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 in predicting COVID-19 cases and deaths across Europe between 08 March 2021 and 07 March 2022.

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 were submitted using standardised quantiles of the predictive distribution. Each week we created an ensemble forecast, where each predictive quantile was calculated as the equally-weighted average (initially the mean and then the median) of all individual models' predictive quantiles. We measured the performance of each model using the relative Weighted Interval Score (WIS), comparing models' forecast accuracy relative to all other models. We retrospectively explored alternative methods for ensemble forecasts, including weighted averages based on models' past predictive performance.

Results Over 52 weeks we collected and combined up to 28 forecast models for 32 countries. We found a weekly ensemble had a strong and consistently reliable performance across countries over time. Across all horizons and locations, the ensemble performed better on relative WIS than 84% of participating models' forecasts of incident cases (with a total N=862), and 92% of participating models' forecasts of deaths (N=746). Across a one to four week time horizon, ensemble performance declined with longer forecast periods when forecasting cases, but remained stable over four weeks for incident death forecasts. In every forecast across 32 countries, the ensemble outperformed most contributing models when forecasting either cases or deaths, frequently outperforming all of its individual component models. Among several choices of ensemble methods we found that the most influential and best choice was to use a median average of models instead of using the mean, regardless of methods of weighting component forecast models.

Conclusions Our results support the use of combining forecasts from individual models into an ensemble in order to improve predictive performance across epidemiological targets and populations during infectious disease epidemics. Our findings further suggest that median ensemble methods yield better predictive performance more than ones based on means. Our findings also highlight that forecast consumers should place more weight on incident death forecasts than incident case forecasts at forecast horizons greater than two weeks.

Code and data availability All data and code are publicly available on Github: covid19-forecast-hub-europe/euro-hub-ensemble.

This document was generated on 2022-04-14

Background

Epidemiological forecasts make quantitative statements about a disease outcome in the near future. 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]. Forecasters use a variety of methods and models for creating and publishing forecasts, varying in both defining the forecast outcome and in reporting the probability distribution of outcomes [2], [3]. Such variation between forecasts makes it difficult to compare predictive performance between forecast models, and consequently derive objective arguments for preferring one forecast over another. This hampers the selection of a forecast as a reliable basis for decisions.

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 [4]. A hub sets a commonly agreed-upon 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 like-for-like basis, which can strengthen public and policy use of disease forecasts [5]. 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 [6];[7]], with similar efforts elsewhere [8]–[10].

Standardising forecasts allows for combining multiple forecasts into a single ensemble with the potential for an improved predictive performance. Evidence from previous efforts in multi-model infectious disease forecasting suggests that forecasts from an ensemble of models can be consistently high performing compared to any one of the component models [11]–[13]. Elsewhere, weather forecasting has a long-standing use of building ensembles of models using diverse methods with standardised data and formatting in order to improve performance [14], [15].

The European COVID-19 Forecast Hub [16] 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 [17], [18] and the German and Polish COVID-19 [19] forecast hubs.

Forecast targets and models

We sought forecasts for two measures of COVID-19 incidence: the total reported number of cases and deaths per week. We considered forecasts for 32 countries in Europe, including all countries of the European Union and European Free Trade Area, and the United Kingdom. We compared forecasts against observed data reported by Johns Hopkins University (JHU, [20]). JHU data included a mix of national and aggregated

subnational data for the 32 countries in the Hub. 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 of the predictive 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. No restrictions were placed on who could submit forecasts, 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 to empirical models, agent-based and statistical models, and ensembles of multiple quantitative or qualitative models (described at https://covid19forecasthub.eu/community.html). We maintain a full project specification with a detailed submissions protocol [21].

Each week, teams submitted forecasts for the following weeks to the hub. Teams submitted at latest two days after the complete dataset for the latest forecasting week became available each Sunday. We implemented an automated validation programme to check that each new forecast conformed to standardised formatting. Forecast validation ensured a monotonic increase of predictions with each increasing quantile, integer-valued non-negative counts of predicted cases, as well as consistent date and location definitions.

Each week we built an ensemble of all validated forecasts. From the first week of forecasting from 8 March 2021, the ensemble method for summarising across forecasts was the arithmetic mean of all models at each predictive quantile for a given location, target, and horizon. We noticed that including highly anomalous forecasts in a mean ensemble produced extremely wide uncertainty. To mitigate this, from 26 July 2021 onwards the ensemble instead used a median of all predictive quantiles.

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 [22]. All forecast and meta data are freely available and held on Zoltar, a platform for hosting epidemiological forecasts [23], [24].

Forecast evaluation

For each model, we evaluated performance in terms of both accuracy (coverage) and overall predictive performance (weighted interval score). 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 [25]. We removed any forecast surrounding (both the week of, and the first week after) a strongly anomalous data point. We defined anomalous as where any subsequent data release revised that data point by over 5%.

We established the accuracy of each model's prediction boundaries as the coverage of the predictive intervals. We calculated coverage at a given interval level k, where $k \in [0, 1]$, 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 underconfident 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 $k = 0.5$ and $k = 0.95$ levels.

We also assessed the overall predictive performance of weekly forecasts using the weighted interval score (WIS) across all available quantiles. The WIS represents a parsimonious approach to scoring forecasts based on uncertainty represented as forecast values across a set of quantiles [26], and is a strictly proper scoring rule, that is, it is optimal 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 [27]. Each forecast for a given location and date is scored based on an observed count of weekly incidence, the median of the predictive distribution and the predictive upper and lower quantiles corresponding to the central predictive interval level.

Not all models provided forecasts for all locations and dates, and we needed to compare predictive performance in the face of various levels of missingness across each forecast target. Therefore we calculated a relative WIS. This is 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). 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, such that smaller values indicate better performance.

We scaled the relative WIS of each model with the relative WIS of a baseline model, for each forecast target, location, date, and horizon. The baseline model assumes case or death counts stay the same as the latest data point over all future horizons, with expanding uncertainty, described previously in [28]. Here we report the relative WIS of each model with respect to the baseline model.

Ensemble methods We retrospectively explored alternative methods for combining 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 [29]

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

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. Setting them all equal so that they sum up to 1, $w_i = 1/n$ results in an arithmetic mean ensemble. To avoid this overrepresentation, we can choose a set of weights to apply to forecasts before they are combined at each quantile level. Numerous options exist for choosing these weights with the aim to maximise predictive performance, including choosing weights to reflect each forecast's past performance (thereby moving from an untrained to a trained ensemble). A straightforward choice is so-called inverse score weighting, which was recently found in the US to outperform unweighted scores during some time periods [30] but not confirmed in a similar study in Germany and Poland Poland [8]. In this case, the weights are calculated as

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

where S_i reflects the forecast skill of forecaster i, normalised so that weights sum to 1.

When constructing ensembles from quantile means, a single outlier can have an oversized effect on the ensemble forecast. Previous research has found that an unweighted median ensemble, where the arithmetic mean of each quantile is replaced by a median, yields competitive performance while maintaining robustness to outlying forecasts [31]. Building on this, created weithed median ensembles using the same weights described above and a Harrel-Davis quantile estimator with a beta function to approximate the weighted percentiles [32]. We then conmpared the performance of unweighted and inverse relative WIS weighted mean and median ensembles.

Results

We collected forecasts submitted weekly in real time over the 52 week period from 08 March 2021 to 07 March 2022. Each week we used all available forecasts to create a weekly real-time ensemble model (referred

to as "the ensemble" from here on) for each of the 256 possible forecast targets: incident cases and deaths in 32 locations over the following one through four weeks. The ensemble model was an unweighted average from March through July 2021 and then an unweighted median. An example of weekly forecasts from the ensemble model is shown in Figure 1.

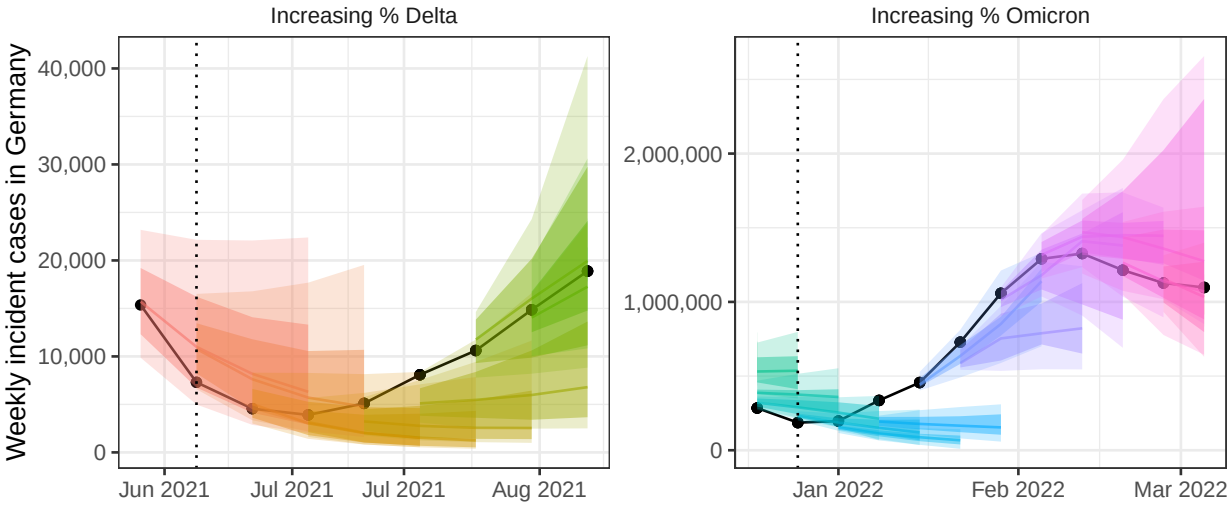


Figure 1: *Ensemble forecasts of weekly incident cases in Germany over periods of increasing SARS-CoV-2 variants Delta (B.1.617.2, left) and Omicron (B.1.1.529, right). Black indicates observed data. Coloured ribbons represent each weekly forecast of 1-4 weeks ahead (showing median, 50%, and 90% probability). For each variant, forecasts are shown over an x-axis bounded by the earliest dates at which 5% and 99% of sequenced cases were identified as the respective variant of concern, while vertical dotted lines indicate the approximate date that the variant reached dominance (>50% sequenced cases).*

The number of models contributing to each ensemble forecast varied over time and by forecasting target (SI Figure 1). Over the whole study period 26 independently participating forecasting teams contributed results from 28 unique forecasting models. While not all modellers created forecasts for all locations, horizons, or variables, no ensemble forecast was composed of less than 3 independent models. At most, 15 models contributed forecasts for cases in Germany at the 1 week horizon, with an accumulated 592 forecasts for that single target over the study period. In contrast, deaths in Finland at the 2 week horizon saw the smallest number of forecasts, with only 6 independent models contributing a total 24 forecasts. Similarly, not all teams forecast across all quantiles of the predictive distribution for each target, with only 23 models providing the full set of 23 quantiles.

Using all models and the ensemble, we created 2106 forecasting scores where each score summarises a unique combination of forecasting model, variable, country, and week ahead horizon (SI Figure 2). We qualitatively reviewed the absolute performance of forecasts in terms of accuracy in predicting numbers of incident cases and deaths. We observed that forecasts were often most accurate in times of stable epidemic behaviour, while struggling to accurately predict at longer horizons around inflection points, for example during rapid changes in population-level behaviour or surveillance. Forecast models varied widely in their ability to predict and account for the introduction of new variants, giving the ensemble forecast over these periods a high level of uncertainty (Figure 1). In this study we focus only on the comparative performance of forecasting models relative to each other.

In relative terms, the ensemble of all models performed well compared to both its component models and the baseline. By relative WIS scaled against a baseline of 1 (where a score <1 indicates outperforming the baseline), the median score for participating models across all submitted forecasts was 1.04, while the median score of forecasts from the ensemble model was 0.71. Across all horizons and locations, the ensemble

performed better on scaled relative WIS than 84% of participating model scores when forecasting cases (with a total N=862), and 92% of participating model scores for forecasts of incident deaths (N=746).

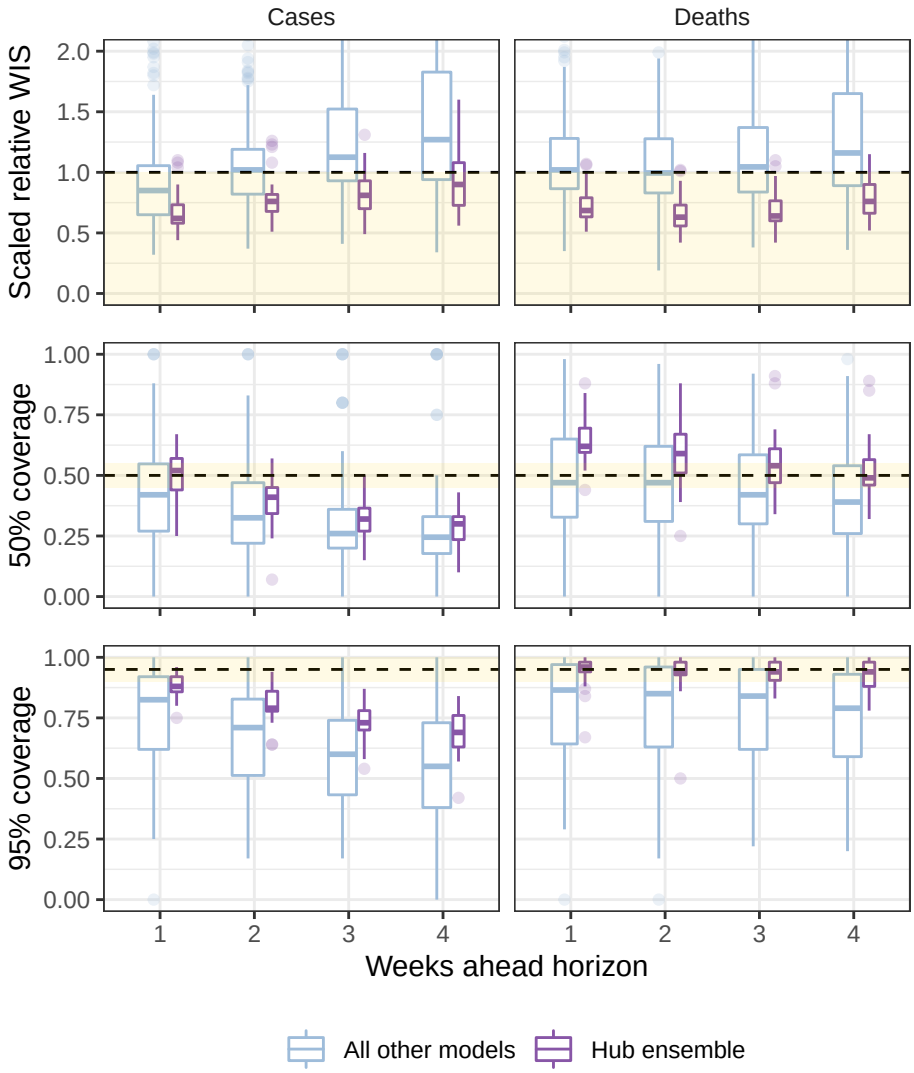


Figure 2: *Performance of short-term forecasts aggregated across all individually submitted models and the Hub ensemble, by horizon, forecasting cases (left) and deaths (right). Performance measured by relative weighted interval score scaled against a baseline (dotted line, 1), and coverage of uncertainty at the 50% and 95% levels. Boxplot, with width proportional to number of observations, show interquartile ranges with outlying scores as faded points. The target range for each set of scores is shaded in yellow.*

The performance of individual and ensemble forecasts varied by length of the forecast horizon (Figure 2). At each horizon, the typical performance of the ensemble outperformed both the baseline model and the aggregated scores of all its component models, although we saw wide variation between individual models in performance across horizons. Both individual models and the ensemble saw a trend of worsening performance at longer horizons when forecasting cases with the median scaled relative WIS of the ensemble across locations worsened from 0.62 for one-week ahead forecasts to 0.9 when forecasting four weeks ahead. Performance for forecasts of deaths was more stable over one through four weeks, with median ensemble performance moving from 0.685 to 0.76 across the four week horizons.

We observed similar trends in performance across horizon 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% and 95% nominal coverage at the 50% and 95% levels respectively). This did not hold at longer forecast horizons as the case forecasts became increasingly over-confident. Meanwhile, the ensemble of death forecasts was well calibrated at the 95% level across all horizons, and the calibration of death forecasts at the 50% level improved with lengthening horizons compared to being underconfident at shorter horizons.

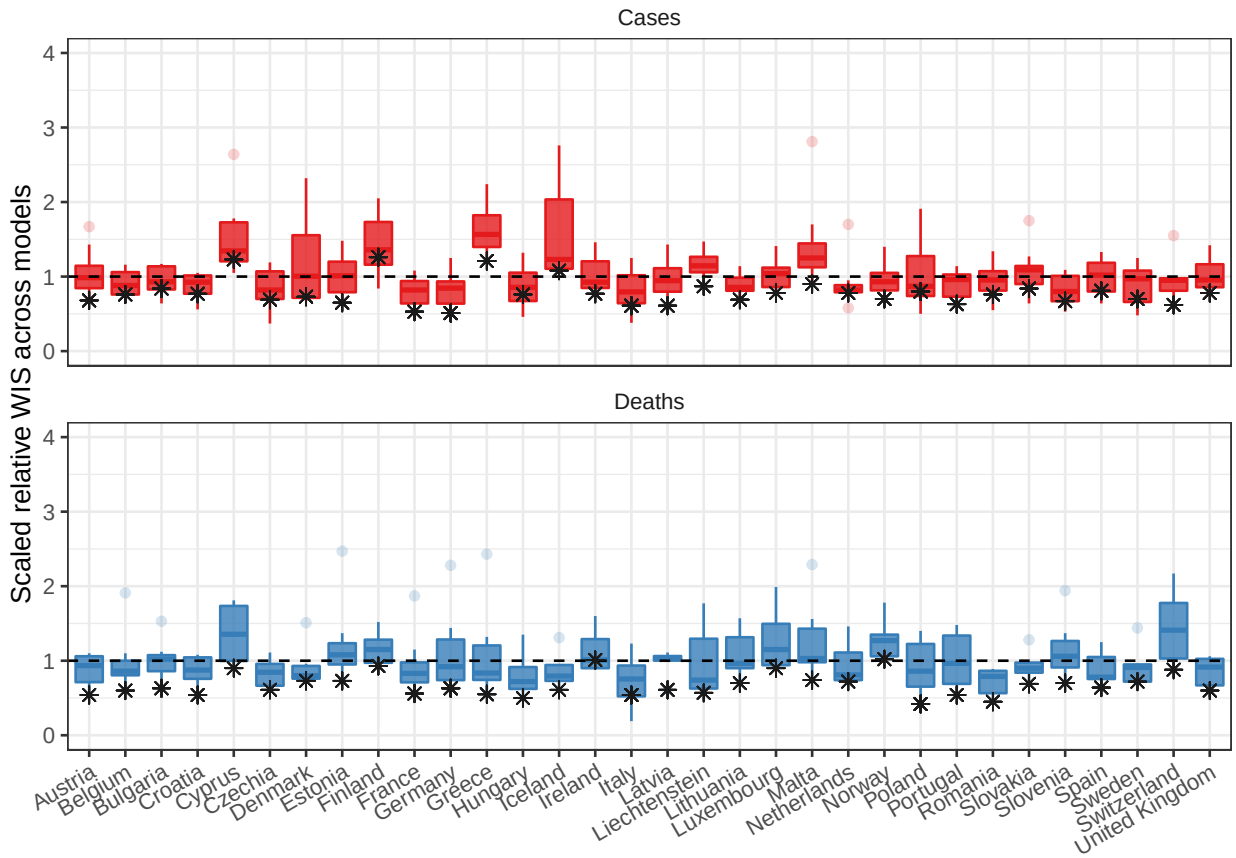


Figure 3: *Performance of short-term forecasts across models and median ensemble (asterisk), by country, forecasting cases (top) and deaths (bottom) for two-week ahead forecasts, according to the relative weighted interval score. Boxplots show interquartile ranges, with outliers as faded points, and the ensemble model performance is marked by an asterisk. y-axis is cut-off to an upper bound of 4 for readability.*

The ensemble also performed consistently well in comparison to individual models when forecasting across countries (Figure 3). Across 32 countries, on aggregate forecasting for one through four weeks, when forecasting cases the ensemble oupterformed 75% of component models in 21 countries, and outperformed all available models in 3 countries. When forecasting deaths, the ensemble outperformed 75% and 100% of models in 30 and 9 countries respectively. Considering only the the two-week horizon shown in Figure 3, the ensemble of case forecasts outperformed 75% models in 24 countries and all models in only 12 countries. At the two-week horizon for forecasts of deaths, the ensemble outperformed 75% and 100% of its component models in 30 and 26 countries respectively.

We considered alternative methods for creating ensembles from the participating forecasts, using either a mean or median to combine either weighted or unweighted forecasts (Table 1). Across locations we observed that the median outperformed the mean across all one through four week horizons and both cases and death targets, for all but cases at the 1 week horizon. This held regardless of whether the component forecasts were

Table 1: Predictive performance of main ensembles, as measured by the scaled relative WIS.

Horizon	Weighted mean	Weighted median	Unweighted mean	Unweighted median
Cases				
1 week	0.59	0.62	0.59	0.61
2 weeks	0.67	0.67	0.67	0.67
3 weeks	0.79	0.70	0.81	0.71
4 weeks	1.06	0.75	1.09	0.79
Deaths				
1 week	0.63	0.59	1.00	0.59
2 weeks	0.57	0.54	0.81	0.53
3 weeks	0.64	0.56	0.83	0.54
4 weeks	0.83	0.64	0.82	0.62

weighted or unweighted by their individual past performance. Between methods of combination, weighting made little difference to the performance of the median ensemble, but slightly improved performance of the mean ensemble.

Discussion

We collated 12 months of forecasts of COVID-19 cases and deaths across 32 countries in Europe, collecting from multiple independent teams and using a principled approach to standardising both forecast targets and the uncertainty around predictions. We combined these into an ensemble forecast and compared the relative performance of forecasts among models, finding that the ensemble forecasts produced among the most consistent predictive performance across countries and horizons over time compared to individual models.

Our results support previous findings that ensemble forecasts are the best or nearly the best performing models with respect to error and appropriate coverage of uncertainty [9], [13], [28]. 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 [8], [33]. 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. We identified the adaptability of an ensemble forecast to changing conditions as a particular benefit from applying our approach to the COVID-19 outbreak in Europe. 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.

In particular, our results suggest the limited value of reporting case forecasts further into the future than the shortest horizons. Previous work has similarly found rapidly declining performance for case forecasts with increasing horizon [28], [34]. 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. In our study's context, this would be a strong assumption with many instances of rapidly changing policies and individual behaviour observed over the period. In contrast, our results highlight the more stable performance of death forecasts over lengthening time horizons. Specifically, we found the ensemble in this study continued to outperform both other models and the baseline at up to four weeks ahead. In general, previous work has found death forecasts perform well with up to six weeks lead time [35]. We could interpret this as due to the longer time lag between infection and death [36], and higher consistency of reporting in surveillance data [37], which allow 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 gradually increasing vaccination rates.

When exploring variations in ensemble methods, we found that the choice of median over means yielded the most consistent improvement in predictive performance, regardless of the method of weighting. Other work has supported the importance of the median in providing a stable forecast that better accounts for outliers than the mean [33]. In contrast, weighing models by past performance did not result in any consistent improvement in performance, in line with 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 [30]. This contrasts with later analyses finding weighted methods to give similar performance to a median average [7], [33]. 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. That said, there is much scope for future resaerch into methods for combining forecasts to improve performance of an ensemble. 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) [30], or using alternative weighting methods such as quantile regression averaging [9]. Exploring these questions would add to our understanding of real time performance, supporting and improving future forecasting efforts.

Over the study period, we saw multiple fundamental changes in viral-, individual-, and population-level factors driving the transmission of COVID-19 across Europe. In early 2021, the introduction of vaccination started to change population-level associations between infections, cases, and deaths [[38], while the Delta variant emerged and became dominant in Europe [39]. Similarly from late 2021 we saw the interaction of individually waning immunity during the emergence and global spread of the Omicron variant [40]. Meanwhile, neither the extent nor timing of these factors were uniform across European countries covered by the Forecast Hub [41]. Future work could explore the impact on forecast models of changing epidemiology at a broad spatial scale 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.

We note several limitations in our approach to assessing the relative performance of an ensemble among forecast models. Our results are the outcome of evaluating forecasts against a specific performance metric and baseline, where multiple options for evaluation exist and the choice reflects the aim of the evaluation process. Further, our choice of baseline model affects the given performance scores in absolute terms, and more generally the choice of appropriate baseline for epidemic forecast models is not obvious when assessing infectious disease forecasts. The model used here is supported by previous work [28], yet previous evaluation in a similar context has suggested that choice of baseline affects relative performance in general [42], and future research should be done on the best choices of baseline models in the context of infectious disease epidemics. Our assessment of forecast performance may further have been inaccurate due to limitations in the observed data against which we evaluated forecasts. We sourced data from a globally aggregated database to maintain compatibility across 32 countries [20]. However, this made it difficult to identify the origin of lags and inconsistencies between national data streams, and to what extent these could bias forecasts for different targets. In particular 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 attempted to mitigate this using by using an automated process for determining data revisions, and excluding forecasts made at a time of missing, unreliable, or heavily revised data.

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 [5], while multi-country efforts have included only single models adapted to country-specific parameters [43]–[45]. By expanding participation to many modelling teams, our work was able to 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.

This study raises many 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 [22], and we recommend its use for further research on forecast performance. In the code developed for this study we hope to provide a worked example of downloading and using both the forecasts and their evaluation scores. We see additional scope to adapt the Hub format to the changing COVID-19 situation across Europe. We have extended the Forecast 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. As the policy focus shifts from immediate response to anticipating changes brought by vaccinations or the geographic spread of new variants [41], we are also separately investigating models for longer term scenarios in addition to the short term forecasts in a similar framework to existing scenario modelling work in the US [46].

In conclusion, we have shown that during a rapidly evolving epidemic spreading through multiple populations, an ensemble forecast performed highly consistently across a large matrix of forecast targets, typically outperforming the majority of its separate component models. In addition, we have demonstrated some limits to predictability, especially for case forecasts, while showing that ensemble methods based on past model performance were unable to reliably improve forecast performance. Our work constitutes a step towards both unifying COVID-19 forecasts and improving our understanding of them.

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, 1, pp. 110-127-110-127, Jul. 2021, doi: 10.23941/ejpe.v14i1.582.

[2] J. Zelner, J. Riou, R. Etzioni, and A. Gelman, "Accounting for uncertainty during a pandemic," *PATTER*, vol. 2, no. 8, Aug. 2021, doi: 10.1016/j.patter.2021.100310.

[3] L. P. James, J. A. Salomon, C. O. Buckee, and N. A. Menzies, "The Use and Misuse of Mathematical Modeling for Infectious Disease Policymaking: Lessons for the COVID-19 Pandemic," *Med Decis Making*, vol. 41, no. 4, pp. 379–385, May 2021, doi: 10.1177/0272989X21990391.

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

[5] CDC, "Coronavirus Disease 2019 (COVID-19)," Feb. 11, 2020. https://www.cdc.gov/coronavirus/2019-ncov/science/forecasting/forecasting.html (accessed Jan. 09, 2022).

[6] E. Y. Cramer *et al.*, "The United States COVID-19 Forecast Hub dataset," p. 2021.11.04.21265886, Nov. 2021, doi: 10.1101/2021.11.04.21265886.

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

[8] J. Bracher *et al.*, "A pre-registered short-term forecasting study of COVID-19 in Germany and Poland during the second wave," *Nat Commun*, vol. 12, no. 1, 1, p. 5173, Aug. 2021, doi: 10.1038/s41467-021-25207-0.

[9] 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, Nov. 2020, doi: 10.1101/2020.11.11.20220962.

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

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

[12] M. A. Johansson *et al.*, "An open challenge to advance probabilistic forecasting for dengue epidemics," *PNAS*, vol. 116, no. 48, pp. 24268–24274, Nov. 2019, doi: 10.1073/pnas.1909865116.

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

[14] R. Buizza, "Introduction to the special issue on '25 years of ensemble forecasting'," *Quarterly Journal of the Royal Meteorological Society*, vol. 145, no. S1, pp. 1–11, 2019, doi: 10.1002/qj.3370.

[15] K. R. Moran *et al.*, "Epidemic Forecasting is Messier Than Weather Forecasting: The Role of Human Behavior and Internet Data Streams in Epidemic Forecast," *J Infect Dis*, vol. 214, pp. S404–S408, Dec. 2016, doi: 10.1093/infdis/jiw375.

[16] European Covid-19 Forecast Hub, *European COVID-19 Forecast Hub*. covid19-forecast-hub-europe, 2021.Available: https://github.com/covid19-forecast-hub-europe/covid19-forecast-hub-europe

[17] E. Cramer *et al.*, "Reichlab/Covid19-forecast-hub: Release for Zenodo, 20210816," Aug. 2021, doi: 10.5281/zenodo.5208210.

[18] S. Y. Wang *et al.*, "Reichlab/covidHubUtils: Repository release for Zenodo," Aug. 2021, doi: 10.5281/zenodo.5207940.

[19] J. Bracher *et al.*, *The German and Polish COVID-19 Forecast Hub*. 2020.Available: https://github.com/KITmetricslab/covid19-forecast-hub-de

[20] E. Dong, H. Du, and L. Gardner, "An interactive web-based dashboard to track COVID-19 in real time," *The Lancet Infectious Diseases*, vol. 20, no. 5, pp. 533–534, May 2020, doi: 10.1016/S1473-3099(20)30120-1.

[21] European Covid-19 Forecast Hub, "Covid19-forecast-hub-europe: Wiki." https://github.com/covid19-forecast-hub-europe/covid19-forecast-hub-europe

[22] European Covid-19 Forecast Hub, "European Covid-19 Forecast Hub." https://covid19forecasthub.eu/index.html

[23] EpiForecasts, "Project: ECDC European COVID-19 Forecast Hub - Zoltar," 2021. https://www.zoltardata.com/project/238

[24] N. G. Reich, M. Cornell, E. L. Ray, K. House, and K. Le, "The Zoltar forecast archive, a tool to standardize and store interdisciplinary prediction research," *Sci Data*, vol. 8, no. 1, 1, p. 59, Feb. 2021, doi: 10.1038/s41597-021-00839-5.

[25] Nikos I Bosse, Sam Abbott, EpiForecasts, and Sebastian Funk, *Scoringutils: Utilities for Scoring and Assessing Predictions*. 2020.Available: https://github.com/epiforecasts/scoringutils

[26] J. Bracher, E. L. Ray, T. Gneiting, and N. G. Reich, "Evaluating epidemic forecasts in an interval format," *PLOS Computational Biology*, vol. 17, no. 2, p. e1008618, Feb. 2021, doi: 10.1371/journal.pcbi.1008618.

[27] T. Gneiting and A. E. Raftery, "Strictly Proper Scoring Rules, Prediction, and Estimation," *Journal of the American Statistical Association*, vol. 102, no. 477, pp. 359–378, Mar. 2007, doi: 10.1198/016214506000001437.

[28] 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, Jan. 2021, doi: 10.1101/2021.02.03.21250974.

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

[30] J. W. Taylor and K. S. Taylor, "Combining Probabilistic Forecasts of COVID-19 Mortality in the United States," *Eur J Oper Res*, Jun. 2021, doi: 10.1016/j.ejor.2021.06.044.

[31] E. L. Ray *et al.*, "Comparing trained and untrained probabilistic ensemble forecasts of COVID-19 cases and deaths in the United States," Jan. 28, 2022. Accessed: Mar. 30, 2022. [Online]. Available: http://arxiv.org/abs/2201.12387

[32] F. E. HARRELL and C. E. DAVIS, "A new distribution-free quantile estimator," *Biometrika*, vol. 69, no. 3, pp. 635–640, Dec. 1982, doi: 10.1093/biomet/69.3.635.

[33] L. Brooks, "Comparing ensemble approaches for short-term probabilistic COVID-19 forecasts in the U.S." 2020. https://forecasters.org/blog/2020/10/28/comparing-ensemble-approaches-for-short-term-probabilistic-covid-19-forecasts-in-the-u-s/ (accessed Jul. 15, 2021).

[34] M. Castro, S. Ares, J. A. Cuesta, and S. Manrubia, "The turning point and end of an expanding epidemic cannot be precisely forecast," *Proceedings of the National Academy of Sciences*, vol. 117, no. 42, pp. 26190–26196, Oct. 2020, doi: 10.1073/pnas.2007868117.

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

[36] R. Jin, "The lag between daily reported Covid-19 cases and deaths and its relationship to age," *J Public Health Res*, vol. 10, no. 3, p. 2049, Mar. 2021, doi: 10.4081/jphr.2021.2049.

[37] M. Català *et al.*, "Robust estimation of diagnostic rate and real incidence of COVID-19 for European policymakers," *PLOS ONE*, vol. 16, no. 1, p. e0243701, Jan. 2021, doi: 10.1371/journal.pone.0243701.

[38] European Centre for Disease Prevention and Control, "Interim guidance on the benefits of full vaccination against COVID-19 for transmission and implications for non-pharmaceutical interventions - 21 April 2021," ECDC, Stockholm, 2021.Available: https://www.ecdc.europa.eu/en/publications-data/interim-guidance-benefits-full-vaccination-against-covid-19-transmission

[39] European Centre for Disease Prevention and Control, "Threat Assessment Brief: Implications for the EU/EEA on the spread of the SARS-CoV-2 Delta (B.1.617.2) variant of concern," ECDC, Stockholm, Jun. 2021.Available: https://www.ecdc.europa.eu/en/publications-data/threat-assessment-emergence-and-impact-sars-cov-2-delta-variant

[40] European Centre for Disease Prevention and Control, "Assessment of the further spread and potential impact of the SARS-CoV-2 Omicron variant of concern in the EU/EEA, 19th update," Jan. 27, 2022. https://www.ecdc.europa.eu/en/publications-data/covid-19-omicron-risk-assessment-further-emergence-and-potential-impact

[41] European Centre for Disease Prevention and Control, "Overview of the implementation of COVID-19 vaccination strategies and deployment plans in the EU/EEA," ECDC, Stockholm, Nov. 2021.Available: https://www.ecdc.europa.eu/en/publications-data/overview-implementation-covid-19-vaccination-strategies-and-deployment-plans

[42] J. Bracher *et al.*, "National and subnational short-term forecasting of COVID-19 in Germany and Poland, early 2021," p. 2021.11.05.21265810, Nov. 2021, doi: 10.1101/2021.11.05.21265810.

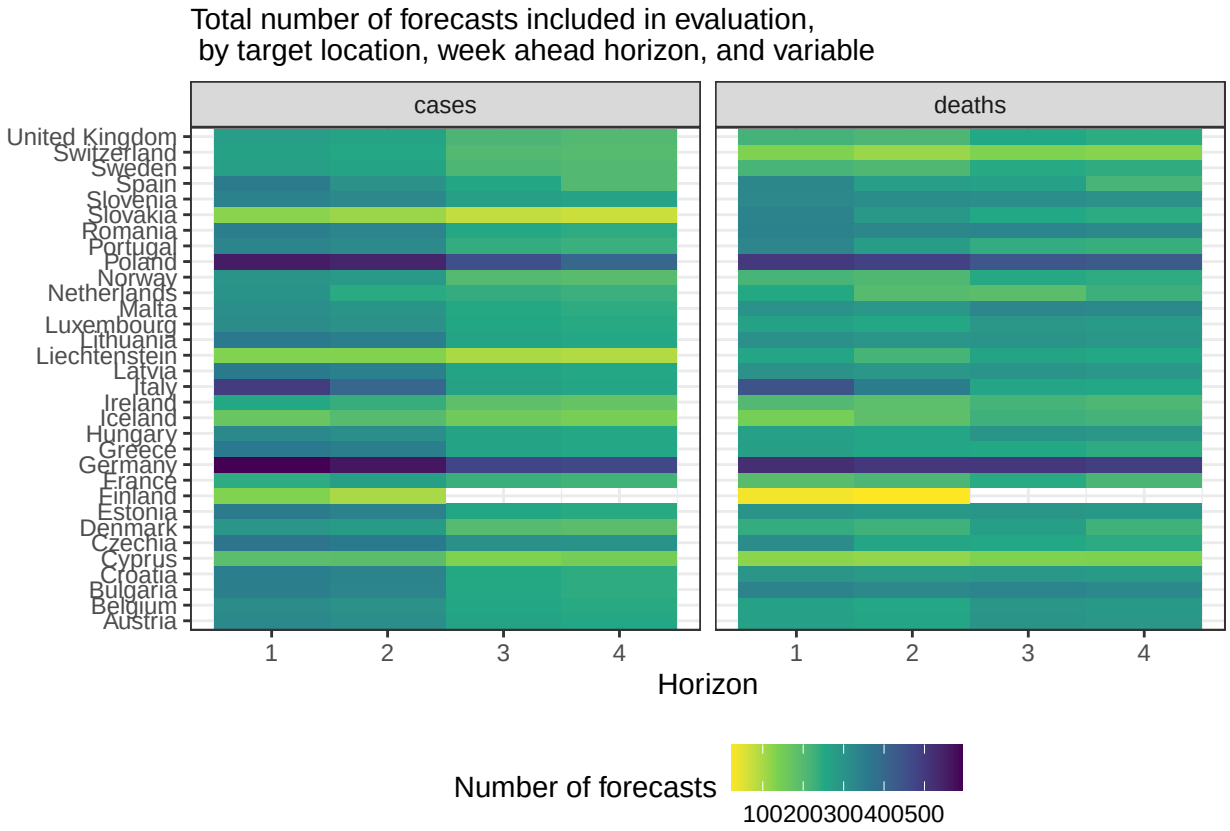
[43] R. Aguas *et al.*, "Modelling the COVID-19 pandemic in context: An international participatory approach," *BMJ Global Health*, vol. 5, no. 12, p. e003126, Dec. 2020, doi: 10.1136/bmjgh-2020-003126.

[44] 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, Mar. 2021, doi: 10.1136/bmjgh-2021-005207.

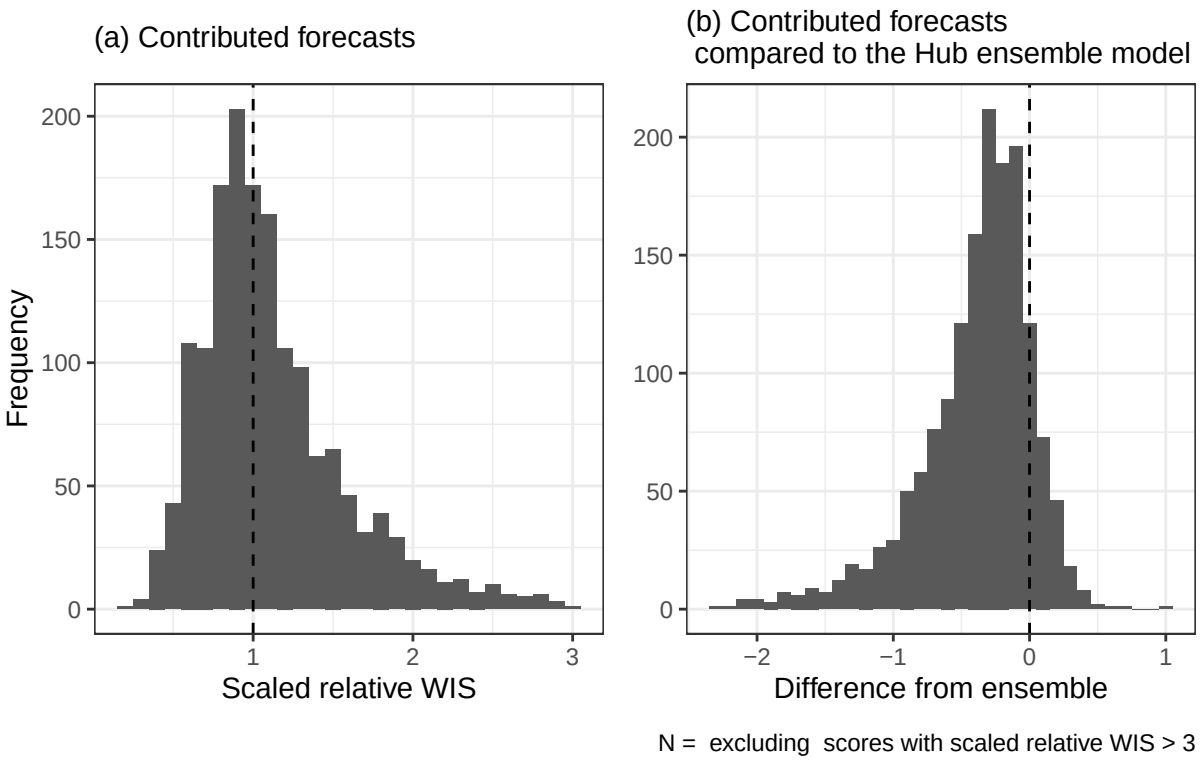
[45] A. Agosto, A. Campmas, P. Giudici, and A. Renda, "Monitoring COVID-19 contagion growth," *Statistics in Medicine*, vol. 40, no. 18, pp. 4150–4160, 2021, doi: 10.1002/sim.9020.

[46] 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 Morb Mortal Wkly Rep*, vol. 70, 2021, doi: 10.15585/mmwr.mm7019e3.

Supplementary information



Comparison of scores between participating model forecasts
and Hub ensemble of all available forecasts for each target



N = excluding scores with scaled relative WIS > 3