Predictive performance of multi-model ensemble forecasts of COVID-19 across European nations

Order tbc; Katharine Sherratt, Hugo Gruson, *Any co-authors*, *Team authors*, *Advisory team authors*, *ECDC authors*, Johannes Bracher, Sebastian Funk

Abstract

Background Short-term forecasts of infectious disease burden can contribute to situational awareness and aid capacity planning. Based on best practice in other fields and recent insights in infectious disease epidemiology, one can maximise the predictive performance of such forecasts if multiple models are combined into an ensemble. Here we report on the performance of ensembles created from over 40 models in predicting COVID-19 cases and deaths across Europe between 08 March and 15 November 2021.

Methods We used open-source tools to develop a public European COVID-19 Forecast Hub. We invited groups globally to contribute weekly forecasts for COVID-19 cases and deaths over the next one to four weeks. Forecasts included quantiles across the predictive distribution. Each week we created an ensemble forecast, where each predictive quantile was calculated as the equally-weighted average of all individual models' predictive quantiles. We retrospectively explored alternative methods for ensemble forecasts, including weighted averages based on models' past predictive performance. The performance of the ensembles was compared to individual models and a baseline model of no change using pairwise comparison with the Weighted Interval Score (WIS).

Results Over 36 weeks we collected and combined 43 forecast models for 32 countries. We found a weekly ensemble had among the most reliable performances across countries over time, with more accurate predictions for reported cases and deaths than a simple baseline for 67% and 91% of possible forecast targets respectively. Ensemble performance declined with increasing forecast time horizon when forecasting cases but remained stable for 4 weeks for incident death forecasts. Among several choices of ensemble methods, we found that the calculation of the average was the most influential choice for performance. Almost any forecast created using a median average performed better than using a mean, regardless of methods for weighting component forecasts.

Conclusions Our results support the use of an ensemble as a reliable way to make real-time forecasts across many populations and epidemiological targets during infectious disease epidemics. We recommend the use of median ensemble methods across many forecasting models in general, and specifically that current policy relevant work using COVID-19 surveillance data across Europe should place more confidence in forecasts of incident death than case counts at longer (more than 2 week) periods into the future.

Background

Epidemiological forecasts make quantitative statements about a disease outcome in the near future. At the most general level, forecasting targets can include measures of prevalent or incident disease and its severity, for some population over a specified time horizon. Researchers, policy makers, and the general public have used such forecasts to understand and respond to the global outbreaks of COVID-19 since early 2020 [1]. However forecasters use a variety of methods and models for creating and publishing forecasts, varying in

both defining the forecast outcome and in reporting the distribution of probability around outcomes **zellnerAccountingUncertaintyPandemic2021?**, **jamesUseMisuseMathematical2021?**. Such variation between forecasts makes it difficult to interpret the likelihood of any one outcome, or to compare the predictive performance between forecast models. These barriers to comparing and evaluating mean there is little objective support for using any one particular forecast for representing or acting on likely outcomes.

A "forecast hub" is a centralised effort to improve the transparency and usefulness of forecasts, by standardising and collating the work of many independent teams producing forecasts [2]. A hub sets a commonly agreed structure for forecast targets, such as type of disease event, spatio-temporal units, or the set of quantiles of the probability distribution to include from probabilistic forecasts. For instance, a hub may collect predictions of the total number of cases reported in a given country for each day in the next two weeks. Forecasters can adopt this format and contribute forecasts for centralised storage in the public domain. This shared infrastructure allows forecasts produced from diverse teams and methods to be visualised and quantitatively compared on a near-exact like-for-like basis, which can strengthen public and policy use of disease forecasts [3]. The underlying approach to creating a forecast hub was pioneered for forecasting influenza in the USA and adapted for forecasts of short-term COVID-19 cases and deaths in the US [4], with similar efforts elsewhere [5], [6], **bracherPreregisteredShorttermForecasting2021?**.

Standardising forecasts allows for combining multiple forecasts into a single ensemble forecast with the potential for a superior predictive performance. Evidence from previous efforts in multi-model infectious disease forecasting suggests that forecasts from an ensemble of many forecast models can be consistently high performing compared to any one of the component models [7]–[9]. Somewhat comparably, weather forecasting has a long standing use of building ensembles of many models using diverse methods with standardised data and formatting [10], **buizzaIntroductionSpecialIssue2019?**.

The European COVID-19 Forecast Hub **europeancovid-19forecasthubEuropeanCOVID19Forecast2021?** is a project to collate short term forecasts of COVID-19 across 32 countries in the European region. The Hub is funded and supported by the European Centre for Disease Prevention and Control (ECDC), with the primary aim to provide reliable information about the near-term epidemiology of the COVID-19 pandemic to the research and policy communities and the general public. Second, the hub aims to create infrastructure for storing and analysing epidemiological forecasts made in real time by diverse research teams and methods across Europe. Third, the hub aims to maintain a community of infectious disease modellers underpinned by open science principles. We started formally collating and combining contributions to the European Forecast Hub in March 2021. Here, we investigate the predictive performance of an ensemble of all forecasts contributed to the hub in real time each week, as well as the performance of variations of ensemble methods created retrospectively.

Methods

We developed infrastructure to host and analyse forecasts, focussing on compatibility with the US **cramerReichlabCovid19forecasthubRelease2021?**, **wangReichlabCovidHubUtilsRepository2021?** and the German and Polish COVID-19 **bracherGermanPolishCOVID192020?** forecast hubs.

Forecast targets and standardisation

We sought forecasts for two measures of COVID-19 incidence: the reported number of deaths and the reported number of cases per week. We considered forecasts for 32 countries in Europe, including all countries of the European Union and European Free Trade Area, and separately the United Kingdom. Incidence was aggregated over the Morbidity and Mortality Weekly Report (MMWR) epidemiological week definition of Sunday through Saturday. When predicting any single forecast target, teams could express uncertainty by submitting predictions across a range of a pre-specified set of 23 quantiles in the probability distribution. Teams could also submit a single point forecast without uncertainty. At the first submission we asked teams to add a single set of metadata briefly describing the forecasting team and methods. Any team

was able to participate, and to increase participation we actively contacted known forecasting teams across Europe and the US and advertised among the ECDC network. Teams submitted a broad spectrum of model types, ranging from mechanistic models, agent-based and statistical models to ensembles of multiple quantitative or qualitative models. We maintain a full project specification with a detailed submissions protocol **europeancovid-19forecasthubCovid19forecasthubeuropeWiki?**.

With the complete dataset for the latest forecasting week available each Sunday, teams typically submitted forecasts to the hub on Monday. We implemented an automated validation programme to check that each new forecast conformed to standardised formatting. The validation step ensured a monotonic increase of predictions with each increasing quantile, integer-valued counts of predicted cases, as well as consistent date and location definitions.

Each week we built an ensemble of all forecasts which was updated each week after all forecasts had been validated. From the first week of forecasting from 8 March 2021, the ensemble method for summarising across forecasts was the mean average of all models at each predictive quantile for a given location, target, and horizon. From 26 July 2021 onwards the ensemble instead used a median average of all predictive quantiles, in order to mitigate the wide uncertainty produced by some highly anomalous forecasts. We created an open and publicly accessible interface to the forecasts and ensemble, including an online visualisation tool allowing viewers to see past data and interact with one or multiple forecasts for each country and target for up to four weeks' horizon **europeancovid-19forecasthubEuropeanCovid19Forecast?**. All forecast and meta data are freely available and held on Zoltar, a platform for hosting epidemiological forecasts (**epiforecastsProjectECDCEuropean2021?**; reichZoltarForecastArchive2021).

Forecast evaluation

We evaluated all previous forecasts against actual observed values for each model, stratified by the forecast horizon, location, and target. We calculated scores using the *scoringutils* R package **nikosibossseScoringutilsUtilitiesScoring2020?** with observed data reported by Johns Hopkins University (JHU, **dongInteractiveWebbasedDashboard2020a?**). JHU data included a mix of national and aggregated subnational data for the 32 countries in the Hub. We removed any forecast surrounding (in the week of or after) a strongly anomalous data point.

For each model, we assessed its calibration as coverage of the predictive intervals and overall predictive performance as the weighted interval score (WIS). For a given interval level $k \in [0, 1]$, the coverage equals the proportion p of all observations (at every location and at every time) that are in the central predictive interval of level k. Coverage at a given interval level k was calculated as the proportion p of observations that fell within the corresponding central predictive intervals across locations and forecast dates. A perfectly calibrated model would have $p = k$ at all 11 levels (corresponding to 22 quantiles excluding the median). An under confident model at level k would have $p > k$, i.e. more observations fall within a given interval than expected.In contrast, an overconfident model at level k would have $p < k$, i.e. fewer observations fall within a given interval than expected. We here focus on coverage at the $k = 50$ and $k = 95$ level.

We assessed weekly forecasts using the WIS (a summary measure of predictive performance), across all quantiles that were being gathered **bracherEvaluatingEpidemicForecasts2021?**. The WIS is a **strictly proper scoring rule**, that is, it is optimised for predictions that come from the data-generating model. As a consequence, the WIS encourages forecasters to report predictions representing their true belief about the future**gneitingStrictlyProperScoring2007?**. The WIS represents a parsimonious approach to scoring forecasts when only quantiles are available:

$$\mathrm{WIS}_{\alpha_1,\ldots,\alpha_K}(y, F) = \frac{1}{K + 0.5}\left(\frac{1}{2}\,|y - m| + \sum_{k=1}^{K}\left(\frac{\alpha_k}{2}\,(u_k - l_k) + (l_k - y)\,\mathbb{I}_{y<l_k} + (y - u_k)\,\mathbb{I}_{y>l_k}\right)\right)$$

Where y is an observed count of weekly incidence for a given location and date, F the forecast, m the median of the predictive distribution, u_k and l_k are the predictive upper and lower quantiles corresponding to the

central predictive interval level k, respectively, $K = 11$ is the number of intervals considered and $1 - \alpha_k$ the width of central predictive interval k.

We compared the performance of each model against the performance of a baseline model. This assumes case or death counts stay the same as the latest data point over all future horizons, with expanding uncertainty, described previously in [11].

Meanwhile, not all models provided forecasts for all locations and dates. In order to be able to compare predictive performance in the face of various levels of missingness, we further calculated a relative WIS, as a measure of forecast performance which takes into account that different teams may not cover the same set of forecast targets (i.e., weeks and locations). Loosely speaking, a relative WIS of x means that averaged over the targets a given team addressed, its WIS was x times higher or lower than the the performance of the baseline model. Smaller values in the relative WIS are thus better and a value below one means that the model has above average performance. The relative WIS is computed using a ?pairwise comparison tournament? where for each pair of models a mean score ratio is computed based on the set of shared targets. The relative WIS of a model with respect to another model is then the ratio of their respective geometric mean of the mean score ratios, where here we report the relative WIS of each model with respect to the baseline model.

Ensemble methods We retrospectively explored alternative methods for ensembling forecasts for each target at each week. A natural way to combine probability distributions available in a quantile format, such as the ones collated in the European COVID-19 Forecast Hub, is[12]

$$F^{-1}(\alpha) = \sum_{i=1}^{n} w_i F_i^{-1}(\alpha)$$

Where $F_1 \dots F_n$ are the cumulative distribution functions of the individual probability distributions (in our case, the predictive distributions of each forecast model i contributed to the hub), w_i are a set of weights in $[0, 1]$; and α are the quantile levels such that

$$F^{-1}(\alpha) = \inf\{t : F_i(t) \geq \alpha\}$$

.

Different ensemble choices then mainly translate to the choice of weights w_i.

The simplest choice of weights w_i is to set them all equal so that they sum up to 1, $w_i = 1/n$, resulting in an unweighted mean ensemble. This can be subject to outlier forecasts that could skew the mean. A choice for the weights w_i that avoids this potential issue is to combine the forecasts at each quantile level in an unweighted median ensemble, where the average is replaced by a median. An unweighted median ensemble has previously been found to yield very competitive performance while maintaining robustness to outlying forecast[13]. By choosing the weights w_i to reflect past performance one can move from an untrained to a trained ensemble. Numerous options exist for choosing the weights with the aim to maximise predictive performance. A straightforward choice is so-called inverse score weighting, which was recently found in the US to outperform unweighted scores during some time periods[14] but not confirmed in a similar study in Germany and Poland **bracherPreregisteredShorttermForecasting2021?**. In this case, the weights are calculated as

$$w_i = \frac{1}{S_i}$$

where S_i reflects the forecast skill of forecaster i.

Here we considered unweighted and inverse relative WIS weighted mean and median ensembles. We additionally considered ensembles where the training period was restricted to the last 10 weeks (for the weighted ensembles), or where only models with relative WIS < 1 (i.e., outperforming the baseline) were included.

Results

We scored all forecasts submitted weekly in real time over the 36 week period from 08 March to 15 November 2021. Each week, forecasts were collated for incident cases and deaths, for 32 locations over the following 4 weeks, creating 256 possible forecast targets. We received 43 unique forecasting models from 36 separate forecasting teams. We added an ensemble model using all available forecasts for each possible target every week.

We used this dataset to create 3998 forecasting scores, each summarising a unique combination of model, variable, country, and week ahead horizon. Not all teams forecast for all targets, nor across all quantiles of the predictive distribution for each target. 37 models provided sufficient quantiles that we could evaluate them using the relative weighted interval score (WIS).

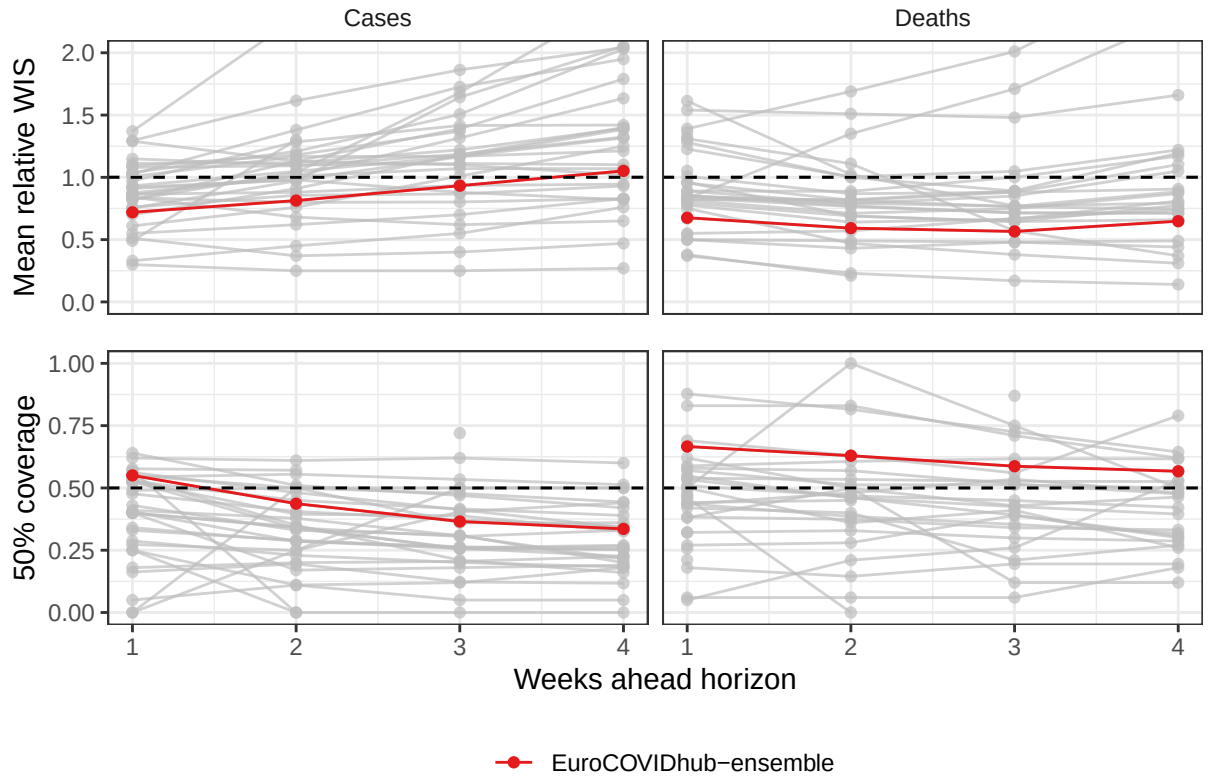


Figure 1: Performance of short-term forecasts, by relative interval score (relative to a baseline forecast, top) and coverage of the central 50% interval (the proportion of observed values that fell within the predicted 50% range, bottom). The scores of each model (grey) and ensemble (red) are averaged across all forecast targets and shown by one- to four-week ahead horizon.

The ensemble model performed well compared to both its component models and the baseline. In ranking all models' scores compared to the baseline, the ensemble performed better on relative WIS than 66% model scores when forecasting cases ($n = 1832$), and 67% of scores for forecasts of incident deaths ($n = 1910$). The ensemble outperformed the baseline model at the one-week ahead horizon for both cases and deaths (Figure 1). For horizons longer than one week, performance depended on the epidemiological target. The ensemble stopped outperforming the baseline at three to four weeks with respect to incidence cases. In contrast, the ensemble outperformed the baseline for deaths at all horizons considered (up to four weeks WIS), with no discernible deterioration in performance.

We observed similar trends in performance when considering how well the ensemble was calibrated with respect to the observed data. At one week ahead the case ensemble was well calibrated (ca. 50% nominal coverage at the 50% level), but this did not hold with increasing forecast horizon. The death ensemble was well calibrated at the 95% level for all horizons and under confident for the 50% level, with only slow deterioration with increasing forecast horizon.

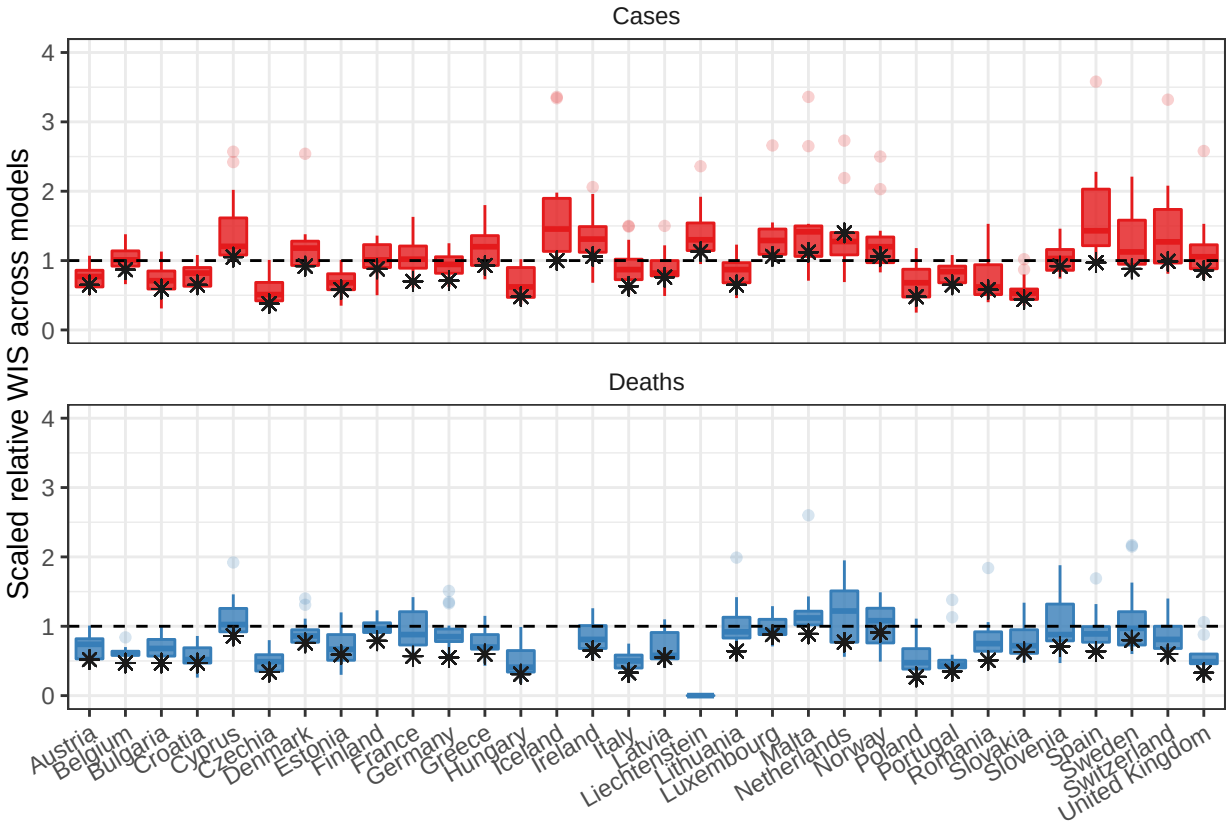
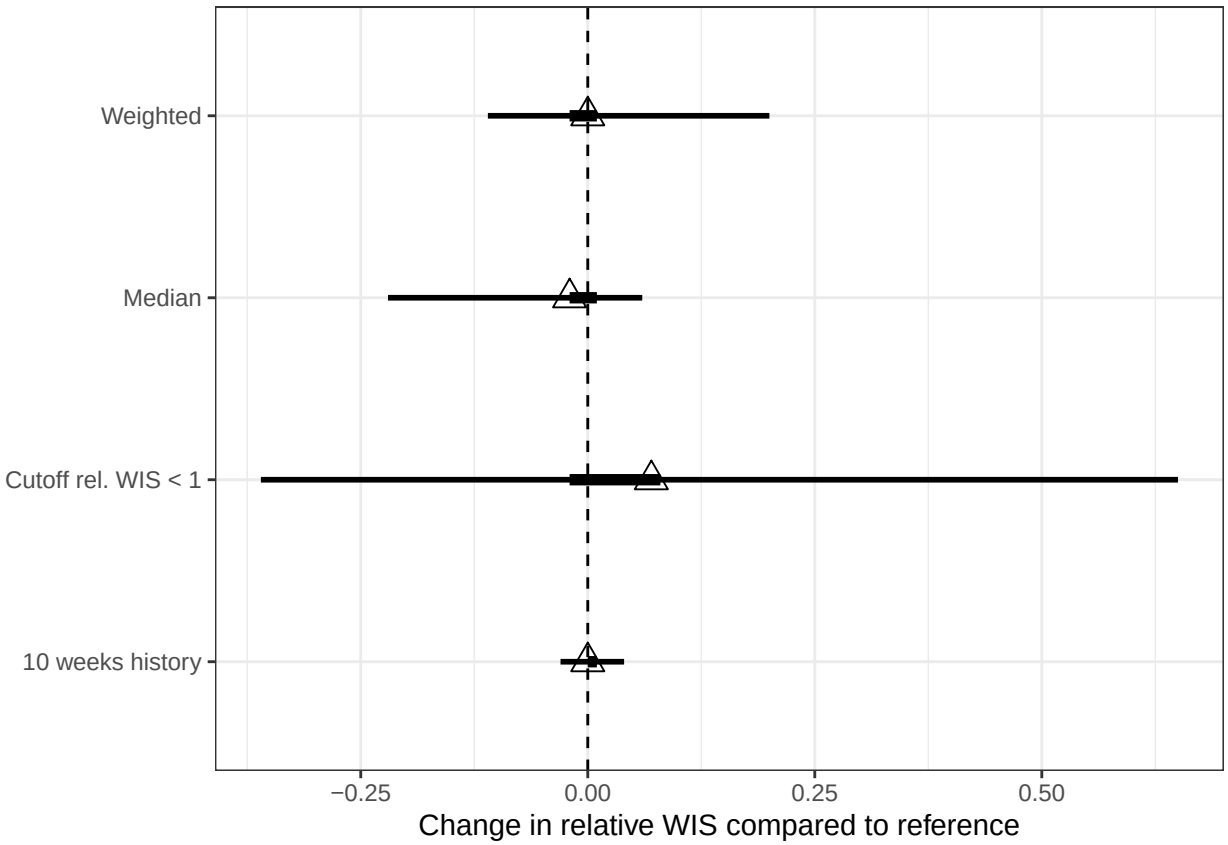


Figure 2: Performance of short-term forecasts across models and median ensemble (asterisk), by country, forecasting cases (top) and deaths (bottom) for two-week ahead forecasts, according to the relative weighted interval score. Boxplots show interquartile ranges, with outliers as faded points, and the ensemble model performance is marked by an asterisk.

The ensemble also performed consistently well when forecasting across countries relative to individual models and the baseline (figure 2, figure SI1). Compared to models that forecast across all 32 countries at the two-week horizon, the ensemble was more consistent in outperforming the baseline across countries compared to any single model forecasting deaths, and all but one model for case forecasts. Considering forecast targets across all 32 countries and over all four horizons (128 targets), the ensemble forecast outperformed the baseline for 67% and 91% of all 128 targets when forecasting cases and deaths, respectively. Comparably, the best individual models forecasting for the same number of targets as the ensemble model outperformed the baseline for 80% ("LANL-GrowthRate" model) for cases and 81% ("RobertWalraven-ESG" model) for death forecasts.



\begin{figure}
\caption{Performance of alternative ensemble methods at 2 week horizon, showing mean difference (triangle) in relative weighted interval score, with 48% and 96% probability (thick and thin line respectively). The difference in WIS is a comparison of scores from forecasts made from all possible combinations of methods, with a single element of ensemble method input changed. Reference categories are: weighted v. unweighted (n=250); median v. mean (n=376); cutoff by WIS v. all models included (n=374); relative WIS measures over 10 weeks of forecast history vs. all forecasts (n=248)} \end{figure}

At the two-week ahead horizon, variations in ensemble methods made little difference to forecast scores (figure 3, figure SI2). Ensembles that weighted forecasts showed no difference in performance to simple unweighted ensemble methods. Similarly, in choosing a method with which to weight forecasts, the choice of whether to use scores across all past forecasts, or scores evaluating only the most recent 10 weeks' forecast scores, made very little difference to the performance of the resulting ensemble (0 mean change in forecast score). The choice to exclude any forecast that scored worse than the baseline forecast ("cut off") affected the performance of the ensemble in both directions, overall slightly worsening performance (0.07 relative WIS). Using the median average was the only variation of ensemble method that typically improved performance, compared to using the mean average across any combination of ensemble method.

Discussion

We collated forecasts from multiple teams making forecasts of COVID-19 cases and deaths across countries over March to November 2021 in Europe, using an open and principled approach to standardising both forecast targets and the uncertainty around predictions. Combining these forecasts into an ensemble we found that the ensemble forecasts outperformed a baseline model at short forecast horizons and produced among the most consistent predictive performance across countries over time.

Our results support previous findings that ensemble forecasts are (or are near the) best performing models

with respect to error and are the most reliably consistent models in terms of appropriate coverage of uncertainty [5], [11]. While the ensemble was consistently high performing, it was not strictly dominant across all forecast targets, with others also seeing this in comparable studies of COVID-19 forecasts [15], **bracherPreregisteredShorttermForecasting2021?**. Our finding suggests the usefulness of an ensemble as a robust summary when forecasting across many spatio-temporal targets, without replacing the importance of communicating the full range of model predictions.

During the COVID-19 outbreak in Europe from March to November 2021, we identified a particular benefit from applying an ensemble forecast approach. The introduction of vaccination changed the associations between infections, cases, and deaths **europeancentrefordiseasepreventionandcontrolInterimGuidanceBenefits2021?**. At the same time, the emergence and subsequent dominance of the delta variant altered transmission dynamics across Europe **europeancentrefordiseasepreventionandcontrolThreatAssessmentBrief2021?**. However, neither of these factors were uniform across countries covered by the Forecast Hub **europeancentrefordiseasepreventionandcontrolOverviewImplementationCOVID192021?**. As epidemic dynamics became increasingly heterogeneous, the forecasting performance of any single model over time and across multiple countries became at least partly dependent on the ability, speed, and precision with which it could adapt to new conditions for each forecast target. This variability in the relative performance of models over time makes using an ensemble, balancing across all models, particularly relevant in rapidly changing epidemic conditions.

Our results also suggest the limited value of reporting case forecasts further than two weeks into the future. Previous work has similarly found rapidly declining performance for case forecasts with increasing horizon [11]. COVID-19 has a typical serial interval of less than a week, which implies that case forecasts of more than two weeks can only hold if rates of transmission and detection remain predictable over the entire period, a strong assumption in the light of the many instances of rapidly changing policies and individual behaviour observed during the pandemic.

In contrast, our ensemble, its component models, and previous work all highlight the more stable performance of death forecasts over time horizon. The ensemble in this study continued to outperform the baseline at four weeks ahead, and other work has found death forecasts perform well with up to six weeks lead time [16]. We could interpret this as due to the longer time lag between infection and death, which allows forecasters to incorporate the effect of changes in transmission. Additionally, the performance of trend-based forecasts may have benefited from the slower changes to trends in incident deaths caused by increasing vaccination rates, in turn supporting the performance of the ensemble.

Exploring variations in ensemble methods we found that the choice of simple mean or median average had the most consistent impact on performance, regardless of methods of weighting and inclusion by performance history. Other work has supported the importance of the median in providing a stable forecast that better accounts for outliers than the mean [15]. However, our results did not show a strong performance benefit for any one methodological choice, joining the existing mixed evidence for any optimal ensemble method for combining short term probabilistic infectious disease forecasts. In similar analyses of US COVID-19 forecasts many methods of combination have performed competitively, including the simple mean and weighted approaches outperforming unweighted or median methods [14]. This contrasts with later analyses finding weighted methods to give similar performance to a median average [4], [15]. We can partly explain this inconsistency if performance of each method depends on the outcome being predicted (cases, deaths), its count (incident, cumulative) and absolute level, the changing disease dynamics, and the varying quality and quantity of forecasting teams over time.

We also identified benefits of our approach beyond the results of this analysis. Open access to visualised forecasts and data is useful for both academics and the public in an emergency setting when forecasts can influence individual to international actions that change epidemic dynamics [1]. Existing participatory modelling efforts for COVID-19 have been useful for policy communication [3], while multi-country efforts have included only single models adapted to country-specific parameters [17], **aguasModellingCOVID19Pandemic2020?**. By expanding participation to many modelling teams, our work can create robust ensemble forecasts across Europe while allowing comparison across forecasts built with different interpretations of current data, on a like for like scale in real time. At the same time, collating time-stamped predictions ensures that we can test true out-of-sample performance of models and avoid retrospective claims

of performance. Testing the limits of forecasting ability with these comparisons forms an important part of communicating any model-based prediction to decision makers.

However, we note several limitations to our approach. First, our assessment of individual model performance may have been inaccurate due to limitations in the data source used. We saw some real time data revised retrospectively, introducing bias in either direction where the data used to create forecasts was not the same as that used to evaluate it. We mitigated this by excluding forecasts made at or for a time of missing, unreliable, or heavily revised data. We used a manual process for determining anomalous data, and if we did not detect where data revisions affected forecasts this could have created inaccurate forecast scores. However, we note that the national data used here are less likely to see revisions than subnational data **dongInteractiveWebbasedDashboard2020a?**.

The results presented also depend on our choice of performance metric and baseline. While other work supports the use of the weighted interval score **gneitingStrictlyProperScoring2007?**, our use of a flat-line comparison meant that it was more difficult for forecasts to perform well in relative terms during periods where incidence was very stable [11]. This may have differentially biased forecast performance where, for equally good forecasts for different targets, models that predicted a change in trend were rewarded with better scores than those that equally accurately predicted a stable continuation. For example, since epidemics in theory follow an exponential growth and decay, an alternative baseline could have modelled a straight line on the log scale. However, previous work in a similar context has suggested that choice of baseline does not substantively affect the result that ensembles outperform individual models **bracherPreregisteredShort-termForecasting2021?**. Further work could also consider how well our results compare when using an alternative baseline suitable for epidemics, for example an exponential growth model.

The result that the ensemble was among the most reliable across countries and over time could also have been influenced by the sample of contributing forecasts. We accepted all modelling teams' participation and teams used a wide variety of methods. Meanwhile, teams may have changed their forecast methods, and entered and exited the hub over time. The ensemble therefore included forecasts based on models with changing assumptions each week, and we did not test how far the stability or methods of component forecasts influenced the resulting ensemble. This could be significant, for example during a time of low incidence, including only compartmental models in an ensemble improved predictive performance relative to including forecasts from a wider variety of methods [14]. However, the same study found the most consistent ensemble over time was that which included all forecasts regardless of method, with performance increasing with the number of forecast models, so our results are unlikely to have changed by excluding any contributing forecasts.

We see additional scope to adapt the hub to the changing COVID-19 situation across Europe. We have recently extended the hub infrastructure to include short term forecasts for hospitalisations with COVID-19, which is a challenging task due to limited data across the locations covered by the hub. We consider it valuable to separately investigate models for longer term scenarios in addition to the short term forecasts, particularly as the policy focus shifts from immediate response to anticipating changes brought by vaccinations or the geographic spread of new variants **europeancentrefordiseasepreventionandcontrolOverviewImplementationCOVID192021?**, which will be further explored in a similar framework to existing work in the US [18].

This study raises further questions which could inform epidemic forecast modellers and users. The dataset created by the European Forecast Hub is an openly accessible, standardised, and extensively documented catalogue of real time forecasting work from a range of teams and models across Europe, and we recommend its use for further research on forecast performance. Future work could explore the impact of changing epidemiology on individual or ensemble models by combining analyses of trends and turning points in cases and deaths with forecast performance, or extending to include data on vaccination, variant, or policy changes over time. There is also a wide range of methods for combining forecasts which could improve performance of an ensemble or continue to demonstrate the value of a simple approach. This includes altering the inclusion criteria of forecast models based on different thresholds of past performance, excluding or including only forecasts that predict the lowest- and highest-values (trimming) [14], or using alternative weighting methods such as quantile regression averaging [5]. Exploring these questions would add to our understanding of real time performance, supporting and improving future forecasting efforts.

We further recommend adapting and using our open-source computational infrastructure elsewhere for applied public health work. The hub structure maximises the transparency and accuracy of real-time forecasts and can reduce reliance on individual models as a basis for action during an epidemic. The benefits of combining multiple models into an ensemble come from individual models' wide variation in forecast performance across varying targets, and this is particularly true during emerging epidemics where forecasters vary in how quickly their models are able to adapt to new information. Setting up the infrastructure for this could be an important component to future epidemic and pandemic preparedness.

In conclusion, we have shown that an ensemble forecast performed reliably well across multiple forecast targets, with good short term predictions during a rapidly evolving epidemic spreading through multiple populations. In addition, we have demonstrated some limits to predictability, especially for case forecasts longer than two weeks, with few methods able to consistently improve forecast performance other than the use of a median rather than mean average. Our results constitute an important step towards unifying, improving and understanding COVID-19 forecasts, which are an essential tool for policy-makers to assess the epidemiological situation in European countries.

References

[1] P. van Basshuysen, L. White, D. Khosrowi, and M. Frisch, "Three Ways in Which Pandemic Models May Perform a Pandemic," *Erasmus Journal for Philosophy and Economics*, vol. 14, no. 1, pp. 110-127-110-127, 2021, doi: 10.23941/ejpe.v14i1.582.

[2] N. G. Reich *et al.*, "A collaborative multiyear, multimodel assessment of seasonal influenza forecasting in the United States," *Proceedings of the National Academy of Sciences*, vol. 116, no. 8, pp. 3146–3154, 2019, doi: 10.1073/pnas.1812594116.

[3] CDC, "Coronavirus Disease 2019 (COVID-19)," *Centers for Disease Control and Prevention*. 2020. Accessed: Jan. 09, 2022. [Online]. Available: https://www.cdc.gov/coronavirus/2019-ncov/science/forecasting/forecasting.html

[4] E. L. Ray *et al.*, "Ensemble Forecasts of Coronavirus Disease 2019 (COVID-19) in the U.S." Cold Spring Harbor Laboratory Press, p. 2020.08.19.20177493, 2020. doi: 10.1101/2020.08.19.20177493.

[5] S. Funk *et al.*, "Short-term forecasts to inform the response to the Covid-19 epidemic in the UK," *medRxiv*, p. 2020.11.11.20220962, 2020, doi: 10.1101/2020.11.11.20220962.

[6] M. Bicher *et al.*, "Supporting COVID-19 Policy-Making with a Predictive Epidemiological Multi-Model Warning System," *medRxiv*, p. 2020.10.18.20214767, 2021, doi: 10.1101/2020.10.18.20214767.

[7] N. G. Reich *et al.*, "Accuracy of real-time multi-model ensemble forecasts for seasonal influenza in the U.S." *PLOS Computational Biology*, vol. 15, no. 11, p. e1007486, 2019, doi: 10.1371/journal.pcbi.1007486.

[8] M. A. Johansson *et al.*, "An open challenge to advance probabilistic forecasting for dengue epidemics," *Proceedings of the National Academy of Sciences*, vol. 116, no. 48, pp. 24268–24274, 2019, doi: 10.1073/pnas.1909865116.

[9] C. Viboud *et al.*, "The RAPIDD ebola forecasting challenge: Synthesis and lessons learnt," *Epidemics*, vol. 22, pp. 13–21, 2018, doi: 10.1016/j.epidem.2017.08.002.

[10] K. R. Moran *et al.*, "Epidemic Forecasting is Messier Than Weather Forecasting: The Role of Human Behavior and Internet Data Streams in Epidemic Forecast," *The Journal of Infectious Diseases*, vol. 214, no. suppl_4, pp. S404–S408, 2016, doi: 10.1093/infdis/jiw375.

[11] E. Y. Cramer *et al.*, "Evaluation of individual and ensemble probabilistic forecasts of COVID-19 mortality in the US," *medRxiv*, p. 2021.02.03.21250974, 2021, doi: 10.1101/2021.02.03.21250974.

[12] C. Genest, "Vincentization Revisited," *The Annals of Statistics*, vol. 20, no. 2, pp. 1137–1142, 1992, Accessed: Jan. 09, 2022. [Online]. Available: https://www.jstor.org/stable/2242003

[13] E. Ray, "Challenges in training ensembles to forecast COVID-19 cases and deaths in the United States," *International Institute of Forecasters*. 2021. Accessed: Aug. 05, 2021. [Online]. Available: https://forecasters.org/blog/2021/04/09/challenges-in-training-ensembles-to-forecast-covid-19-cases-and-deaths-in-the-united-states/

[14] J. W. Taylor and K. S. Taylor, "Combining Probabilistic Forecasts of COVID-19 Mortality in the United States," *European Journal of Operational Research*, 2021, doi: 10.1016/j.ejor.2021.06.044.

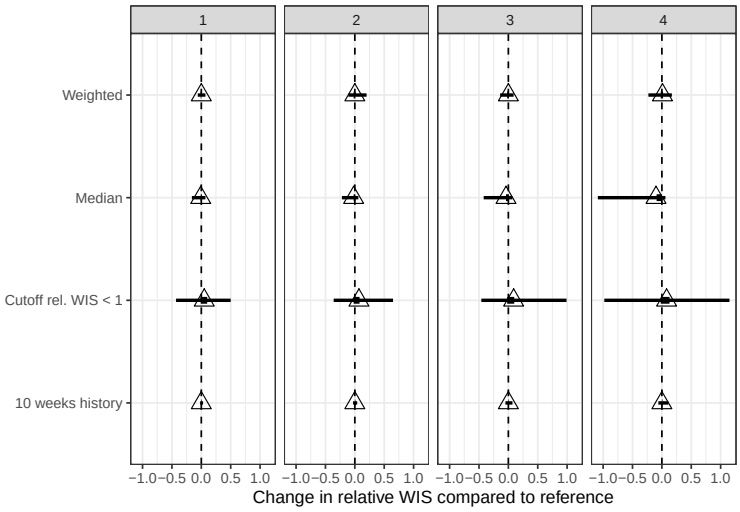
[15] L. Brooks, "Comparing ensemble approaches for short-term probabilistic COVID-19 forecasts in the U.S." *International Institute of Forecasters*. 2020. Accessed: Jul. 15, 2021. [Online]. Available: https://forecasters.org/blog/2020/10/28/comparing-ensemble-approaches-for-short-term-probabilistic-covid-19-forecasts-in-the-u-s/

[16] J. Friedman *et al.*, "Predictive performance of international COVID-19 mortality forecasting models," *Nature Communications*, vol. 12, no. 1, p. 2609, 2021, doi: 10.1038/s41467-021-22457-w.

[17] K. Adib *et al.*, "A participatory modelling approach for investigating the spread of COVID-19 in countries of the Eastern Mediterranean Region to support public health decision-making," *BMJ Global Health*, vol. 6, no. 3, p. e005207, 2021, doi: 10.1136/bmjgh-2021-005207.

[18] R. K. Borchering, "Modeling of Future COVID-19 Cases, Hospitalizations, and Deaths, by Vaccination Rates and Nonpharmaceutical Intervention Scenarios — United States, April–September 2021," *MMWR. Morbidity and Mortality Weekly Report*, vol. 70, 2021, doi: 10.15585/mmwr.mm7019e3.

Supplementary information



\begin{figure} \caption{Figure SI2: Performance of alternative ensemble methods at each weekly horizon (1-4), showing mean difference (triangle) in relative weighted interval score, with 48% and 96% probability (thick and thin line respectively). The difference in WIS is a comparison of scores from forecasts made from all possible combinations of methods, with a single element of ensemble method input changed. Reference categories are: weighted v. unweighted; median v. mean; cutoff by WIS v. all models included;, relative WIS measures over 10 weeks of forecast history vs. all forecasts} \end{figure}

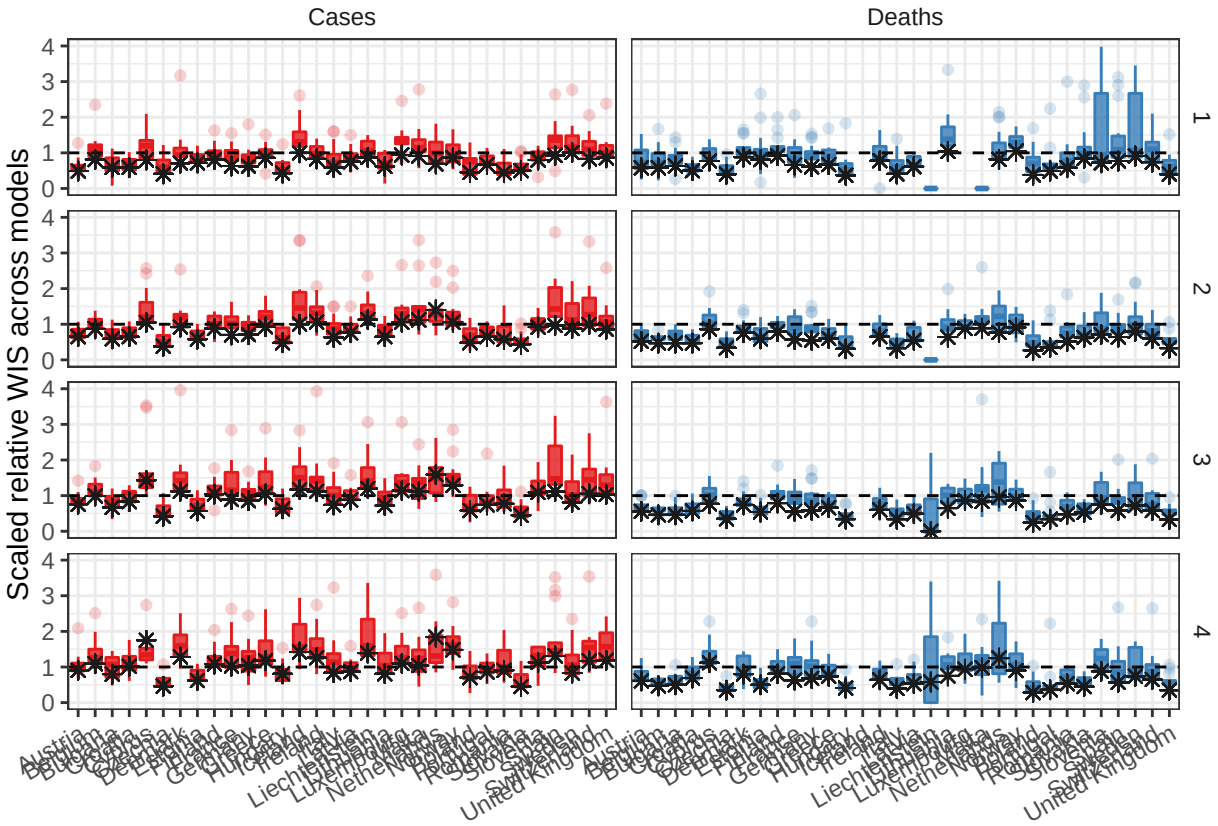


Figure 3: Figure SI1: Performance of short-term forecasts across models and median ensemble (asterisk), by country, forecasting cases (left) and deaths (right) for one-week (top) through four-week (bottom) ahead forecasts, according to the relative weighted interval score. Boxplots show interquartile ranges, with outliers as faded points, and the ensemble model performance is marked by an asterisk.