Main Manuscript for

Assessing the Performance of COVID-19 Forecasting Models in the U.S.

Kyle J. Colonna[a]*, Roger M. Cooke[b,c], and John S. Evans[a]

[a]Environmental Health Department, Harvard T.H. Chan School of Public Health, Harvard University, Boston, MA 02115
[b]Resources for the Future, Washington, DC 20036
[c]Department of Mathematics, Delft University of Technology, Delft, The Netherlands 2628 XE

***Corresponding Author:** Kyle J. Colonna

Address: 665 Huntington Ave., Building 1, Room 1301, Boston, MA, 02115 USA

Phone Number: 484-832-4003

Email: kcolonna@g.harvard.edu

Author Contributions: *Concept and design:* Colonna, Evans
Acquisition, analysis, or interpretation of data: Colonna, Cooke
Drafting of the manuscript: Colonna, Evans
Critical revision of the manuscript for important intellectual content: All authors
Supervision: Evans

Competing Interest Statement: The authors declare they have no competing interests that might be perceived to influence the results and/or discussion reported in this manuscript. There has also been no prior discussion with an editor.

Classification: Physical Sciences, Statistics; Social Sciences, Social Sciences

Keywords: COVID-19; COVID-19 Decision-Making; Forecasting; Uncertainty Analysis; Cooke's Classical Model

Preprint Server: medRxiv - https://doi.org/10.1101/2020.12.09.20246157

This PDF file includes:

Main Text
Figure 1
Tables 1 and 2

Abstract

Dozens of coronavirus disease 2019 (COVID-19) forecasting models have been created, however, little information exists on their performance. Here we examined the performance of nine commonly-used COVID-19 forecasting models, as well as equal- and performance-weighted ensembles, based on their predictive accuracy and precision, and their probabilistic 'statistical accuracy (aka calibration)' and 'information' scores (measures commonly employed in the evaluation of expert judgment) (Cooke, 1991). Data on observed COVID-19 mortality in eight states, selected to reflect differences in racial demographics and COVID-19 case rates, over eight weeks in the summer of 2020 and eight weeks in the winter of 2021, provided the basis for evaluating model forecasts and exploring the stability/robustness of the results. Two models exhibited superior performance with both predictive and probabilistic measures during both pandemic phases. Models that performed poorly reflected 'overconfidence' with tight forecast distributions. Models also systematically under-predicted mortality when cases were rising and over-predicted when cases were falling. Performance-weighted ensembles consistently outperformed the equal-weighted ensemble, with the Cooke's Classical Model-weighted ensemble outperforming the predictive-performance-weighted ensemble. Model performance depended on the time-frame of interest and racial composition, with better predictive forecasts in the near-term and for states with relatively high proportions of non-Hispanic Blacks. Performance also depended on case rate, with better predictive forecasts for states with relatively low case rates but better probabilistic forecasts for states with relatively high case rates. Both predictive and probabilistic performance are important, and both deserve consideration by model developers and those interested in using these models to inform policy.

Significance Statement

Coronavirus disease 2019 (COVID-19) forecasting models can provide critical information for decision-making; however, there has been little published information on their performance. We examined the COVID-19 mortality forecasting performance of nine commonly-used and oft-cited models, as well as density-averaged equal- and performance-weighted ensembles of these models, during two phases of the pandemic. Only two of the nine models demonstrated superior predictive and probabilistic performance during both phases of the pandemic. Most of the models exhibited overconfidence, with overly-narrow forecast intervals. Performance-weighted ensembles demonstrated some advantages over the equal-weighted ensemble, primarily in probabilistic performance. As might be anticipated, predictions of near term mortality (next week) were better than predictions of mortality four weeks later.

Main Text

1. Introduction

Effective non-pharmaceutical interventions (NPIs), or community mitigation strategies, are crucial in combating the spread of contagious illnesses like coronavirus disease 2019 (COVID-19) (1). Community NPIs, such as social distancing guidelines, restrictions, closures, and lockdowns, can effectively delay and diminish an epidemic peak (1-3), also known as flattening the epidemic curve (2). However, these NPIs can also have immediate educational and economic consequences (4-6). To make decisions on the implementation of community NPIs amidst the COVID-19 pandemic, the relevant stakeholders (e.g. government officials, community leaders, school administrators etc.) may desire estimates of the number of coronavirus cases, hospitalizations, and deaths likely to occur in their region of interest within the coming weeks.

Information about, and forecasts from, dozens of COVID-19 forecasting models have been made available via the University of Massachusetts Amherst Reich Lab's COVID-19 Forecast Hub (7). All of the modeling groups that have participated have provided predictions (or, central estimates), and most also quantitatively characterize the uncertainty in their predictions – often giving an interquartile range (25% LCL to 75% UCL) and a *90%* confidence interval (5% LCL to 95% UCL) for each prediction (8).

Unfortunately, little peer-reviewed information about the predictive accuracy (i.e., how close are predictions to the actual observed values) or precision (i.e., how close are predictions to each other) of the forecasting models is widely available. Thus, stakeholders lack a scientific basis for deciding which models to trust and how much confidence to place in the forecasts they provide. Quite recently, one study in pre-print has become available (9). It compared the median absolute percent error of eight models stratified across world region, month of model estimation, and weeks of extrapolation (9). Collectively, the eight models released in July over a twelve week forecasting range had a median average percent error of about *25%*, with errors tending to increase with longer forecasts and the best performing model varying by region (9). While this information is quite useful and provides a sense of the typical bias of model predictions, other aspects of model performance may also be of interest. Users may also care about the precision of predictions and about the modeler's ability to properly characterize the uncertainty in their predictions.

Some may wonder whether better forecasts might be obtained by averaging the forecasts of two or more individual models. The aforementioned Reich Lab has created one such 'ensemble' model (10), which has been prominently featured by the Centers for Disease Control and Prevention (CDC) (11). It involves an equal-weighted quantile averaged combination of individual model forecasts and is not performance-weighted (10). In the expert judgment literature it has been clearly demonstrated that performance-weighted combinations of expert opinion consistently outperform equally weighted combinations (12,13). Additionally, ensembles that average forecast quantiles have been shown to perform worse than those that average densities, producing overconfident results (i.e., tight forecast distributions that do not adequately capture the realization), and the results from averaging quantiles do not reflect the distributions provided by each included model (12, 14). These issues deserve consideration in the development of any ensemble forecasting models.

The analysis presented below compares model forecasts with subsequent observations using several measures that reflect predictive (i.e., accuracy and precision) and probabilistic (i.e., statistical accuracy and information) performance. We also construct three ensemble models, one equal-weighted and two performance-weighted, all of which average forecast densities, and compare their performance with each other and with the best performing individual models. Lastly, we explore whether the available models tend to provide better forecasts under certain circumstances (i.e., differing case rates, racial demographics, and forecast time periods/phases of the pandemic).

2. Results

2.1. Performance Criteria

Two aspects of performance were evaluated for each model/ensemble– (i) predictive performance (i.e., the accuracy and precision of the model/ensemble's central estimates, or predictions); and (ii) probabilistic performance (the modeler's ability to characterize the uncertainty in their predictions, reflected by their reported confidence intervals).

First, to evaluate the performance of the models' central estimates, each observation, O_j, was divided by the corresponding point prediction, $P_{m,,j}$, of model m, to obtain the accuracy ratio, $R_{m,j}$ – where j is an index reflecting the date, state, and time interval:

$$R_{m,j} = O_j / P_{m,j}$$

The distribution of the resulting accuracy ratios was then examined. For each model, m, the geometric mean ($GM_m = \exp(\text{mean}_m(\ln(O_j / P_{m,j}))$ taken over all calibration variables, j, and geometric standard deviation ($GSD_m = \exp(SD_m(\ln(O_j / P_{m,j}))$, also over all j, were computed and used as respective measures of the observed accuracy and precision of the model's central estimates.

Second, to evaluate the models' ability to characterize the uncertainty in their predictions, the probabilistic performance of each model was assessed using the Cooke's Classical Model (CM) method (15). This

method was initially designed for the evaluation of the performance of formally-elicited structured expert judgment (SEJ) – where an expert's ability to meaningfully characterize the uncertainty in his or her estimates is arguably as important as the predictions they provide (15) – and has been employed in many studies (16-18). We believe the CM can also be applied to the forecasts provided by the models, as the true observations are unknown at the time of forecasting and the modeling groups may serve as the 'expert' while their forecasts may serve as their 'judgments'.

The CM assesses 'statistical accuracy', or 'calibration', C, using Shannon's relative information statistic, I_s, which compares the assessed and observed probabilities of calibration variables falling within various inter-quantile ranges (15, 19). When assessing 5 quantiles, 5%, 25%, 50%, 75%, 95%, there are 6 inter quantile ranges for each model, m. The calibration score for each model, C_m, on each variable takes the following formula:

$$C_m = 1 - \chi^2 (2n * I_s), \text{ where } I_s = \Sigma_{k=1...6} (s_k * \ln(s_k / p_k))$$

Where χ^2 is the chi-square distribution function with five degrees of freedom, s_k is the percentage of the N realizations falling in the inter-quantile interval, k, for m, and p_k is the probability which should apply to k. As the divergence between stated and observed probabilities increases, I_s increases and C moves toward 0.

The CM assesses 'information', I, by comparing the width of the confidence intervals given by each model with the 'intrinsic range' of each calibration variable (15, 19). The intrinsic range for a variable is defined as the difference between the largest forecasted or observed value and the smallest forecasted or observed value (15, 19). This range is expanded slightly by multiplying it by a user-defined expansion factor, 1+ F, where F is typically a small fraction (for our data, the default 10% was used) (15, 19). Using this framework, the information score for each model, I_m, on each variable is defined as:

$$I_m = \Sigma_{k=1,...6} (p_k * \ln(p_k / r_k))$$

Where p_k are the probabilities provided by the model, m, and r_k are the probabilities from a uniform (or log-uniform) probability density function over the intrinsic range. Models which concentrate their forecasts in a narrow range will have high information scores.[*] From these two scores, together with the theory of proper scoring rules, the CM calculates performance weights as the product of calibration and information scores and then normalizes these so that they sum to 1 across all models (15, 19). See ***SI appendix***, **Notes 1** & **2**, for more details.

2.2. Individual Model Performance

The predictive and probabilistic performance results for each individual forecasting model are summarized in **Fig. 1** for both the summer 2020 and winter 2021 time periods. More quantitative detail on the performance results is provided in **Table S1** of the ***SI appendix***.

Focusing initially on the predictive performance of the models during the summer 2020 period, we see that – (i) all models, except model C, have a GM greater than 1, indicating systemic underprediction of COVID-19 mortality, (ii) typically model predictions have little bias (≤ 35%) with precision within a factor of 2, (iii) models A, D, and I have excellent accuracy (GM = 1.01, 1.11, and 1.09, respectively) and good precision (GSD = 1.57, 1.71, and 1.47, respectively), and (iv) models B, C, and E have substantial bias (> 0.35%) and models B, E, G, and H have far worse precision (> 2). When focusing on the probabilistic performance of models during this period, only three (or perhaps, four) of the nine models considered perform at all well – models A, D, F, and to a lesser extent, G. The models that have the highest information scores (models B, E, H, and I) all have extremely low calibration scores (<< 0.01), suggesting

[*] It is important to note that, while a precision score of 1 is optimal (i.e., all predictions are equally close to each other relative to the observation), the higher the information score the better (i.e., the tighter the forecast distribution, the more information the forecast provides).

‘overconfidence’ – i.e., that their stated confidence intervals are far too narrow while simultaneously poorly capturing the true value.

Looking at the predictive performance during the winter 2021 period, we see that – (i) models F, G, and H no longer provide forecasts, (ii) all six remaining models, except model E, have a GM less than 1, indicating systemic overprediction of mortality, (iii) the remaining models again have little bias (< 30%) with precision within a factor of 2, except model C, (iv) model D exhibits even better accuracy (GM = 0.99) and precision (GSD = 1.33) than it did in the summer 2020 period, and model E now has excellent accuracy (GM = 1.04) and better precision (GSD = 1.52), (v) model A has good precision (GSD = 1.53) but so-so accuracy (GM = 0.76), and (vi) the models generally have better accuracy and precision than the summer 2020 period. For the probabilistic model performance, models A and D exhibit the best calibration scores (0.17 and 0.21, respectively), and again models B, E, and I are informative but their calibration is incredibly poor (<<0.01). Interestingly, unlike accuracy and precision, model calibration and information scores generally seem to be worse during the winter period.

The random expert hypothesis was tested to assess whether the putative differences in performance between models is due to noise. This hypothesis has been rejected for both periods of assessment. More detail on this analysis can be found in ***SI appendix***, **Note 3**.

2.3. Ensemble Model Designs

It is possible that better and more robust estimates might be obtained by producing an ensemble based on weighted combinations of the forecasts provided by the individual models. Two approaches of potential interest are – (i) equal-weighted combinations, and (ii) performance-weighted combinations. In addition to a density-averaged equal-weighted ensemble, two density-averaged performance-weighted ensembles are considered, based on – (a) ‘predictive-performance’ weights and (b) CM weights.

For ‘predictive-performance’ weighting, each random variable from each individual model is weighted in inverse proportion to the model’s predictive variance ($\text{Weight}_m \propto 1 / \text{Var}(O_j / P_{m,j})$; i.e., inverse-variance weighting the distribution of ratios of observations to predictions for each individual model). This weighting method is based on, but is not to be confused with, the inverse-variance weighting method commonly used in meta-analysis (20).

The equal-weighted ensemble is established by equally-weighting the probability densities, not quantiles, given by the individual model forecasts. The predictive-performance-weighted ensemble and the CM-weighted ensemble both apply performance-weights to the probability densities of the individual model forecasts. The CM-weighted ensemble also uses a threshold significance level of 0.05 for its normalized CM-weights (calibration*information), to gain robustness without significant loss in performance (18, 19). More detail on why this cutoff was used is available in ***SI appendix***, **Note 2** and **Note 4**.

2.4. Ensemble Model Performance

The predictive and probabilistic performance results for each of the density-averaged ensemble models are summarized in **Table 1** for both the summer 2020 and winter 2021 time periods.

Both performance-weighted ensembles would be expected to outperform the equal-weighted ensemble (12, 13) and they do fulfill this expectation on almost all measures of performance, however, the differences in predictive performance are often small. The accuracy of the performance-based combinations is appreciably better than the equal-weighted combination in the summer 2020 data, but this advantage is not seen in the winter 2021 data. Very small differences in precision are observed in both data sets.

To explore whether these differences in accuracy between the summer 2020 and winter 2021 periods were due to the fact that three of the nine models included in the summer 2020 analysis were not reflected in the winter 2021 data, we reassessed model performance in the summer of 2020 using only data from the six models that were available during both seasons. The differences in accuracy seen in the

summer of 2020 between the equal-weighted and performance-weighted combinations were even greater (equal weight = 1.23, predictive weight = 0.94, CM weight = 1.02) than those observed in the full data set. More detail on this analysis is provided in ***SI appendix***, **Table S2**.

There are substantial differences in probabilistic performance. The calibration scores of the CM-weighted ensemble (0.54 in summer 2020; 0.44 in winter 2021) are far better than those of the equal-weighted ensemble (0.03 during the summer, 0.17 during the winter) or the predictive-performance-weighted ensemble (0.04 during the summer, 0.25 during the winter). There are also small differences in information scores, where again the CM-weighted ensemble slightly outperforms the other two ensembles – resulting in substantially higher CM weights during both periods.

2.5. Performance by Domain

It is also interesting to compare the performance of the individual models across three domains of interest – (i) race (i.e., states which are heavily non-Hispanic White vs. states with relatively large non-Hispanic Black populations), (ii) COVID-19 case rates (i.e., states with a relatively low amount of weekly cases per 100,000 population vs. states with a relatively high amount of weekly cases per 100,000 population), and (iii) forecast period (i.e., forecasts of mortality for the upcoming week vs. forecasts of mortality for the week ending four weeks from the date on which the forecast was made). **Table 2** evaluates the performance of the equal-weighted density-averaged ensemble stratified by these three different domains (i.e., eight forecasts per domain).

Both the accuracy and precision of predictions for states with relatively high non-Hispanic Black populations are better than for states with a largely non-Hispanic White population. This difference appears to be stable and is seen in both the summer 2020 and winter 2021 data. However, there is no evidence of any stable differences in the calibration and information scores of forecasts depending on the racial composition of the state.

The summer 2020 data suggests that predictions are more accurate and precise in states with low case rates, and this remains true in the more recent winter 2021 data, but the difference in accuracy is not as substantial. On the other hand, it would appear that model calibration scores are somewhat better in states with high case rates. This difference appears to be stable and is seen with both the summer 2020 and winter 2021 data.

Lastly, as expected, the predictions of deaths in the near-term appear to be more accurate and precise than those in the mid-term, especially for the summer 2020 data. This remains true for the more recent winter 2020 data, but the differences in both accuracy and precision are smaller. There is some evidence that calibration scores are better for near-term forecasts in the summer data, but this does not persist in the winter data.

3. Discussion

The results suggest that, when considering both predictive (i.e., accuracy and precision) and probabilistic (i.e. statistical accuracy and information) forecast performance during both independent eight-week time periods (i.e., summer 2020 and winter 2021), two forecasting models, A and D, clearly outperform the other seven models, with model D demonstrating the most consistent dominance in all performance measures.

The fact that all but three of the nine models considered (models A, D, and F) have such low calibration scores for both time periods and that most of those also have high information scores indicates pervasive overconfidence. It is understandably difficult to predict changes in human behavior, thus, it is clear from these results that many modelers need to re-evaluate how they are quantifying uncertainty in their predictions.

We found that the tendency of models to underpredict in the summer of 2020 was replaced by a tendency to overpredict in the winter of 2021 – suggesting that perhaps, in general, model predictions lag the actual changes in disease rates, which were increasing in the summer of 2020 and decreasing in the winter of 2021.

All of the assessed individual models are variations of a susceptible-exposed-infectious-removed (SEIR) compartmental model (21-29). They all forecast at the U.S. state level, and some also forecast at the national and county levels (8). Every model provides forecasts at the daily scale, which may then be aggregated to the weekly scale for at least four weeks ahead from forecast date (8).

However, while all models use mortality data to inform compartment transition rates, they varied in how they incorporated case, hospitalization, demographic, and mobility data (21-29). Both models A and D utilize case and mortality data, and model A utilizes demographic data, but neither model incorporates hospitalization or mobility data (21, 24). Using case and mortality data, models A and D extrapolate deaths from lagged case counts, with a 14-day lag for model A (21) and a varying 15- to 30-day lag for model D (24). Additionally, while models A and D, like most other models, incorporate time-varying COVID-19 transmission rates which may indirectly reflect social distancing measures, neither model directly incorporates changes in social distancing measures for forecasts in the near-term (although, model A may incorporate changes in social distancing measures for forecasts past three weeks) (21, 24).

The density-averaged performance-weighted ensembles outperform the density-averaged equal-weighted ensemble, but the differences in predictive performance are modest. In contrast, when evaluated in terms of probabilistic performance there are clear advantages to performance weighting using the CM weights.

Across the three domains and two performance attributes considered, consistent and substantial performance differences were seen only in a few instances – and these differed depending on whether one focuses on predictive or probabilistic performance. Predictive performance is generally better for forecasts in states with relatively high non-Hispanic Black populations and relatively low case rates, but the observed differences are often modest. Probabilistic performance is consistently better when case rates are high. Finally, as might be expected, near-term predictions (i.e., deaths in the next week) outperform mid-term predictions (i.e., deaths in the week ending four weeks from forecast date).

Our sense is that these results are more suggestive than conclusive, because – (i) due to the selection criteria, only nine models are examined; (ii) they are based on model performance in one eight-week period in the summer of 2020 and one eight-week period in the winter of 2021; (iii) they are limited to eight states; (iv) they consider only forecasts of mortality and no other outcomes, such as cases or hospitalizations, that may be of interest for decision-makers; (v) our analysis does not attempt to explain what model design choices lead to better or worse performance.

Conversely, our analysis has several strengths – (i) it evaluates the performance of a set of leading models which have been used to forecast COVID-19 mortality in the US; (ii) it considers both predictive (i.e., accuracy and precision) and probabilistic (i.e., statistical accuracy and information) performance; (iii) it assesses several approaches for constructing ensemble models; (iv) it examines performance differences across states selected to reflect differences in racial composition and variations in COVID-19 case rate; and (v) the analysis considers data from two distinct phases of the pandemic.

4. Materials and Methods

Our analysis involves a comparison of model forecasts with subsequent observations of weekly deaths from COVID-19 in four states (Idaho, Louisiana, New York and Maine) over an eight-week period during the summer of 2020 and four other states (Georgia, Maryland, Vermont, and Wyoming) over a subsequent eight-week period during the winter of 2021.

The states considered in our analysis were selected on the basis of recent case rates of COVID-19 (cases/100,000 population within the previous week) (30) and racial composition (majority non-Hispanic Black vs. majority non-Hispanic White) (31). Racial composition was of interest as the COVID-19 mortality rate for non-Hispanic Black Americans at the time of analysis was more than twice that of non-Hispanic White Americans (32). With these two domains in mind, our goal was to assess two states for each time period with relatively high case rates (Idaho and Louisiana; Georgia and Vermont); two with relatively low case rates (Maine and New York; Maryland and Wyoming); two with a relatively high fraction of population reported as non-Hispanic Black (Louisiana and New York; Georgia and Maryland); and two with a relatively high fraction of population reported as non-Hispanic White (Idaho and Maine; Vermont and Wyoming). This was done to assess how models perform forecasting for states under varying circumstances. More detail on how the case rates and racial composition for states were determined, as well as how states were selected, is available in **SI appendix**, **Note 5** and **Tables S3** & **S4**.

We were interested in the models' ability to forecast COVID-19 deaths in both the near-term and the medium-term. Near-term performance was gauged using projected COVID-19 deaths in the week immediately after the forecast was made. Medium-term performance was gauged using projected COVID-19 deaths in the week ending four weeks after the forecast was made.

Our evaluations of model performance for the eight states and the two forecast periods of interest (week ending one week in the future and week ending four weeks in the future) were examined twice within each time period – once for forecasts made on June 13th, 2020, or January 10th, 2021, and a second time for forecasts made on July 11th, 2020, or February 7th, 2021 (no overlap in forecasts). In total, 16 comparisons of model forecasts with observed deaths were made for each time period and each model.

Of the many models providing data to the COVID-19 Forecast Hub's data repository (8), only the models that provided weekly COVID-19 mortality forecasts made on the same day of the week, with forecasts at the weekly time-scale for all four of the initially assessed states and were without missing forecasts were selected for this analysis. These include: OliverWyman-Navigator (Model A) (21), MOBS-GLEAM_COVID (Model B) (22), JHU_IDD-CovidSP (Model C) (23), UMass-MechBayes (Model D) (24), UCLA-SuEIR (Model E) (25), YYG-ParamSearch (Model F) (26), UT-Mobility (Model G) (27), USACE-ERDC_SEIR (Model H) (28), and Covid19Sim-Simulator (Model I) (29).

Code Availability

Data was analyzed using Microsoft Excel and EXCALIBUR (a software package for using Cooke's Classical Model) (33).

Data Availability

Observed state COVID-19 mortality and case data was gathered from the Centers for Disease Control and Prevention (CDC) (30). State population and racial composition data was collected from one-year estimates from the Census Bureau's 2018 and 2019 American Community Survey (ACS) (31). **Tables S3** & **S4** in the ***SI appendix*** provide the racial composition statistics and case rate data.

Model forecasting data was gathered from the COVID-19 Forecast Hub's publicly available structured data storage repository on GitHub (8).**Tables S5** & **S6** in the ***SI appendix*** provide the model and ensemble predictions, their uncertainty distributions, and the subsequent observations of COVID-19 mortality.

ACKNOWLEDGMENTS

We want to thank Willy Aspinall, Jouni Tuomisto, and Jacqueline Macdonald for contributing to our thoughts about this and for their feedback on our early drafts of the paper.

Funding Sources

Kyle J. Colonna's involvement was funded by the Harvard Population Health Sciences PhD scholarship. Roger M. Cooke's involvement was pro bono. John S. Evans' involvement was funded by the Department of Environmental Health and the Harvard Cyprus Initiative at the T.F. Chan School of Public Health.

References

1. Centers for Disease Control and Prevention, Nonpharmaceutical Interventions (NPIs). *Centers for Disease Control and Prevention* (2020). https://www.cdc.gov/nonpharmaceutical-interventions/index.html (accessed 3 May 2021).
2. R. M. Anderson, H. Heesterbeek, D. Klinkenberg, T. D. Hollingsworth, How will country-based mitigation measures influence the course of the COVID-19 epidemic? *Lancet* **395**, 931–934 (2020).
3. S. Flaxman, *et al.*, Estimating the effects of non-pharmaceutical interventions on COVID-19 in Europe. *Nature* **584**, 257–261 (2020).
4. International Monetary Fund Research Dept., World Economic Outlook, April 2020 : The Great Lockdown. *International Monetary Fund* (2020) (accessed 3 May 2021).
5. The United Nations Educational, Scientific and Cultural Organization, Adverse consequences of school closures. *The United Nations Educational, Scientific and Cultural Organization* (2020). https://en.unesco.org/covid19/educationresponse/consequences (accessed 3 May 2021).
6. M. Nicola, *et al.*, The socio-economic implications of the coronavirus pandemic (COVID-19): A review. *Int J Surg* **78**, 185–193 (2020).
7. University of Massachusetts, Reich Lab, The COVID-19 Forecast Hub. *University of Massachusetts* (2021). *https://covid19forecasthub.org/* (accessed 3 May 2021).
8. University of Massachusetts, Reich Lab, Data from "reichlab/covid19-forecast-hub." Github. Available at https://github.com/reichlab/covid19-forecast-hub. Deposited 3 May 2021.
9. J. Friedman, *et al.*, Predictive performance of international COVID-19 mortality forecasting models. *medRxiv* [Preprint] (2020). https://doi.org/10.1101/2020.07.13.20151233 (accessed 3 May 2021)
10. E. L. Ray, *et al.*, Ensemble Forecasts of Coronavirus Disease 2019 (COVID-19) in the U.S. *medRxiv* [Preprint] (2020). https://doi.org/10.1101/2020.08.19.20177493 (accessed 3 May 2021)
11. Centers for Disease Control and Prevention, Forecasts of COVID-19 Deaths. *Centers for Disease Control and Prevention* (2020). https://www.cdc.gov/coronavirus/2019-ncov/covid-data/forecasting-us.html (accessed 3 May 2021).
12. A. R. Colson, R. M. Cooke, Cross validation for the classical model of structured expert judgment. *Reliability Engineering & System Safety* **163**, 109–120 (2017).
13. A. R. Colson, R. M. Cooke, Expert Elicitation: Using the Classical Model to Validate Experts' Judgments. *Rev Environ Econ Policy* **12**, 113–132 (2018).
14. J. Bamber, W. Aspinall, R. Cooke, A commentary on "how to interpret expert judgment assessments of twenty-first century sea-level rise" by Hylke de Vries and Roderik SW van de Wal. *Climatic Change* **137**, 321–328 (2016).
15. R. M. Cooke, *Experts in Uncertainty: Opinion and Subjective Probability in Science* (Oxford University Press, 1991).
16. R. M. Cooke, *et al.*, A Probabilistic Characterization of the Relationship between Fine Particulate Matter and Mortality: Elicitation of European Experts. *Environ. Sci. Technol.* **41**, 6598–6605 (2007).
17. T. Hald, *et al.*, World Health Organization Estimates of the Relative Contributions of Food to the Burden of Disease Due to Selected Foodborne Hazards: A Structured Expert Elicitation. *PLoS One* **11** (2016).
18. J. L. Bamber, M. Oppenheimer, R. E. Kopp, W. P. Aspinall, R. M. Cooke, Ice sheet contributions to future sea-level rise from structured expert judgment. *PNAS* **116**, 11195–11200 (2019).
19. R. Cooke, L. Goossens, Procedures Guide for Structured Expert Judgment. *European Communities, Luxembourg, EUR* (2000).
20. J. Hartung, G. Knapp, B. K. Sinha, *Statistical Meta-Analysis with Applications* (Wiley, 2008).

21. Oliver Wyman, Oliver Wyman COVID-19 Pandemic Navigator. *Oliver Wyman* (2021). https://pandemicnavigator.oliverwyman.com/ (accessed 3 May 2021).
22. Laboratory fo the Modeling of Biological + Socio-Technical Systems, COVID-19 Mobility. *The Gleam Project* (2021). https://covid19.gleamproject.org/mobility (accessed 3 May 2021).
23. Infectious Disease Dynamics, Projects | COVID-19. *Johns Hopkins University* (2021). http://www.iddynamics.jhsph.edu/projects/covid-19 (accessed 3 May 2021).
24. D. Sheldon, C. Gibson, N. Reich, Data from "*dsheldon/covid*." Github. Available at https://github.com/dsheldon/covid . Deposited on 12 April 2021.
25. Statistical Machine Learning Lab, UCLAML Combating COVID-19. *University of California, Los Angeles* (2021). https://covid19.uclaml.org/index.html (accessed 3 May 2021).
26. Y. Gu, COVID-19 Projections Using Machine Learning. *Youyang Gu* (2021). https://covid19-projections.com/ (accessed on 3 May 2021).
27. COVID-19 Modeling Consortium, COVID-19 Mortality Projections for US States. *The University of Texas at Austin* (2021). https://covid-19.tacc.utexas.edu/dashboards/us/ (accessed 3 May 2021).
28. Engineer Research and Development Center, Data from "*erdc-cv19/seir-model*." Github. Available at https://github.com/erdc-cv19/seir-model. Deposited on 3 August 2020.
29. Massachusetts General Hospital Institute for Technology Assessment, COVID-19 Simulator. *Massachusetts General Hospital Institute for Technology Assessment* (2021). https://covid19sim.org/ (accessed 3 May 2021).
30. Centers for Disease Control and Prevention, COVID-19 Response, COVID-19 Case Surveillance Public Data Access, Summary, and Limitations. *Centers for Disease Control and Prevention* (2021). https://www.census.gov/programs-surveys/acs/data.html (accessed 3 May 2021).
31. U.S. Census Bureau, American Community Survey Data. *The United States Census Bureau* (2021). https://data.cdc.gov/Case-Surveillance/United-States-COVID-19-Cases-and-Deaths-by-State-o/9mfq-cb36 (accessed 3 May 2021).
32. Centers for Disease Control and Prevention, Hospitalization and Death by Race/Ethnicity. *Centers for Disease Control and Prevention* (2020). https://www.cdc.gov/coronavirus/2019-ncov/covid-data/investigations-discovery/hospitalization-death-by-race-ethnicity.html (accessed 18 August 2020).
33. LightTwist Software, Excalibur. *LightTwist Software.* https://lighttwist-software.com/excalibur/ (accessed 3 May 2021).

Figures and Tables

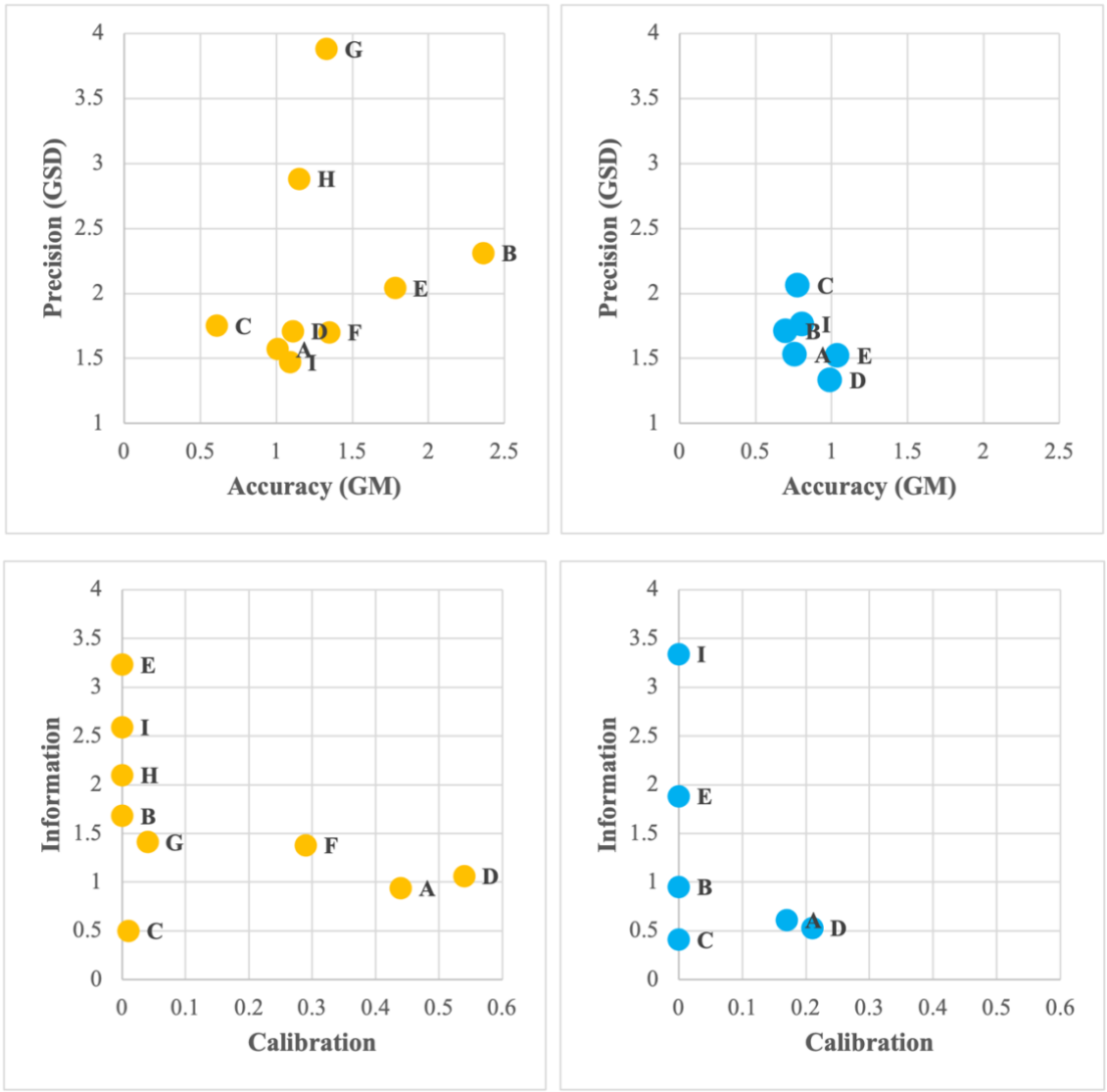

Fig. 1. Accuracy (GM) vs. precision (GSD) of the central estimates provided by the individual models for summer 2020 (*Top Left*) and winter 2021 (*Top Right*) periods. Calibration vs. information scores based on the forecast densities provided by the individual models for the summer 2020 (*Bottom Left*) and winter 2021 (*Bottom Right*) periods.

Table 1. Ensemble model performance for summer 2020 (*Left*) and winter 2021 (*Right*)

Ensemble	Summer 2020					Winter 2021				
	Accuracy - GM	Precision - GSD	Calibration	Information	CM Weight (Unnormalized)	Accuracy - GM	Precision - GSD	Calibration	Information	CM Weight (Unnormalized)
Equal Weight	1.24	1.60	0.03	0.72	0.02	0.84	1.55	0.17	0.36	0.06
Performance – Predictive Weight	0.91	1.41	0.04	0.55	0.02	0.84	1.49	0.25	0.35	0.09
Performance – CM Weight†	1.12	1.57	0.54	0.81	0.43	0.85	1.45	0.44	0.43	0.19

† A hypothesis rejection significance level of 0.05 was used for this ensemble. More detail on why this was used is available in ***SI appendix***, **Note 2** and **Note 4**.

Table 2: Equal-weighted ensemble model performance for summer 2020 (*Left*) and winter 2021 (*Right*) stratified by different domains

Sub-Domain	Summer 2020					Winter 2021				
	Accuracy - GM	Precision – GSD	Calibration	Information	CM Weight (Unnormalized)	Accuracy - GM	Precision – GSD	Calibration	Information	CM Weight (Unnormalized)
High % non-Hispanic White (ID, ME; VT, WY)	1.30	1.86	0.60	0.86	0.52	0.69	1.56	0.14	0.36	0.05
High % non-Hispanic Black (LA, NY; GA, MD)	1.19	1.34	0.05	0.59	0.03	1.02	1.42	0.73	0.37	0.27
High Case Rate (ID, LA; VT, GA)	1.42	1.77	0.32	0.67	0.21	0.81	1.76	0.66	0.41	0.27
Low Case Rate (ME, NY; WY. MD)	1.09	1.39	0.14	0.78	0.11	0.87	1.34	0.16	0.31	0.05
One Week Ahead	0.95	1.41	0.27	0.53	0.14	0.88	1.51	0.02	0.35	0.01
Four Weeks Ahead	1.62	1.56	0.05	0.92	0.05	0.80	1.62	0.06	0.37	0.02

Note: Analogues of **Table 2** for the CM- and predictive-performance-weighted ensembles are also provided in ***SI appendix***, **Table S7** & **S8**, respectively.