Evaluation of individual and ensemble probabilistic forecasts of COVID-19 mortality in the United States

Short-term probabilistic forecasts of the trajectory of the COVID-19 pandemic in the United States have served as a visible and important communication channel between the scientific modeling community and both the general public and decision-makers. Forecasting models provide specific, quantitative, and evaluable predictions that inform short-term decisions such as healthcare staffing needs, school closures, and allocation of medical supplies. Starting in April 2020, the US COVID-19 Forecast Hub (https://covid19forecasthub.org/) collected, disseminated, and synthesized tens of millions of specific predictions from more than 90 different academic, industry, and independent research groups. A multimodel ensemble forecast that combined predictions from dozens of groups every week provided the most consistently accurate probabilistic forecasts of incident deaths due to COVID-19 at the state and national level from April 2020 through October 2021. The performance of 27 individual models that submitted complete forecasts of COVID-19 deaths consistently throughout this year showed high variability in forecast skill across time, geospatial units, and forecast horizons. Two-thirds of the models evaluated showed better accuracy than a naïve baseline model. Forecast accuracy degraded as models made predictions further into the future, with probabilistic error at a 20-wk horizon three to five times larger than when predicting at a 1-wk horizon. This project underscores the role that collaboration and active coordination between governmental public-health agencies, academic modeling teams, and industry partners can play in developing modern modeling capabilities to support local, state, and federal response to outbreaks.

forecasting | COVID-19 | ensemble forecast | model evaluation

Effective pandemic response requires federal, state, and local leaders to make timely decisions in order to reduce disease transmission. During the COVID-19 pandemic, surveillance data on the number of cases, hospitalizations, and disease-associated deaths were used to inform response policies (1, 2). While these data provide insight into recent trends in the outbreak, they only present a partial, time-lagged picture of transmission and do not show if and when changes may occur in the future.

Anticipating outbreak change is critical for optimal resource allocation and response. Forecasting models provide quantitative, evaluable, and probabilistic predictions about the epidemic trajectory for the near-term future. Forecasts can inform operational decisions about allocation of healthcare supplies (e.g., personal protective equipment, therapeutics, and vaccines), staffing needs, and school closures (3). Providing prediction uncertainty is critical for such decisions, as it allows stakeholders to assess the most likely outcomes and plausible worst-case scenarios (3).

Academic research groups, government agencies, industry teams, and individuals produced COVID-19 forecasts at an unprecedented scale starting in March 2020. Publicly available forecasts reflect varied approaches, data sources, and assumptions. Some models had mechanisms that allowed them to incorporate an estimated impact of current or potential future policies on human behavior and COVID-19 transmission. Other models assumed that currently observed trends would continue into the future without considering external data on policies in different jurisdictions.

To leverage these forecasts for the COVID-19 response, the US Centers for Disease Control and Prevention (CDC) partnered with the Reich Laboratory at the University of Massachusetts Amherst to create the COVID-19 Forecast Hub (https://covid19forecasthub.org/) (4). Launched in early April 2020, the Forecast Hub facilitated the collection, archiving, evaluation, and synthesis of forecasts. Teams were explicitly asked to submit "unconditional" forecasts of the future, in other words, predictions that integrate across all possible changes in future dynamics. In practice, most individual models made predictions that were conditional on explicit or implicit assumptions of how policies, behaviors, and pathogens would evolve in the coming weeks. From these forecasts, a multimodel ensemble was developed, published weekly in real time, and used by the CDC in official public communications about the pandemic (https://www.cdc.gov/coronavirus/2019-ncov/covid-data/mathematical-modeling.html)

Significance

This paper compares the probabilistic accuracy of short-term forecasts of reported deaths due to COVID-19 during the first year and a half of the pandemic in the United States. Results show high variation in accuracy between and within stand-alone models and more consistent accuracy from an ensemble model that combined forecasts from all eligible models. This demonstrates that an ensemble model provided a reliable and comparatively accurate means of forecasting deaths during the COVID-19 pandemic that exceeded the performance of all of the models that contributed to it. This work strengthens the evidence base for synthesizing multiple models to support public-health action.

Author contributions: E.Y.C., E.L.R., V.K.L., J. Bracher, T.G., K.H.H., Y.H., K.L., J.N., J. White, R.B.S., M.A.J., M. Biggerstaff, and N.G.R. designed research; E.Y.C., E.L.R., J. Bracher, A.J.C.R., A.G., K.H.H., Y.H., D.J., A.H.K., A.K., K.L., A.M., J.N., A. Shah, A. Stark, Y. Wang, N.W., M.W.Z., and N.G.R. performed research; J. Bracher, A. Brennen, T.G., Y.G., S.J., N.B., A.D., M. Kulkarni, S. Merugu, A.R., S. Shingi, A. Tiwari, J. White, N.F.A., S. Woody, M. Dahan, S. Fox, K.G., M.L., L.A.M., J.G.S., M.T., A. Srivastava, G.E.G., J.C.C., I.D.D., W.P.E., M.W.F., R.H.H., B.L., I.L., M.L.M., M.D.P., M.A.R., B.D.T., Y.Z.-J., S. Chen, S.V.F., J. Hess, C.P.M., A. Salekin, D. Wang, S.M.C., T.M.B., M.C.E., K.F., Y.H., E.T.M., E.M., R.L.M., T.S., D. Sheldon, G.C.G., R.Y., Liyao Gao, Y.M., D. Wu, X.Y., X.J., Y.-X.W., Y.C., Lihong Guo, Y.Z., Q.G., J. Chen, Lingxiao Wang, P.X., W.Z., D. Zou, H.B., J. Lega, S. McConnell, V.P.N., S.L.G., C.H.-L., S.D.T., Y.S., X.B., R. Walraven, Q.-J.H., S.K., A. van de Walle, J.A.T., M.B.-N., S. Riley, P.R., U.K., D.D., P.F., B.H., C.K., H.L., J. Milliken, M. Moloney, J. Morgan, N.N., G.O., N. Piwonka, M. Ravi, C. Schrader, E.S., D. Siegel, R. Spatz, C. Stiefeling, B.W., A.W., S. Cavany, G.E., S. Moore, R.O., A. Perkins, D. Kraus, A.K., Z. Gao, J. Bian, W.C., J.L.F., C. Li, T.-Y.L., X. Xie, S. Zhang, S. Zheng, A.V., M.C., J.T.D., K.M., A.P., X. Xiong, A.Z., J. Baek, V.F., A. Georgescu, R.L., D. Sinha, J. Wilde, G.P., M.A.B., D.N.-N., D. Singhvi, I.S., L.T., A. Tsiourvas, A. Sarker, A.J., D. Shah, N.D.P., L.A.C., S. Sundar, R.W., D.O., L.C., G.F., I.M., D.K., M. Kinsey, L.C.M., K.R.-L., L.S., K.T., S. Wilson, E.C.L., J. Dent, K.H.G., A.L.H., J.K., K.K., L.T.K., S.A.L., J.C.L., J.L., H.R.M., J.P.-S., S. Shah, C.P.S., S.A.T., J. Wills, M. Marshall, L. Gardner, K.N., J.C.B., Lily Wang, L.G., Z. Gu, M. Kim, X.L., G.W., Y. Wang, S.Y., R.C.R., R.B., E.G., S.I.H., S.L., C.M., D.P., H.L.G., P.B., S.A.S., B.T.S., B.A.P., B.A., J. Cui, A. Rodríguez, A.T., J. Xie, P.K., J.A., A. Baxter, B.E.O., N. Serban, S.O.A., M. Dusenberry, A.E., E.K., L.T.L., C.-L.L., T.P., D. Sava, R.S., T.T., N.Y., J.Y., L.Z., S.A., N.I.B., S. Funk, J. Hellewell, S.R.M., K.S., M.Z., R.K., T.K.Y., S.P., J. Shaman, M.L.L., D.B., O.S.L., S. Soni, H.T.B., T.A., M.A., J. Chhatwal, O.O.D., M.A.L., B.P.L., P.M., J. Xiao, Y. Wang, Q.W., S.X., D. Zeng, A. Green, J. Bien, L.B., A.J.H., M.J., D.M., B.N., C.P., S. Rajanala, A. Rumack, N. Simon, R.J.T., R.T., V.V., L. Wasserman, E.B.O., J.M.D., R.P., Q.T.T., L.S.T.H., H.H., and N.G.R. contributed new reagents/analytic tools; E.Y.C., E.L.R., J. Bracher, A. Gerding, Y.H., K.L., and N.G.R. analyzed data; and E.Y.C., E.L.R., V.K.L., J. Bracher, and N.G.R. wrote the paper.

Competing interest statement: A.V., M.C., and A.P.P. report grants from Metabiota Inc. outside the submitted work.

This article is a PNAS Direct Submission.

The findings and conclusions in this report are those of the authors and do not necessarily represent the views of the Centers for Disease Control and Prevention.

[1] To whom correspondence may be addressed. Email: nick@umass.edu.

This article contains supporting information online at http://www.pnas.org/lookup/suppl/doi:10.1073/pnas.2113561119/-/DCSupplemental.

Published April 8, 2022.

STATISTICS

POPULATION BIOLOGY

https://doi.org/10.1073/pnas.2113561119

(5). Forecasts were generated for the outcomes of reported cases, hospitalizations, and deaths due to COVID-19. This paper focuses on evaluating forecasts of reported deaths.

Ensemble models incorporate the information and uncertainties from multiple forecasts, each with their own perspectives, strengths, and limitations, to create accurate predictions with well-calibrated uncertainty (6–11). Synthesizing multiple models removes the risk of overreliance on any single approach for accuracy or stability. It is challenging for individual models to make calibrated predictions of the future when the behavior of the system being studied is nonstationary due to continually changing policies and behaviors. Ensemble approaches have previously demonstrated superior performance compared with single models in forecasting influenza (12–14), Ebola (15), and dengue fever outbreaks (16). Preliminary research suggested that COVID-19 ensemble forecasts were also more accurate and precise than individual models in the early phases of the pandemic (17, 18).

Predicting the trajectory of a novel pathogen outbreak such as COVID-19 is subject to many challenges. These include the role of human behavior and decision-making in outbreak trajectories, and the fact that epidemic forecasts may play a role in a "feedback loop" when and if the forecasts themselves have the ability to impact future societal or individual decision-making (19). There are also a host of data irregularities, especially in the early stages of the pandemic.

It is important to systematically and rigorously evaluate forecasts designed to predict real-time changes to the outbreak in order to identify strengths and weaknesses of different approaches and to understand the extent to which the forecasts are a reliable input to public-health decisions. Knowledge of what leads to more or less accurate and well-calibrated forecasts can inform their development and their use within outbreak science and public policy. In this analysis, we sought to evaluate the accuracy of individual and ensemble probabilistic forecasts submitted to the Forecast Hub, focusing on forecasts of reported weekly incident deaths.

Results

Summary of Models. Forecasts evaluated in this analysis are based on submissions in a continuous 79-wk period starting in late April 2020 and ending in late October 2021 (Fig. 1 and *Methods*). Forecasts were evaluated at 55 locations including all 50 states, four jurisdictions and territories (Guam, US Virgin Islands, Puerto Rico, and the District of Columbia), and the US national level. The evaluation period captured the decline of the spring 2020 wave, a late summer 2020 increase in several locations, a large late-fall/early-winter surge in 2020/2021, and the rise and fall of the Delta variant in the summer and fall of 2021 (Fig. 1*B*).

The number of models that submitted forecasts of incident deaths to the Forecast Hub and were screened for inclusion in this analysis increased from four models at the beginning of the evaluation period to an average of 41.2 models per week during the first 10 months of 2021 (Fig. 1*C* and *SI Appendix*, Fig. S1). A total of 28 models met inclusion criteria, yielding 1,791 submission files with 556,050 specific predictions for unique combinations of forecast dates, targets (horizons forecasted), and locations.

The evaluated forecasts used different data sources and made varying assumptions about future transmission patterns (*SI Appendix*, Table S1). All evaluated models, other than CEID-Walk, the COVIDhub-baseline, the COVIDhub-ensemble, and PSI-Draft, used case data as inputs to their forecast models. A total of 10 models included data on COVID-19 hospitalizations, 10 models incorporated demographic data, and 9 models used mobility data. Of the 28 evaluated models, 7 made explicit assumptions that social distancing and other behavioral patterns would change over the prediction period. Two naıve models were included. The COVIDhub-baseline is a neutral model built with median predicted incidence equal to the number of reported deaths in the most recent week with uncertainty around the median based on weekly differences in previous observations (see *Methods*). CEID-Walk is a random walk model with simple outlier handling (*SI Appendix*, Table S1).

Overall Model Accuracy. To evaluate probabilistic accuracy, the primary metric used was the weighted interval score (WIS), a nonnegative metric, which measures how consistent a collection of prediction intervals is with an observed value (20). For WIS, a lower value represents smaller error (see *Methods* and *SI Appendix*, *SI Text*).

Led by the ensemble model, a majority of the evaluated models achieved better accuracy than the baseline model in forecasting incident deaths (Table 1). The COVIDhub-ensemble achieved a relative WIS of 0.61, which can be interpreted as achieving, on average, 39% less probabilistic error than the baseline forecast in the evaluation period, adjusting for the difficulty of the specific predictions made. An additional seven models achieved a relative WIS of less than or equal to 0.75. In total, 18 models had a relative WIS of less than 1, indicating lower probabilistic forecast error than the baseline model, and 10 models (including the baseline) had a relative WIS of 1 or greater (Table 1). Patterns in relative point forecast error were similar, with 18 models having equal or lower mean absolute error (MAE) than the baseline (Table 1). Values of relative WIS and rankings of models were robust to changing thresholds for submission inclusion criteria and to the inclusion or exclusion of individual outlying or revised observations (*SI Appendix*, Tables S3 and S4). When stratified by phase of the pandemic, different models showed the highest accuracy overall (*SI Appendix*, Fig. S5).

The degree to which individual models provided calibrated predictions varied (Table 1). We measured the probabilistic calibration of model forecasts using the empirical coverage rates of prediction intervals (PIs). Across 1- through 4-wk-ahead horizons, 79 wk, and 50 states, only the ensemble model achieved near-nominal coverage rates for both the 50% and 95% PIs. Eight models achieved coverage rates within 5% of the desired coverage level for the 50% PI, and only the COVIDhub-ensemble and UMass-MechBayes achieved coverage rates within 5% for the 95% PI. Typically, observed coverage rates were lower than the nominal rate (Table 1 and *SI Appendix*, Fig. S2). Three models had very low coverage rates (less than 50% for the 95% PI or less than 15% for the 50% PI). In general, models were penalized more for underpredicting the eventually observed values than overpredicting (*SI Appendix*, Fig. S7).

Among the top-performing models, there was variation in data sources used, indicating that the inclusion of additional data sources was not a sufficient condition for high accuracy. Of the seven top individual models with a relative WIS less than or equal to 0.75 (Table 1), four used data beyond the epidemiological hospitalization, case, and death surveillance data from the Center for System Science and Engineering (CSSE) (*SI Appendix*, Table S1). A total of 10 of the 18 individual models that performed better than the baseline used data other than epidemiological surveillance data (e.g., demographics or mobility). The top performers consisted of both models with mechanistic components and mostly phenomenological ones.

Model Accuracy Rankings Are Highly Variable. The COVIDhub-ensemble was the only model that ranked in the top half of all models (standardized rank > 0.5) for more than 75% of the observations it forecasted, although it made the single best forecast less frequently than any other model (Fig. 2). We ranked models based on relative WIS for each combination of

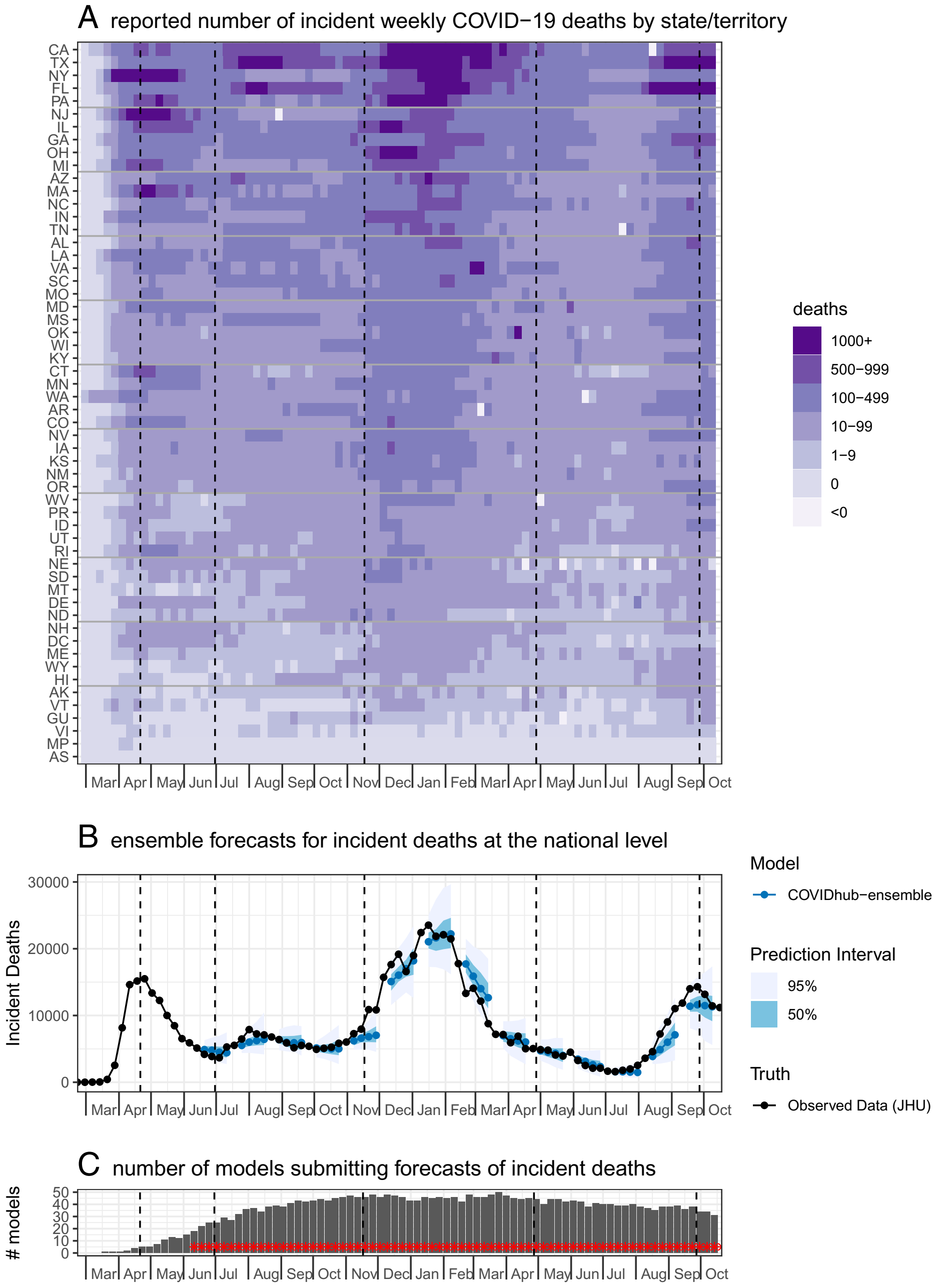

Fig. 1. Overview of the evaluation period included in the paper. Vertical dashed lines indicate "phases" of the pandemic analyzed separately in *SI Appendix*. (*A*) The reported number of incident weekly COVID-19 deaths by state or territory, per JHU CSSE reports. Locations are sorted by the cumulative number of deaths as of October 30th, 2021. (*B*) The time series of weekly incident deaths at the national level overlaid with example forecasts from the COVID-19 Forecast Hub ensemble model. (*C*) The number of models submitting forecasts for incident deaths each week. Weeks in which the ensemble was submitted are shown with a red asterisk.

1- through 4-wk-ahead horizons, 79 wk, and 55 locations, contributing to 17,006 possible predicted observations for each model (Fig. 2). All models showed large variability in relative skill, with each model having observations for which it had the best (lowest) WIS and thereby a standardized rank of 1. Some models, such as JHUAPL-BUCKY and PSI-DRAFT, show a bimodal distribution of standardized rank, with one mode in the top quartile of models and another in the bottom quartile. In these cases, the models frequently made overconfident predictions (*SI Appendix*, Fig. S6) resulting in either lower scores (indicating better performance) in instances in which their predictions were very close to the truth or higher scores (indicating worse performance) when their predictions were far from the truth. Similar patterns in ranking and relative model performance were seen when stratifying ranks by pandemic phase (*SI Appendix*, Fig. S3).

Observations on Accuracy in Specific Weeks. Forecasts from individual models showed variation in accuracy by forecast week and horizon (Fig. 3). The COVIDhub-ensemble model showed better average WIS than both the baseline model and the average error of all models across the entire evaluation period, except for 3 wk during which the baseline had lower 1-wk-ahead error than the ensemble. The COVIDhub-ensemble 1-wk-ahead forecast for EW02-2021 yielded its highest average WIS across all weeks (average WIS = 72.7), and 9 out of 26

Table 1. Summary accuracy metrics for all submitted forecasts from 28 models meeting inclusion criteria, aggregated across locations (50 states only), submission week, and 1- through 4-wk forecast horizons

Model	No. forecasts	95% PI Coverage	50% PI Coverage	Relative WIS	Relative MAE
BPagano-RtDriven	10,864	0.72	0.36	0.77	0.80
CEID-Walk	12,161	0.78	**0.45**	1.00	1.03
CMU-TimeSeries	10,456	0.77	0.42	0.78	0.80
Covid19Sim-Simulator	11,770	0.34	0.11	1.02	0.85
CovidAnalytics-DELPHI	11,064	0.82	**0.46**	0.99	1.01
COVIDhub-baseline	15,460	0.88	**0.51**	1.00	1.00
COVIDhub-ensemble	14,260	**0.90**	**0.53**	**0.61**	**0.66**
CU-select	13,710	0.72	0.43	0.92	0.89
DDS-NBDS	12,261	0.86	0.43	1.25	2.19
epiforecasts-ensemble1	12,204	0.87	**0.46**	3.17	2.74
GT-DeepCOVID	13,585	0.84	0.41	0.75	0.82
IHME-SEIR	11,116	0.59	0.25	0.79	0.82
JHU_CSSE-DECOM	10,190	0.80	0.35	0.75	0.80
JHU_IDD-CovidSP	14,170	0.82	0.33	0.99	1.04
JHUAPL-Bucky	11,664	0.63	0.29	1.05	1.06
Karlen-pypm	13,060	0.86	**0.47**	0.64	0.70
LANL-GrowthRate	13,560	0.83	0.38	0.85	0.91
MOBS-GLEAM_COVID	15,452	0.71	0.37	0.77	0.78
OliverWyman-Navigator	10,548	0.82	**0.45**	0.72	0.76
PSI-DRAFT	13,209	0.34	0.15	1.51	1.27
RobertWalraven-ESG	13,430	0.51	0.28	1.13	0.97
SteveMcConnell-CovidComplete	12,063	0.8	**0.45**	0.74	0.77
UA-EpiCovDA	13,710	0.72	0.41	0.98	0.94
UCLA-SuEIR	10,549	0.31	0.09	1.37	1.21
UCSD_NEU-DeepGLEAM	11,664	**0.91**	0.7	0.83	0.78
UMass-MechBayes	14,660	**0.93**	0.56	0.63	0.67
UMich-RidgeTfReg	11,394	0.63	0.34	1.18	1.08
USC-SI_kJalpha	9,660	0.52	0.22	0.75	0.72

The "No. forecasts" column refers to the number of individual location/target/week combinations. Empirical prediction interval (PI) coverage rates calculate the fraction of times the 50% or 95% PIs covered the eventually observed value. Values within 5% coverage of the nominal rates are highlighted in boldface text. The "relative WIS" and "relative MAE" columns show the relative mean WIS and relative MAE, which compare each model to the baseline model while adjusting for the difficulty of the forecasts the given model made for state-level forecasts (see *Methods*). The baseline model is defined to have a relative score of 1. Models with relative WIS or MAE values lower than 1 had "better" accuracy relative to the baseline model (best score in bold).

other models that submitted for the same locations outperformed it that week. The 4-wk-ahead COVIDhub-ensemble forecasts were worse in EW49-2020 than in any other week during the evaluation period (average WIS = 111.7), and 15 out of the 26 models outperformed the ensemble that week at a forecast horizon of 4 wk.

There was high variation among the individual models in their forecast accuracy during periods of increasing deaths and near peaks (i.e., forecast dates in July through early August of 2020, November through March, and August through October of 2021; Fig. 3). High errors in the baseline model tended to be associated with large outliers in observed data for a particular week (e.g., times when a state reported a large backfill of deaths in the most recent week) (*SI Appendix, SI Text*). In general, other models did not show unusual errors in their forecasts originating from these anomalous data, suggesting that their approaches, including possible adjustments to recent observations, were robust to anomalies in how data were reported.

Model Performance in Specific Pandemic Waves. In addition to evaluating performance in aggregate across the entire evaluation period and separate phases, we evaluated model performance during important moments during the pandemic. To assess the impact of rapidly changing trends on incident death forecasting accuracy, we ran an analysis restricted to specific locations and time periods that experienced high rates of change during four different waves of the pandemic (Fig. 4): 1) the summer 2020 waves in the south and southwest, 2) the late-fall 2020 rise in deaths in the upper Midwest, 3) the wave driven by the Alpha SARS-CoV-2 variant in Michigan in in March/April 2021, and 4) the Delta variant wave in summer 2021 throughout most states in the United States.

Forecast performance varied substantially in these examples. Models in general systematically underpredicted the mortality curve as trends were rising and overpredicted as trends were falling. In some of the selected waves (e.g., North Dakota and Florida), the ensemble forecast showed inappropriate levels of uncertainty, with the 95% PIs covering the eventual observations less than 80% of the time. However, during other waves (e.g., Louisiana and Michigan), the ensemble forecast, while systematically biased first below and then above the eventually reported counts of deaths, did cover the observations at or above 95% of the time, although PIs were very wide. In general, lower-than-expected coverage rates and bias were more pronounced at a 4-wk horizon than a 1-wk horizon. These four examples appeared to be representative of trends observed when looking across a larger number of waves (Dataset S2).

Individual Model Forecast Performance Varies Substantially by Location. Forecasts from individual models showed large variation in accuracy by location when aggregated across all weeks and targets (Fig. 5). Only the ensemble model showed superior accuracy when compared to baseline in all locations. Ensemble forecasts of incident deaths showed the largest relative accuracy improvements in New York, New Jersey, Indiana (relative WIS = 0.4), California, Massachusetts, and at the national level (relative WIS = 0.5) and the lowest relative accuracy in

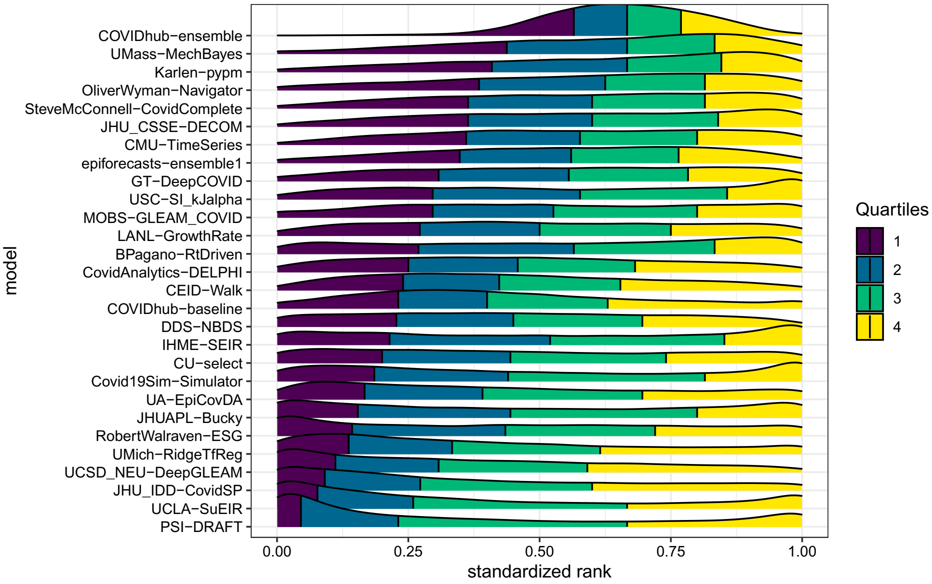

Fig. 2. A comparison of each model's distribution of standardized rank of WIS for each location/target/week observation. A standardized rank of 1 indicates that the model had the best WIS for that particular location, target, and week, and a value of 0 indicates it had the worst WIS. The density plots show interpolated distributions of the standardized ranks achieved by each model for every observation that model forecasted. The quartiles of each model's distribution of standardized ranks are shown in different colors: yellow indicates the top quarter of the distribution and purple indicates the bottom quarter of the distribution. The models are ordered by the first quartile of the distribution, with models that rarely had a low rank near the top.

Vermont, Guam, and The Virgin Islands (relative WIS = 0.9). The COVIDhub-ensemble was the only model to outperform the baseline in every location when eligible in a specific pandemic phase (*SI Appendix*, Fig. S6).

Forecast Performance Degrades with Increasing Horizons. Averaging across all states and weeks in the evaluation period, forecasts from all models showed lower accuracy and higher variance at a forecast horizon of 4 wk ahead compared to a horizon of 1 wk ahead; however, models generally showed improved performance relative to the naıve baseline model at larger horizons (*SI Appendix*, Fig. S4). A total of 11 models showed a lower average WIS (range: 24.9 to 34.3) than the baseline at a 1-wk horizon (average WIS = 35.8). At a 4-wk-ahead horizon, 19 models had a lower average WIS (range: 39.9 to 65.2) than baseline (average WIS = 70.1). Across all models except one, the average WIS was higher than the median WIS, indicative of outlying forecasts impacting the mean value.

When averaging across locations and stratifying by phase of the pandemic, there was variation in the top-performing models (*SI Appendix*, Fig. S5). Four models had a lower mean WIS than baseline for both 1- and 4-wk-ahead targets in at least three out of four phases (COVIDhub-ensemble, GT-DeepCOVID, Karlen-pypm, and UMass-MechBayes). Additionally, UMass-MechBayes and COVIDhub-ensemble were the only models to appear in the top three models in three of the four phases analyzed (*SI Appendix*, Fig. S5). In contrast to average WIS, PIs coverage rates did not change substantially across the 1- to 4-wk horizons for most models (*SI Appendix*, Fig. S2).

While many teams submitted only short-term (1- to 4-wk-ahead) forecasts, a smaller number of teams consistently submitted longer-term predictions with up to a 20-wk horizon for all 50 states (*SI Appendix*, Fig. S8). Across all teams submitting forecasts for the 50 states, 4-wk-ahead forecasts had around 76% more error (based on relative WIS) than 1-wk-ahead forecasts, a relationship that was consistent across the entire evaluation period. Longer-term forecasts showed less accuracy on average than 1- and 4-wk-ahead forecasts. There were no clear overall differences in probabilistic model accuracy between 8- and 20-wk horizons, although in early summer 2020, late spring 2021, and fall of 2021, average WIS at 8-wk horizons were slightly lower than at longer horizons (*SI Appendix*, Fig. S8*B*). For the two teams who made 20-wk-ahead forecasts for all 50 states, the average WIS was 2.9 to 4 times higher at a 20-wk horizon than it was at a 1-wk horizon. The increased WIS at longer prediction horizons for these models were due to larger dispersion (i.e., wider predictive distributions representing increased uncertainty) as well as larger penalties for underprediction and overprediction (*SI Appendix*, Fig. S9). The biggest increases in WIS were from increased penalties for underprediction, suggesting that the model forecasts did not accurately capture the possibility of increases in incidence at long horizons. Coverage rates for 95% PIs tended to be stable or decline as the horizon increased (*SI Appendix*, Fig. S8*C*).

Discussion

Given the highly visible role that forecasting has played in the response to the COVID-19 pandemic, it is critical that model consumers, such as decision-makers, the general public, and modelers themselves, understand how reliable models are. This paper provides a comprehensive and comparative look at the probabilistic accuracy of different modeling approaches for forecasting COVID-19–related deaths during the COVID-19 pandemic in the United States from April 2020 through October 2021. This work illustrates the tension between the desire for long-term forecasts, which would be helpful for public-health practitioners, and the decline in forecast accuracy at longer horizons shown by all forecasting methods.

As seen in prior epidemic forecasting projects, ensemble forecasts simplify the information provided to model consumers and can provide a stable, accurate, and low-variance forecast (3, 14–16). The results presented here show high variation in accuracy between and within stand-alone models but more consistent accuracy from an ensemble forecast. This supports prior results and confirms that an ensemble model can provide a reliable and comparatively accurate forecast that exceeds the performance of most, if not all, of the models that contribute to it.

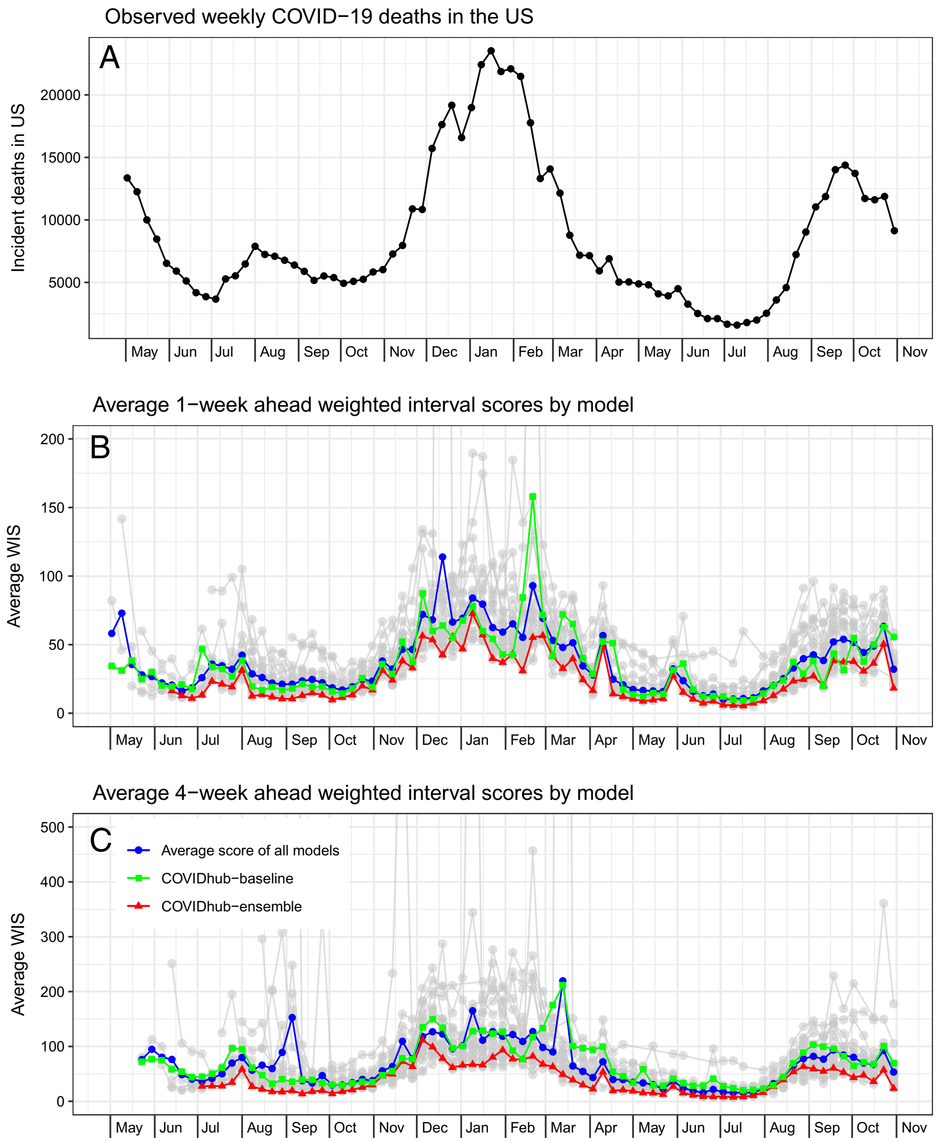

Fig. 3. Average WIS by the target forecasted week for each model across all 50 states. *A* shows the observed weekly COVID-19 deaths based on the CSSE-reported data as of May 25, 2021. *B* shows the average 1-wk-ahead WIS values per model (in gray). For all 21 wk in which the ensemble model (red triangle) is present, this model has lower WIS values than the baseline model (green square) and the average score of all models (blue circle). *C* shows the average 4-wk-ahead WIS values per model (in gray). For all 21 wk in which the ensemble model (red triangle) is present, this model has lower WIS values than the baseline model (green square) and the average score of all models (blue circle). The *y*-axes are truncated in *B* and *C* for readability of the majority of the data.

The ensemble approach was the only model that 1) outperformed the baseline forecast in every location, 2) had better overall 4-wk-ahead accuracy than the baseline forecast in every week, and 3) ranked in the top half of forecasts for more than 75% of the forecasts it made. It achieved the best overall measures of point and probabilistic forecast accuracy for forecasting deaths. However, during key moments in the pandemic, while the ensemble outperformed many models, it often showed lower than desired accuracy (Figs. 3 and 4). These results strengthen the evidence base for synthesizing multiple models for public-health decision support.

We summarize the key findings of the work as follows.

- The performance of all individual models forecasting COVID-19 mortality was highly variable even for short-term targets (Figs. 2 and 3). One source of variation was data inputs. Further investigation is needed to determine in what settings additional data can yield measurable improvements in forecast accuracy or add valuable diversity to a collection of models that are combined.
- A simple ensemble forecast that combined all submitted models each week was consistently the most accurate model when aggregating performance across forecast targets, weeks, or locations (Fig. 3 and 5 and *SI Appendix*, Fig. S4). Although rarely the “most accurate” model for individual predictions, the ensemble was consistently one of the top few models for any single prediction (Fig. 2). For public-health agencies concerned with using a model that shows dependably accurate performance, this is a desirable feature of a model.
- The high variation in ranks of models for each location/target/week suggests that all models, even those that are not as accurate on average, have observations for which they are the most accurate (Fig. 2).
- The post hoc evaluation of models during forecasting waves in select states showed poor accuracy of the ensemble model’s

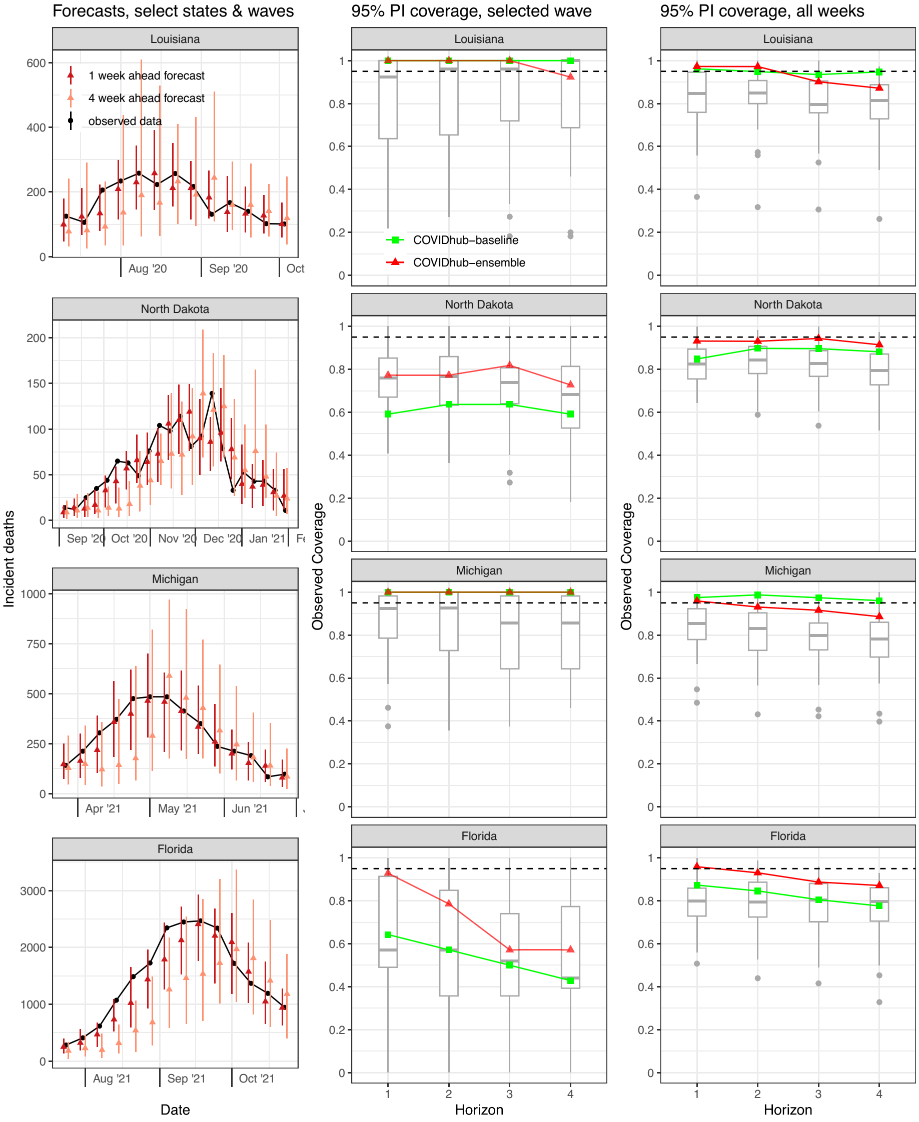

Fig. 4. Forecasts for selected states and pandemic waves, with PIs coverage. The first column shows every 1- and 4-wk-ahead forecast with 95% PIs made by the ensemble during the selected evaluation period. The second and third columns of plots show evaluations of PIs across 1- through 4-wk horizons (*x*-axis). The red line with triangle points corresponds to the coverage rates of the COVIDhub-ensemble forecasts, and green squares refer to the COVIDhub-baseline model. The boxplots represent the distribution of coverage rates from all component models. The second column evaluates only forecasts made for the dates shown in the first column. The third column evaluates forecasts across all weeks in the evaluation period. In the last two columns, the expected coverage rate (95%) is shown by the dashed line.

point forecasts for 1 and 4 wk ahead. During periods of increasing incident deaths, the ensemble tended to underpredict while tending to overpredict during periods of decreasing incident deaths. PIs coverage during these times varied (Fig. 4).

- Forecast accuracy and calibration were substantially degraded as forecast horizons increased, largely due to underestimating the possibility of increases in incidence at long horizons (*SI Appendix*, Figs. S8 and S9).

Model performance should be assessed both in aggregate (to compare models that showed the best overall performance) and in specific important moments during the pandemic. It is of public-health interest to evaluate how well models are able to predict points at which the observed trends change. However, we note that a post hoc evaluation that focuses only on times at which a specific type of trend was observed raises conceptual challenges. Extreme turning points in the pandemic are relatively rare compared with the many weeks during which trends continue or only slightly change from previous weeks. A post hoc evaluation that focuses exclusively on these "change-points" may reward models that may regularly predict extreme changes even when they do not occur at other times (21). Adapting proper scoring rules to weigh good performance in both kinds of situations is difficult.

Rigorous evaluation of forecast accuracy faces many limitations in practice. The large variation and correlation in forecast errors across targets, submission weeks, and locations makes it difficult to create rigorous comparisons of models. Forecast comparison is also challenging because teams have submitted forecasts for different lengths of time, different locations, and for different numbers of horizons (*SI Appendix*, Figs. S1 and S8). Some teams have also changed their models over time (*SI Appendix*, Tables S1 and S2 and Fig. S1). To account for some of this variability, we implemented specific inclusion criteria. However, those criteria may exclude valuable approaches that

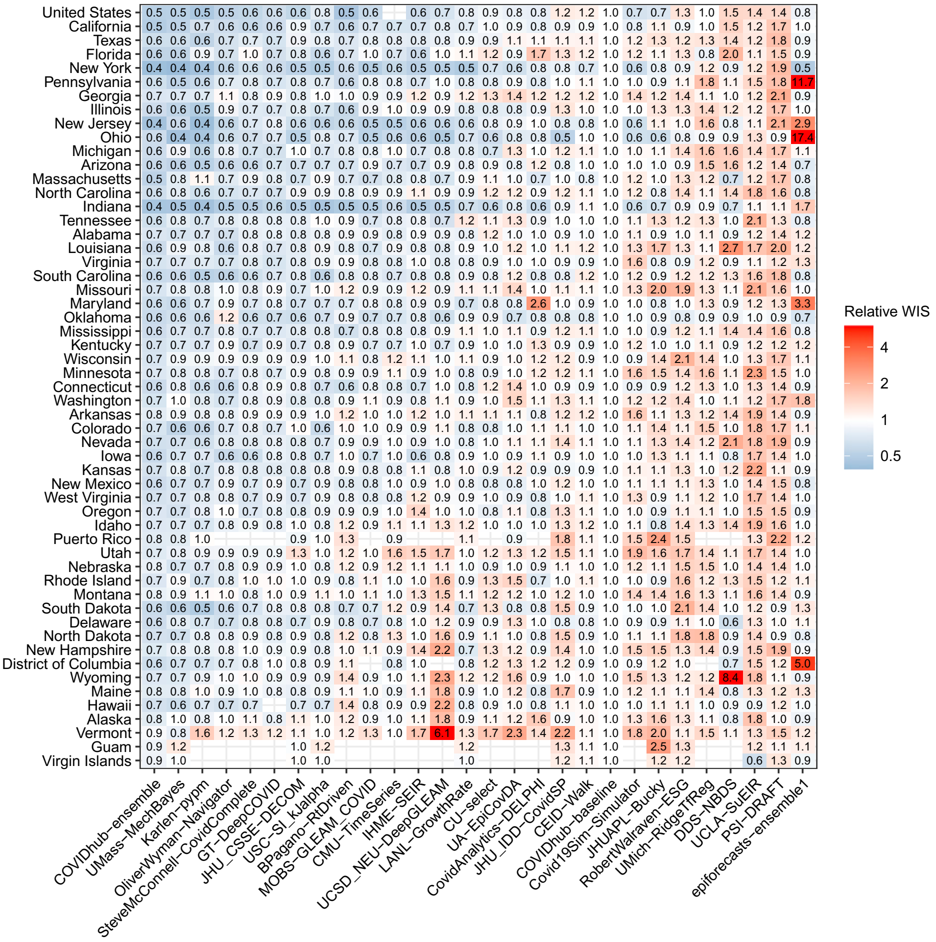

Fig. 5. Relative WIS by location for each model across all horizons and submission weeks. The value in each box represents the relative WIS calculated from 1- to 4-wk-ahead targets available for a model at each location. Boxes are colored based on the relative WIS compared to the baseline model. Blue boxes represent teams that outperformed the baseline, and red boxes represent teams that performed worse than the baseline. Locations are sorted by cumulative deaths as of the end of the evaluation period (October 30, 2021). Teams are listed on the horizontal axis in order from the lowest to highest relative WIS values (Table 1).

were not applied to a large fraction of locations or weeks (see *Methods*).

Forecast performance may be affected by ground-truth data and forecast target. Ground-truth data are not static. They can be later revised as more data become available (Dataset S1). There are also instances in which data are not revised but rather left with large peaks or dips due to reporting effects, especially around holidays. Different sources for ground truth data can also have substantial differences that impact model performance. Lastly, because this evaluation focuses on incident death forecasts, it cannot speak to model performance for incident cases or hospitalizations. Deaths may serve as a lagging indicator of COVID-19, thus making it more predictable than hospitalization and case targets (22).

While the Hub has provided many insights into what has and has not been predictable in the COVID-19 pandemic, it also has left many important questions unanswered. Due to the operational, real-time orientation of the project, the Hub did not collect data on experimental modeling studies for which certain features can be included or left out to explicitly test what features of a model increase predictive accuracy. An observational study could be conducted with forecasts collected by the Hub, but any such analysis would be confounded by other factors about how the model was built and validated. Other research in this area has shown small but measurable improvements in predictive accuracy by including other data streams available in real time (23). Continued research is needed to evaluate how behavioral, mobility, variant prevalence, or other data streams might enhance predictive modeling

Short-term forecasts of COVID-19 mortality have informed public-health response and risk communication for the pandemic. The number of teams and forecasts contributing to the COVID-19 ensemble forecast model has exceeded forecasting activity for any prior epidemic or pandemic. These forecasts are only one component of a comprehensive public-health data and modeling system needed to help inform outbreak response. Preparedness for future pandemics could be facilitated by creating resources for creating and maintaining model submission formats. This project underscores the role that collaboration and active coordination between governmental public-health agencies, academic modeling teams, and industry partners can play in developing modeling capabilities to support local, state, and federal response to outbreaks.

Methods

Surveillance Data. Early in the COVID-19 pandemic, the Johns Hopkins CSSE developed a publicly available data-tracking system and dashboard that was widely used (24). CSSE collected daily data on cumulative reported deaths due to COVID-19 at the county, state, territorial, and national levels and made these data available in a standardized format beginning in March 2020.

Incident deaths were inferred from this time series as the difference in reported cumulative deaths on successive days. Throughout the real-time forecasting exercise described in this paper, the Forecast Hub stated that forecasts of deaths would be evaluated using the CSSE data as the ground truth and encouraged teams to train their models on CSSE data.

Like data from other public-health systems, the CSSE data occasionally exhibited irregularities due to reporting anomalies. CSSE made attempts to redistribute large "backlogs" of data to previous dates in instances in which the true dates of deaths, or dates when the deaths would have been reported, were known. However, in some cases, these anomalous observations were left in the final dataset (*SI Appendix*, *SI Text*). All updates were made available in a public GitHub repository (https://github.com/CSSEGISandData/COVID-19/tree/master/csse_covid_19_data#data-modification-records). Weekly incidence values were defined and aggregated based on daily totals from Sunday through Saturday, according to the standard definition of epidemiological weeks (EW) used by the CDC (25).

Forecast Format. Research teams from around the world developed forecasting models and submitted their predictions to the COVID-19 Forecast Hub, a central repository that collected forecasts of the COVID-19 pandemic in the United States beginning in April 2020. The Forecast Hub submission process has been described in detail elsewhere (26). Incident death forecasts, the focus of this evaluation, could be submitted with predictions for horizons of 1 to 20 wk after the week in which a forecast was submitted.

A prediction for a given target (e.g., "1-week ahead incident death") and location (e.g., "California") was specified by one or both of a point forecast (a single number representing the prediction of the eventual outcome) and a probabilistic forecast. Probabilistic forecasts were represented by a set of 23 quantiles at probability levels 0.01, 0.025, 0.05, 0.10, 0.15, …, 0.95, 0.975, 0.99.

Forecast Model Eligibility and Evaluation Period. To create a set of standardized comparisons between forecasts, only models that met specific inclusion criteria were included in the analysis. For the 79 wk beginning in EW17-2020 and ending with EW42-2021, a model's weekly submission was determined to be eligible for evaluation if the forecast

1. was designated as the "primary" forecast model from a team (groups who submitted multiple parameterizations of similar models were asked to designate prospectively a single model as their scored forecast);
2. contained predictions for at least 25 out of 51 focal locations (national level and states);
3. contained predictions for each of the 1- through 4-wk-ahead targets for incident deaths; and
4. contained a complete set of quantiles for all predictions.

A model was included in the evaluation if it had submitted an eligible forecast for at least 60% ($n = 47$) of the submission weeks during the continuous 79-wk period (*SI Appendix*, Fig. S1). Based on the eligibility criteria, we compared 28 models that had at least 47 eligible weeks during this time period.

Aggregated Forecast Evaluation of Pandemic Phases. In a secondary analysis, forecasts were evaluated based on model submissions during four different phases of the pandemic. A model was eligible for inclusion in a given phase if it met the eligibility criteria listed in the *Methods* section: *Forecast model eligibility and evaluation period*, and had forecast submissions for at least 60% of the weeks during that phase. For the spring phase, models had to submit eligible forecasts for at least 6 out of 10 wk starting EW16-2020 and ending EW26-2020. For summer eligibility, a model required submissions for at least 12 out of 20 submission weeks between EW27-2020 and EW46-2020. For winter eligibility, a model required submissions for at least 14 out of 23 submission weeks between EW47-2020 and EW16-2021. For delta phase eligibility, a model required submissions for at least 16 out of 26 submission weeks between EW17-2021 and EW42-2021. These phases were determined based on the waves of deaths at the national level during pandemic (Fig. 1*B*). Each phase includes a period of increasing and decreasing incident deaths, although forecasts for the spring phase did not begin early enough to capture the increase in many locations.

Forecasts were scored using CSSE data available as of November 16, 2021. We did not evaluate forecasts on data first published in the 2 wk prior to this date due to possible revisions to the data.

Disaggregated Forecast Evaluation by Pandemic Wave. In a post hoc secondary analysis, we evaluated forecasts made in selected locations during selected pandemic waves. We used the following criteria in selecting locations and waves to represent this analysis (Fig. 4 and Dataset S2).

- We selected states that had unusually severe waves or whose waves "led" the overall wave. Locations for which data for weekly deaths during the wave had been substantially revised after the initial report were excluded from consideration.
- We picked an initial date near the start of the first increase at the start of the wave and a last date at the end of the steep decline of the wave.

To compare the forecasts during the waves, we plotted 1- and 4-wk-ahead forecasts and calculated 95% PIs coverage rates of forecasts made for the given location both during the wave of interest over all weeks. Coverage rates were computed for models that were included in the overall analysis (see eligibility in *Methods* section: *Forecast model eligibility and evaluation period*) and, for inclusion in the coverage calculations for each wave, the model additionally had to have made forecasts for at least 3 wk in the selected wave.

Forecast Locations. Forecasts were submitted for 57 locations including all 50 states, six jurisdictions and territories (American Samoa, Guam, the Northern Mariana Islands, US Virgin Islands, Puerto Rico, and the District of Columbia), and a US national-level forecast. Because American Samoa and the Northern Mariana Islands had no reported COVID-19 deaths and one reported COVID-19 death, respectively, during the evaluation period, we excluded these locations from our analysis.

In analyses for which measures of forecast skill were aggregated across locations, we typically only included the 50 states in the analysis. Including these territories in raw score aggregations would favor models that had forecasted for these regions because models were often accurate in predicting low or zero deaths each week, thereby reducing their average error. The national-level forecasts were not included in the aggregated scores because the large magnitude of scores at the national level strongly influences the averages. However, in analyses for which scores were stratified by location, we included forecasts for all US states, including territories and at the national level.

This evaluation used the CSSE COVID-19 surveillance data as ground truth when assessing forecast performance. We did not score observations when ground-truth data showed negative values for weekly incident deaths (due to changes in reporting practices from state/local health agencies [e.g., removing "probable" COVID-19 deaths from cumulative counts]). This occurred 11 times.

Forecast Models. For the primary evaluation, we compared 28 models that submitted eligible forecasts for at least 47 of the 79 wk considered in the overall model eligibility period (Fig. 1). Teams that submitted to the COVID-19 Forecast Hub used a wide variety of modeling approaches and input data (*SI Appendix*, Tables S1 and S2). Two of the evaluated models are from the COVID-19 Forecast Hub itself: a baseline model and an ensemble model.

The COVIDhub-baseline model was designed to be a neutral model to provide a simple reference point of comparison for all models. This baseline model forecasted a predictive median incidence equal to the number of reported deaths in the most recent week (y_t), with uncertainty around the median based on changes in weekly incidence that were observed in the past of the time series (details in *SI Appendix*, *SI Text*).

The COVIDhub-ensemble model combined forecasts from all models that submitted a full set of 23 quantiles for 1- through 4-wk-ahead forecasts for incident deaths. The ensemble for incident weekly deaths was first submitted in the week ending June 6, 2020 (EW23). For submission from EW23 through EW29 (week ending July 18, 2020), the ensemble took an equally weighted average of forecasts from all models at each quantile level. For submissions starting in EW30 (week ending July 25, 2020), the ensemble computed the median across forecasts from all models at each quantile level (27). We evaluated more complex ensemble methods, and while they did show modest improvements in accuracy, they also displayed undesirable increases in variability in performance during this evaluation period (28, 29).

Forecast Submission Timing. Of the 3,555 forecast submissions we included in the evaluation, 230 (6%) were either originally submitted or updated more than 24 h after the submission deadline. In all of these situations, modeling teams attested (via annotation on the public data repository) to the fact that they were correcting inadvertent errors in the code that produced the forecast and that the forecast used as input only data that would have been available before the original submission due date. In these limited instances, we evaluated the most recently submitted forecasts.

Evaluation Methodology. We evaluated aggregate forecast skill using a range of scores that assessed both point and probabilistic accuracy. These scores were aggregated over time and locations for near-term forecasts (4 wk or less into the future) and, in a single analysis, for longer-term projections (5 to 20 wk into the future).

Point forecast error was assessed using the MAE defined for a set of observations $y_{1:N}$ and each model's designated point predictions $\hat{y}_{1:N}$ as $MAE = \frac{1}{N}\sum_{i=1}^{N} | y_i - \hat{y}_i |$.

To assess probabilistic forecast accuracy, we used two scores that are easily computable from the quantile representation for forecasts described in the *Methods* section, *Forecast Format*. Briefly, the WIS is a proper score that combines a set of interval scores for probabilistic forecasts that provide quantiles of the predictive forecast distribution (20):

$$\begin{aligned} IS_{\alpha}(F,y) &= (u-l) + \frac{2}{\alpha}\cdot\ (l-y)\cdot 1(y<l) + \frac{2}{\alpha}\cdot\ (y-u)\cdot 1(y>u) \\ WIS_{\alpha_{0:K}}(F,y) &= \frac{1}{K+1/2}\cdot\left(w_0\cdot|y-m| + \sum_{k=1}^{K} w_k\cdot IS_{\alpha_k}(F,y)\right). \end{aligned}$$

An individual interval score for a single prediction and uncertainty level can be broken into three additive components. These components—dispersion, underprediction, and overprediction as they appear, respectively, in the preceding *IS* equation—represent contributions to the score. As an example, say a 50% PIs ($\alpha = 0.\ 5$) is (40, 60) and the observation is 30. The $IS_{\alpha=0.5}(\{40, 60\}, 30) = 20 + 40 + 0 = 60$, where the dispersion is 20, the penalty for underprediction is 40, and there is no penalty for overprediction. Similarly, the WIS, which is computed as a weighted sum of interval scores across all available uncertainty levels as shown in the preceding equation, can be split into contributions from each of these components. These then can be used to summarize the average performance of a model in terms of the width of its intervals and the average penalties it receives for intervals missing below or above the observation. Proper scores promote "honest" forecasting by not providing forecasters with incentives to report forecasts that differ from their true beliefs about the future (30).

We also evaluated the PIs coverage, the proportion of times a PIs of a certain level covered the observed value, to assess the degree to which forecasts accurately characterized uncertainty about future observations. Computational details for PIs coverage are provided in *SI Appendix*, *SI Text*.

Forecast Comparisons. Comparative evaluation of the considered models 1, …, M is hampered by the fact that not all of them provide forecasts for the same set of locations and time points. To adjust for the level of difficulty of each model's set of forecasts, we computed 1) a standardized rank between 0 and 1 for every forecasted observation relative to other models that made the same forecast and 2) an adjusted relative WIS and MAE.

To compute the WIS standardized rank score for model m and observation i ($sr_{m,i}$), we computed the number of models that forecasted that observation (n_i) and the rank of model m among those n_i models ($r_{m,i}$). The model with the best (i.e., lowest) WIS received a rank of 1 and the worst received a rank of n_i. The standardized rank then rescaled the ranks to between 0 and 1, where 0 corresponded to the worst rank and 1 to the best (31–33), as follows:

$$sr_{m,i} = 1 - \frac{r_{m,i} - 1}{n_i - 1}.$$

This metric is not dependent on the scale of the observed data. If all models were equally accurate, distributions of standardized ranks would be approximately uniform.

A procedure to compute a measure of relative WIS, which evaluates the aggregate performance of one model against the baseline model is described in *SI Appendix*, *SI Text*. To adjust for the relative difficulty of beating the baseline model on the covered set of forecast targets, the chosen measure also takes into account the performance of all other available models. The same procedure was used to compute a relative MAE.

Data Availability. The forecasts from models used in this paper are available from the COVID-19 Forecast Hub GitHub repository (https://github.com/reichlab/covid19-forecast-hub) (4, 34) and the Zoltar forecast archive (https://zoltardata.com/project/44) (35). These are both publicly accessible. The code used to generate all figures and tables in the manuscript is available in a public repository (https://github.com/reichlab/covid19-forecast-evals). All analyses were conducted using the R language for statistical computing (version 4.0.2) (36). We followed the EPIFORGE 2020 guidelines for reporting results from epidemiological forecasting studies (*SI Appendix*, Table S5) (37).

ACKNOWLEDGMENTS. We report the funding and disclosures below for all teams. The content is solely the responsibility of the authors and does not necessarily represent the official views of any of the funding agencies. CMU-TimeSeries: CDC Center of Excellence, gifts from Google and Facebook; CU-select: NSF DMS-2027369 and a gift from the Morris-Singer Foundation; COVIDhub: US CDC (1U01IP001122); National Institute of General Medical Sciences (NIGMS) (R35GM119582); Helmholtz Foundation (SIMCARD Information & Data Science Pilot Project); Klaus Tschira Foundation. Columbia_UNC-SurvCon*:* GM124104. DDS-NBDS: NSF III-1812699. EPIFORECASTS-ENSEMBLE1: Wellcome Trust (210758/Z/18/Z);_GT_CHHS-COVID19: William W. George Endowment, Virginia C. and Joseph C. Mello Endowments, NSF DGE-1650044, NSF MRI 1828187, CDC and Council of state and Territorial Epidemiologists (CSTE) NU38OT000297, The Partnership for Advanced Computing Environment (PACE) at Georgia Tech. Andrea Laliberte, Joseph C. Mello, Richard "Rick" E. and Charlene Zalesky, and Claudia and Paul Raines; GT-DeepCOVID: CDC Modeling Infectious Diseases in Healthcare (MinD-Healthcare) U01CK000531-Supplement; NSF (Expeditions CCF-1918770, CAREER IIS-2028586, Rapid Response Research (RAPID) IIS-2027862, Medium IIS-1955883, and NSF Research Traineeship (NRT) DGE-1545362), CDC MInD program, Oak Ridge National Laboratory (ORNL) and funds/computing resources from Georgia Tech and Georgia Tech Research Institute (GTRI); Institute for Health Metrics and Evaluation (IHME): The Bill & Melinda Gates Foundation; the state of Washington and NSF (FAIN: 2031096); IowaStateLW-STEM: Iowa State University Plant Sciences Institute Scholars Program, NSF DMS-1916204, NSF CCF-1934884, Laurence H. Baker Center for Bioinformatics and Biological Statistics; Johns Hopkins University (JHU) CSSE: NSF RAPID (2108526 and 2028604); JHU_IDD-CovidSP: State of California, US Health and Human Services (HHS), US Department of Health Services (DHS), US Office of Foreign Disaster Assistance, Johns Hopkins Health System, Office of the Dean Johns Hopkins Bloomberg School of Public Health (JHBSPH), Johns Hopkins University Modeling and Policy Hub, CDC (5U01CK000538-03), University of Utah Immunology, Inflammation, & Infectious Disease Initiative (26798 Seed Grant); LANL-GrowthRate: Los Alamos National Lab (LANL) Laboratory Directed Research and Development (LDRD) 20200700ER; MOBS-GLEAM_COVID: COVID Supplement CDC-HHS-6U01IP001137-01; CSTE Cooperative Agreement No. NU38OT000297; NotreDame-mobility and NotreDame-FRED: NSF RAPID Division of Environmental Biology (DEB) 2027718; PSI-DRAFT: NSF RAPID Grant No. 2031536; *UA-EpiCovDA*: NSF RAPID DMS 2028401; UCSB-ACTS: NSF RAPID Division of Information and Intelligent Systems (IIS) 2029626; UCSD-NEU: Google Faculty Award, Defense Advanced Research Projects Agency (DARPA) W31P4Q-21-C-0014, COVID Supplement CDC-HHS-6U01IP001137-01; UMass-MechBayes: NIGMS R35GM119582, NSF 1749854; UMich-RidgeTfReg: University of Michigan Physics Department and Office of Research; Covid19Sim-Simulator: NSF awards 2035360 and 2035361; and Gordon and Betty Moore Foundation and Rockefeller Foundation to support the work of the Society for Medical Decision Making COVID-19 Decision Modeling Initiative.

Estee Y. Cramer[a], Evan L. Ray[a], Velma K. Lopez[b], Johannes Bracher[c,d], Andrea Brennen[e], Alvaro J. Castro Rivadeneira[a], Aaron Gerding[a], Tilmann Gneiting[d,f], Katie H. House[a], Yuxin Huang[a], Dasuni Jayawardena[a], Abdul H. Kanji[a], Ayush Khandelwal[a], Khoa Le[a], Anja Mühlemann[g], Jarad Niemi[h], Apurv Shah[a], Ariane Stark[a], Yijin Wang[a], Nutcha Wattanachit[a], Martha W. Zorn[a], Youyang Gu[i], Sansiddh Jain[j], Nayana Bannur[j], Ayush Deva[j], Mihir Kulkarni[j], Srujana Merugu[j], Alpan Raval[j], Siddhant Shingi[j], Avtansh Tiwari[j], Jerome White[j], Neil F. Abernethy[k], Spencer Woody[l], Maytal Dahan[m], Spencer Fox[l], Kelly Gaither[m], Michael Lachmann[n], Lauren Ancel Meyers[l], James G. Scott[o], Mauricio Tec[p], Ajitesh Srivastava[q], Glover E. George[r], Jeffrey C. Cegan[s], Ian D. Dettwiller[r], William P. England[r], Matthew W. Farthing[r], Robert H. Hunter[r], Brandon Lafferty[r], Igor Linkov[s], Michael L. Mayo[r], Matthew D. Parno[t], Michael A. Rowland[r], Benjamin D. Trump[s], Yanli Zhang-James[u], Samuel Chen[v], Stephen V. Faraone[u], Jonathan Hess[u], Christopher P. Morley[w], Asif Salekin[x], Dongliang Wang[w], Sabrina M. Corsetti[y], Thomas M. Baer[z], Marisa C. Eisenberg[aa,bb,cc], Karl Falb[y], Yitao Huang[y], Emily T. Martin[cc], Ella McCauley[y], Robert L. Myers[y], Tom Schwarz[y], Daniel Sheldon[dd], Graham Casey Gibson[ee], Rose Yu[ff,gg], Liyao Gao[hh], Yian Ma[ii], Dongxia Wu[ff], Xifeng Yan[jj], Xiaoyong Jin[jj], Yu-Xiang Wang[jj], YangQuan Chen[kk], Lihong Guo[ll], Yanting Zhao[mm], Quanquan Gu[nn], Jinghui Chen[nn], Lingxiao Wang[nn], Pan Xu[nn], Weitong Zhang[nn], Difan Zou[nn], Hannah Biegel[oo], Joceline Lega[oo], Steve McConnell[pp], V. P. Nagraj[qq], Stephanie L. Guertin[qq], Christopher Hulme-Lowe[rr], Stephen D. Turner[qq], Yunfeng Shi[ss], Xuegang Ban[tt], Robert Walraven[uu], Qi-Jun Hong[vv,ww], Stanley Kong[xx], Axel van de Walle[ww], James A. Turtle[yy], Michal Ben-Nun[yy], Steven Riley[zz], Pete Riley[yy], Ugur Koyluoglu[aaa], David DesRoches[bbb], Pedro Forli[ccc], Bruce Hamory[ddd], Christina Kyriakides[eee], Helen Leis[fff], John Milliken[aaa], Michael Moloney[aaa], James Morgan[aaa], Ninad Nirgudkar[ggg], Gokce Ozcan[aaa], Noah Piwonka[fff], Matt Ravi[ggg], Chris Schrader[fff], Elizabeth Shakhnovich[fff], Daniel Siegel[aaa], Ryan Spatz[ggg], Chris Stiefeling[hhh], Barrie Wilkinson[iii], Alexander Wong[eee], Sean Cavany[jjj], Guido España[jjj], Sean Moore[jjj], Rachel Oidtman[jjj,kkk], Alex Perkins[jjj], David Kraus[lll], Andrea Kraus[lll], Zhifeng Gao[mmm], Jiang Bian[mmm], Wei Cao[mmm], Juan Lavista Ferres[mmm], Chaozhuo Li[mmm], Tie-Yan Liu[mmm], Xing Xie[mmm], Shun Zhang[mmm], Shun Zheng[mmm], Alessandro Vespignani[nnn,ooo], Matteo Chinazzi[ooo], Jessica T. Davis[ooo], Kunpeng Mu[ooo], Ana Pastore y Piontti[ooo], Xinyue Xiong[ooo], Andrew Zheng[ppp], Jackie Baek[ppp], Vivek Farias[qqq], Andreea Georgescu[ppp], Retsef Levi[qqq], Deeksha Sinha[ppp], Joshua Wilde[ppp], Georgia Perakis[rrr], Mohammed Amine Bennouna[rrr], David Nze-Ndong[rrr], Divya Singhvi[sss], Ioannis Spantidakis[rrr], Leann Thayaparan[rrr], Asterios Tsiourvas[rrr], Arnab Sarker[ttt], Ali Jadbabaie[ttt], Devavrat Shah[ttt], Nicolas Della Penna[uuu], Leo A. Celi[uuu], Saketh Sundar[vvv], Russ Wolfinger[www], Dave Osthus[xxx], Lauren Castro[yyy], Geoffrey Fairchild[yyy], Isaac Michaud[xxx], Dean Karlen[zzz,aaaa], Matt Kinsey[bbbb], Luke C. Mullany[bbbb], Kaitlin Rainwater-Lovett[bbbb], Lauren Shin[bbbb], Katharine Tallaksen[bbbb], Shelby Wilson[bbbb], Elizabeth C. Lee[cccc], Juan Dent[cccc], Kyra H. Grantz[cccc], Alison L. Hill[dddd], Joshua Kaminsky[cccc], Kathryn Kaminsky[eeee], Lindsay T. Keegan[ffff], Stephen A. Lauer[cccc], Joseph C. Lemaitre[gggg], Justin Lessler[cccc], Hannah R. Meredith[cccc], Javier Perez-Saez[cccc], Sam Shah[hhhh], Claire P. Smith[cccc], Shaun A. Truelove[cccc,iiii,jjjj], Josh Wills[kkkk], Maximilian Marshall[llll], Lauren Gardner[llll], Kristen Nixon[llll], John C. Burant[mmmm], Lily Wang[h], Lei Gao[nnnn], Zhiling Gu[h], Myungjin Kim[h], Xinyi Li[oooo], Guannan Wang[pppp], Yueying Wang[h], Shan Yu[qqqq], Robert C. Reiner[rrrr], Ryan Barber[rrrr], Emmanuela Gakidou[rrrr], Simon I. Hay[rrrr], Steve Lim[rrrr], Chris Murray[rrrr], David Pigott[rrrr], Heidi L. Gurung[ssss], Prasith Baccam[ssss], Steven A. Stage[tttt], Bradley T. Suchoski[ssss], B. Aditya Prakash[uuuu], Bijaya Adhikari[vvvv], Jiaming Cui[uuuu], Alexander Rodríguez[uuuu], Anika Tabassum[wwww], Jiajia Xie[uuuu], Pinar Keskinocak[xxxx], John Asplund[yyyy], Arden Baxter[xxxx], Buse Eylul Oruc[xxxx], Nicoleta Serban[xxxx], Sercan O. Arik[zzzz], Mike Dusenberry[zzzz], Arkady Epshteyn[zzzz], Elli Kanal[zzzz], Long T. Le[zzzz], Chun-Liang Li[zzzz], Tomas Pfister[zzzz], Dario Sava[zzzz], Rajarishi Sinha[zzzz], Thomas Tsai[aaaaa], Nate Yoder[zzzz], Jinsung Yoon[zzzz], Leyou Zhang[zzzz], Sam Abbott[bbbbb], Nikos I. Bosse[bbbbb], Sebastian Funk[bbbbb], Joel Hellewell[bbbbb], Sophie R. Meakin[bbbbb], Katharine Sherratt[bbbbb], Mingyuan Zhou[ccccc], Rahi Kalantari[ddddd], Teresa K. Yamana[eeeee], Sen Pei[eeeee], Jeffrey Shaman[eeeee], Michael L. Li[ppp], Dimitris Bertsimas[qqq], Omar Skali Lami[ppp], Saksham Soni[ppp], Hamza Tazi Bouardi[ppp], Turgay Ayer[xxxx,fffff], Madeline Adee[ggggg], Jagpreet Chhatwal[ggggg], Ozden O. Dalgic[hhhhh], Mary A. Ladd[ggggg], Benjamin P. Linas[iiiii], Peter Mueller[ggggg], Jade Xiao[xxxx], Yuanjia Wang[jjjjj,kkkkk], Qinxia Wang[jjjjj], Shanghong Xie[jjjjj], Donglin Zeng[lllll], Alden Green[mmmmm], Jacob Bien[nnnnn], Logan Brooks[mmmmm], Addison J. Hu[mmmmm], Maria Jahja[mmmmm], Daniel McDonald[ooooo], Balasubramanian Narasimhan[ppppp,qqqqq], Collin Politsch[rrrrr], Samyak Rajanala[qqqqq], Aaron Rumack[rrrrr], Noah Simon[sssss], Ryan J. Tibshirani[mmmmm], Rob Tibshirani[qqqqq], Valerie Ventura[mmmmm], Larry Wasserman[mmmmm], Eamon B. O'Dea[ttttt], John M. Drake[ttttt], Robert Pagano[uuuuu], Quoc T. Tran[vvvv], Lam Si Tung Ho[wwwww], Huong Huynh[xxxxx], Jo W. Walker[b], Rachel B. Slayton[b], Michael A. Johansson[b], Matthew Biggerstaff[b], and Nicholas G. Reich[a,1]

Edited by Kenneth Wachter, University of California, Berkeley, CA; received July 24, 2021; accepted January 24, 2022

[a]Department of Biostatistics and Epidemiology, University of Massachusetts, Amherst, MA 01003; [b]COVID-19 Response, Centers for Disease Control and Prevention; Atlanta, GA 30333; [c]Chair of Econometrics and Statistics, Karlsruhe Institute of Technology, 76185 Karlsruhe, Germany; [d]Computational Statistics Group, Heidelberg Institute for Theoretical Studies, 69118 Heidelberg, Germany; [e]IQT Labs, In-Q-Tel, Waltham, MA 02451; [f]Institute of Stochastics, Karlsruhe Institute of Technology, 69118 Karlsruhe, Germany; [g]Institute of Mathematical Statistics and Actuarial Science, University of Bern, CH-3012 Bern, Switzerland; [h]Department of Statistics, Iowa State University, Ames, IA 50011; [i]Unaffiliated, New York, NY 10016; [j]Wadhwani Institute of Artificial Intelligence, Andheri East, Mumbai, 400093, India; [k]University of Washington, Seattle, WA 98109; [l]Department of Integrative Biology, University of Texas at Austin, Austin, TX 78712; [m]Texas Advanced Computing Center, Austin, TX 78758; [n]Santa Fe Institute, Santa Fe, NM 87501; [o]Department of Information, Risk, and Operations Management, University of Texas at Austin, Austin, TX 78712; [p]Department of Statistics and Data Sciences, University of Texas at Austin, Austin, TX 78712; [q]Ming Hsieh Department of Computer and Electrical Engineering, University of Southern California, Los Angeles, CA 90089; [r]US Army Engineer Research and Development Center, Vicksburg, MS 39180; [s]US Army Engineer Research and Development Center, Concord, MA 01742; [t]US Army Engineer Research and Development Center, Hanover, NH 03755; [u]Department of Psychiatry and Behavioral Sciences, State University of New York Upstate Medical University, Syracuse, NY 13210; [v]School of Medicine, State University of New York Upstate Medical University, Syracuse, NY 13210; [w]Department of Public Health & Preventive Medicine, State University of New York Upstate Medical University, Syracuse, NY 13210; [x]Department of Electrical Engineering and Computer Science, Syracuse University, Syracuse, NY 13207; [y]Department of Physics, University of Michigan, Ann Arbor, MI, 48109; [z]Department of Physics, Trinity University, San Antonio, TX 78212; [aa]Department of Complex Systems, University of Michigan, Ann Arbor, MI 48109; [bb]Department of Mathematics, University of Michigan, Ann Arbor, MI 48109; [cc]School of Public Health, Department of Epidemiology, University of Michigan, Ann Arbor, MI 48109; [dd]College of Information and Computer Sciences, University of Massachusetts, Amherst, MA 01003; [ee]School of Public Health and Health Sciences, University of Massachusetts, Amherst, MA 01003; [ff]Department of Computer Science and Engineering, University of California, San Diego, CA 92093; [gg]Khoury College of Computer Sciences, Northeastern University, Boston, MA 02115; [hh]Department of Statistics, University of Washington, Seattle, WA 98185; [ii]Halıcıoğlu Data Science Institute, University of California, San Diego, CA 92093; [jj]Department of Computer Science, University of California, Santa Barbara, CA 93106; [kk]Mechatronics, Embedded Systems and Automation Lab, Department of Mechanical Engineering, University of California, Merced, CA 95301; [ll]Jilin University, Changchun City, Jilin Province, 130012, People's Republic of China; [mm]University of Science and Technology of China, Heifei, Anhui, 230027, People's Republic of China; [nn]Department of Computer Science, University of California, Los Angeles, CA 90095; [oo]Department of Mathematics, University of Arizona, Tucson, AZ 85721; [pp]Construx, Bellevue, WA 98004; [qq]Quality Assurance and Data Science, Signature Science, LLC, Charlottesville, VA 22911; [rr]Quality Assurance and Data Science, Signature Science, LLC, Austin, TX 78759; [ss]Department of Materials Science and Engineering, Rensselaer Polytechnic Institute, Troy, NY 12309; [tt]Department of Civil and Environmental Engineering, University of Washington, Seattle, WA 98195; [uu]Unaffiliated, Davis, CA 95616; [vv]School for Engineering of Matter, Transport and Energy, Arizona State University, Tempe, AZ 85287; [ww]School of Engineering, Brown University, Providence, RI 02912; [xx]Manhasset Secondary School, Manhasset, NY 11030; [yy]Infectious Disease Group, Predictive Science, Inc, San Diego, CA 92121; [zz]Medical Research Council Centre for Global Infectious Disease Analysis, Department of Infectious Disease Epidemiology, School of Public Health, Imperial College, W2 1PG London, United Kingdom; [aaa]Financial Services, Oliver Wyman, New York, NY 10036; [bbb]Oliver Wyman Digital, Oliver Wyman, Boston, MA 02110; [ccc]Oliver Wyman Digital, Oliver Wyman, Sao Paolo, Brazil 04711-904; [ddd]Health & Life Sciences, Oliver Wyman, Boston, MA 02110; [eee]Oliver Wyman Digital, Oliver Wyman, New York, NY 10036; [fff]Health & Life Sciences, Oliver Wyman, New York, NY 10036; [ggg]Core Consultant Group, Oliver Wyman, New York, NY 10036; [hhh]Financial Services, Oliver Wyman Digital, Toronto, ON, Canada M5J 0A1; [iii]Financial Services, Oliver Wyman, London, UK W1U 8EW; [jjj]Department of Biological Sciences, University of Notre Dame, Notre Dame, IN 46556; [kkk]Department of Ecology and Evolution, University of Chicago, Chicago, IL 60637; [lll]Department of Mathematics and Statistics, Masaryk University, 61137 Brno, Czech Republic; [mmm]Microsoft, Redmond, WA 98029; [nnn]Institute for Scientific Interchange Foundation, Turin, 10133, Italy; [ooo]Laboratory for the Modeling of Biological and Socio-technical Systems, Northeastern University, Boston, MA 02115; [ppp]Operations Research Center, Massachusetts Institute of Technology; Cambridge, MA 02139; [qqq]Sloan School of Management, Massachusetts Institute of Technology, Cambridge, MA 02142; [rrr]Massachusetts Institute of Technology, Cambridge, MA 02142; [sss]Technology, Operations and Statistics (TOPS) group, Stern School of Business, New York University, New York, NY 10012; [ttt]Institute for Data, Systems, and Society, Massachusetts Institute of Technology, Cambridge, MA 02139; [uuu]Laboratory for Computational Physiology, Massachusetts Institute of Technology, Cambridge, MA 02139; [vvv]River Hill High School, Clarksville, MD 21029; [www]JMP Life Sciences, SAS Institute Inc, Cary, NC 27513; [xxx]Statistical Sciences Group, Los Alamos National Laboratory, Los Alamos, NM 87545; [yyy]Information Systems and Modeling Group, Los Alamos National Laboratory, Los Alamos, NM 87545; [zzz]Department of Physics and Astronomy, University of Victoria, Victoria, BC, V8W 2Y2, Canada; [aaaa]Physical Sciences Division, TRIUMF, Vancouver, BC, V8W 2Y2, Canada; [bbbb]Johns Hopkins University Applied Physics Laboratory, Laurel, MD 20723; [cccc]Department of Epidemiology, Johns Hopkins Bloomberg School of Public Health, Baltimore, MD 21215; [dddd]Institute for Computational Medicine, Johns Hopkins Bloomberg School of Public Health, Baltimore, MD 21218; [eeee]Unaffiliated, Baltimore, MD 21205; [ffff]Division of Epidemiology, Department of Internal Medicine, University of Utah, Salt Lake City, UT 84108; [gggg]Laboratory of Ecohydrology, School of Architecture, Civil and Environmental Engineering, École Polytechnique Fédérale de Lausanne, 1015 Lausanne, Switzerland; [hhhh]Unaffiliated, San Francisco, CA 94122; [iiii]International Vaccine Access Center, Johns Hopkins University, Baltimore, MD 21231; [jjjj]Department of International Health, Johns Hopkins Bloomberg School of Public Health, Baltimore, MD 21231; [kkkk]Unaffiliated, San Francisco, CA 94122; [llll]Department of Civil and Systems Engineering, Johns Hopkins University, Baltimore, MD 21218; [mmmm]Unaffiliated, Amsterdam, The Netherlands; [nnnn]Department of Finance, Iowa State University, Ames, IA 50011; [oooo]School of Mathematical and Statistical Sciences, Clemson University, Clemson, SC 29634; [pppp]Department of Mathematics, College of William & Mary, Williamsburg, VA 23187; [qqqq]Department of Statistics, University of Virginia, Charlottesville, VA 22904; [rrrr]Institute for Health Metrics and Evaluation, University of Washington, Seattle, WA 98195; [ssss]Emerging Technologies, IEM, Inc., Bel Air, MD 21015; [tttt]Emerging Technologies, IEM, Inc., Baton Rouge, LA 70809; [uuuu]College of Computing, Georgia Institute of Technology, Atlanta, GA 30308; [vvvv]Department of Computer Science, University of Iowa, Iowa City, IA 52242; [wwww]Department of Computer Science, Virginia Tech, Falls Church, VA 22043; [xxxx]H. Milton Stewart School of Industrial and Systems Engineering, Georgia Institute of Technology, Atlanta, GA, 30332; [yyyy]Advanced Data Analytics, Metron, Inc., Reston, VA 20190; [zzzz]Google Cloud, Sunnyvale, CA 94089; [aaaaa]Department of Health Policy and Management, Harvard University, Cambridge, MA 02138; [bbbbb]Centre for Mathematical Modelling of Infectious Diseases, London School of Hygiene & Tropical Medicine, WC1E 7HT London, United Kingdom; [ccccc]McCombs School of Business, The University of Texas at Austin, Austin, TX 78712; [ddddd]Department of Electrical and Computer Engineering, The University of Texas at Austin, Austin, TX 78712; [eeeee]Department of Environmental Health Sciences, Columbia University, New York, NY 10032; [fffff]Winship Cancer Institute, Emory University Medical School, Atlanta, GA 30322; [ggggg]Radiology-Institute for Technology Assessment, Massachusetts General Hospital, Boston, MA 02114; [hhhhh]Health Economic Modeling, Value Analytics Labs, 34776 İstanbul, Turkey; [iiiii]Department of Medicine, Section of Infectious Diseases, Boston University School of Medicine, Boston, MA 02118; [jjjjj]Department of Biostatistics, Columbia University, New York, NY 10032; [kkkkk]Department of Psychiatry, Columbia University, New York, NY 10032; [lllll]Department of Biostatistics, University of North Carolina at Chapel Hill, Chapel Hill, NC 27599; [mmmmm]Department of Statistics, Carnegie Mellon University, Pittsburgh, PA 15213; [nnnnn]Marshall School of

STATISTICS

POPULATION BIOLOGY

Business, Department of Data Sciences and Operations (DSO), University of Southern California, Los Angeles, CA 90089; [ooooo]Department of Statistics, University of British Columbia, Vancouver, BC, V6T 1Z4, Canada; [ppppp]Department of Biomedical Data Sciences, Stanford University, Stanford, CA 94305; [qqqqq]Department of Statistics, Stanford University, Stanford, CA 94305; [rrrrr]Machine Learning Department, Carnegie Mellon University, Pittsburgh, PA 15213; [sssss]Department of Biostatistics, University of Washington, Seattle, WA 98195; [ttttt]Odum School of Ecology, University of Georgia, Athens, GA 30602; [uuuuu]Unaffiliated, Tucson, AZ 85704; [vvvvv]Catalog Data Science, Walmart Inc., Sunnyvale, CA 94085; [wwwww]Department of Mathematics and Statistics, Dalhousie University, Halifax, NS, B3H 4R2, Canada; and [xxxxx]Virtual Power System Inc, Milpitas, CA 95035

1. S. E. Davies, J. R. Youde, *The Politics of Surveillance and Response to Disease Outbreaks: The New Frontier for States and Non-state Actors* (Routledge, 2016).
2. J. A. Polonsky *et al.*, Outbreak analytics: A developing data science for informing the response to emerging pathogens. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* **374**, 20180276 (2019).
3. C. S. Lutz *et al.*, Applying infectious disease forecasting to public health: A path forward using influenza forecasting examples. *BMC Public Health* **19**, 1659 (2019).
4. E. Cramer *et al.*, COVID-19 Forecast Hub: 4 December 2020 snapshot. https://zenodo.org/record/4305938#.Yf1TQOrMI2x (Accessed 11 December 2020).
5. CDC, COVID-19 Forecasting and Mathematical Modeling. Centers for Disease Control and Prevention. https://www.cdc.gov/coronavirus/2019-ncov/science/forecasting/mathematical-modeling.html (Accessed 2 March 2022).
6. J. M. Bates, C. W. J. Granger, The combination of forecasts. *J. Oper. Res. Soc.* **20**, 451–468 (1969).
7. T. N. Krishnamurti *et al.*, Improved weather and seasonal climate forecasts from multimodel superensemble. *Science* **285**, 1548–1550 (1999).
8. T. Gneiting, A. E. Raftery, Atmospheric science. Weather forecasting with ensemble methods. *Science* **310**, 248–249 (2005).
9. M. Leutbecher, T. N. Palmer, Ensemble forecasting. *J. Comput. Phys.* **227**, 3515–3539 (2008).
10. R. Polikar, Ensemble based systems in decision making. *IEEE Circuits Syst. Mag.* **6**, 21–45 (2006).
11. B. Lakshminarayanan, A. Pritzel, C. Blundell, Simple and scalable predictive uncertainty estimation using deep ensembles. arXiv [Preprint] (2017). https://arxiv.org/abs/1612.01474 (Accessed 24 December 2020).
12. C. J. McGowan *et al.*; Influenza Forecasting Working Group, Collaborative efforts to forecast seasonal influenza in the United States, 2015-2016. *Sci. Rep.* **9**, 683 (2019).
13. N. G. Reich *et al.*, A collaborative multiyear, multimodel assessment of seasonal influenza forecasting in the United States. *Proc. Natl. Acad. Sci. U.S.A.* **116**, 3146–3154 (2019).
14. N. G. Reich *et al.*, Accuracy of real-time multi-model ensemble forecasts for seasonal influenza in the U.S. *PLOS Comput. Biol.* **15**, e1007486 (2019).
15. C. Viboud *et al.*; RAPIDD Ebola Forecasting Challenge group, The RAPIDD ebola forecasting challenge: Synthesis and lessons learnt. *Epidemics* **22**, 13–21 (2018).
16. M. A. Johansson *et al.*, An open challenge to advance probabilistic forecasting for dengue epidemics. *Proc. Natl. Acad. Sci. U.S.A.* **116**, 24268–24274 (2019).
17. S. Funk *et al.*, Short-term forecasts to inform the response to the Covid-19 epidemic in the UK. medRxiv [Preprint] (2020). https://www.medrxiv.org/content/10.1101/2020.11.11.20220962v2 (Accessed 2 December 2020).
18. K. S. Taylor, J. W. Taylor, A comparison of aggregation methods for probabilistic forecasts of COVID-19 mortality in the United States. arXiv [Preprint] (2020). https://arxiv.org/abs/2007.11103 (Accessed 2 December 2020).
19. 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.* **214** (suppl. 4), S404–S408 (2016).
20. J. Bracher, E. L. Ray, T. Gneiting, N. G. Reich, Evaluating epidemic forecasts in an interval format. *PLOS Comput. Biol.* **17**, e1008618 (2021).
21. S. Lerch, T. L. Thorarinsdottir, F. Ravazzolo, T. Gneiting, Forecaster's dilemma: Extreme events and forecast evaluation. *SSO Schweiz. Monatsschr. Zahnheilkd.* **32**, 106–127 (2017).
22. J. Bracher *et al.*; List of Contributors by Team, A pre-registered short-term forecasting study of COVID-19 in Germany and Poland during the second wave. *Nat. Commun.* **12**, 5173 (2021).
23. D. J. McDonald *et al.*, Can auxiliary indicators improve COVID-19 forecasting and hotspot prediction? bioRxiv [Preprint] (2021). https:/doi.org/10.1101/2021.06.22.21259346.
24. E. Dong, H. Du, L. Gardner, An interactive web-based dashboard to track COVID-19 in real time. *Lancet Infect. Dis.* **20**, 533–534 (2020).
25. Department of Health, NM-IBIS - MMWR week description and corresponding calendar dates (2006–2025). https://ibis.health.state.nm.us/resource/MMWRWeekCalendar.html (Accessed 13 January 2021).
26. E. Y. Cramer *et al.*, The United States COVID-19 Forecast Hub hub dataset. medRxiv [Preprint] (2021). https://doi.org/10.1101/2021.11.04.21265886 (Accessed 4 December 2021).
27. E. L. Ray *et al.*, Ensemble forecasts of coronavirus disease 2019 (COVID-19) in the U.S. medRxiv [Preprint] (2020). https://www.medrxiv.org/content/10.1101/2020.08.19.20177493v1 (Accessed 2 December 2020).
28. L. C. Brooks *et al.*, *Comparing Ensemble Approaches for Short-term Probabilistic COVID-19 Forecasts in the U.S.* (International Institute of Forecasters, 2020).
29. E. L. Ray *et al.*, *Challenges in Training Ensembles to Forecast COVID-19 Cases and Deaths in the United States* (International Institute of Forecasters, 2021).
30. T. Gneiting, A. E. Raftery, Strictly proper scoring rules, prediction, and estimation. *J. Am. Stat. Assoc.* **102**, 359–378 (2007).
31. S. R. Soloman, S. S. Sawilowsky, Impact of rank-based normalizing transformations on the accuracy of test scores. *J. Mod. Appl. Stat. Methods* **8**, 448–462 (2009).
32. S. Wu, F. Crestani, Y. Bi, "Evaluating score normalization methods in data fusion" in Information Retrieval Technology, Lecture notes in computer science., (Springer Berlin Heidelberg, 2006), pp. 642–648.
33. M. E. Renda, U. Straccia, "Web metasearch: Rank vs. score based rank aggregation methods" in *Proceedings of the 2003 ACM Symposium on Applied Computing*, SAC '03. (Association for Computing Machinery, 2003)), pp. 841–846.
34. E. Y. Cramer *et al.*, COVID-19 Forecast Hub. GitHub. https://github.com/reichlab/covid19-forecast-hub/tree/master/data-processed. Accessed 17 November 2021.
35. N. G. Reich, M. Cornell, E. L. Ray, K. House, K. Le, The Zoltar forecast archive, a tool to standardize and store interdisciplinary prediction research. *Sci. Data* **8**, 59 (2021).
36. R Core Team, *R: A Language and Environment for Statistical Computing* (R Foundation for Statistical Computing, 2020).
37. S. Pollett *et al.*, Recommended reporting items for epidemic forecasting and prediction research: The EPIFORGE 2020 guidelines. *PLoS Med.* **18**, e1003793 (2021).