International Journal of Forecasting 39 (2023) 606–622

Contents lists available at ScienceDirect

International Journal of Forecasting

journal homepage: www.elsevier.com/locate/ijforecast

Testing the predictive accuracy of COVID-19 forecasts[☆]

Laura Coroneo [a,*], Fabrizio Iacone [b,a], Alessia Paccagnini [c,d], Paulo Santos Monteiro [a]

[a] University of York, United Kingdom
[b] Università degli Studi di Milano, Italy
[c] University College Dublin, Ireland
[d] CAMA, Australia

ARTICLE INFO

Keywords:
Forecast evaluation
Forecasting tests
Epidemic
COVID-19 Infectious diseases
Mortality Time series methods

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 for 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. 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.

1. Introduction

Forecasting the evolution of an epidemic is of utmost importance for policymakers and healthcare providers. Timely and reliable forecasts are necessary to help health authorities and the community at large cope 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 a rapidly growing number of research teams have developed 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 4 weeks. The forecasting teams that submit their forecasts to the CDC include data scientists, epidemiologists, and statisticians. They 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

☆ We thank the editor, Pierre Pinson, and three anonymous reviewers for comments that helped improve the paper. We are also grateful to Valentina Corradi, Paul Levine, Massimiliamo Marcellino, Elmar Mertens, Barbara Rossi and participants to the seminars at the University of Durham, University of Surrey, the University of York, University College Dublin, Bank of Greece, the 41st International Symposium on Forecasting 2021, the International Association for Applied Econometrics Annual Conference 2021, the 2021 North American Summer Meeting of the Econometric Society, 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.

* Correspondence to: Department of Economics and Related Studies, University of York, YO22 1DL, York, United Kingdom.
E-mail address: laura.coroneo@york.ac.uk (L. Coroneo).

https://doi.org/10.1016/j.ijforecast.2022.01.005
0169-2070/© 2022 Published by Elsevier B.V. on behalf of International Institute of Forecasters.

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 & Mariano, 1995; Giacomini & White, 2006) adopts a model-free perspective to compare competing forecasts, imposing assumptions only on the forecast error 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 & 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 autocovariance 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 are available (see Coroneo & Iacone, 2020; Harvey, Leybourne, & 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 from 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 around the second half of the sample. Accordingly, we perform our forecast evaluation separately on two sub-samples of equal size: the first evaluation sub-sample (from June 20, 2020 to October 31, 2020) and the second evaluation sub-sample (from November 7, 2020 to March 20, 2021). This implies that for each evaluation sub-sample 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 that combines 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 on 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 are 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, including 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 weeks ahead), the forecasters become more competitive, and some statistically outperform the simple benchmark, especially in the first evaluation sub-sample. 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 evaluation sub-sample, but even in the second sub-sample 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 modeling 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. Finally, in the Online Appendix, we perform a Monte Carlo simulation exercise to study the size and power properties of the two tests of equal predictive ability with fixed-smoothing asymptotics for the sample sizes considered in our empirical study, and consider several additional experiments, including some alternative benchmark forecasts.

2. Forecast evaluation

We consider the time series of cumulative weekly deaths $\{y_1, \ldots, y_T\}$, with T the sample size for which forecasts are available. 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 $i = 1, 2$ denotes 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 $L\left(e^{(i)}_{t|t-h}\right) = \left(e^{(i)}_{t|t-h}\right)^2$ and an absolute value loss would be $L\left(e^{(i)}_{t|t-h}\right) = \left|e^{(i)}_{t|t-h}\right|$. The null hypothesis of equal predictive ability of the two forecasts is

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

Denote the time-t loss differential between the two forecasts as

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

and the sample mean of the loss differential as

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

where $w \equiv \max(w_1, w_2)$.

When a large sample T is available, standard asymptotic theory may provide 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 T has only a moderate size, as it is indeed the case when comparing the 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),

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

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

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

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

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

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 tradeoff.

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

$$\sqrt{T}\frac{\overline{d}}{\widehat{\sigma}_{DAN,m}} \to_d t_{2m}, \quad (3)$$

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 standardization 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 tradeoff 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). The 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 $M = \left\lfloor T^{1/2}\right\rfloor$ 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 $m = \left\lfloor T^{1/3}\right\rfloor$ for the weighted periodogram estimate of the long-run variance using the Daniell kernel.

In Appendix A, we perform a Monte Carlo simulation exercise to investigate the empirical size and power of the two tests for sample sizes that match the ones in our empirical study. We simulate forecast errors as in Clark (1999), Coroneo and Iacone (2020), Diebold and Mariano (1995). Our findings indicate that both tests are, in general, correctly sized, even when only 20 observations are available and in presence of autocorrelation of the loss differential, although the test with WCE and fixed-b asymptotics can be slightly oversized in short samples and in presence of strong autocorrelation. On the other hand, the test with WPE and fixed-m asymptotics trails slightly behind the test with WCE in terms of power.

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

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 modeling methodology and whether the model considers a change in the assumptions about policy interventions. In the fourth column, "yes" means that the modeling 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.

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 are 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 weeks ahead), July 7 (3 weeks ahead), and July 12 (4 weeks ahead). Realized values are also taken 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 modeling 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 EN) and one computed by us (that combines the individual forecasts from the eight teams listed in Table 1 and that we call the core ensemble, CE).[3]

Combining forecasts is an effective procedure when there is uncertainty about the model and 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 forecasts is useful, as it helps to diversify risk and to pool information (see Bates & 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 & 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 Claeskens, Magnus, Vasnev, & Wang, 2016; Clemen, 1989; Smith & Wallis, 2009).

In epidemiology, forecast combination has proved its ability to improve on the performance of individual competing models. For example, Reich et al. (2019) found

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

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

[3] The CDC ensemble forecast is produced in collaboration with several research groups who form part of the COVID-19 Forecast Hub consortium (see https://covid19forecasthub.org/ for a detailed description).

that ensemble forecasting for influenza performed better on average against the constituting models; similar results were also obtained by Chowell et al. (2020) in the Ebola Forecasting Challenge. Both of 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) also considered 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 treats equally all the available forecasts. Specifically, it is obtained by combining the forecasts of all the teams that submitted to the CDC, as long as they submitted forecasts up to 4 weeks ahead and these forecasts were at least equal to the level observed on the day the forecast was submitted. The weekly composition of the pool of models contributing to the CDC ensemble forecast changes, and it includes, in general, a larger number of teams than the one we consider in our evaluation exercise.[4] 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 a 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. This experience may give them an edge relative to 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:

$$y_t = \beta_0 + \beta_1 t + \beta_2 t^2 + u_t, \quad (4)$$

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 (2021), 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; Li and Linton (2021) 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 by combining all the forecasts of the teams in Table 1, and the forecast of the polynomial benchmark (PO), described above.

Fig. 1 plots all the forecasts considered for the 1- to 4-week-ahead forecasting horizons, alongside the realized data. Comparing the graphs in Fig. 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.

Fig. 2 plots the forecast errors for each forecasting team, the ensembles, and the benchmark (computed as the difference between the realization and the point forecast). The figure indicates that most forecasting teams seem to have systematically under-predicted the target, in particular in the second part of the sample. This is, of course, relevant for policymakers if the costs of over-prediction and under-prediction are different. Fig. 2 also shows that the size of the forecast errors is increasing with the forecasting horizon, suggesting that there is more uncertainty about the evolution of the epidemic in the long run (4 weeks ahead), compared to the short run (1 week ahead).

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 autocorrelations up to order 4 (AC(1), AC(2), AC(3), AC(4)). With the exception of the benchmark, the average of the forecast errors is positive for all forecasts, meaning that the forecasters tend to under-predict the fatalities.

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

[4] In July 2020, the COVID-19 Forecast Hub changed the way it constructed the CDC ensemble forecast (We thank an anonymous reviewer for bringing this to our attention). Up until the week ending on July 18, 2020, the CDC ensemble forecast was obtained from an equally weighted average of forecasts across all eligible models. After that date, the methodology changed and the ensemble was obtained from the median forecast across all the eligible models (see the COVID-19 Hub documentation, and, in particular, Ray et al., 2020).

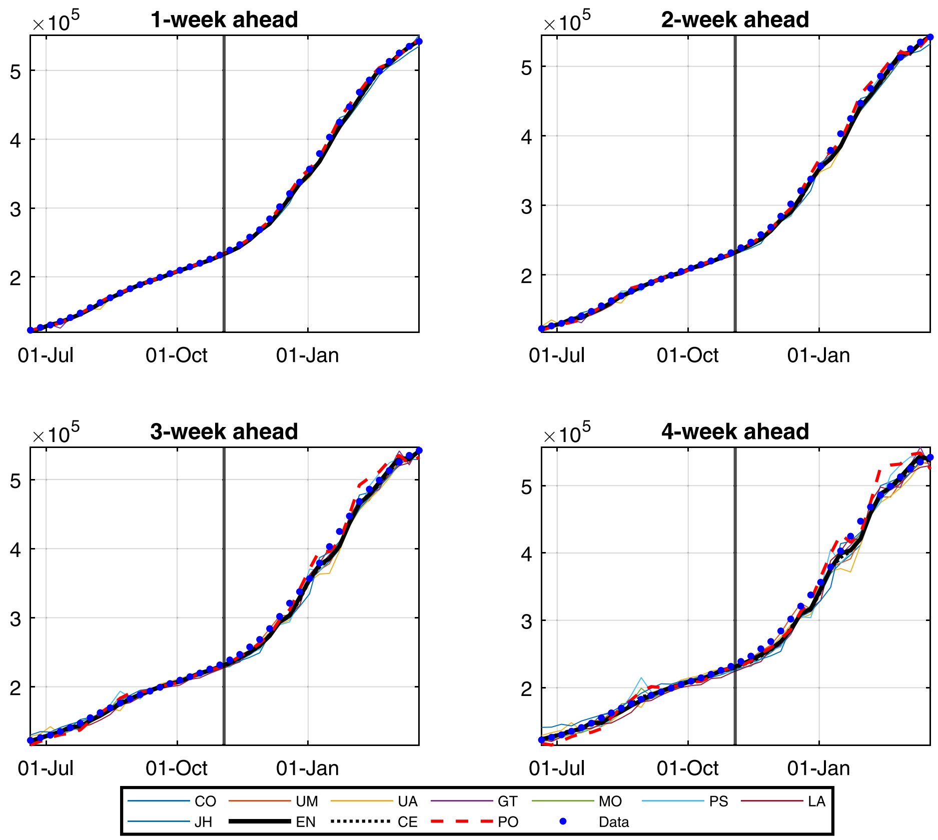

Fig. 1. Cumulative deaths in the US, observed vs. forecasted.
Note: Forecasts at forecasting horizons from 1 to 4 weeks, along with realized cumulative fatalities. Weekly observations from June 20, 2020 to March 20, 2021. The vertical line indicates November 3, 2020 and delimits the two sub-samples. 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 by combining all the forecasts of the teams in Table 1, and PO denotes the polynomial benchmark.

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 superior to simple statistical benchmarks.

From Table 2 we can also observe that the forecast errors are autocorrelated, as documented in 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 the 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 forecasts.

One explanation of this higher-order autocorrelation in the forecast errors and the fact that the forecasting teams systematically under-predict 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.

Finally, Fig. 2 shows a break in the volatility of the forecast errors across the first and the second halves of the sample (as illustrated by the vertical line in each diagram of Figs. 1–2). This is also shown in Tables 3–4, where we report summary statistics for the forecast errors in the two sub-samples. In particular, we note that the volatility of the forecast errors is considerably higher in the second sub-sample. Such decline in the quality of the forecasts 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 at the end of 2020, with specific mutations in the spike protein of SARS-CoV-2 resulting in increased transmissibility. Consistent

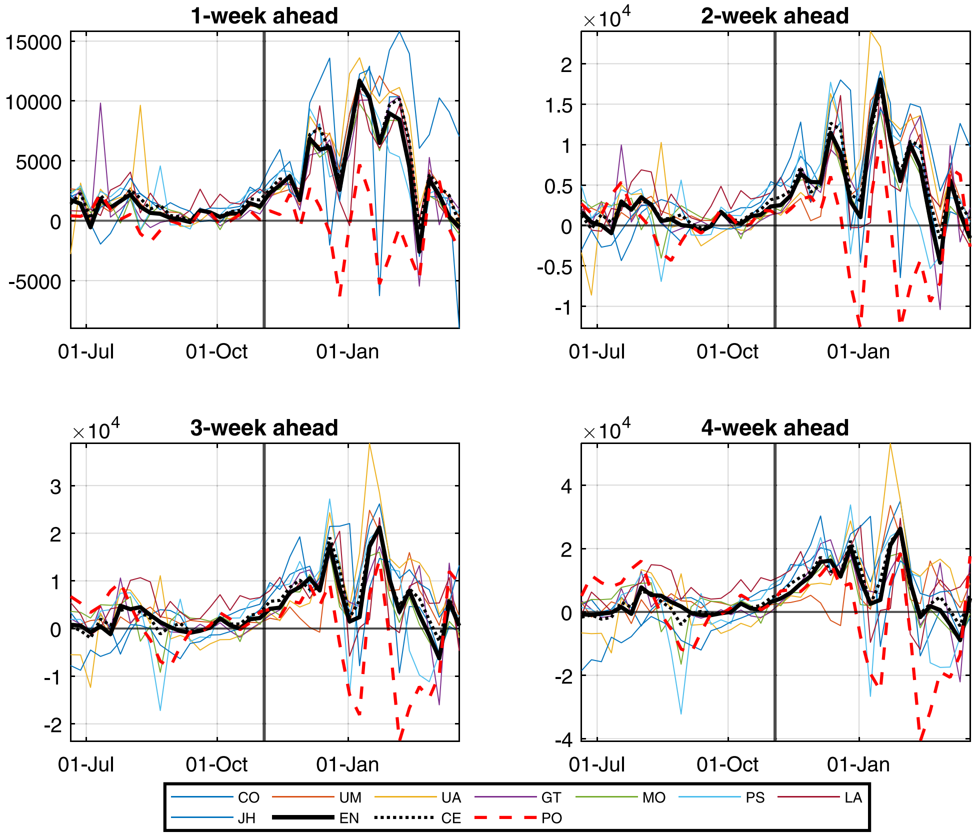

Fig. 2. Forecast errors.
Note: Forecast errors at forecasting horizons from 1 to 4 weeks. Weekly observations from June 20, 2020 to March 20, 2021. The vertical line indicates November 3, 2020 and delimits the two sub-samples. The names of the forecasting teams are as in Table 1; EN denotes the CDC ensemble forecast, CE denotes the core ensemble, and PO the polynomial benchmark. Forecast errors are defined as the realized value minus the forecast.

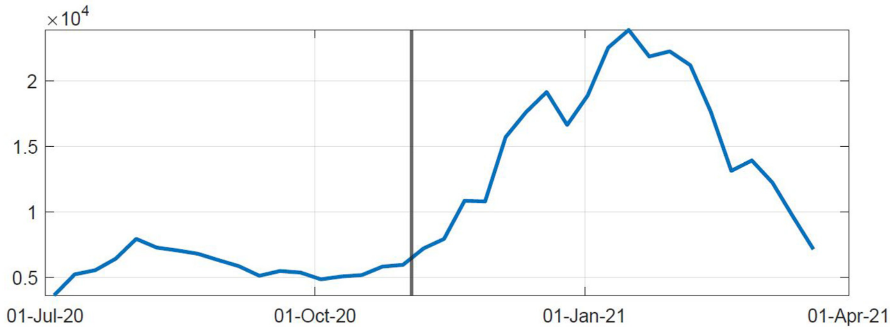

Fig. 3. Incident deaths in the US over the evaluation period.
Note: Weekly observations from June 20, 2020 to March 20, 2021. The vertical line indicates November 3, 2020 and delimits our two evaluation sub-samples.

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 et al., 2021). This is illustrated in Fig. 3, showing the incident deaths across the overall evaluation period. Thus the increased forecasting errors may be driven by the sudden heightened incident deaths and associated increased slope of the cumulative deaths curve.

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 weeks 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 weeks 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 weeks 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 the 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.

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

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 5.

We conduct the analysis separately for the two evaluation sub-samples identified in Fig. 2. This yields 20 observations for each sub-sample, underlying the importance of alternative asymptotics in evaluating predictive ability.[5] In the baseline analysis, we evaluate forecast

[5] In Appendix C, we conduct the test of equal predictive ability on the full sample.

Table 3
Summary statistics of forecast errors – first evaluation sub-sample.

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 weeks 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 weeks 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 weeks 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 the 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.

errors relying on the absolute value loss function. We present the test statistics using both the weighted covariance estimator with the Bartlett kernel (WCE) and the weighted periodogram estimator with the 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 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 bandwidth 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 first consider the upper-panel of Table 5, which reports results for the first sub-sample from June 20, 2020 to October 31, 2020. The results indicate that no forecasting scheme predicts better than the benchmark at the 1-week forecasting horizon. In fact, we find that the benchmark often significantly outperforms the forecasting teams. On the other hand, at the 2-week horizon, the sign of some test statistics turns from negative to positive,

Table 4
Summary statistics of forecast errors – second evaluation sub-sample.

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
UA	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 weeks 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 weeks 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 weeks 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 the 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.

reflecting a smaller relative loss by the forecasting teams. The relative performance of the forecasters improves further at longer horizons (3 and 4 weeks ahead), and we observe statistically significant relative gains in performance for some forecasting teams and the ensemble forecasts.

The MAEs for the first sub-sample reported in Table 6 indicate that, at 3 and 4 weeks ahead, both ensemble forecasts (the EN ensemble provided by the CDC and the CE core ensemble constructed by combining all the forecasts of the teams in Table 1) performed better than the individual forecasting teams. This finding is consistent with the consensus in the literature about the advantages of forecast combination (see Stock & Watson, 1998; Timmermann, 2006).

Turning to the second sub-sample from November 7, 2020 to March 20, 2021, we can see from Table 5 that the benchmark still significantly outperforms some teams and the ensemble forecasts at the shortest horizon. However, in this sub-sample, the forecasting teams and the ensembles fail to significantly outperform the benchmark even at the longer horizons. The MAEs for the second sub-sample reported in Table 6 indicate that, also in this sub-sample, at 3- and 4-week horizons, the two ensembles performed better than most of the individual teams.

Finally, notice that the performance of the ensemble forecast published by the CDC is similar to the one of the core ensemble constructed by combining all the forecasts of the teams in Table 1. However, the MAEs for the

Table 5
Tests for equal predictive ability.

	1 week ahead		2 weeks ahead		3 week s ahead		4 weeks ahead	
First evaluation sub-sample: Jun 20, 2020–Oct 31, 2020								
	WCE	WPE	WCE	WPE	WCE	WPE	WCE	WPE
CO	−2.829**	−2.063	−0.340	−0.294	1.452	1.254	2.147*	1.812
UM	−2.231*	−2.362*	1.663	2.237*	2.939**	2.782**	2.953**	2.636*
UA	−1.993	−1.867	−1.378	−2.157*	−0.510	−1.636	0.689	0.791
GT	−1.149	−0.954	1.204	1.040	3.619**	3.147**	4.057**	3.579**
MO	−2.248*	−1.860	−0.645	−0.879	1.412	1.317	1.587	1.479
PS	−4.815**	−3.758**	−3.018**	−2.801**	−0.523	−0.458	0.086	0.073
LA	−3.527**	−3.292**	−1.862	−1.509	−0.925	−0.746	−0.241	−0.199
JH	−2.646**	−2.280*	1.194	1.119	0.714	0.544	−0.307	−0.243
EN	−1.840	−1.701	1.881	1.911	3.074**	2.798**	3.445**	3.014**
CE	−2.992**	−2.521*	1.381	1.381	2.937**	2.789**	3.472**	3.044**
Second evaluation sub-sample: Nov 7, 2020 – Mar 20, 2021								
	1 week ahead		2 weeks ahead		3 weeks ahead		4 weeks ahead	
	WCE	WPE	WCE	WPE	WCE	WPE	WCE	WPE
CO	−3.985**	−3.235**	−1.163	−0.972	0.055	0.048	0.322	0.269
UM	−2.704**	−2.186*	−0.747	−0.625	1.732	1.493	2.450*	2.097
UA	−2.845**	−2.279*	−1.875	−1.563	−0.765	−0.664	0.061	0.053
GT	−3.084**	−2.501*	−1.833	−1.797	0.546	0.572	0.995	0.930
MO	−2.504*	−2.016	0.183	0.157	1.674	1.532	2.073	1.890
PS	−2.641**	−2.207*	−0.900	−0.813	−0.643	−0.583	−0.413	−0.366
LA	−2.327*	−1.845	−0.211	−0.199	0.754	0.694	0.730	0.656
JH	−6.675**	−6.429**	−5.431**	−6.656**	−0.884	−0.775	0.357	0.306
EN	−2.756**	−2.212*	−0.639	−0.562	1.102	1.023	1.486	1.342
CE	−3.249**	−2.616*	−1.036*	−0.902	0.709	0.641	1.221	1.066

Note: The table reports 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 five observations. The forecast errors are evaluated using the absolute value loss function. 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 * respectively indicate two-sided significance at the 5% and 10% levels using fixed-b asymptotics for WCE and fixed-m asymptotics for WPE. and respectively indicate two-sided significance at the 5% and 10% levels 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).

ensemble are in all cases smaller 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 small.[6]

Results are similar overall, regardless of the type of estimator of the long-run variance. We also notice that findings 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.

4.1. Alternative loss functions

The absolute value loss function that we use in the baseline forecast evaluation reported in Table 5 is a common choice in forecast evaluation. In particular, the null hypothesis is the equality of the mean absolute 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. For these reasons, in this section we consider alternative loss functions.[7]

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:

- Quadratic: $L(e_t) = (e_t)^2$;
- Absolute percentage: $L(e_t) = |e_t|/(y_t - y_{t-1})$;
- Linex: $L(e_t) = \exp(e_t/(y_t - y_{t-1})) - e_t/(y_t - y_{t-1}) - 1$.

The quadratic loss function is a popular measure whereby large forecast errors are penalized more: in this case,

6 The equal predictive ability of the CDC ensemble forecast relative to the forecasting teams and the core ensemble is formally tested in Appendix E.

7 The teams submitting forecasts to the CDC were advised that their point forecasts would be evaluated with the mean absolute error loss function. The predictive median minimizes the mean absolute error and should, therefore, correspond to the optimal point forecast. However, if forecasters put greater weight on under-prediction and, thus, seek to minimize a linex-type loss function, Christoffersen and Diebold (1997) show that such a loss function implies an optimal point forecast that is a weighted sum of the mean and variance.

Table 6
MAE across sub-samples.

MAE								
	1st sub-sample				**2nd sub-sample**			
	1 week	2 weeks	3 weeks	4 weeks	1 week	2 weeks	3 weeks	4 weeks
CO	1304	1979	2486	3045	6013	7347	9740	13744
UM	1063	1330	1700	2559	5681	6399	7354	8640
UA	1544	2482	4118	5870	6201	8821	11790	14907
GT	1144	1430	1847	2512	5009	6846	9013	11601
MO	1034	1957	3221	5026	4516	5300	6182	7797
PS	1671	2827	4357	6367	4879	7036	11414	16466
LA	1583	3003	4983	7048	4260	5739	8275	12443
JH	1372	1431	3318	7016	8879	10453	12004	13382
EN	945	1168	1535	1899	4943	6274	7518	9422
CE	1169	1340	1636	2204	5365	6752	8318	10137
PO	654	1830	3848	6541	2415	5515	9896	15132
Relative MAE								
	1st sub-sample				**2nd sub-sample**			
	1 week	2 weeks	3 weeks	4 weeks	1 week	2 weeks	3 weeks	4 weeks
CO	1.99	1.08	0.65	0.47	2.49	1.33	0.98	0.91
UM	1.63	0.73	0.44	0.39	2.35	1.16	0.74	0.57
UA	2.36	1.36	1.07	0.90	2.57	1.60	1.19	0.99
GT	1.75	0.78	0.48	0.38	2.07	1.24	0.91	0.77
MO	1.58	1.07	0.84	0.77	1.87	0.96	0.62	0.52
PS	2.56	1.54	1.13	0.97	2.02	1.28	1.15	1.09
LA	2.42	1.64	1.29	1.08	1.76	1.04	0.84	0.82
JH	2.10	0.78	0.86	1.07	3.68	1.90	1.21	0.88
EN	1.44	0.64	0.40	0.29	2.05	1.14	0.76	0.62
CE	1.79	0.73	0.43	0.34	2.22	1.22	0.84	0.67
PO	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00

Notes: The table reports the MAE of forecast errors for each team, the ensemble (EN), the core ensemble (CE), and the polynomial (PO) forecasts. The top panel shows the MAE level, and the bottom panel shows the MAE relative to the MAE of the benchmark model. The first evaluation sub-sample is from June 20, 2020 to October 31, 2020, and the second evaluation sub-sample from November 7, 2020 to March 20, 2021.

it seems natural to interpret it as giving more weight to fatalities that happen when the epidemic is less predictable. The absolute percentage loss considers the scale of the number of fatalities occurring in the period, thus allowing for different evaluations of the same forecast error when only a few fatalities occur, as opposed to when there are 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, by imposing a lockdown when it is not necessary (on the precautionary principle in public health, see Goldstein, 2001).

The findings are summarized in Figs. 4 and 5 for the first evaluation sub-sample, and in Figs. 6 and 7 for the second evaluation sub-sample. The results for the absolute value loss function are also included, to facilitate the comparison. The dotted, dashed, and continuous red horizontal lines respectively denote 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 whether adopting a WCE or WPE estimator, as already documented in Table 5. Figs. 4–7 also show that changing the loss function to quadratic or absolute percentage does not have great impact on the evaluation of predictive ability. On the other hand, the results are different if the linex function is used: in this case, for forecast horizons larger than 1 week ahead, the null hypothesis is almost never rejected at the 5% significance level.

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 moves to 3 and 4 weeks, the teams improve their performance with respect to the polynomial benchmark. In the first evaluation sub-sample, the Georgia Institute of Technology, Deep Outbreak Project (GT), and the University of Massachusetts, Amherst (UM) teams, and also the CDC ensemble and the core ensemble, outperform the benchmark at almost any level of statistical significance when the quadratic and the absolute percentage loss functions are used.

In the second evaluation sub-sample, when we use the quadratic and the absolute percentage loss functions, we still document more accurate predictions for several teams, for example for the University of Massachusetts, Amherst (UM), for 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,

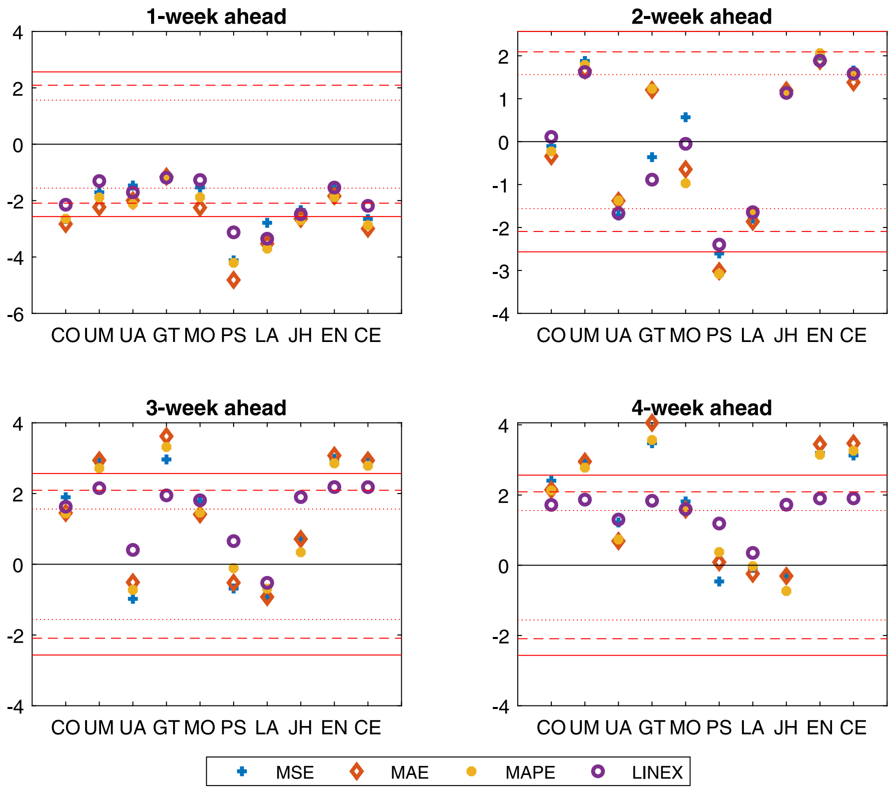

Fig. 4. Forecast evaluation with WCE – first evaluation sub-sample.
This figure reports the test statistic for the test of equal predictive accuracy using the weighted covariance estimator (WCE) of the long-run variance and fixed-*b* asymptotics. The benchmark is a second-degree polynomial model fitted on a rolling window of five 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: a plus sign refers to a quadratic loss function, a diamond to the absolute loss function, a filled circle to the absolute percentage loss function, and an empty circle to the asymmetric loss function. The dotted, dashed, and continuous red horizontal lines respectively denote the 20%, 10%, and 5% significance levels. The forecast horizons are 1, 2, 3, and 4 weeks ahead. The evaluation sample is from June 20, 2020 to October 31, 2020.

neither the forecasting teams nor the ensemble forecasts outperform the benchmark significantly when 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 viewed as offering support for 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. Notice, however, that the teams submitting forecasts to the CDC were advised that their point forecasts would be evaluated with the absolute value loss function, and therefore it is fair to conjecture that their predictions do not optimize the quadratic, the absolute percentage, or the linex loss functions. Had the teams been told that an alternative loss function was to be used to evaluate the forecast accuracy, then their point predictions might have been different. For example, if the linex loss function were used, the predictions would in general be larger.

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 evaluation sub-sample, when losses are evaluated using the absolute value, quadratic, or absolute percentage loss function. Even when the outperformance is not significant, ensemble predictions perform relatively well, in the sense of not underperforming the benchmark, even during the second evaluation sub-sample when the forecast errors are larger. 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.

4.2. Comparing the two evaluation sub-samples

Comparing the results of the tests for equal predictive ability across the two sub-samples, we note that the null hypothesis is more difficult to reject during the second evaluation sub-sample. In particular, whereas during the first sub-sample (Figs. 4 and 5), 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 (Figs. 6 and 7).

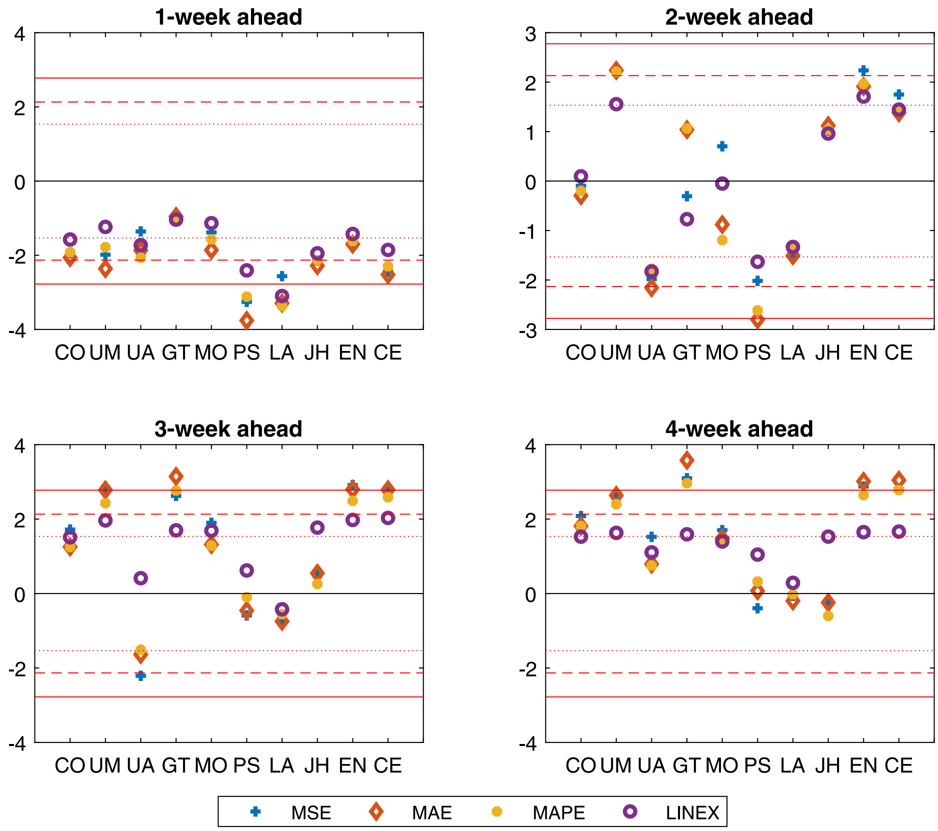

Fig. 5. Forecast evaluation with WPE – first evaluation sub-sample.
This figure reports the test statistic for the test of equal predictive accuracy using the weighted periodogram estimator (WPE) of the long-run variance and fixed-*m* asymptotics. The benchmark is a second-degree polynomial model fitted on a rolling window of five 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: a plus sign refers to a quadratic loss function, a diamond to the absolute loss function, a filled circle to the absolute percentage loss function, and an empty circle to the asymmetric loss function. The dotted, dashed, and continuous red horizontal lines respectively denote the 20%, 10%, and 5% significance levels. The forecast horizons are 1, 2, 3, and 4 weeks ahead. The evaluation sample is from June 20, 2020 to October 31, 2020.

Heightened incident deaths, associated with the increased transmissibility of the new strains of the virus that emerged in late 2020 and in 2021, may have affected the statistical performance of the tests of equal predictive ability in two ways: by making the task of forecasting more difficult (as evident by the larger MAEs in the second sub-sample in Table 6), and by inflating the long-run variance of the test statistic (thus reducing the power to detect a significant difference for given MAE differential). Of course, the purpose of the tests of equal predictive ability is not to compare predictive ability across different periods, so a more pronounced failure to reject the null hypothesis during the second evaluation sub-sample is not evidence that the models were less valuable in this period.

To examine further the reasons for the apparent change in the forecastability of the epidemic, in the bottom panel of Table 6 we report the ratio of the MAE of each forecasting team and the MAE of the benchmark. This enables us to compare the relative performance of each forecasting team and the benchmark across the two evaluation sub-samples. Considering for example the 4-week forecasting horizon, we can notice that for some forecasting teams and the ensembles, the ratios are considerably smaller for the first evaluation sub-sample compared to the second evaluation sub-sample. This is true for all the loss functions considered (MAE, RMSE, MAPE, and MLinex), and suggests that the forecasting environment was different across the first and the second evaluation sub-samples.[8]

4.3. Additional experiments

In the Online Appendix, we consider several additional experiments and empirical exercises. In particular, in Appendix C, we perform the tests of equal predictive ability for the full sample, instead of considering each evaluation sub-sample separately. In Appendix D, we use an alternative benchmark model, obtained based on fitting an AR(1) model to the log incidence of weekly deaths. Finally, in Appendix E, we use the CDC ensemble forecast as the benchmark model, to test formally the null of equal predictive ability of the forecasting teams and the ensemble forecast.

8 In Appendix B we report the RMSE, MAPE, and MLinex of the forecast errors across the two different evaluation sub-samples.

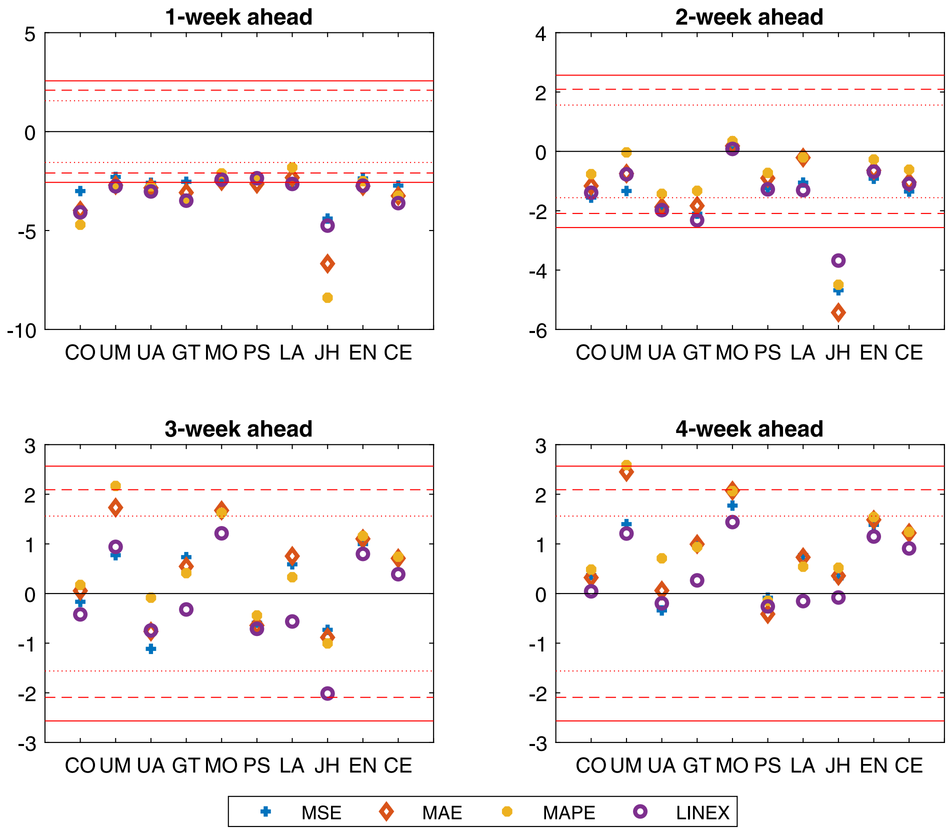

Fig. 6. Forecast evaluation WCE – second evaluation sub-sample.
This figure reports the test statistic for the test of equal predictive accuracy using the weighted covariance estimator (WCE) of the long-run variance and fixed-*b* asymptotics. The benchmark is a second-degree polynomial model fitted on a rolling window of five 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: a plus sign refers to a quadratic loss function, a diamond to the absolute loss function, a filled circle to the absolute percentage loss function, and an empty circle to the asymmetric loss function. The dotted, dashed, and continuous red horizontal lines respectively denote the 20%, 10%, and 5% significance levels. The forecast horizons are 1, 2, 3, and 4 weeks ahead. The evaluation sample is from November 7, 2020 to March 20, 2021.

5. Conclusion

We evaluated the relative predictive accuracy of real-time forecasts for the number of COVID-19 fatalities in the US produced by several competing forecasting teams for two evaluation periods: June 20, 2020 to October 31, 2020 (first evaluation sub-sample); and November 7, 2020 to March 20, 2021 (second evaluation sub-sample). Ensemble forecasts that combine all available forecasts using an equal-weights scheme were also included. Since the sample sizes are small, we used fixed-smoothing asymptotics for the limit distribution of the scaled expected loss differential between two competing forecasts. We found that none of the forecasting teams outperformed a simple statistical benchmark at the 1-week horizon; however, at longer forecasting horizons some teams showed superior predictive ability.

The ensemble forecasts delivered some of the most competitive predictions. Whilst they did not yield the best forecasts overall, they were competitive in the sense of delivering predictions that significantly outperformed the benchmark at longer horizons during the first evaluation sub-sample, and they never performed statistically worse than the benchmark, even in the second evaluation sub-sample. In this sense, the ensemble forecast may be seen as a robust choice. We also documented that the broader ensemble published by the CDC was more accurate than the core ensemble that only pooled forecasts from the teams that we included in our exercise.

Overall, our 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.

A natural extension of our analysis is to evaluate the interval forecasts with different levels of coverage submitted by the forecasting teams to the CDC. This would require choosing an appropriate loss function, for example the weighted interval score (see Bracher, Ray, Gneiting, & Reich, 2021), and applying the same alternative asymptotics we used here. In particular, recent work by Coroneo, Iacone, and Profumo (2019) shows that fixed-smoothing asymptotics may also be employed successfully to evaluate density forecasts.

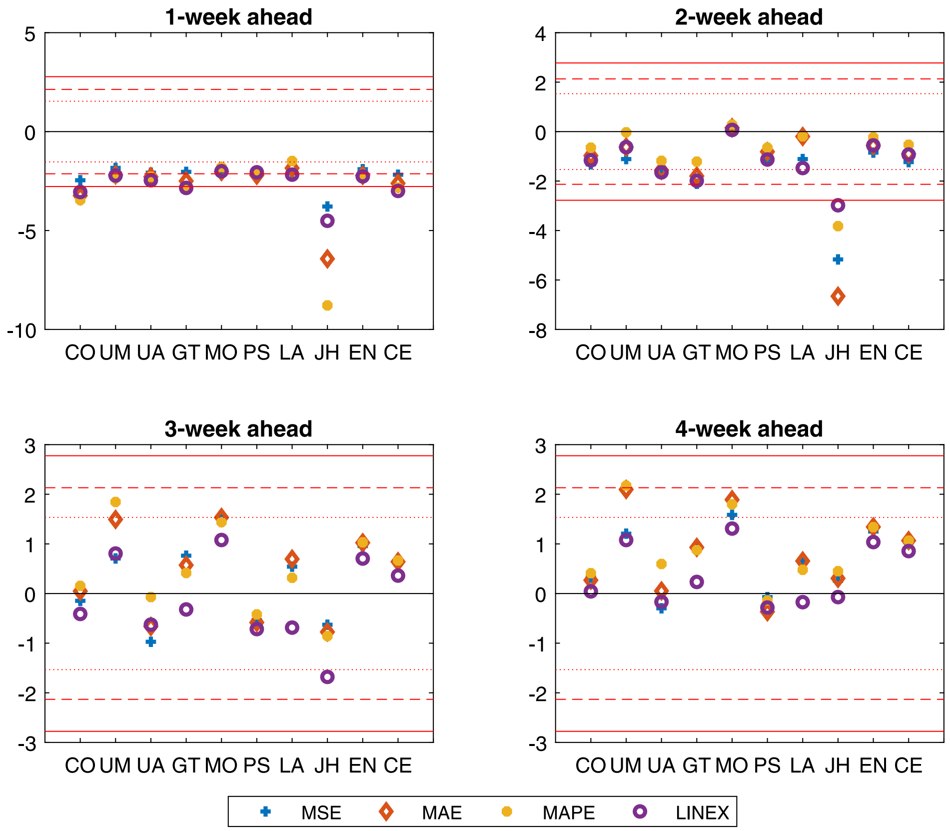

Fig. 7. Forecast evaluation WPE - second evaluation sub-sample.
This figure reports the test statistic for the test of equal predictive accuracy using the weighted periodogram estimator (WPE) of the long-run variance and fixed-*m* asymptotics. The benchmark is a second-degree polynomial model fitted on a rolling window of five 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: a plus sign refers to a quadratic loss function, a diamond to the absolute loss function, a filled circle to the absolute percentage loss function, and an empty circle to the asymmetric loss function. The dotted, dashed, and continuous red horizontal lines respectively denote the 20%, 10%, and 5% significance levels. The forecast horizons are 1, 2, 3, and 4 weeks ahead. The evaluation sample is from November 7, 2020 to March 20, 2021.

Another interesting extension to the current work is to consider the predictive accuracy for a panel data of forecasts (since the forecasting teams predict not only the national spread of the disease but also the regional evolution of the epidemic). Timmermann and Zhu (2019) propose methods for testing predictive accuracy for panel forecasts. In particular, they develop a panel-data Diebold–Mariano test for equal predictive accuracy that pools information across both the time-series and cross-sectional dimensions. Our analysis could, therefore, be extended to the evaluation of a panel of regional predictions.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Appendix A. Supplementary data

The Online Appendix includes further detail about: A) Monte Carlo study, B) Comparing forecast errors across sub-samples, C) Full sample testing results, D) AR(1) benchmark, E) Ensemble benchmark.

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.ijforecast.2022.01.005.

References

Bates, J. M., & Granger, C. W. J. (1969). The combination of forecasts. *OR*, *20*(4), 451–468.
Bracher, Johannes, Ray, Evan L., Gneiting, Tilmann, & Reich, Nicholas G. (2021). Evaluating epidemic forecasts in an interval format. *PLoS Computational Biology*, *17*(2), Article e1008618.
Choi, Hwan-sik, & Kiefer, Nicholas M. (2010). Improving robust model selection tests for dynamic models. *The Econometrics Journal*, *13*(2), 177–204.
Chowell, G., Luo, R., Sun, K., Roosa, K., Tariq, A., & Viboud, C. (2020). Real-time forecasting of epidemic trajectories using computational dynamic ensembles. *Epidemics*, *30*, Article 100379.
Christoffersen, Peter F., & Diebold, Francis X. (1997). Optimal prediction under asymmetric loss. *Economic Theory*, *13*(6), 808–817.
Claeskens, Gerda, Magnus, Jan R., Vasnev, Andrey L., & Wang, Wendun (2016). The forecast combination puzzle: A simple theoretical explanation. *International Journal of Forecasting*, *32*(3), 754–762.
Clark, Todd E. (1999). Finite-sample properties of tests for equal forecast accuracy. *Journal of Forecasting*, *18*(7), 489–504.

Clark, Todd E., & McCracken, Michael W. (2013). Advances in forecast evaluation. In *Handbook of economic forecasting, Vol. 2* (pp. 1107–1201). Elsevier.

Clemen, Robert T. (1989). Combining forecasts: A review and annotated bibliography. *International Journal of Forecasting*, *5*(4), 559–583.

Coroneo, Laura, & Iacone, Fabrizio (2020). Comparing predictive accuracy in small samples using fixed-smoothing asymptotics. *Journal of Applied Econometrics*, *35*(4), 391–409.

Coroneo, Laura, Iacone, Fabrizio, & Profumo, Fabio (2019). *A real-time density forecast evaluation of the ECB Survey of Professional Forecasters*: *Technical report*.

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

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

Galloway, Summer E., Paul, Prabasaj, MacCannell, Duncan R., Johansson, Michael A., Brooks, John T., MacNeil, Adam, et al. (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, & White, Halbert (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, & Vogelsang, Timothy J. (2011). Block bootstrap HAC robust tests: The sophistication of the naive bootstrap. *Economic Theory*, 745–791.

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

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

Jiang, Feiyu, Zhao, Zifeng, & Shao, Xiaofeng (2020). Time series analysis of COVID-19 infection curve: A change-point perspective. *Journal of Econometrics*, (in press) http://dx.doi.org/10.1016/j.jeconom.2020.07.039.

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

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

Li, Shaoran, & Linton, Oliver (2021). When will the Covid-19 pandemic peak? *Journal of Econometrics*, *220*(1), 130–157.

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

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

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

Ray, Evan L, Wattanachit, Nutcha, Niemi, Jarad, Kanji, Abdul Hannan, House, Katie, Cramer, Estee Y, et al. (2020). Ensemble forecasts of coronavirus disease 2019 (COVID-19) in the US. *MedRXiv*.

Reich, N. G., McGowan, C. J., Yamana, T. K., Tushar, A., Ray, E. L., Osthus, D., 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, & Wallis, Kenneth F. (2009). A simple explanation of the forecast combination puzzle. *Oxford Bulletin of Economics and Statistics*, *71*(3), 331–355.

Stock, James H., & Watson, Mark W. (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.

Timmermann, Allan, & Zhu, Yinchu (2019). Comparing forecasting performance with panel data. *Technical report*, CEPR Discussion Paper.