UNIVERSITY of York

Discussion Papers in Economics

No. 20/10

Testing the predictive accuracy of COVID-19 forecasts

Laura Coroneo, Fabrizio Iacone,
Alessia Paccagnini and Paulo Santos Monteiro

Department of Economics and Related Studies
University of York
Heslington
York, YO10 5DD

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

2nd October 2020

Abstract

We test the predictive accuracy of forecasts for the number of COVID-19 fatalities produced by several forecasting teams and collected by the United States Centers for Disease Control and Prevention (CDC), both at the national and state levels. We find three main results. First, at short-horizon (1-week ahead) no forecasting team outperforms a simple time-series benchmark. Second, at longer horizons (3 and 4-weeks 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 forecasts using uniform weights. In view of these results, collecting a wide range of forecasts and combining them in an ensemble forecast may be a safer 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.

[*] Updated testing results are available at https://sites.google.com/view/lauracoroneo/covid-19.

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. It is not surprising, therefore, that 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 deaths, hospitalizations, and cases produced by different institutions and research groups. 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 use different models and methods (e.g. structural time-dependent SEIR models, nonlinear Bayesian hierarchical models, Ridge 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 is 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 by b) is constant, as recommended by Kiefer and Vogelsang (2005). With this alternative asymptotics, denoted 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 (US), and also at the state level (focusing on New York, Florida, Texas, and California, the most populated States). We compare the predictive accuracy of forecasts for each team taken from the CDC portfolio relative to the forecasts of a simple benchmark model. The benchmark forecasts are obtained from fitting a second-order polynomial to the last five available observations, using a rolling window. An Ensemble forecast obtained combining all the available forecasts (with equal weights), is also considered.

Throughout, we test the hypothesis of equal predictive accuracy against the two-sided alternative.

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 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, as opposed to 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. We therefore evaluate the predictive accuracy of each forecasting team using a number of 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 penalises more under-prediction than over-prediction).

Our main findings are as follows. First, the simple benchmark outperforms the forecasters at the short horizon (one week ahead), often significantly so. Second, at longer horizons (up to four weeks ahead), the forecasters become more competitive and some statistically outperform the simple benchmark. This suggests that forecasters can successfully help inform forward looking policy decisions. Thirdly, the Ensemble forecast is among the best performing forecasts, underlining the virtues of model averaging when uncertainty prevails. This result 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.

Repeating the same exercise for each individual state, we document heterogeneity across states. When confronting forecasts at the state level, the benchmark model is even more competitive, and on occasions the forecasters are outperformed by the simple time-series benchmark even at longer horizons. On the other hand, the performance of the Ensemble forecast is at least comparable with the benchmark. These findings may indicate that a one-size-fits-all approach, in which one single model is used to produce forecasts at the national level and for each state is inappropriate, strengthening the case for an approach that does not depend too heavily on just one model. The good performance of the Ensemble forecast once again demonstrates the robustness of forecast combinations in the presence of model uncertainty. These results appear robust to changes in the loss function. In view of these results, we recommend that health authorities should not rely on a few selected forecasts: collecting many forecasts and combining them in an ensemble forecast seems a safer strategy.

The remainder of the paper is organized as follows. In Section 2 we lay out the methodology to implement the test of equal predictive accuracy. In Section 3 we describe the data and the models. Results are reported and discussed in Section 4, Finally, in Section 5 we conclude.

2 Forecast Evaluation

We consider the time series of cumulative daily deaths $\{y_1, ..., y_T\}$, for which we want to compare two h-weeks 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 $e^{(i)}_{t|t-h}\left(\widehat{\theta}^{(i)}_{w_i}\right) = y_t - \widehat{y}^{(i)}_{t|t-h}\left(\widehat{\theta}^{(i)}_{w_i}\right)$ and the associated loss is $L^{(i)}_{t|t-h}\left(\widehat{\theta}^{(i)}_{w_i}\right) \equiv L\left(e^{(i)}_{t|t-h}\left(\widehat{\theta}^{(i)}_{w_i}\right)\right)$, for example, a quadratic loss would be $L\left(e^{(i)}_{t|t-h}\left(\widehat{\theta}^{(i)}_{w_i}\right)\right) = \left(e^{(i)}_{t|t-h}\left(\widehat{\theta}^{(i)}_{w_i}\right)\right)^2$. The null hypothesis of equal predictive ability of the two forecasts is

$$H_0 : E\left[L\left(e^{(1)}_{t|t-h}\left(\widehat{\theta}^{(1)}_{w_1}\right)\right) - L\left(e^{(2)}_{t|t-h}\left(\widehat{\theta}^{(2)}_{w_2}\right)\right)\right] = 0. \tag{1}$$

Let

$$d_t \equiv d_t\left(\widehat{\theta}^{(1)}_{w_1}, \widehat{\theta}^{(2)}_{w_2}, h\right) = L\left(e^{(1)}_{t|t-h}\left(\widehat{\theta}^{(1)}_{w_1}\right)\right) - L\left(e^{(2)}_{t|t-h}\left(\widehat{\theta}^{(2)}_{w_2}\right)\right),$$

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 prediction 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 Choi and Kiefer (2010), Harvey et al. (2017), Coroneo and Iacone (2020).

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

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

where $\widehat{\sigma}_{BART,M}$ 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, Hualde and Iacone (2017) show that under the null in (1), we have

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

where $\widehat{\sigma}_{DAN,m}$ 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

Table 1: Forecasting Teams

Columbia University (Columbia)
Johns Hopkins University, Infectious Disease Dynamics Lab (JHU)
Los Alamos National Laboratory (LANL)
Massachusetts Institute of Technology, Operations Research Center (MIT-ORC)
Northeastern University, Laboratory for the Modeling of Biological and Socio-technical Systems (MOBS)
University of California, Los Angeles (UCLA)
University of Massachusetts, Amherst (UMass-MB)
Youyang Gu (COVID-Projections) (YYG)

Notes: The table reports the forecasting teams included in the forecast evaluation exercise. A forecasting team is included if it submitted their predictions for all the weeks in our sample.

(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 the national public health institute in the United States. It is a federal agency and its main goal is to protect public health through the control and prevention of disease. The CDC 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 at least three reasons. First, the number of fatalities is more reliable than the number of cases and hospitalisations, since the latter ignore asymptomatic cases and other undetected infections. Second, the measurement of

[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 2: Methods and Assumptions

Team	Method	Change Assumptions
Columbia	Metapopulation SEIR model	yes
JHU	Metapopulation SEIR model	yes
LANL	Statistical dynamical growth model	no
MIT-ORC	SEIR Model	no
MOBS	Metapopulation, age structured SLIR model	no
UCLA	Modified SEIR model	yes
UMass – MB	Mechanistic Bayesian compartment model	no
YYG	SEIR model	yes

Notes: 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 third column, "yes" means that the modeling team makes assumptions about how levels of social distancing will change in the future. Otherwise, "no" means modeling teams assumes that existing measures will continue through the projected four-week time period.

fatalities is less subject to spatial and temporal distortions, which may result from variation across regions and time in the intensity of testing and contact tracing. Third, the number of fatalities is arguably the primary objective of health authorities and the main concern of the public when facing a pandemic.

Our sample includes projections for national and state COVID-19 cumulative deaths made from May 18, 2020, to August 24, 2020, by eight forecasting teams. The deadline for teams to submit their weekly forecasts is on 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 May 18 had as targets the cumulative number of deaths as of May 23 (1-week ahead), May 30 (2-weeks ahead), June 6 (3-weeks ahead), and June 13 (4-weeks ahead). Realised values are also taken from the CDC website and span the period from May 23, 2020, to September 19, 2020. Therefore the sample sizes available for forecast evaluation are equal to 15 observations for all forecasting horizons. With such a small sample size, fixed-smoothing asymptotics is crucial to obtain correctly sized tests for equal predictive accuracy.

The eight forecasting teams selected are those that submitted their predictions for all the weeks in our sample. The teams are listed in Table 1. 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). In Table 2, we summarize the main features of the selected models.[2]

3.2 Ensemble forecast

The CDC also publishes an Ensemble forecast that combines individual forecasts. The Ensemble forecast produced by the CDC is naïve, as 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 their forecasts up to four weeks and these forecasts are at least equal to the level observed on the day in which the forecast is submitted. The Ensemble forecast is a combination of 4 to 20 models. 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 the evaluation exercise of the individual forecasts.

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 forecast is useful as it helps to diversify risk and pooling 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 be derived. However, in many practical applications, optimal weights schemes

[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.

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 naïve 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, as indeed it is the situation that we are analysing. In fact, the inclusion of models in the COVID-19 Ensemble forecast published by the CDC depends solely on availability, and this may even change week by week.

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 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.

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 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

Plots of all the forecasts of the number of cumulative deaths at the national level for one to four weeks ahead, alongside the realised data, are in Figure 1. From the comparison of the graphs at each horizon, it is apparent that the heterogeneity in forecasts grows with the forecasting horizon and, concurrently, that forecasts are less precise as the horizon is moved towards the four weeks. This simple observation may make the case for the Ensemble forecast at longer horizons more compelling.

Some forecasts tend to systematically overpredict the target (JHU and Columbia over the first part of the sample being the most conspicuous examples); while others tend to underpredict it. Such differences are of course relevant for decision makers if the costs of overprediction and underprediction are different. This result adds weights to the importance of considering asymmetric loss functions to test for equal predictive accuracy.

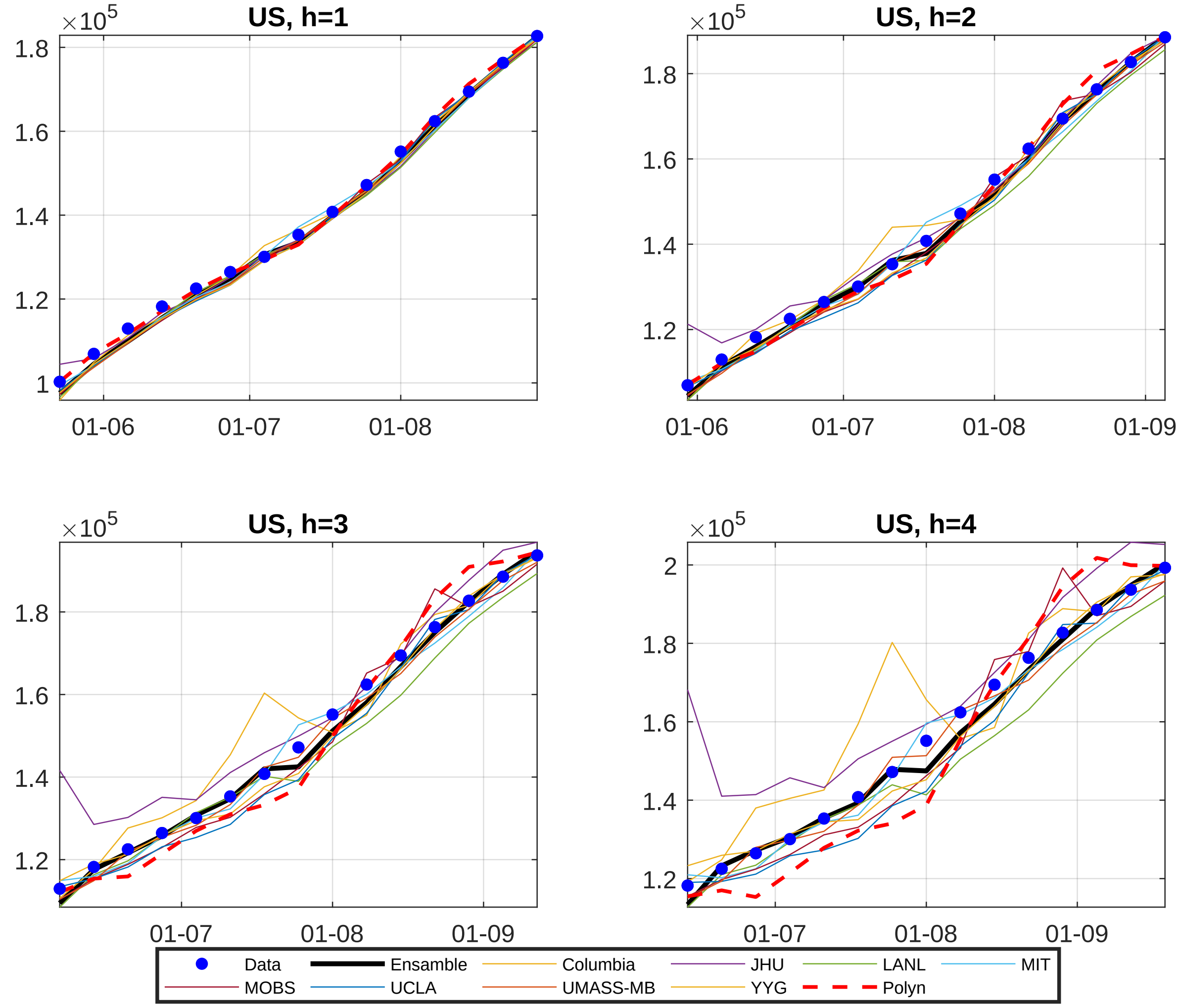
In Table 3, we report summary statistics of the forecast errors, defined as realised value minus the forecast, as in Section 2. The table reports for each forecasting horizon and forecasting scheme (Ensemble, team or polynomial) the sample mean, the median, the standard deviation, the skewness, the maximum and minimum forecast error recorded, and the first and second-order sample autocorrelation coefficients (in the columns AC(1) and AC(2), respectively). In most cases (Columbia and JHU at medium and long horizons being the only exceptions) the average of the forecast errors is positive, meaning that most forecasters (and the polynomial benchmark) tend to underpredict the number of fatalities.

Table 3: Summary Statistics of Forecast Errors

1-week ahead	Mean	Median	Std	Max	Min	Skew	AC(1)	AC(2)
Ensemble	1506	1571	1025	3037	-471	-0.27	0.50	0.42
Columbia	881	932	1609	4390	-2661	-0.10	0.57	0.13
JHU	1214	1477	1704	3482	-4165	-2.14	0.28	-0.24
LANL	2040	2321	1186	3770	-647	-0.48	0.42	0.20
MIT-ORC	1224	1510	1502	3317	-1880	-0.61	0.63	0.08
MOBS	1622	1467	1305	3320	-894	-0.33	0.49	0.64
UCLA	1669	2076	1328	3405	-382	-0.36	0.63	0.61
UMass-MB	1792	1462	1140	3486	46	-0.01	0.59	0.39
YYG	1906	2329	1057	3129	56	-0.52	0.58	0.47
Polyn	272	309	1012	2346	-1854	-0.18	0.53	0.08
2-weeks ahead	Mean	Median	Std	Max	Min	Skew	AC(1)	AC(2)
Ensemble	1289	1520	1291	3431	-884	0.01	0.35	0.12
Columbia	-389	75	3143	4300	-8673	-1.09	0.50	-0.18
JHU	-1765	-900	3948	2533	-14348	-2.19	0.65	0.34
LANL	2885	3078	2208	6466	-639	-0.26	0.66	0.36
MIT-ORC	942	1538	2122	3360	-4387	-1.13	0.47	-0.15
MOBS	1850	2483	1966	3571	-4169	-2.13	0.20	0.30
UCLA	2198	2576	1896	4789	-1415	-0.48	0.45	0.32
UMass-MB	1779	1665	1045	3515	-368	-0.20	0.31	-0.01
YYG	1860	1930	1396	3996	-475	-0.05	0.62	0.27
Polyn	785	1112	2639	5368	-4425	-0.36	0.73	0.30
3-weeks ahead	Mean	Median	Std	Max	Min	Skew	AC(1)	AC(2)
Ensemble	1338	714	2015	4711	-1207	0.47	0.45	0.25
Columbia	-2688	-2752	6460	7427	-19574	-1.01	0.48	-0.19
JHU	-6022	-4962	7042	774	-28663	-2.29	0.74	0.54
LANL	4466	4460	3541	9643	-1309	-0.03	0.75	0.49
MIT-ORC	1056	2281	2593	3870	-5444	-1.04	0.27	-0.30
MOBS	2129	2815	3858	6628	-9262	-1.84	0.21	0.22
UCLA	3309	3337	3030	7835	-1863	-0.24	0.56	0.37
UMass-MB	1840	1970	1404	4395	-1601	-0.57	0.05	0.03
YYG	1746	650	2565	6325	-1931	0.35	0.73	0.43
Polyn	1640	2814	5205	9855	-8197	-0.42	0.81	0.45
4-weeks ahead	Mean	Median	Std	Max	Min	Skew	AC(1)	AC(2)
Ensemble	1571	-147	2880	7683	-1085	0.80	0.42	0.25
Columbia	-6030	-6134	10558	10940	-33042	-0.89	0.46	-0.18
JHU	-11720	-8998	11630	-1539	-49961	-2.52	0.75	0.60
LANL	6682	6833	4845	13801	238	0.19	0.75	0.54
MIT-ORC	1557	2282	2727	4607	-4514	-0.83	0.15	-0.48
MOBS	2423	3959	6636	9108	-16519	-1.67	0.32	0.24
UCLA	4986	4250	4551	12890	-2083	0.02	0.63	0.45
UMass-MB	1972	3124	2437	5688	-3747	-0.92	-0.10	-0.07
YYG	1611	862	4178	9925	-5064	0.31	0.76	0.49
Polyn	2909	5500	8886	16414	-13277	-0.43	0.83	0.51

Notes: The table reports summary statistics of forecast errors for the ensemble, the teams, and the polynomial forecasts. The table reports mean, median, standard deviation (std), maximum (max), minimum (min), skewness (skew), first and second-order autocorrelation coefficients (AC(1) and AC(2)).

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



Note: The top-left figure reports 1-week ahead forecasts and realised values, weekly observations from 23 May 2020 to 29 Aug 2020. The top-right figure reports 2-weeks ahead forecasts and realised values, weekly observations from 30 May 2020 to 5 Sept 2020. The bottom-left figure reports 3-weeks ahead forecasts and realised values, weekly observations from 6 June 2020 to 12 Sept 2020. The bottom-right figure reports 4-weeks ahead forecasts and realised values, weekly observations from 13 June 2020 to 19 Sept 2020. Polynomial denotes the quadratic function in (4), estimated using a rolling window of 5 observations.

At one week horizon, the polynomial forecast outperforms all the competitors, with a much smaller average error and smaller dispersion. This is not surprising, 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, the situation is quite reversed: the performance of the polynomial model in terms of sample average and dispersion of the errors being now worse than most competitors. For example, at four week horizon, all the teams except two provide forecasts with lower dispersion, confirming that the epidemiological models may produce forecasts that are stable in the long-run.

In Table 3 we can also observe that the Ensemble forecast performs very well in terms of sample average and dispersion. Although two teams (MIT-ORC and UMass-MB) may produce forecasts with slightly better sample statistics, the Ensemble forecast seems a safe choice when compared to other forecasts.

Finally, we notice that the forecast errors are autocorrelated, as documented in the columns AC(1) and AC(2): this is the case even for one-step-ahead forecasts, where the first order sample autocorrelation may be as high as 0.63. This is interesting because optimal h-step-ahead forecast errors should be at most MA($h-1$): so one-step-ahead forecast errors should be Martingale Differences, two-step-ahead 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 the modelling teams. This may seem reasonable, as these models are primarily designed to forecast medium to long term dynamics, but it is important to mention it here nonetheless as it suggests that none of these models yields an optimal forecast and that short term forecasts may be improved upon.

4 Forecast Evaluation Results

4.1 Baseline Analysis

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 4. For our baseline analysis, we consider the US national level forecasts, and evaluate forecast errors using the quadratic loss function. We

Table 4: Tests for Equal Predictive Ability (US National Level)

	1-week		2-weeks		3-weeks		4-weeks	
	WCE	WPE	WCE	WPE	WCE	WPE	WCE	WPE
Ensemble	-1.790	-1.483	1.758	1.803	3.262**	3.566**	3.926**	3.913**
Columbia	-1.439	-1.447	-0.502	-0.473	-0.772	-0.684	-0.890	-0.789
JHU	-2.336*	-2.625*	-0.736	-0.704	-0.925	-0.853	-1.016	-0.932
LANL	-2.898**	-2.425*	-1.190	-1.028	-0.271	-0.238	0.507	0.446
MIT-ORC	-2.332*	-1.830	1.521	1.376	3.469**	3.641**	3.925**	4.026**
MOBS	-1.893	-1.630	0.044	0.043	1.783	2.132*	2.069	2.167*
UCLA	-2.159*	-1.814	-0.404	-0.405	1.202	1.446	2.263*	2.788**
UMass-MB	-2.001	-1.671	1.090	1.103	3.038**	3.329**	3.770**	3.925**
YYG	-2.669**	-2.331*	0.839	0.952	3.047**	3.786**	3.838**	3.918**

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).

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 from the model in question is more accurate than the polynomial benchmark.

We report two-sided significance at the 5% and 10% level, using fixed smoothing asymptotics (fixed-b for WCE and fixed-m for WPE) to establish the critical values. 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).

No forecasting model predicts better than the benchmark at one-week forecasting horizon. In fact, the benchmark often significantly outperforms the forecasting teams. Instead, differences at the two-weeks horizon are 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 models are the MIT-ORC and the UMass-MB models. The Ensemble forecast is also among the best performing models, and it outperforms the benchmark's at 5% using the WCE estimator, and at 10% using WPE estimator. This finding is consistent with the consensus in the literature about the advantages of forecast combination.

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, the evidence against equal predictive ability at one week horizon is slightly weaker using the WCE (but not the WPE). Overall the outcomes 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).

4.2 Alternative Loss Functions

The quadratic loss function that we use in the baseline forecast evaluation reported in Table 4 is a common choice in forecast evaluation. Thus, 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 initial number of infections is small, compared to when it is large.

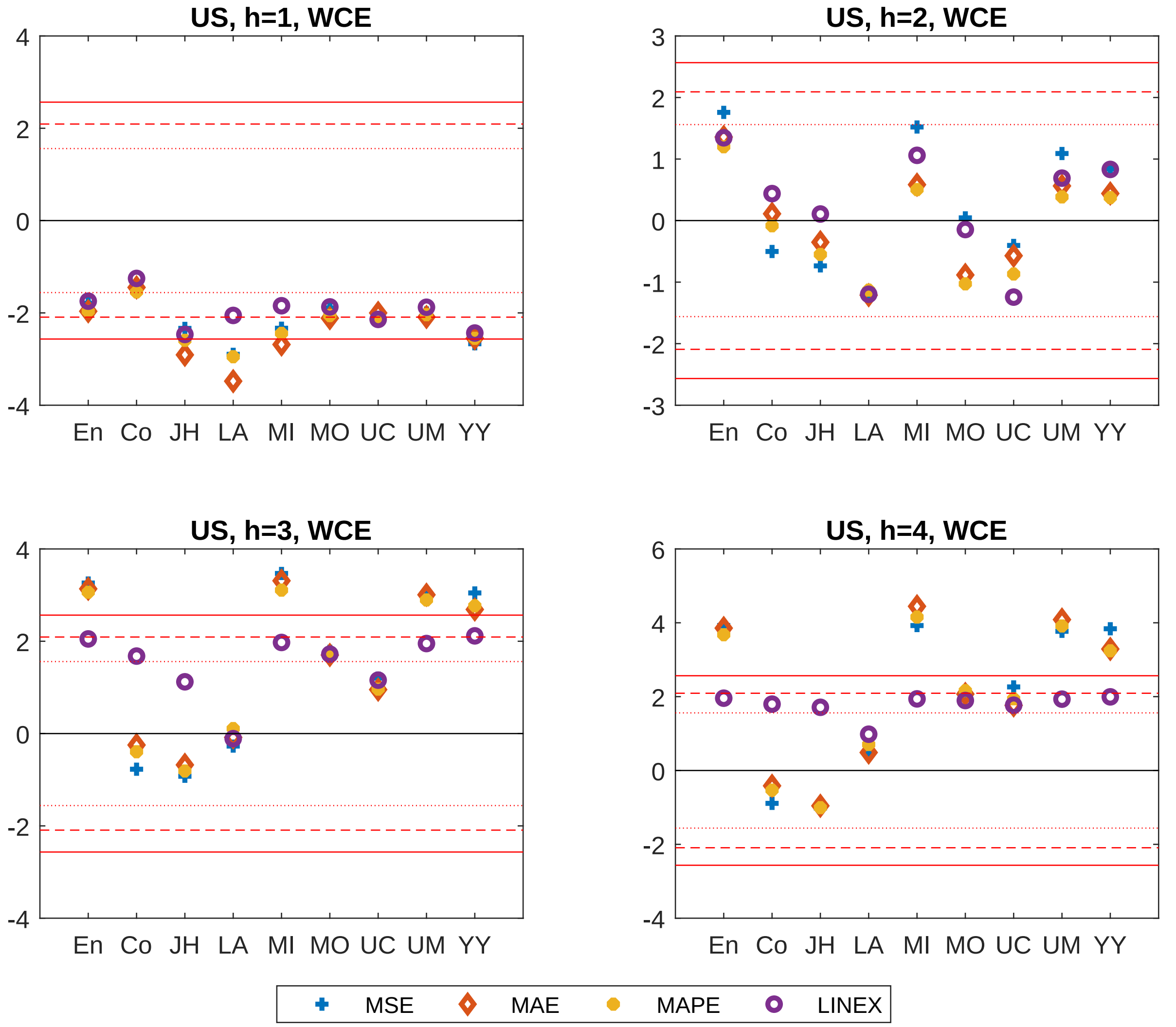
The DM test can be applied directly to alternative loss functions. Thus, we consider three alternative loss functions that provide alternative criteria for forecast comparison. Denoting e_t as the forecast error (thus abbreviating in this way $e_{t|t-h}^{m_i(\widehat{\theta}_i,w_i)}$), the alternative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

- Absolute: $L(e_t) = |e_t|$;
- Absolute Percentage: $L(e_t) = |e_t|/y_t$;
- Linex: $L(e_t) = \exp(ae_t/y_t) - ae_t/y_t - 1$ with $a = 50$.

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 being predicted, 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 underprediction than to overprediction. 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 believed 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).

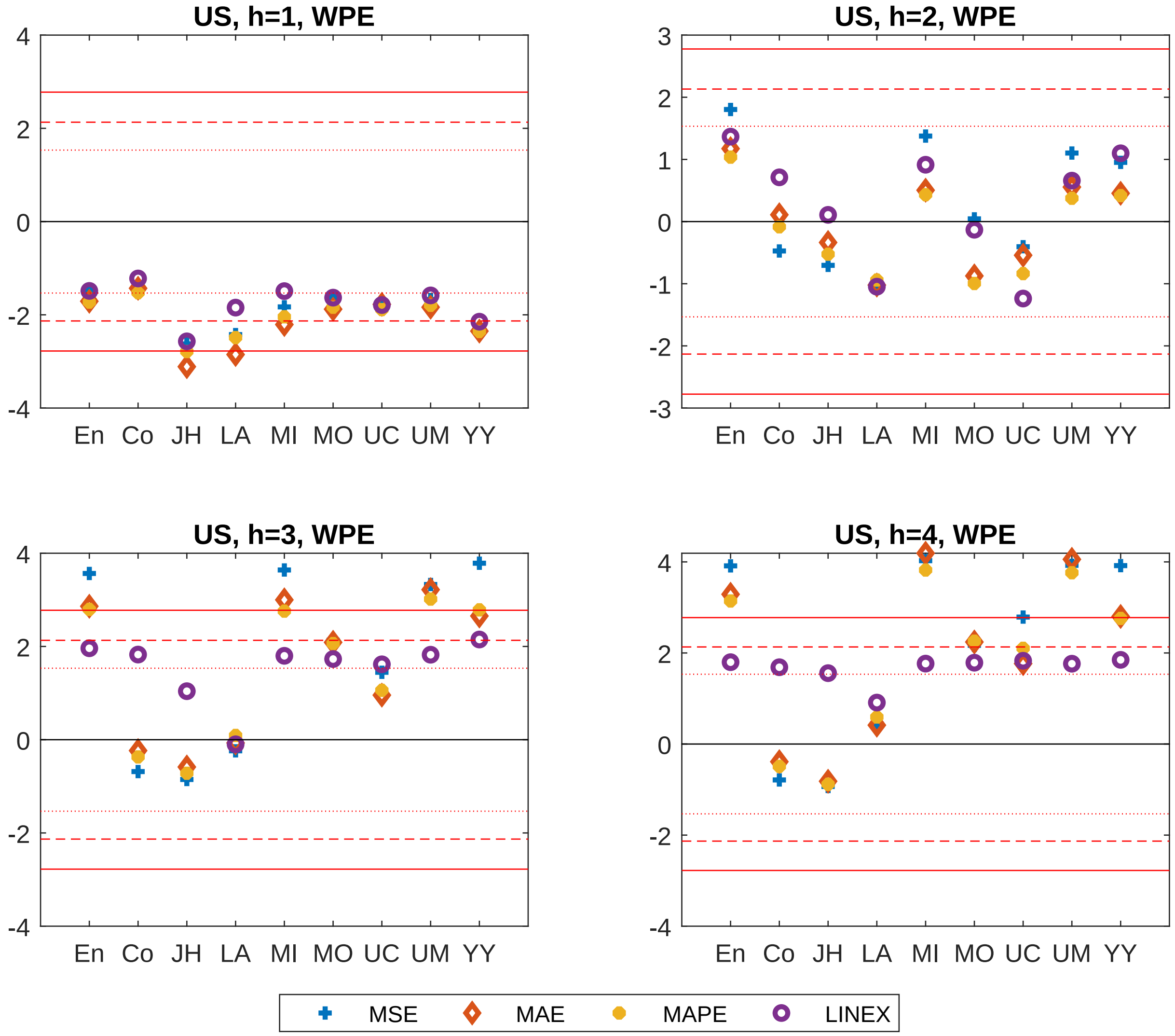
Results at the national level are summarised in Figures 2 and 3 (results for the quadratic loss function are plotted as well, to facilitate comparisons). We find that changing the loss function from quadratic to absolute or absolute percentage has very little consequence on the evaluation of predictive ability. On the other hand, the results are very different if the linex function is used, as in this case the null hypothesis is never rejected at the 5% significance level, although at one week horizon we still find that the simple benchmark outperforms the JHU and YYG teams if we use the 10% threshold. As the forecasting horizon is moved to three and four weeks, the polynomial benchmark tends to underestimate the target more.

Figure 2: Forecast evaluation with WCE - US



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 3: Forecast evaluation WPE - US



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.

However, these differences are not statistically significant, and the null hypothesis of equal predictive accuracy cannot be rejected.

The Ensemble forecast is one of the best performing alternatives when losses are evaluated using the quadratic loss function. However, the good relative performance of the Ensemble forecast is less salient when the test is constructed using the linex function. Elliott and Timmermann (2004) showed that the equal weights ensemble is less appropriate in presence of asymmetries in the loss function and in the distribution of the errors, and our empirical findings may be evidence of this fact.

It is interesting to notice also that the performance of the Columbia and JHU, the two forecasting teams that on average overpredicted fatalities, is comparable to the performance of the other teams (and the benchmark) under the linex measure. Thus, these two forecasting teams may have rationally chosen to adopt a more prudent approach to forecasting. At any rate, using the linex loss function, no forecast seems to stand out as superior.

4.3 State level results

In Figures 4 to 7 we report the tests of equal predictive ability for, in turn, New York, California, Florida and Texas. In the interest of brevity, we only present results for the WCE case, the outcomes when the WPE estimator is used are qualitatively very similar and available upon request.

Regional results exhibit more heterogeneity, and the forecasting teams tend to perform less well against the benchmark than at national level. In particular, professional forecasters never beat the benchmark for New York, Texas or California, and some forecasts are in fact significantly worse than the benchmark, especially at the one week horizon. The case of New York is certainly noteworthy as all the teams are outperformed at the 10% significance level by the benchmark, when the absolute loss or the absolute percentage loss function are used.

Occasionally, the benchmark predicts better than the teams even at longer horizons. For example, for California the JHU performed significantly worse than the benchmark at the 5% significance level, and Columbia at the 10% when using the absolute loss function, even at the three and four weeks horizons; for Florida, the MOBS model had significantly less predictive power at four weeks horizon with the absolute and absolute percentage loss functions. On the other hand, YYG performs significantly better than the benchmark, so in this sense the results for Florida are more mixed.

In all these instances the Ensemble forecast provides a valuable alternative, as it often performs as well as the best constituent models and avoids the worst adverse results. For one remarkable example, consider again the forecast evaluation at one week for New York: all the teams are significantly outperformed by the benchmark, but the Ensemble forecast is not (recall, however, that the Ensemble forecast is obtained using also forecasts from models not included in this exercise because their time series is not sufficiently long).

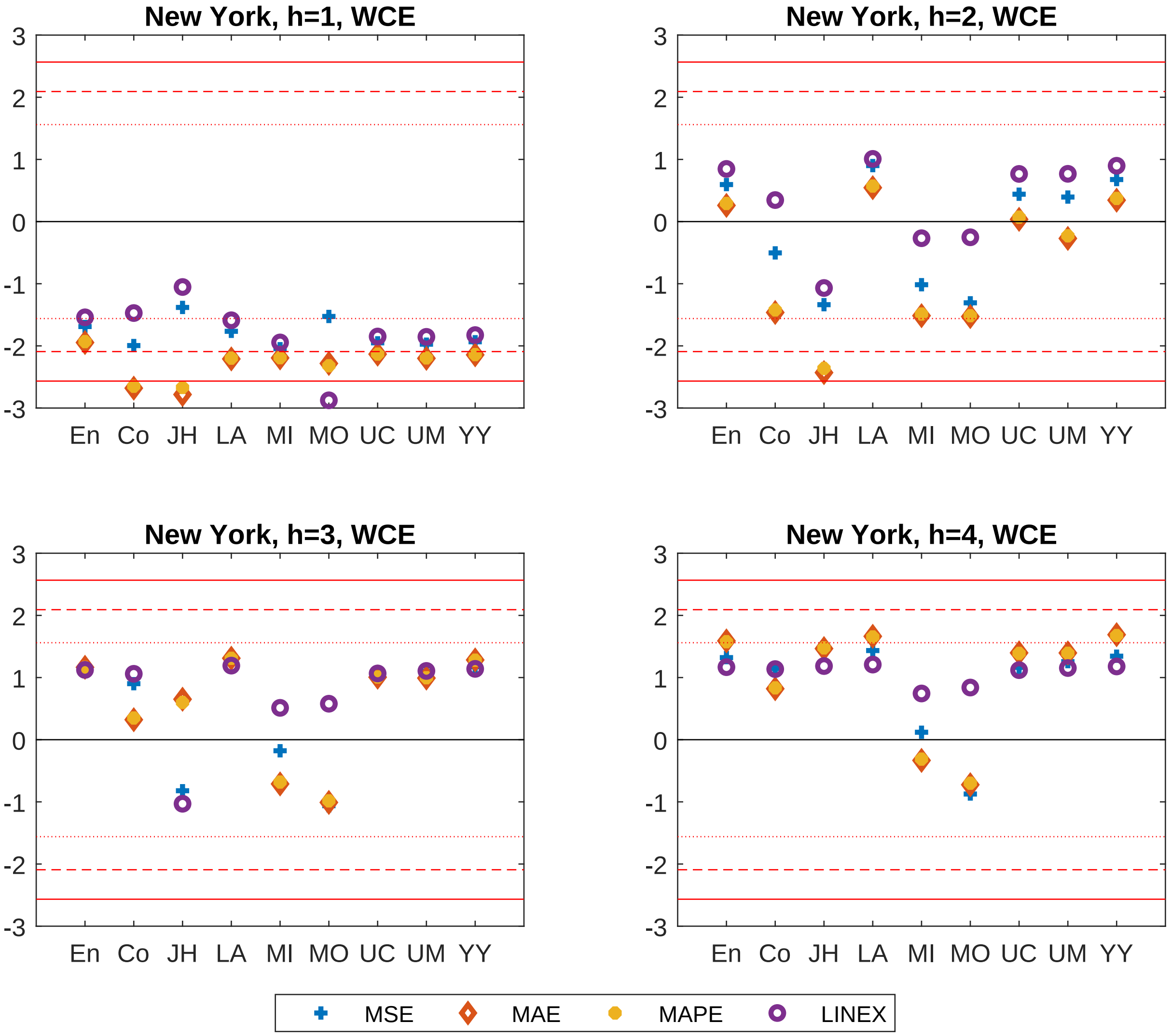
5 Conclusion

We evaluate the relative predictive accuracy of forecasts for the COVID-19 fatality level, produced by several competing forecasting teams, for the US and for four main States (California, New York, Florida, and Texas). An Ensemble forecast is also included, that combines all available forecasts using an equal weights scheme. Since sample sizes are small, we use alternative (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 one-week horizon; however, at longer forecasting horizons some teams show superior predictive ability, especially at the national level. We also document heterogeneous results across states. Overall, these results indicate that forecasts of the COVID-19 epidemic need to be used with caution.

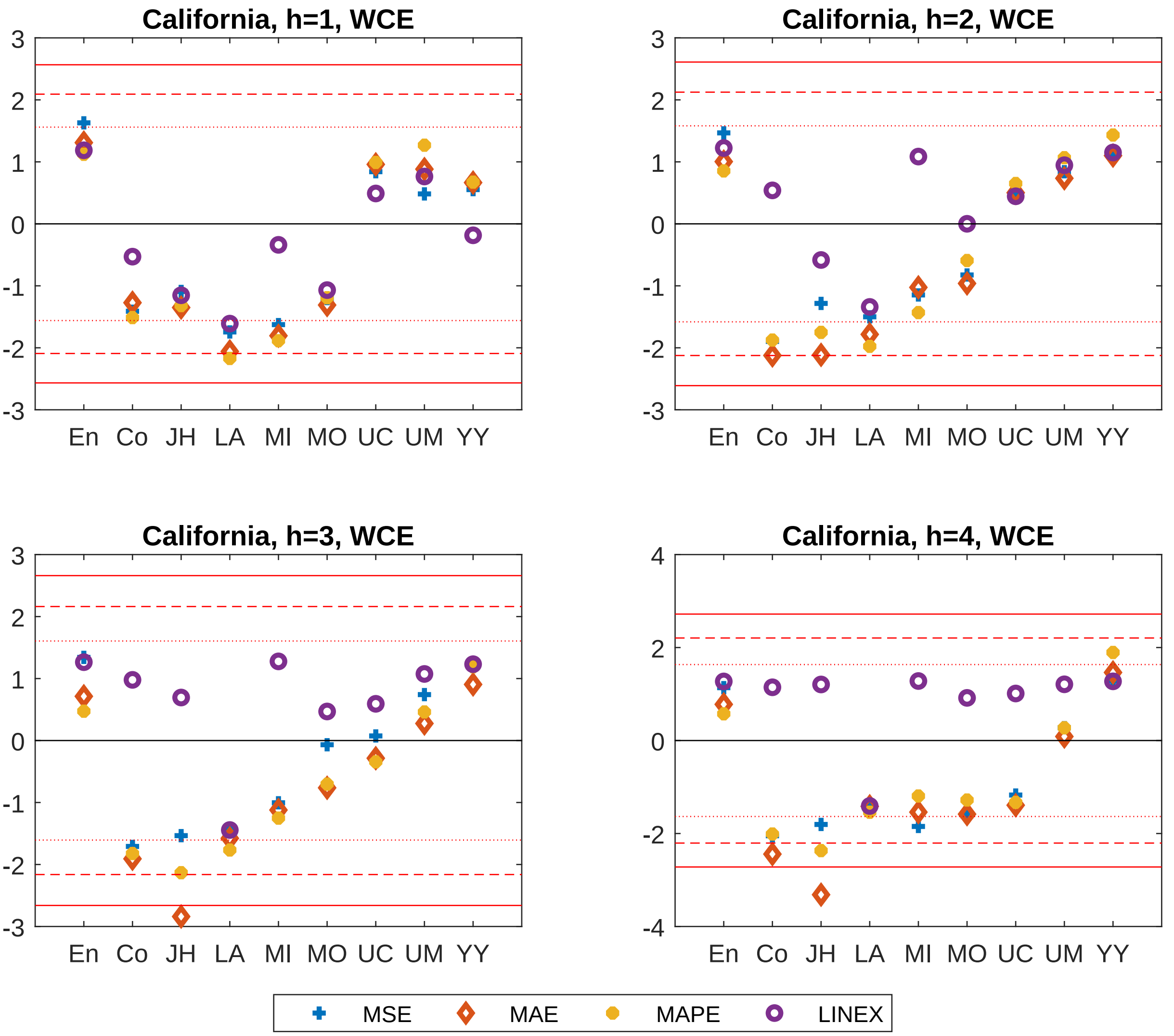
One of the best performing forecasts is the Ensemble forecast that combines several forecasts. Whilst the Ensemble does not yield the best forecast, it is competitive in the sense of being one of the few forecasts performing statistically better than the benchmark at longer horizons (4 weeks ahead), and it never performs worse than the benchmark. In this sense, the Ensemble forecast may be seen as a robust choice. Thus, health authorities should not 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 use an ensemble forecast.

Figure 4: Forecast evaluation - New York



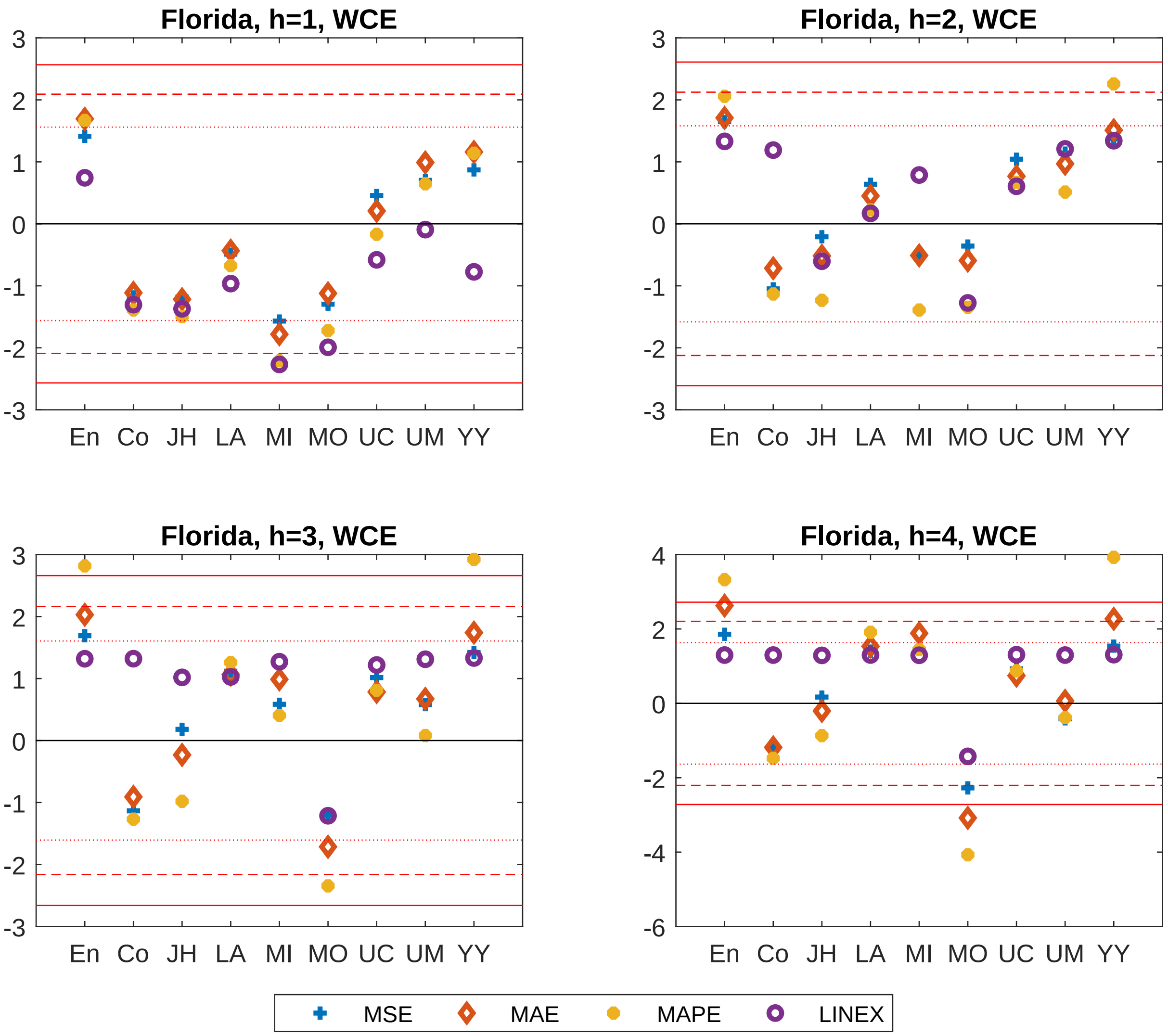
This figure reports the test statistic for the test of equal predictive accuracy using the weighted covariance estimator (WCE) of the long-run variance and fixed-b asymptotics. The benchmark is a second degree polynomial model fitted on a rolling window of 5 observations. A positive value of the test 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 5: Forecast evaluation - California



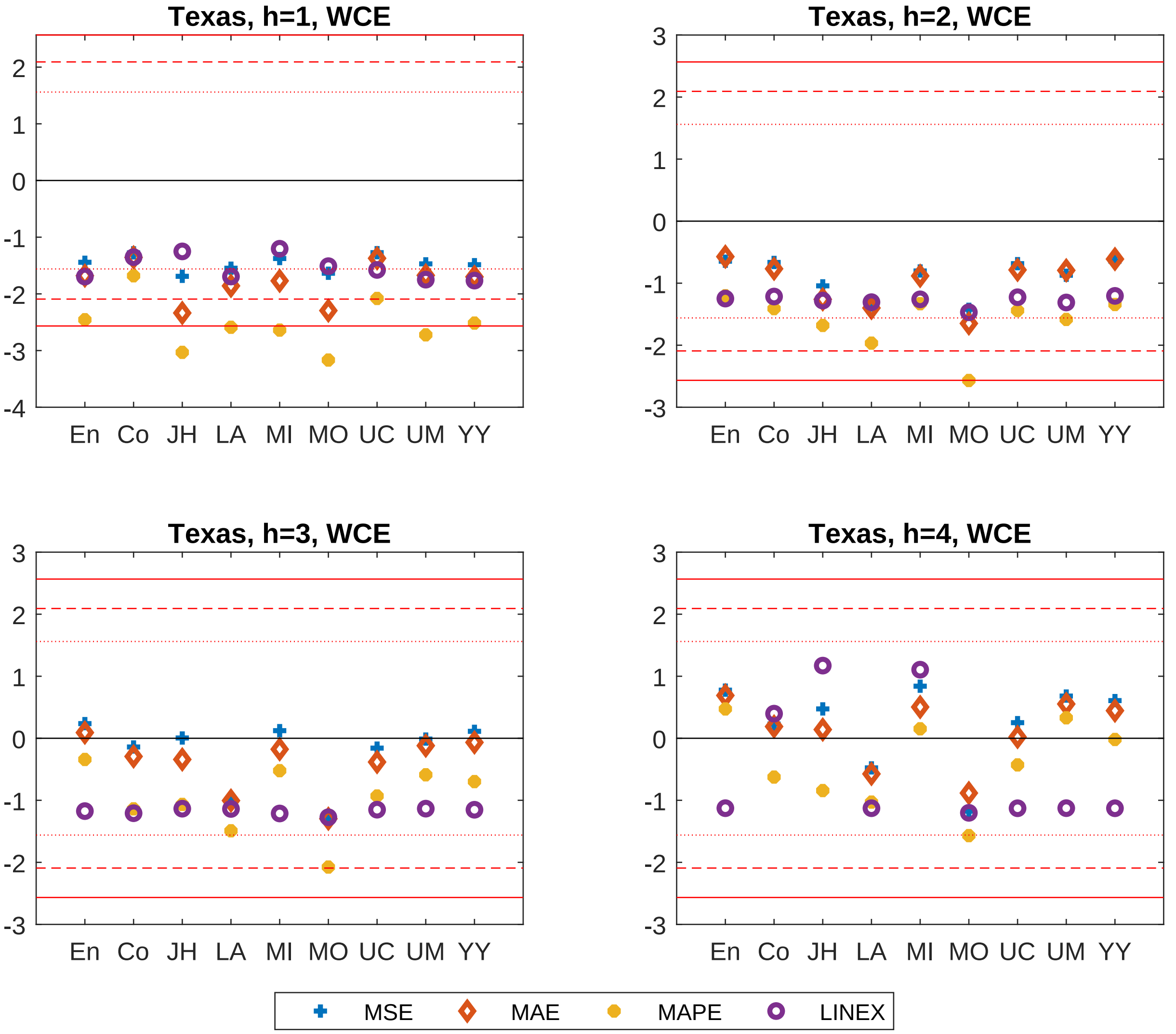
This figure reports the test statistic for the test of equal predictive accuracy using the weighted covariance estimator (WCE) of the long-run variance and fixed-b asymptotics. The benchmark is a second degree polynomial model fitted on a rolling window of 5 observations. A positive value of the test 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 - Florida



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 7: Forecast evaluation - Texas



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.

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

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

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