1969) 'The combination of forecasts.' *OR* 20(4), 451–468

Choi, Hwan-sik, and Nicholas M. Kiefer (2010) 'Improving robust model selection tests for dynamic models.' *The Econometrics Journal* 13(2), 177–204

Chowell, G., R. Luo, K. Sun, K. Roosa, A. Tariq, and C. Viboud (2020) 'Real-time forecasting of epidemic trajectories using computational dynamic ensembles.' *Epidemics* 30, 100379

Claeskens, Gerda, Jan R. Magnus, Andrey L. Vasnev, and Wendun Wang (2016) 'The forecast combination puzzle: A simple theoretical explanation.' *International Journal of Forecasting* 32(3), 754 – 762

Clark, Todd E, and Michael W McCracken (2013) 'Advances in forecast evaluation.' In 'Handbook of Economic Forecasting,' vol. 2 (Elsevier) pp. 1107–1201

Clemen, Robert T (1989) 'Combining forecasts: A review and annotated bibliography.' *International Journal of Forecasting* 5(4), 559–583

Coroneo, Laura, and Fabrizio Iacone (2020) 'Comparing predictive accuracy in small samples using fixed-smoothing asymptotics.' *Journal of Applied Econometrics*

Diebold, Francis X, and Roberto S Mariano (1995) 'Comparing predictive accuracy.' *Journal of Business & Economic Statistics* pp. 253–263

Elliott, Graham, and Allan Timmermann (2004) 'Optimal forecast combinations under general loss functions and forecast error distributions.' *Journal of Econometrics* 122(1), 47–79

Galloway, Summer E, Prabasaj Paul, Duncan R MacCannell, Michael A Johansson, John T Brooks, Adam MacNeil, Rachel B Slayton, Suxiang Tong, Benjamin J Silk, and Gregory L Armstrong (2021) 'Emergence of sars-cov-2 b. 1.1. 7 lineage—united states, december 29, 2020–january 12, 2021.' *Morbidity and Mortality Weekly Report* 70(3), 95

Giacomini, Raffaella, and Halbert White (2006) 'Tests of conditional predictive ability.' *Econometrica* 74(6), 1545–1578

Goldstein, Bernard D (2001) 'The precautionary principle also applies to public health actions.' *American Journal of Public Health* 91(9), 1358–1361

Gonçalves, Sílvia, and Timothy J Vogelsang (2011) 'Block bootstrap hac robust tests: The sophistication of the naive bootstrap.' *Econometric Theory* pp. 745–791

Harvey, David I, Stephen J Leybourne, and Emily J Whitehouse (2017) 'Forecast evaluation tests and negative long-run variance estimates in small samples.' *International Journal of Forecasting* 33(4), 833–847

Hualde, Javier, and Fabrizio Iacone (2017) 'Fixed bandwidth asymptotics for the studentized mean of fractionally integrated processes.' *Economics Letters* 150, 39–43

Jiang, Feiyu, Zifeng Zhao, and Xiaofeng Shao (2020) 'Time series analysis of covid-19 infection curve: A change-point perspective.' *Journal of Econometrics*

Kiefer, Nicholas M, and Timothy J Vogelsang (2005) 'A new asymptotic theory for heteroskedasticity-autocorrelation robust tests.' *Econometric Theory* 21(6), 1130–1164

Lazarus, Eben, Daniel J Lewis, James H Stock, and Mark W Watson (2018) 'HAR inference: recommendations for practice.' *Journal of Business & Economic Statistics* 36(4), 541–559

Li, Shaoran, and Oliver Linton (2020) 'When will the covid-19 pandemic peak?' *Journal of Econometrics*

Manski, Charles F (2020) 'Forming covid-19 policy under uncertainty.' *Journal of Benefit-Cost Analysis* pp. 1–20

Patton, Andrew J, and Allan Timmermann (2007) 'Properties of optimal forecasts under asymmetric loss and nonlinearity.' *Journal of Econometrics* 140(2), 884–918

Politis, Dimitris N, and Joseph P Romano (1994) 'The stationary bootstrap.' *Journal of the American Statistical Association* 89(428), 1303–1313

Reich, N.G., C.J. McGowan, T.K. Yamana, A. Tushar, E.L. Ray, and D. Osthus *et al.* (2019) 'Accuracy of real-time multi-model ensemble forecasts for seasonal influenza in the U.S.' *PLoS Computational Biology* 15(11), 1–19

Smith, Jeremy, and Kenneth F. Wallis (2009) 'A simple explanation of the forecast combination puzzle.' *Oxford Bulletin of Economics and Statistics* 71(3), 331–355

Stock, James H, and Mark W Watson (1998) 'A comparison of linear and nonlinear univariate models for forecasting macroeconomic time series.' Technical Report, National Bureau of Economic Research

Sun, Yixiao (2013) 'A heteroskedasticity and autocorrelation robust f test using an orthonormal series variance estimator.' *The Econometrics Journal* 16(1), 1–26

Timmermann, Allan (2006) 'Forecast combinations.' *Handbook of Economic Forecasting* 1, 135–196

A Additional results by subsample

Table A.1: Summary Statistics of Forecast Errors - First Wave

1-week ahead	Mean	Median	Std	Skew	AC(1)	AC(2)	AC(3)	AC(4)
CO	965.10	1060.50	1238.58	-0.59	0.61	-0.13	-0.57	-0.64
UM	1059.70	740.00	799.46	0.99	0.33	-0.09	-0.02	-0.12
UA	1149.25	718.50	2377.89	2.20	0.01	0.20	-0.03	0.01
GT	961.88	340.84	2235.86	3.32	-0.14	0.20	0.18	-0.09
MO	860.03	804.30	914.07	0.10	-0.01	0.18	0.19	-0.15
PS	1461.63	1469.50	1388.68	0.12	0.12	-0.17	0.12	0.20
LA	1583.01	1261.77	1034.43	0.68	0.21	0.13	-0.26	-0.26
JH	1314.88	1135.98	1028.01	0.21	0.60	0.16	-0.13	-0.35
EN	878.69	781.00	702.74	-0.08	0.22	0.10	0.15	-0.23
CE	1169.43	1072.08	739.99	0.24	0.24	0.33	0.20	-0.11
PO	184.05	240.70	863.85	-0.01	0.52	0.10	0.11	-0.18
2-week ahead	Mean	Median	Std	Skew	AC(1)	AC(2)	AC(3)	AC(4)
CO	1248.00	1622.00	2072.00	-0.76	0.60	0.05	-0.24	-0.39
UM	1286.20	1245.50	933.21	0.70	0.33	0.05	-0.22	-0.32
UA	535.85	-223.00	3737.17	0.23	0.23	0.32	0.01	-0.13
GT	1099.22	436.94	2391.34	2.65	-0.14	0.34	-0.01	0.08
MO	1334.24	1707.04	1769.66	-1.58	0.06	0.18	-0.16	-0.09
PS	1766.03	2284.75	2986.82	-1.13	0.25	-0.27	-0.08	0.17
LA	2873.01	2637.79	2117.38	0.18	0.41	0.11	-0.20	-0.45
JH	-24.23	-308.80	1874.64	0.38	0.68	0.32	-0.18	-0.40
EN	988.15	653.50	1227.02	0.41	0.37	0.14	-0.23	-0.33
CE	1264.79	1296.37	1069.34	0.25	0.40	0.24	-0.11	-0.15
PO	665.81	1050.20	2286.05	-0.30	0.72	0.31	0.13	-0.11
3-week ahead	Mean	Median	Std	Skew	AC(1)	AC(2)	AC(3)	AC(4)
CO	1733.15	2588.00	2487.69	-1.11	0.49	-0.04	-0.10	-0.31
UM	1543.35	1401.00	1212.55	0.03	0.15	0.16	-0.07	-0.26
UA	-307.65	-1466.00	5456.38	0.26	0.36	0.40	0.01	-0.20
GT	1179.99	869.56	2787.10	2.08	-0.07	0.45	-0.04	-0.01
MO	1761.96	2198.23	3470.14	-1.61	0.17	0.19	-0.22	-0.25
PS	1303.08	2203.00	5607.86	-1.79	0.40	-0.19	-0.18	-0.08
LA	4844.44	4477.22	3177.37	0.13	0.51	0.15	-0.25	-0.50
JH	-2536.57	-2703.05	3367.77	-0.01	0.80	0.49	0.16	-0.08
EN	1151.16	648.45	1792.83	0.71	0.50	0.25	-0.20	-0.38
CE	1190.22	948.76	1733.00	0.16	0.48	0.16	-0.15	-0.38
PO	1546.59	1622.03	4607.05	-0.36	0.81	0.46	0.20	-0.05
4-week ahead	Mean	Median	Std	Skew	AC(1)	AC(2)	AC(3)	AC(4)
CO	2495.20	3046.00	2746.48	-0.53	0.46	-0.14	-0.11	-0.24
UM	1985.65	2532.00	2120.85	-0.99	-0.09	-0.01	0.18	-0.05
UA	-862.95	-3495.00	7232.63	0.73	0.47	0.43	0.14	-0.22
GT	1346.80	293.71	3724.08	1.70	0.04	0.44	-0.02	-0.02
MO	2238.44	3443.73	5868.38	-1.70	0.29	0.22	-0.25	-0.33
PS	200.98	1934.75	9607.37	-1.97	0.48	-0.08	-0.23	-0.22
LA	6953.06	6431.77	4444.55	0.16	0.56	0.32	0.01	-0.32
JH	-6589.80	-5408.15	5803.66	-0.38	0.85	0.66	0.28	0.07
EN	1309.30	368.50	2490.15	1.14	0.52	0.37	-0.01	-0.34
CE	970.92	597.7	2736.02	0.36	0.56	0.25	-0.13	-0.56
PO	2672.81	2938.39	7805.47	-0.35	0.83	0.53	0.25	-0.04

Notes: The table reports summary statistics of forecast errors for the teams, the Ensemble (EN), the Core Ensemble (CE) and the polynomial (PO) forecasts. The table reports mean, median, standard deviation (std), skewness (skew), and autocorrelation coefficients up to order 4 (AC(1), AC(2), AC(3), AC(4)). Weekly observations from June 20, 2020 to October 31, 2021.

Table A.2: Summary Statistics of Forecast Errors - Second Wave

1-week ahead	Mean	Median	Std	Skew	AC(1)	AC(2)	AC(3)	AC(4)
CO	3986.60	4375.00	5825.61	-0.56	0.01	-0.13	0.01	0.38
UM	5668.05	5257.00	4023.67	0.22	0.71	0.50	0.28	0.04
UM	5668.05	5257.00	4023.67	0.22	0.71	0.50	0.28	0.04
GT	4428.55	4503.65	4006.28	-0.38	0.38	0.20	0.16	0.14
MO	4181.28	4291.58	3378.33	-0.13	0.58	0.30	0.24	0.27
PS	4654.20	4090.50	3395.22	0.15	0.65	0.27	0.17	0.05
LA	4138.55	3610.50	3544.55	0.43	0.33	-0.05	0.00	0.24
JH	8878.81	9467.99	4306.70	-0.21	0.43	-0.01	-0.05	0.21
EN	4629.95	4108.50	3705.17	0.05	0.56	0.25	0.22	0.27
CE	5264.19	4893.07	3520.45	0.10	0.56	0.18	0.18	0.28
PO	-379.36	-446.50	3019.97	-0.34	0.19	-0.51	-0.22	0.36
2-week ahead	Mean	Median	Std	Skew	AC(1)	AC(2)	AC(3)	AC(4)
CO	6227.75	6081.00	6569.00	-0.01	0.21	-0.19	0.05	0.34
UM	6375.15	4424.50	5275.00	0.57	0.54	0.09	0.26	0.31
UA	8637.60	7281.00	7038.68	0.63	0.63	0.15	0.20	0.21
GT	5562.74	5781.50	5831.40	-0.84	0.15	-0.12	0.22	0.15
MO	4674.44	4429.20	4508.98	0.13	0.48	-0.01	0.16	0.36
PS	5752.08	5342.25	6493.53	0.08	0.52	0.04	-0.01	0.00
LA	5383.28	4634.00	5386.77	0.77	0.01	-0.35	-0.17	0.39
JH	9996.60	10425.31	5488.58	-0.81	0.12	-0.48	-0.28	0.13
EN	5657.25	5428.00	5277.33	0.31	0.44	-0.12	0.11	0.26
CE	6576.20	5324.99	4684.47	0.40	0.42	-0.19	0.12	0.35
PO	-844.56	1196.90	6733.56	-0.33	0.37	-0.37	-0.25	0.11
3-week ahead	Mean	Median	Std	Skew	AC(1)	AC(2)	AC(3)	AC(4)
CO	8210.65	8926.50	8696.68	-0.02	0.49	0.06	0.03	0.24
UM	7277.40	5187.00	7173.65	1.06	0.36	-0.07	0.10	0.43
UA	11597.15	11219.50	9903.23	1.18	0.56	0.04	0.03	0.25
GT	6796.22	8185.65	7739.23	-1.30	0.04	0.00	0.19	0.17
MO	5231.13	4332.77	6034.96	0.33	0.50	0.15	-0.05	0.22
PS	6956.88	9382.50	10975.21	-0.38	0.49	0.03	-0.02	-0.05
LA	7140.59	7402.15	7479.55	0.30	-0.10	-0.34	-0.23	0.39
JH	11223.98	10597.45	7117.63	-0.59	0.13	-0.29	-0.27	0.17
EN	6751.10	6560.00	6618.05	0.41	0.45	-0.01	-0.12	0.15
CE	8054.25	7246.77	5988.86	0.58	0.43	-0.11	-0.03	0.34
PO	-1491.33	3231.80	11592.63	-0.44	0.52	-0.12	-0.26	-0.04
4-week ahead	Mean	Median	Std	Skew	AC(1)	AC(2)	AC(3)	AC(4)
CO	10694.65	14709.50	12328.52	-0.25	0.67	0.21	0.07	0.12
UM	7941.60	4073.50	9640.82	1.21	0.34	-0.20	0.00	0.34
UA	14906.85	12236.50	12611.04	1.62	0.56	-0.08	-0.06	0.24
GT	8404.01	9258.78	10673.80	-1.12	0.26	-0.16	0.24	0.21
MO	5525.68	4118.23	8338.64	0.14	0.63	0.26	-0.01	0.11
PS	7915.83	12412.50	16483.96	-0.70	0.45	0.06	0.01	-0.05
LA	9593.30	10939.26	11087.67	-0.26	0.09	-0.15	-0.12	0.45
JH	12129.44	11167.02	10763.97	0.02	0.34	0.01	0.05	0.30
EN	7972.85	7476.50	8910.65	0.23	0.57	0.06	-0.04	0.13
CE	9638.92	9422.56	8009.63	0.31	0.53	-0.04	0.01	0.28
PO	-3275.21	6527.39	17741.72	-0.58	0.58	0.05	-0.09	0.03

Notes: The table reports summary statistics of forecast errors for the teams, the Ensemble (EN), the Core Ensemble (CE) and the polynomial (PO) forecasts. The table reports mean, median, standard deviation (std), skewness (skew), and autocorrelation coefficients up to order 4 (AC(1), AC(2), AC(3), AC(4)). Weekly observations from November 7, 2020 to March 20, 2021.