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

Estee Y Cramer[1], Evan L Ray[1], Velma K Lopez[2], Johannes Bracher[3,4], Andrea Brennen[5], Alvaro J Castro Rivadeneira[1], Aaron Gerding[1], Tilmann Gneiting[4,6], Katie H House[1], Yuxin Huang[1], Dasuni Jayawardena[1], Abdul H Kanji[1], Ayush Khandelwal[1], Khoa Le[1], Anja Mühlemann[7], Jarad Niemi[8], Apurv Shah[1], Ariane Stark[1], Yijin Wang[1], Nutcha Wattanachit[1], Martha W Zorn[1], Youyang Gu[9], Sansiddh Jain[10], Nayana Bannur[10], Ayush Deva[10], Mihir Kulkarni[10], Srujana Merugu[10], Alpan Raval[10], Siddhant Shingi[10], Avtansh Tiwari[10], Jerome White[10], Spencer Woody[11], Maytal Dahan[12], Spencer Fox[11], Kelly Gaither[12], Michael Lachmann[13], Lauren Ancel Meyers[11], James G Scott[11], Mauricio Tec[11], Ajitesh Srivastava[14], Glover E George[15], Jeffrey C Cegan[15], Ian D Dettwiller[15], William P England[15], Matthew W Farthing[15], Robert H Hunter[15], Brandon Lafferty[15], Igor Linkov[15], Michael L Mayo[15], Matthew D Parno[15], Michael A Rowland[15], Benjamin D Trump[15], Yanli Zhang-James[16], Samuel Chen[16], Stephen V Faraone[16], Jonathan Hess[16], Christopher P Morley[16], Asif Salekin[17], Dongliang Wang[16], Sabrina M Corsetti[18], Thomas M Baer[19], Marisa C Eisenberg[18], Karl Falb[18], Yitao Huang[18], Emily T Martin[18], Ella McCauley[18], Robert L Myers[18], Tom Schwarz[18], Daniel Sheldon[1], Graham Casey Gibson[1], Rose Yu[20,21], Liyao Gao[22], Yian Ma[20], Dongxia Wu[20], Xifeng Yan[23], Xiaoyong Jin[23], Yu-Xiang Wang[23], YangQuan Chen[24], Lihong Guo[25], Yanting Zhao[26], Quanquan Gu[27], Jinghui Chen[27], Lingxiao Wang[27], Pan Xu[27], Weitong Zhang[27], Difan Zou[27], Hannah Biegel[28], Joceline Lega[28], Steve McConnell[29], VP Nagraj[30], Stephanie L Guertin[30], Christopher Hulme-Lowe[30], Stephen D Turner[30], Yunfeng Shi[31], Xuegang Ban[22], Robert Walraven[9], Qi-Jun Hong[32,33], Axel van de Walle[32], Stanley Kong[34], James A Turtle[35], Michal Ben-Nun[35], Pete Riley[35], Steven Riley[36], Ugur Koyluoglu[37], David DesRoches[37], Pedro Forli[37], Bruce Hamory[37], Christina Kyriakides[37], Helen Leis[37], John Milliken[37], Michael Moloney[37], James Morgan[37], Ninad Nirgudkar[37], Gokce Ozcan[37], Noah Piwonka[37], Matt Ravi[37], Chris Schrader[37], Elizabeth Shakhnovich[37], Daniel Siegel[37], Ryan Spatz[37], Chris Stiefeling[37], Barrie Wilkinson[37], Alexander Wong[37], Sean Cavany[38], Guido España[38], Sean Moore[38], Rachel Oidtman[38,39], Alex Perkins[38], Zhifeng Gao[40], Jiang Bian[40], Wei Cao[40], Juan Lavista Ferres[40], Chaozhuo Li[40], Tie-Yan Liu[40], Xing Xie[40], Shun Zhang[40], Shun Zheng[40], Alessandro Vespignani[41,42], Matteo Chinazzi[41], Jessica T Davis[41], Kunpeng Mu[41], Ana Pastore y Piontti[41], Xinyue Xiong[41], Andrew Zheng[43], Jackie Baek[43], Vivek Farias[44], Andreea Georgescu[43], Retsef Levi[44], Deeksha Sinha[43], Joshua Wilde[43], Arnab Sarker[45], Ali Jadbabaie[45], Devavrat Shah[45], Nicolas D Penna[46], Leo A Celi[46], Saketh Sundar[47], Russ Wolfinger[48], Dave Osthus[49], Lauren Castro[49], Geoffrey Fairchild[49], Isaac Michaud[49], Dean Karlen[50,51], Matt Kinsey[52], Katharine Tallaksen[52], Shelby Wilson[52], Lauren Shin[52], Luke C Mullany[52], Kaitlin Rainwater-Lovett[52], Elizabeth C Lee[53], Juan Dent[53], Kyra H Grantz[53], Joshua Kaminsky[53], Kathryn Kaminsky[9], Lindsay T Keegan[54], Stephen A Lauer[53], Joseph C Lemaitre[55], Justin Lessler[53], Hannah R Meredith[53], Javier Perez-Saez[53], Sam Shah[9], Claire P Smith[53], Shaun A Truelove[53], Josh Wills[9], Maximilian Marshall[53], Lauren Gardner[53], Kristen Nixon[53], John C Burant[9], Lily Wang[8], Lei Gao[8], Zhiling Gu[8], Myungjin Kim[8], Xinyi Li[56], Guannan Wang[57], Yueying Wang[8], Shan Yu[58], Robert C Reiner[22], Ryan Barber[22], Emmanuela Gaikedu[22], Simon Hay[22], Steve Lim[22], Chris Murray[22], David Pigott[22], Heidi L Gurung[59], Prasith Baccam[59], Steven A Stage[59], Bradley T Suchoski[59], B Aditya Prakash[60], Bijaya Adhikari[61], Jiaming Cui[60], Alexander Rodríguez[60], Anika Tabassum[60,62], Jiajia Xie[60], Pinar Keskinocak[60], John Asplund[63], Arden Baxter[60], Buse Eylul Oruc[60], Nicoleta Serban[60], Sercan O Arik[64], Mike Dusenberry[64], Arkady Epshteyn[64], Elli Kanal[64], Long T Le[64], Chun-Liang Li[64], Tomas Pfister[64], Dario Sava[64], Rajarishi Sinha[64], Thomas Tsai[65], Nate Yoder, Jinsung Yoon[64], Leyou Zhang[64], Sam Abbott[66], Nikos I Bosse[66], Sebastian Funk[66], Sophie R Meakin[66], Katherine Sherratt[66], Mingyuan Zhou[11], Rahi Kalantari[11], Teresa K Yamana[67], Sen Pei[67], Jeffrey Shaman[67], Michael L Li[43], Dimitris Bertsimas[44], Omar Skali Lami[43], Saksham Soni[43], Hamza Tazi Bouardi[43],Turgay Ayer[60,68], Madeline Adee[69], Jagpreet Chhatwal[69], Ozden O Dalgic[70], Mary A Ladd[69], Benjamin P Linas[71], Peter Mueller[69], Jade Xiao[60] Yuanjia Wang[67], Qinxia Wang[67], Shanghong Xie[67], Donglin Zeng[72], Alden Green[73], Jacob Bien[14], Logan Brooks[73], Daniel McDonald[74], Addison J Hu[73], Maria Jahja[73], Balasubramanian Narasimhan[75], Collin Politsch[73], Samyak Rajanala[75], Aaron Rumack[73], Noah Simon[22], Ryan Tibshirani[73], Rob Tibshirani[75], Valerie Ventura[73], Larry Wasserman[73], Eamon B O'Dea[76], John M Drake[76], Robert Pagano[9], Neil F Abernethy[22], Jo W Walker[2], Rachel B Slayton[2], Michael Johansson[2], Matthew Biggerstaff[2], Nicholas G Reich[1]

Affiliations

1 University of Massachusetts, Amherst
2 Centers for Disease Control and Prevention
3 Chair of Econometrics and Statistics, Karlsruhe Institute of Technology
4 Computational Statistics Group, Institute of Technology (KIT)
5 In-Q-Tel
6 Institute for Stochastics, Karlsruhe Affiliations Heidelberg Institute for Theoretical Studies
7 Institute of Mathematical Statistics and Actuarial Science, University of Bern
8 Iowa State University
9 No affiliation
10 Wadhwani Institute for Artificial Intelligence
11 The University of Texas at Austin
12 Texas Advanced Computing Center
13 Santa Fe Institute
14 University of Southern California
15 US Army Engineer Research and Development Center
16 State University of New York Upstate Medical University
17 Syracuse University
18 University of Michigan - Ann Arbor
19 Trinity University, San Antonio
20 University of California, San Diego
21 Northeastern University
22 University of Washington
23 University of California at Santa Barbara
24 University of California, Merced
25 Jilin University
26 University of Science and Technology of China
27 University of California, Los Angeles
28 University of Arizona
29 Construx
30 Signature Science, LLC
31 Rensselaer Polytechnic Institute
32 Brown University
33 Arizona State University
34 Manhasset Secondary School
35 Predictive Science, Inc
36 Imperial College, London
37 Oliver Wyman
38 University of Notre Dame
39 University of Chicago
40 Microsoft
41 Laboratory for the Modeling of Biological and Socio-technical Systems, Northeastern University
42 Institute for Scientific Interchange Foundation
43 Operations Research Center, Massachusetts Institute of Technology
44 Sloan School of Management, Massachusetts Institute of Technology
45 Institute for Data, Systems, and Society, Massachusetts Institute of Technology
46 Laboratory for Computational Physiology, Massachusetts Institute of Technology
47 River Hill High School
48 SAS Institute Inc
49 Los Alamos National Laboratory
50 University of Victoria
51 TRIUMF
52 Johns Hopkins University Applied Physics Lab
53 Johns Hopkins Bloomberg School of Public Health
54 University of Utah
55 École Polytechnique Fédérale de Lausanne
56 Clemson University
57 College of William & Mary
58 University of Virginia

59 IEM, Inc
60 Georgia Institute of Technology
61 University of Iowa
62 Virginia Tech
63 Metron, Inc
64 Google Cloud
65 Harvard University
66 London School of Hygiene & Tropical Medicine
67 Columbia University
68 Emory University Medical School
69 Massachusetts General Hospital
70 Value Analytics Labs
71 Boston University School of Medicine
72 University of North Carolina Chapel Hill
73 Carnegie Mellon University
74 University of British Columbia
75 Stanford University
76 University of Georgia

***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.

Abstract

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 COVID-19 Forecast Hub (https://covid19forecasthub.org/) collected, disseminated, and synthesized tens of millions of specific predictions from more than 80 different academic, industry, and independent research groups. A multi-model ensemble forecast that combined predictions from dozens of different research groups every week provided the most consistently accurate probabilistic forecasts of incident mortality due to COVID-19 at the state and national level from April 2020 through April 2021. The performance of 27 individual models that submitted complete forecasts consistently throughout this year showed high variability in forecast skill across time, geospatial units, and forecast horizons. Slightly more than half 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-week horizon 3-5 times larger than when predicting at a 1-week 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.

Introduction

Effective pandemic responses require 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 effective resource allocation and response. Forecasting models provide specific, quantitative, evaluable, and often probabilistic predictions about the epidemic trajectory for the near-term future. Typically provided for a horizon of up to one or two months, 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 policy makers to assess the most likely outcomes and plausible worst-case scenarios (3). Forecasts occupy a unique niche in infectious disease modeling, as they provide opportunities to concretely evaluate the accuracy of different modeling approaches, often in real-time.

With a great need to understand how the COVID-19 epidemic would progress over time in the United States, academic research groups, government agencies, industry groups, and individuals produced COVID-19 forecasts at an unprecedented scale starting in March 2020. Publicly accessible forecasts reflect varied approaches, data sources, and assumptions. For example, forecasts were created from statistical or machine learning models, mechanistic models that incorporated disease transmission dynamics, and combinations of these approaches. 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 United States Centers for Disease Control and Prevention (CDC) partnered with the Reich Lab 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 was designed to facilitate 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 futures. In practice most individual models make predictions that are conditional on explicit or implicit assumptions of how policies, behaviors, and pathogens will evolve in the coming weeks. From these forecasts, a multi-model ensemble was developed, published weekly in real-time, and used by CDC in official public communications about the pandemic (https://www.cdc.gov/coronavirus/2019-ncov/covid-data/mathematical-modeling.html).

It is challenging for individual models to make calibrated predictions of the future when the behavior of the system being studied is non-stationary due to continually changing policies and behaviors. In the case of COVID-19, these included but were not limited to school closures,

lockdowns, masking, social distancing, and emergence of new strains. With a sufficiently diverse set of models, an ensemble will incorporate uncertainties from a broader range of perspectives than most individual models. Indeed, ensemble approaches have previously demonstrated superior performance compared with single models in forecasting influenza (5–7), Ebola (8), and dengue fever outbreaks (9). Preliminary research suggested that COVID-19 ensemble forecasts were also more accurate and precise than individual models in the early phases of the pandemic (10, 11). As has been seen in research across disciplines, ensemble approaches are able to draw and incorporate the information from multiple forecasts, each with their own strengths and limitations, to create accurate predictions with well-calibrated uncertainty (12–17). Additionally, synthesizing multiple models removes the risk of over-reliance on any single approach for accuracy or stability.

While forecasts provide important information to policy makers for the COVID-19 response, predicting the trajectory of a novel pathogen outbreak is subject to many challenges. First, due to the role of human behavior and decision-making in outbreak trajectories, epidemic forecasts must account for both biological and societal trends. Furthermore, 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 (18). Moreover, there is inherent uncertainty in many critical parameters needed to model future trends in transmission, and this uncertainty grows quickly as models try to look further into the future. There are also a host of data irregularities, especially in the early stages of the pandemic, including delayed reporting of data, revisions to existing data, and outlying values. Models trained using historical data may lack sufficient characterization of all underlying uncertainties (19).

It is important to systematically and rigorously evaluate COVID-19 forecasts designed to predict real-time changes to the outbreak in order to identify strengths and weaknesses of different approaches. Understanding 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 53-week period starting late April 2020 and ending in late April 2021 (Figure 1 and Methods). Forecasts were evaluated at 55 locations including all 50 states, 4 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 a spring wave, a late summer increase in several locations, and a large late-fall/early-winter surge (Figure 1B).

The number of models that submitted forecasts of incident deaths to the Forecast Hub and were screened for eligibility increased from 4 models at the beginning of the evaluation period to over 35 for much of the first five months of 2021 (Figure 1C). Not all models submitted forecasts

every week, and some models submitted forecasts for varying numbers of locations or different quantiles in different weeks (Supplemental Figure 1). Twenty-seven models met our inclusion criteria (see Methods), yielding 1,150 submission files with 356,616 specific predictions for unique combinations of targets (horizons forecasted) and locations. Five of these models (including the baseline model, see Methods) submitted forecasts for 50 or more of the 53 evaluated weeks.

The evaluated forecasts used different data sources and made varying assumptions about future transmission patterns (Table 1). All models other than the COVIDhub-ensemble incorporated data on prior deaths to create forecasts; the COVIDhub-ensemble did not directly use surveillance data to create forecasts as it is a combination of submitted forecasts. Additionally, all evaluated models other than the COVIDhub-ensemble, the COVIDhub-baseline, CEID-Walk, PSI-Draft, RPI_IW-Mob-Collision, and UT-Mobility used case data as inputs to their forecast models. Nine models included data on COVID-19 hospitalizations, ten models incorporated demographic data, and eleven models used mobility data. Of the 27 models evaluated, 5 made explicit assumptions that social distancing and other behavioral patterns would change over the prediction period.

Overall model accuracy

To evaluate probabilistic accuracy, the primary metric used was a weighted interval score (WIS) which measures how closely a collection of prediction intervals is consistent with an observed value (20). For WIS, a lower value (closer to 0) represents smaller error (see Methods).

Led by the ensemble model, which showed the best average probabilistic accuracy of all models across the evaluation period, a narrow majority of the evaluated models achieved better accuracy than the baseline model (see Methods) in forecasting incident deaths (Table 2). The COVIDhub-ensemble model 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 ease or difficulty of the specific predictions made. An additional four models achieved a relative WIS of less than 0.75. In total, 15 models had a relative WIS of less than 1, indicating lower probabilistic forecast error than the baseline model, and 12 had a relative WIS of 1 or greater (Table 2). Patterns in relative point forecast error were similar, with 17 models having equal or lower mean absolute error (MAE) than the baseline (Table 2). Values of relative WIS and rankings of models were robust to different sets of models being included or excluded and to individual outlying or revised observations being included or excluded from the analysis (Supplemental Tables 2 and 3).

The degree to which individual models provided calibrated predictions varied widely (Table 2). We measured the probabilistic calibration of model forecasts using the empirical coverage rates of prediction intervals (PIs). Across 1 through 4 week-ahead horizons, 53 weeks, and 50 states, few models achieved near nominal coverage rates for both the 50% and 95% PIs. Five models achieved coverage rates within 5% of the desired coverage level for the 50% PI, but only one model, UMass-MechBayes, achieved coverage rates within 5% for the 95% PI. In general, coverage rates were lower than the nominal rate (Table 2, Supplemental Figure 2). Six 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 (Supplemental Figure 6).

Models with simple data inputs were some of the most accurate stand-alone models. Of the top five individual models based on relative WIS (UMass-MechBayes, Karlen-pypm, OliverWyman-Navigator, SteveMcConnell-CovidComplete, and GT-DeepCOVID) only two used data beyond the epidemiological hospitalization, case, and death surveillance data from CSSE (Table 1). However, other models that use the same data inputs were not as accurate, so merely including only these inputs was not a sufficient condition for high accuracy. The top five performers consisted of both models with mechanistic components (UMass-MechBayes, Karlen-pypm, OliverWyman-Navigator) and mostly statistical ones (SteveMcConnell-CovidComplete and GT-DeepCOVID). Eight of the 15 individual models that performed better than the baseline used data other than epidemiological surveillance data (e.g., demographics or mobility).

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 some of the other models (Figure 2). We ranked models based on relative WIS for each combination of 1 through 4 week-ahead horizons, 53 weeks, and 55 locations, contributing to 11,726 possible predicted observations for each model. 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 COVID19Sim-Simulator and IowaStateLW-STEM 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 (i.e., too narrow prediction intervals, Supplemental Table 6) resulting in either lower scores (indicating better performance) in instances when 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 (Supplemental Figure 3).

Forecast accuracy relative to the baseline improves as short-term forecast horizon increases

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 weeks ahead compared to a horizon of 1 week ahead; however, models generally showed improved performance relative to the naive baseline model at larger horizons (Figure 3). Ten models showed a lower average WIS (range: 26.9 - 38.9) than the baseline at a 1-week horizon (average WIS = 39.8). At a 4-week ahead horizon, 14 models had a lower average WIS (range: 45.9 - 72.6) than baseline (average WIS = 78.9) . For a 1-week ahead horizon, the model UMass-MechBayes had the lowest average WIS values. At a 4-week horizon, the COVIDhub-ensemble model had the lowest average WIS values. The COVIDhub-ensemble model outperformed the COVIDhub-baseline model at all forecast horizons. Across all models except two, the average

WIS is higher than the median WIS. This is indicative of outlying forecasts impacting the mean value.

When averaging across locations and stratifying by phase of the pandemic (see Methods), there was variation in the top performing models. Seven models outperformed the baseline for both 1- and 4-week ahead targets in at least 2 out of 3 phases (COVIDhub-ensemble, CMU-TimeSeries, GT-DeepCOVID, Karlen-pypm, OliverWyman-Navigator, UMass-MechBayes, and YYG-ParamSearch). Additionally, YYG-ParamSearch, UMass-MechBayes, and COVIDhub-ensemble were the only models to appear in the top three models in two of the three phases analyzed (Supplemental Figure 4).

In contrast to average WIS, prediction interval coverage rates did not change substantially across the 1- to 4-week horizons for most models (Supplemental Figure 2).

Observations on accuracy in specific weeks

Forecasts from individual models showed variation in accuracy by forecast week and horizon (Figure 4). 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 one week where the baseline had lower 1-week ahead error than the ensemble. In weeks where the COVIDhub-ensemble forecast showed its worst probabilistic accuracy, other models also showed lower predictive performance. The COVIDhub-ensemble 1-week ahead forecast for EW 02-2021 (ending January 9, 2021) yielded its highest average WIS across all weeks (average WIS = 76.3), and 8 out of 20 other models that submitted for the same locations outperformed it. The 4-week ahead COVIDhub-ensemble forecasts were worse in EW49-2020 (ending December 5, 2020) than in any other week during the evaluation period (average WIS = 112.4), and 12 out of the 22 models outperformed the ensemble that week at a forecast horizon of 4 weeks. During the holiday period, the reporting patterns may not have reflected the true underlying trends at the time, thus making accurate forecasting particularly difficult.

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 and November through March 2021, Figure 4). 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 (Supplemental File 1). 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.

Individual model forecast performance varies substantially by location

Forecasts from individual models also showed large variation in accuracy by location when aggregated across all weeks and targets (Figure 5). Only the ensemble model showed superior or equivalent accuracy when compared to baseline in all locations. Ensemble forecasts of incident deaths showed the largest relative accuracy improvements in New York (relative WIS =

0.4), California, New Jersey, Ohio, Massachusetts, Indiana, and at the national level (all with relative WIS = 0.5), and the lowest relative accuracy in Vermont (relative WIS = 1.0).

When stratified by phase of the pandemic, YYG-ParamSearch had the lowest relative WIS overall in both the spring and summer phases of the pandemic; that model did not contribute forecasts during the winter phase (Supplemental Figure 5). The COVIDhub-ensemble was still the only model to outperform the baseline in every single location when eligible (summer and winter phases). When stratified by phase, it is more apparent that the quality of data impacted the relative performance of models. For instance, in winter in Ohio, there were large data revisions (Supplemental File 1), which led to a more inaccurate baseline model. Therefore, nearly every model outperformed the baseline during this period. Additional locations in which large data revisions may have impacted baseline accuracy were observed in New York, New Jersey, and Indiana (Supplemental File 1).

Forecast performance at long horizons

While many teams submitted only short-term (1- to 4-week horizon) forecasts, a smaller number of teams consistently submitted longer-term predictions with up to a 20-week horizon for all 50 states (Figure 6). The trends over time from all teams submitting forecasts for the 50 states showed that 4-week ahead forecasts had around 70% more average error than 1-week ahead forecasts, a relationship that was consistent across the entire evaluation period. Longer-term forecasts showed less accuracy on average than 1- and 4-week ahead forecasts. There were not clear overall differences in probabilistic model accuracy between 8- and 20-week horizons, although in early summer 2020 and late spring 2021 average WIS at 8-week horizons were slightly lower than at longer horizons (Figure 6B). For the two teams who made 20-week ahead forecasts for all 50 states, average WIS was 3 to 4.5 times higher at a 20-week horizon than it was at a 1-week 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 (Supplemental Figure 7). 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. No model made forecasts for horizons of 8 or greater that were calibrated at the 95% level, and coverage rates for each model tended to be stable or decline as the horizon increased (Figure 6C).

Discussion

Given the highly visible role that forecasting has played in the response to the COVID-19 pandemic, it is critical that consumers of models, 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 during the COVID-19 pandemic in the US from April 2020 through April 2021. These evaluations were adjusted for regions, time periods, and horizons, and multiple metrics were used.

As has been shown in prior epidemic forecasting projects, ensemble forecasts streamline and simplify the information provided to model consumers, and can provide a stable, accurate, and low-variance forecast (3, 7–9). The results presented here, which show high variation in accuracy between and within stand-alone models but consistent accuracy from an ensemble forecast, support these prior results and confirm that an ensemble model can provide a reliable and comparatively accurate means of forecasting that exceeds the performance of most, if not all, of the models that contribute to it. The ensemble approach was the only model that (a) outperformed the baseline forecast in every location, (b) had better overall 4-week-ahead accuracy than the baseline forecast in every week, and (c) ranked in the top half of forecasts for more than 75% of the forecasts it made. Additionally, it achieved the best overall measures of point and probabilistic forecast accuracy. These results continue to 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 (Figures 2 and 3). However, some consistent patterns of which models were more accurate on average do emerge. Stand-alone models with few data inputs were among the most accurate (Tables 1 and 2). This is consistent with findings from earlier infectious disease forecasting challenges (5, 9, 21). 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 being combined together.
- A simple ensemble forecast that combined all submitted models each week was consistently the most accurate model when performance was aggregated by forecast target (Figure 3), weeks (Figure 4), or locations (Figure 5). Although it was rarely the “most accurate” model for individual predictions, the ensemble was consistently one of the top few models for any single prediction (Figure 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 (Figure 2).
- Forecast accuracy and calibration were substantially degraded at horizons longer than 4 weeks into the future, largely due to underestimating the possibility of increases in incidence at long horizons (Figure 6).

Rigorous evaluation of forecast accuracy faces many limitations in practice. The large variation and correlation in forecast errors across targets, submission weeks, and locations (Figure 3) makes it difficult to create simple and 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 (Figure 6, Supplemental Figure 1). Some teams have also changed their models over time (Table 1, Supplemental Table 1, Supplemental Figure 1). To account for some of this variability, we implemented specific inclusion criteria. However,

those criteria may exclude valuable approaches that were not applied to a large fraction of locations or weeks (see Methods).

Additionally, ground truth data are not static. They can be later revised as more data become available (Supplemental File 1). There are also instances where data is not revised but rather left with large peaks or dips due to reporting effects. This is seen especially around US holidays in which reporting patterns may not reflect the true underlying trends in deaths. It is particularly difficult to model these holiday periods in the context of an emerging infectious disease with new reporting systems and no historical data on how those systems are impacted by holidays from previous years. 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 incident hospitalizations. Deaths may serve as a lagging indicator of COVID-19, thus making it more predictable than hospitalization and case targets.

A key achievement of the COVID-19 Forecast Hub has been providing an ensemble forecast to the US Centers for Disease Control and Prevention in real-time since April 2020. Updated forecasts were featured on the CDC website, an interactive feature on the FiveThirtyEight data-journalism website, and in numerous mass media articles (22, 23). The Forecast Hub website, on average, received more than 40,000 views per month during May – December 2020.

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. However, 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 template infrastructure and resources for arriving at 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 modern modeling capabilities to support local, state, and federal response to outbreaks.

Methods

Surveillance data

During the COVID-19 pandemic in the US, data on cases and deaths were collected by state and local governmental health agencies and aggregated into standardized, sharable formats by third-party data tracking systems. Early in the pandemic, the Johns Hopkins Center for Systems Science and Engineering (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 successive reports of cumulative deaths. 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. For instance, if a public health agency changed the criteria used to classify COVID-19 cases or deaths, it could have resulted in a large number of cases entered in a single day, a negative difference in cumulative counts, or a revision upward or downward in previously reported values. CSSE made attempts to redistribute large “backlogs” of data to previous dates in instances where the true dates of deaths, or dates when the deaths would have been reported, were known, but in some cases, these anomalous observations were left in the final dataset (Supplemental File 1). All major 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 US beginning in April 2020 (26). Any team was permitted to submit a model as long as they provided data in the specified format and included a description of the methods used to generate the forecasts. The descriptions could be updated at any time if a team adopted new methods. The deadline for weekly forecast submission was 6:00 PM ET each Monday.

Submitted forecasts could include predictions for any of the following targets: COVID-19 weekly cumulative deaths, weekly incident deaths, weekly incident cases, and daily incident hospitalizations. Incident death forecasts, the focus of this evaluation, could be submitted with predictions for horizons of 1- 20 weeks after the week in which a forecast was submitted.

As an example of a forecast and the corresponding observation, forecasts submitted between Tuesday, October 6, 2020 (day 3 of EW41-2020) and Monday, October 12, 2020 (day 2 of EW42-2020) contained a “1-week ahead” forecast of incident deaths that corresponded to the total number of deaths observed in EW42-2020, a 2-week ahead forecast corresponded to the total number of deaths in week EW43-2020, and so on. In this paper, we refer to the “forecast week” of a submitted forecast as the week corresponding to a “0-week ahead” target. In the example above, the forecast week would be EW41-2020.

A prediction for a given target (e.g., “1-week ahead incident deaths”) 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

Because forecasts were made for non-stationary processes in locations with different population sizes and scales of observed deaths, forecast accuracy measures that depend on the scale of the observed data are not comparable across time without appropriate normalization. To create a set of standardized comparisons between forecasts, we only included models in our analyses that met specific inclusion criteria. For the 53 weeks beginning in EW17-2020 and ending with EW16-2021, 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-week ahead targets for incident deaths; and
4. contained a complete set of quantiles for all predictions.

Based on the eligibility criteria, we compared 27 models that had at least 32 eligible weeks during this time period (Supplemental Figure 1).

Forecast evaluation period

Forecasts were evaluated based on submissions in a continuous 53-week period starting late April 2020 and ending in late April 2021 (EW17-2020 – EW16-2021, Figure 1). Forecasts were scored using CSSE data available as of May 25, 2021. We did not evaluate forecasts on data first published in the 2 weeks prior to this date due to possible revisions to the data.

In a secondary analysis, forecasts were evaluated based on model submissions during three different phases of the pandemic. A model was eligible for inclusion in a given phase if it met the eligibility criteria outlined above 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 seven out of 11 weeks starting April 21st, 2020 and ending June 29th, 2020. A model was eligible for inclusion in the summer if it had submissions for at least 13 out of 21 submission weeks between June 30th, 2020 and November 17th, 2020. A model was eligible for inclusion in the winter phase if it had submissions for at least 15 out of 24 submission weeks between November 18th, 2020 and April 26th, 2021. These phases were determined based on the waves of deaths at the national level during pandemic (Figure 1b). Each phase includes a period of increasing and decreasing incident deaths, although forecasts for the spring phase were not started early to capture the increase in many locations.

Forecast locations

Forecasts were submitted for 57 locations including all 50 states, 6 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 during the evaluation period, we excluded these locations from our analysis.

In analyses where measures of forecast skill were aggregated across locations, we typically only included the 50 states in the analysis. Other territories and jurisdictions were not included in aggregations because they had relatively few deaths, and very few teams made forecasts for some of these locations (for example, only 11 models submitted forecasts for the Virgin Islands). 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 where scores were stratified by location, we included forecasts for all US states, including territories and the national level.

This evaluation used the CSSE COVID-19 surveillance data as ground truth when assessing forecast performance. Because of the potential impact COVID-19 surveillance data reporting anomalies could have on forecast evaluation, 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 6 times: New Jersey during EW35 of 2020, Arkansas during EW05-2021, West Virginia during EW17-2021, Guam during EW20-2021, and Nebraska during EW152021 and EW20-2021.

We identified two main types of anomalies in the CSSE data: data revisions and outliers. We defined “data revisions” as observations that, after first being reported, were later substantially revised to a new value. A substantial revision was defined as one where (a) the absolute value of the observed difference between the original and updated observation was greater than 20, and (b) the relative difference was greater than or equal to 50%. We defined “outliers” as points that lay substantially far away from the reported data in nearby weeks based on review from two data experts. The goal was to identify observations that models should not be expected to predict accurately, either because the input data at a given time was not reliable due to later revisions or because the target data was not evaluable due to it being a substantial outlier. Supplemental analyses showed that excluding revised or outlying observations did not substantially change the ordering of the models or the overall conclusions of the analysis (Supplemental Table 3).

Forecast models

For the primary evaluation, we compared 27 models that submitted eligible forecasts for at least 32 of the 53 weeks considered in the overall model eligibility period (Figure 1). Teams that submitted to the COVID-19 Forecast Hub used a wide variety of modeling approaches and input data (Table 1, Supplemental Table 1). 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. This predictive distribution was created by collecting, for a particular location, the first differences and their negatives from the previously observed time series (i.e., $y_t - y_{t-1}$ and $-(y_t - y_{t-1})$ for all past times t). To obtain a smoother distribution of values to sample, we formed a distribution of possible differences based on a piecewise linear approximation to the empirical cumulative distribution function of the observed differences. We then obtained a Monte Carlo approximation of the distribution for incident deaths at forecast horizon *h* by independently sampling 100,000 changes in incidence at each week 1, 2, …, *h*, and adding sequences of *h* differences to the most recent observed incident deaths. Quantiles are reported for each horizon, with the median forced to be equal to the last observed value (to adjust for any noise introduced from the sampling process) and the distribution truncated so that it has no negative values.

The COVIDhub-ensemble model combined forecasts from all models that submitted a full set of 23 quantiles for 1- through 4-week ahead forecasts for incident deaths. The ensemble for incident weekly deaths was first submitted in the week ending June 06, 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 (26). We evaluated more complex ensemble methods, and while they did show modest improvements in accuracy, they also displayed undesirable increases in variability in performance (27, 28).

Forecast submission timing

Because this was a real-time forecasting project, forecasts were occasionally submitted late and/or resubmitted. Of the 2,277 forecast submissions we included in the evaluation, 136 (6%) were either originally submitted or updated more than 24 hours after the submission deadline. In all of these situations, modeling teams attested publicly (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 metrics that assessed both point and probabilistic accuracy. Metrics were aggregated over time and locations for near-term forecasts (4 weeks or less into the future) and, in a single analysis, for longer-term projections (5-20 weeks into the future).

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

.

To assess probabilistic forecast accuracy, we used two metrics that are easily computable from the quantile representation for forecasts described above. The weighted interval score (WIS) is a proper score that combines a set of interval scores for probabilistic forecasts that provide quantiles of the predictive forecast distribution. Proper scores promote "honest" forecasting by not providing forecasters with incentives to report forecasts that differ from their true beliefs about the future (29).

Given quantiles of a forecast distribution F, an observation y and an uncertainty level α, a single interval score is defined as

$$IS_{\alpha}(F,y) = (u-l) + \frac{2}{\alpha}\cdot(l-y)\cdot 1(y<l) + \frac{2}{\alpha}\cdot(y-u)\cdot 1(y>u)$$

where $1(\cdot)$is the indicator function and l and u are the $\frac{\alpha}{2}$and $1-\frac{\alpha}{2}$quantiles of F (i.e., the lower and upper end of a central 1 -α prediction interval). Given a set of central prediction intervals, a weighted sum of interval scores can be computed to summarize accuracy across the entire predictive distribution. We define the WIS as a particular linear combination of K interval scores, as

$$WIS_{\alpha_{0:K}}(F,y) = \frac{1}{K+1/2}\cdot\left(w_0\cdot|y-m| + \sum_{k=1}^{K} w_k\cdot IS_{\alpha_k}(F,y)\right)$$

where $w_k = \frac{\alpha_k}{2}$ for $k = 1,\ldots, K$ and $w_0 = 1/2$. In our setting, we used $K = 11$ interval scores, for α =0.02, 0.05, 0.1, 0.2, …, 0.9.

This particular choice of weights for WIS is equivalent to the pinball loss used in quantile regression and has been shown to approximate the commonly used continuous ranked probability score (CRPS) (20). As such, it can be viewed as a distributional generalization of the absolute error, with smaller values of WIS corresponding to forecasts that are more consistent with the observed data (20, 29). WIS can be interpreted as a measure of how close the entire distribution is to the observation, in units on the scale of the observed data. We note that some alternative scores that are commonly used such as CRPS and the logarithmic score cannot be exactly calculated if only a set of quantiles of the predictive distribution are available.

The interval score can be broken into three additive components, in order as they appear in the IS equation above: dispersion, underprediction and overprediction. The WIS can similarly be split into contributions from each of these components, which 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.

We also evaluated prediction interval coverage, the proportion of times a prediction interval of a certain level covered the observed value, to assess the degree to which forecasts accurately characterized uncertainty about future observations. While prediction interval coverage is not a proper score and only assesses one feature of a full predictive distribution, it does provide a clear and interpretable measure of forecast calibration. We compute prediction interval coverage

for a set of observations $(y_i,\ i\ =\ 1,\ \ldots,\ N)$ and prediction interval bounds with an uncertainty level $1\ -\ \alpha,\ (l_{\alpha,i},\ u_{\alpha,i}),\ i\ =\ 1,\ \ldots,\ N$ as

$$\text{prediction interval coverage} = \frac{1}{N}\sum_{i=1}^{N} 1(l_{\alpha,\,i} \le y_i \le u_{\alpha,i}).$$

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 (a) a standardized rank between 0 and 1 for every forecasted observation relative to other models that made the same forecast, and (b) 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,(30–32) as follows:

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

Evaluating a model's standardized ranks across many observations provides a way to evaluate the relative long-run performance of a given model that is not dependent on the scale of the observed data. If all models were equally accurate, distributions of standardized ranks would be approximately uniform.

The following describes a procedure to compute a measure of relative WIS, which evaluates the aggregate performance of one model against the baseline model. 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 procedure described below was also used to compute a relative MAE.

For each pair of models m and m', we computed the pairwise relative WIS skill

$$\theta_{mm'} = \frac{average\ WIS\ of\ model\ m}{average\ WIS\ of\ model\ m'}$$

based on the available overlap of forecast targets. Subsequently, we computed for each model the geometric mean of the results achieved in the different pairwise comparisons, denoted by

$$\theta_m = \left(\prod_{m'=1}^{M} \theta_{mm'}\right)^{1/M}.$$

Then,θ_m is a measure of the relative skill of model m with respect to the set of all other models 1, …, *M* , including the baseline. The central assumption here is that performing well relative to

individual models 1, …, *M* is similarly difficult for each week and location so that no model can gain an advantage by focusing on just some of them. As is, θ_m is a comparison to a hypothetical "average" model. Because we consider a comparison to the baseline model more straightforward to interpret, we rescaled θ_m and reported

$$\theta^*_m = \frac{\theta_m}{\theta_B},$$

where θ_B is the geometric mean of the results achieved by the baseline model in pairwise comparisons to all other models. The quantity θ^*_m then describes the relative performance of model *m*, adjusted for the difficulty of the forecasts model *m* made, and scaled so the baseline model has a relative performance of 1. For simplicity, we refer to θ^*_m as the "relative WIS" or "relative MAE" throughout the manuscript. A value of $0 < \theta^*_m < 1$ means that model *m* is better than the baseline, a value of $\theta^*_m > 1$ means that the baseline is better.

Data and code 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) and the Zoltar forecast archive (https://zoltardata.com/project/44). 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 (v 4.0.2) (33).

Citations

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* (2020) https:/doi.org/10.5281/zenodo.4305938.

5. C. J. McGowan, *et al.*, Collaborative efforts to forecast seasonal influenza in the United States, 2015–2016. *Sci. Rep.* **9**, 683 (2019).

6. N. G. Reich, *et al.*, From the Cover: PNAS Plus: A collaborative multiyear, multimodel assessment of seasonal influenza forecasting in the United States. *Proc. Natl. Acad. Sci. U. S. A.* **116**, 3146 (2019).

7. 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).

8. C. Viboud, *et al.*, The RAPIDD ebola forecasting challenge: Synthesis and lessons learnt. *Epidemics* **22**, 13–21 (2018).

9. 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).

10. S. Funk, *et al.*, Short-term forecasts to inform the response to the Covid-19 epidemic in the UK. *medRxiv*, 2020.11.11.20220962 (2020).

11. K. S. Taylor, J. W. Taylor, A Comparison of Aggregation Methods for Probabilistic Forecasts of COVID-19 Mortality in the United States. *arXiv:2007.11103 [stat]* (2020) (December 2, 2020).

12. J. M. Bates, C. W. J. Granger, The Combination of Forecasts. *J. Oper. Res. Soc.* (2017) https:/doi.org/10.1057/jors.1969.103 (December 24, 2020).

13. T. N. Krishnamurti, *et al.*, Improved Weather and Seasonal Climate Forecasts from Multimodel Superensemble. *Science* **285**, 1548–1550 (1999).

14. T. Gneiting, A. E. Raftery, Weather Forecasting with Ensemble Methods. *Science* **310**, 248–249 (2005).

15. M. Leutbecher, T. N. Palmer, Ensemble forecasting. *J. Comput. Phys.* **227**, 3515–3539

(2008).

16. R. Polikar, Ensemble based systems in decision making. *IEEE Circuits and Systems Magazine* **6**, 21–45 (2006).

17. B. Lakshminarayanan, A. Pritzel, C. Blundell, Simple and Scalable Predictive Uncertainty Estimation using Deep Ensembles. *arXiv:1612.01474 [cs, stat]* (2017) (December 24, 2020).

18. 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**, S404–S408 (2016).

19. J. Friedman, *et al.*, Predictive performance of international COVID-19 mortality forecasting models. *medRxiv*, 2020.07.13.20151233 (2020).

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. Y. Del Valle, *et al.*, Summary results of the 2014-2015 DARPA Chikungunya challenge. *BMC Infect. Dis.* **18**, 245 (2018).

22. R. B. J. Boice, Where The Latest COVID-19 Models Think We're Headed — And Why They Disagree. *FiveThirtyEight* (2020) (January 3, 2021).

23. CDC, COVID-19 Forecasts: Deaths (2021) (January 13, 2021).

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. ,MMWR Weeks. *CDC* (January 13, 2020).

26. E. L. Ray, *et al.*, Ensemble Forecasts of Coronavirus Disease 2019 (COVID-19) in the U.S. *medRxiv*, 2020.08.19.20177493 (2020).

27. L. C. Brooks, *et al.*, Comparing ensemble approaches for short-term probabilistic COVID-19 forecasts in the U.S. *International Institute of Forecasters* (2020) (January 13, 2021).

28. 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) (June 20, 2021).

29. T. Gneiting, A. E. Raftery, Strictly Proper Scoring Rules, Prediction, and Estimation. *J. Am. Stat. Assoc.* **102**, 359–378 (03/2007).

30. 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).

31. S. Wu, F. Crestani, Y. Bi, *Evaluating Score Normalization Methods in Data Fusion* (2006).

32. 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.

33. R Core Team, R: A Language and Environment for Statistical Computing (2020).

34. E. O'Dea, *e3bo/random-walks* (2021) (January 8, 2021).

35. ,COVID-19 Findings, Simulations | Shaman Group (January 13, 2021).

36. S. Pei, J. Shaman, "Initial Simulation of SARS-CoV2 Spread and Intervention Effects in the Continental US" (Epidemiology, 2020) (January 8, 2021).

37. S. Pei, S. Kandula, J. Shaman, Differential effects of intervention timing on COVID-19 spread in the United States. *Science Advances* **6**, eabd6370 (2020).

38. A. Rodríguez, *et al.*, DeepCOVID: An Operational Deep Learning-driven Framework for Explainable Real-time COVID-19 Forecasting in *Proceedings of the 35th AAAI Conference on Artificial Intelligence*.

39. L. Wang, *et al.*, Spatiotemporal Dynamics, Nowcasting and Forecasting of COVID-19 in the United States. *arXiv:2004.14103 [stat]* (2020) (January 7, 2021).

40. J. C. Lemaitre, *et al.*, A scenario modeling pipeline for COVID-19 emergency planning. *medRxiv*, 2020.06.11.20127894 (2020).

41. ,Case studies and reports (January 13, 2021).

42. D. Karlen, Characterizing the spread of CoViD-19. *arXiv:2007.07156 [physics, q-bio, stat]* (2020) (January 7, 2021).

43. LANL COVID-19 Cases and Deaths Forecasts (January 8, 2021).

44. M. Chinazzi, *et al.*, The effect of travel restrictions on the spread of the 2019 novel coronavirus (COVID-19) outbreak. *Science* **368**, 395–400 (2020).

45. EpiGro & EpiCovDA (January 13, 2021).

46. D. Wu, *et al.*, DeepGLEAM: A hybrid mechanistic and deep learning model for COVID-19 forecasting. *arXiv [cs.LG]* (2021).

47. D. Wu, *et al.*, Quantifying Uncertainty in Deep Spatiotemporal Forecasting in *In Proceedings of the 27th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining.*, (2021).

48. G. C. Gibson, N. G. Reich, D. Sheldon, REAL-TIME MECHANISTIC BAYESIAN FORECASTS OF COVID-19 MORTALITY. *medRxiv*, 2020.12.22.20248736 (2020).

49. P. Keskinocak, B. E. Oruc, A. Baxter, J. Asplund, N. Serban, The impact of social distancing on COVID19 spread: State of Georgia case study. *PLoS One* **15**, e0239798 (2020).

50. A. Baxter, B. E. Oruc, P. Keskinocak, J. Asplund, N. Serban, Evaluating scenarios for school reopening under COVID19. *bioRxiv* (2020)

https:/doi.org/10.1101/2020.07.22.20160036.

51. J. Baek, *et al.*, The Limits to Learning an SIR Process: Granular Forecasting for Covid-19. *arXiv:2006.06373 [cs, stat]* (2020) (January 7, 2021).

52. M. A. Rowland, *et al.*, COVID-19 infection data encode a dynamic reproduction number in response to policy decisions with secondary wave implications. *Sci. Rep.* **11**, 10875 (2021).

53. Y. Zhang-James, *et al.*, A seq2seq model to forecast the COVID-19 cases, deaths and reproductive R numbers in US counties. *medRxiv* (2021) https:/doi.org/10.1101/2021.04.14.21255507.

54. A. Srivastava, T. Xu, V. K. Prasanna, Fast and Accurate Forecasting of COVID-19 Deaths Using the SIkJα Model. *arXiv:2007.05180 [physics, q-bio]* (2020) (January 8, 2021).

55. A. Srivastava, V. K. Prasanna, Data-driven Identification of Number of Unreported Cases for COVID-19: Bounds and Limitations. *arXiv:2006.02127 [cs, q-bio]* (2020) (January 8, 2021).

56. COVID-19 Projections Using Machine Learning (January 13, 2021).

Funding

For teams that reported receiving funding for their work, we report the sources and disclosures below.

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: This work has been supported by the US Centers for Disease Control and Prevention (1U01IP001122) and the National Institutes of General Medical Sciences (R35GM119582). The content is solely the responsibility of the authors and does not necessarily represent the official views of CDC, NIGMS or the National Institutes of Health. Johannes Bracher was supported by the Helmholtz Foundation via the SIMCARD Information & Data Science Pilot Project. Tilmann Gneiting gratefully acknowledges support by the 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 CSTE NU38OT000297, research cyberinfrastructure resources and services provided by the Partnership for an Advanced Computing Environment (PACE) at Georgia Tech, and the following benefactors at Georgia Tech: Andrea Laliberte, Joseph C. Mello, Richard “Rick” E. & Charlene Zalesky, and Claudia & Paul Raines. Council of State and Territorial Epidemiologists and the Centers for Disease Control and Prevention. The content is solely the responsibility of the authors and does not necessarily represent the official views of the CSTE, CDC, or the universities employing the researchers.
GT-DeepCOVID: CDC MInD-Healthcare U01CK000531-Supplement. NSF (Expeditions

CCF-1918770, CAREER IIS-2028586, RAPID IIS-2027862, Medium IIS-1955883, NRT DGE-1545362), CDC MInD program, ORNL and funds/computing resources from Georgia Tech and GTRI.
IHME: This work was supported by the Bill & Melinda Gates Foundation, as well as funding from the state of Washington and the National Science Foundation (award no. 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.
JHU CSSE: National Science Foundation (NSF) RAPID "Real-time Forecasting of COVID-19 risk in the USA". 2021-2022. Award ID: 2108526. National Science Foundation (NSF) RAPID "Development of an interactive web-based dashboard to track COVID-19 in real-time". 2020. Award ID: 2028604
JHU_IDD-CovidSP: State of California, US Dept of Health and Human Services, US Dept of Homeland Security, US Office of Foreign Disaster Assistance, Johns Hopkins Health System, Office of the Dean at Johns Hopkins Bloomberg School of Public Health, Johns Hopkins University Modeling and Policy Hub, Centers for Disease Control and Prevention (5U01CK000538-03), University of Utah Immunology, Inflammation, & Infectious Disease Initiative (26798 Seed Grant).
LANL-GrowthRate: LANL LDRD 20200700ER.
MOBS-GLEAM_COVID: COVID Supplement CDC-HHS-6U01IP001137-01; Cooperative Agreement number NU38OT000297 from The Centers for Disease Control and Prevention (CDC) and CSTE
NotreDame-mobility and *NotreDame-FRED*: NSF RAPID DEB 2027718
PSI-DRAFT: NSF RAPID Grant # 2031536.
UA-EpiCovDA: NSF RAPID DMS 2028401.
UCSB-ACTS: NSF RAPID IIS 2029626.
UCSD-NEU: Google Faculty Award, DARPA W31P4Q-21-C-0014, COVID Supplement CDC-HHS-6U01IP001137-01.
UMass-MechBayes: NIGMS R35GM119582, NSF 1749854.
UMich-RidgeTfReg: The University of Michigan Physics Department and the University of Michigan Office of Research.

Tables and Figures

Table 1: List of models evaluated, including sources for case, hospitalization, death, demographic and mobility data when used as inputs for the given model. We evaluated 27 models contributed by 26 teams. The COVIDhub team submitted two models including the baseline model and the ensemble model. A brief description is included for each model, with a reference where available. The last column indicates whether the model made assumptions about how and whether social distancing measures were assumed to change during the period for which forecasts were made.

	Data Sources Included					Model Information	
Team-Model	**Cases**	**Hosp.**	**Deaths**	**Demog.**	**Mob.**	**Description**	**Assumes social distancing measures change in the future**
CEID-Walk			J			Random walk model starting from the most recent observation with a dispersion based on the spread of the last 5 observations (34)	No
CMU-TimeSeries	J		J			A basic autoregressive-type time series model fit using case counts and deaths as features	No
COVIDhub-baseline			J			Median prediction at all future horizons is equal to the most recent observed incidence	No
COVIDhub-ensemble						Unweighted average or median of submitted forecasts to the COVID-19 Forecast Hub (26)	No
Covid19Sim-Simulator	J	CTP	J			SEIR model accounting for undiagnosed infections	No
CU-select	J, UF	CTP, HHS	J, UF	Cen	SG, Cen	Metapopulation county-level SEIR model (35–37)	Yes
DDS-NBDS	J		J			Negative binomial distribution based generalized linear dynamical system	No
epiforecasts-ensemble1	J		J			Mean ensemble of three models: an Rt-based forecast, a timeseries forecast using deaths only and a timeseries forecast using deaths and cases	No
GT-DeepCOVID	CTP	CTP, HHS, CN	J		G,A	Data-driven approach based on deep learning for forecasting mortality and hospitalizations (38)	No

IHME-SEIR[a]	J, CTP	CTP, HHS	J, CTP	GBD	SG, G, USDT, FB	Ensemble spline model to estimate past infections combined with covariate-driven deterministic SEIR model	Yes
IowaStateLW-STEM[b]	J,NYT		J,NYT	Cen	USDT	Nonparametric space-time disease transmission model (39)	No
JHUAPL-Bucky	J	HHS	J	Cen	SG, PIQ	Spatial compartment model using public mobility data and local parameters	Yes
JHU_IDD-CovidSP[c]	J,UF		J,UF	Cen	Cen	Metapopulation model with commuting, nonpharmaceutical interventions, and stochastic SEIR disease dynamics (40)	No
Karlen-pypm	J	HHS	J			Finite time difference equations implemented as a general-purpose population modelling framework (41, 42)	No
LANL-GrowthRate[d]	J		J			Statistical dynamical growth model accounting for population susceptibility (43)	No
MOBS-GLEAM_COVID	J	HHS	J	Cen	G	Metapopulation, age-structured SLIR model with mobility and nonpharmaceutical interventions (44)	Yes
OliverWyman-Navigator	J		J	Cen		Compartmental formulation with non-stationary transition rates	Blended. (No for immediate term up to next 3 weeks. Yes for longer term.)
PSI-DRAFT			J	Cen		Age-stratified compartmental SEIRX model with time-dependent reproduction number	No
RobertWalraven-ESG	J		J			Multiple skewed gaussian mathematical fit	No
RPI_UW-Mob_Collision			J		G	A mobility-informed simplified SIR model	No

						motivated by collision theory.	
SteveMcConnell-CovidComplete	CTP		J, CTP	Cen		Multiple proxy-based forecast models with positive tests and past deaths used as proxies for future deaths; ongoing accuracy evaluation of each model; voting algorithms based on past performance used to select specific forecast models each week, selected state by state; ; most forecasts are error-corrected based on errors in past forecasts	No
UA-EpiCovDA[e]	CTP, J		CTP, J			SIR mechanistic model with data assimilation (45)	No
UCLA-SuEIR	J	CTP	J			SEIR model variant considering both untested and unreported cases	Yes
UCSD_NEU-DeepGLEAM	J	HHS	J	Cen	G	Combines the signal of a discrete stochastic epidemic computational model with a deep learning spatiotemporal forecasting framework (46, 47)	Yes
UMass-MechBayes	J		J			Bayesian compartmental model with observations on incident case counts and incident deaths (48)	No
UMich-RidgeTfReg[f]	J		J		G	Ridge regression model using confirmed case and death reports to generate predictions	No
UT-Mobility			J		SG	Bayesian multilevel negative binomial regression model	No

A = Apple mobility (https://covid19.apple.com/mobility), Cen = US Cen (https://www.census.gov/), CN = Coronavirus Disease 2019 (COVID-19)-Associated Hospitalization Surveillance Network (COVID-NET) (https://www.cdc.gov/coronavirus/2019-ncov/covid-data/covid-net/purpose-methods.html), CTP = COVID Tracking Project (https://covidtracking.com/), DL= Descartes Labs (https://github.com/descarteslabs/DL-COVID-19), FB =

Facebook (https://visualization.covid19mobility.org/), G = Google mobility (https://www.google.com/covid19/mobility/), GBD = Global Burden of Disease project (http://www.healthdata.org/gbd/2019), HHS = Health and human services hospitalizations (https://protect-public.hhs.gov/pages/covid19-module), J = JHU CSSE (https://github.com/CSSEGISandData/COVID-19)(24), NYT = New York Times (https://github.com/nytimes/covid-19-data), SEIR = Susceptible-Exposed-Infectious-Recovered compartmental model, SG = SafeGraph mobility (https://www.safegraph.com/), SIR = Susceptible-Infectious-Recovered compartmental model, UF = USA Facts (https://usafacts.org/visualizations/coronavirus-covid-19-spread-map/), USDT = U.S. Department of Transportation Bureau of Transportation Statistics (https://www.transportation.gov/connect/available-datasets)

[a]The IHME-SEIR model on 2020-06-24 switched from curve fitting for past infections and SEIR model for infection projections to using an ensemble spline model to estimate past infections combined with covariate-driven deterministic SEIR model

[b]The IowaStateLW-STEM model on 2020-07-27 switched from using the NYT data to JHU CSSE data and started incorporating mobility data.

[c]The JHU_IDD-CovidSP model on 2020-12-14 switched to using JHU CSSE data only for cases and deaths.

[d]The LANL-GrowthRate model on 2020-10-28 switched from a Bayesian hierarchical approach to share information between states to fitting each state separately for improved computational time.

[e]The UA-EpiCovDA model on 2020-07-05 switched the way the initial conditions were being estimated. After March 8, 2021, forecasts were updated using JHU CSSE instead of CTP.

[f]The UMich-RidgeTfReg model on 2020-11-30 started to incorporate social mobility data.

Table 2: Summary accuracy metrics for all submitted forecasts from 27 models meeting inclusion criteria, aggregated across locations (50 states only), submission week, and 1- through 4-week forecast horizons. The '# 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. If the model is approximately well calibrated, the values in these columns should be close to 0.50 and 0.95, respectively (values within 5% coverage of the nominal rates are highlighted in boldface text). The "relative WIS" and "relative MAE" columns show the relative mean weighted interval score (WIS) and relative mean absolute error (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).

Model	# forecasts	95% PI Cov.	50% PI Cov.	Relative WIS	Relative MAE
CEID-Walk	7135	0.81	**0.46**	0.95	1.00
CMU-TimeSeries	7840	0.72	0.39	0.79	0.80
Covid19Sim-Simulator	8886	0.27	0.08	1.01	0.81
COVIDhub-baseline	10284	0.84	0.44	1.00	1.00
COVIDhub-ensemble	9084	0.87	**0.47**	**0.61**	**0.66**
CU-select	8684	0.68	0.34	0.96	0.93
DDS-NBDS	7485	0.84	0.40	1.16	1.52
epiforecasts-ensemble1	7028	0.86	**0.45**	3.88	3.17
GT-DeepCOVID	8684	0.82	0.37	0.77	0.85
IHME-CurveFit	6937	0.64	0.27	0.77	0.80
IowaStateLW-STEM	7761	0.44	0.18	1.03	0.92
JHU_IDD-CovidSP	9623	0.80	0.36	0.88	0.99
JHUAPL-Bucky	6488	0.53	0.24	1.10	1.09
Karlen-pypm	7884	0.84	0.44	0.66	0.72
LANL-GrowthRate	8884	0.89	0.40	0.80	0.89
MOBS-GLEAM_COVID	10276	0.67	0.35	0.81	0.80
OliverWyman-Navigator	9060	0.83	0.44	0.70	0.74
PSI-DRAFT	8272	0.34	0.14	1.47	1.24
RobertWalraven-ESG	8254	0.37	0.21	1.26	1.03

RPI_UW-Mob_Collision	4157	0.55	0.22	1.34	1.29
SteveMcConnell-CovidComplete	6887	0.79	**0.49**	0.74	0.78
UA-EpiCovDA	8684	0.64	0.33	0.93	0.88
UCLA-SuEIR	8191	0.24	0.08	1.28	1.08
UCSD_NEU-DeepGLEAM	6488	0.87	**0.55**	0.81	0.79
UMass-MechBayes	9884	**0.94**	0.56	**0.61**	**0.66**
UMich-RidgeTfReg	6823	0.45	0.23	1.30	1.13
UT-Mobility	7143	0.67	0.30	2.66	2.45

Figure 1: Overview of the evaluation period included in the paper. Vertical dashed lines indicate "phases" of the pandemic analyzed separately in the supplement. (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 May 1, 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.

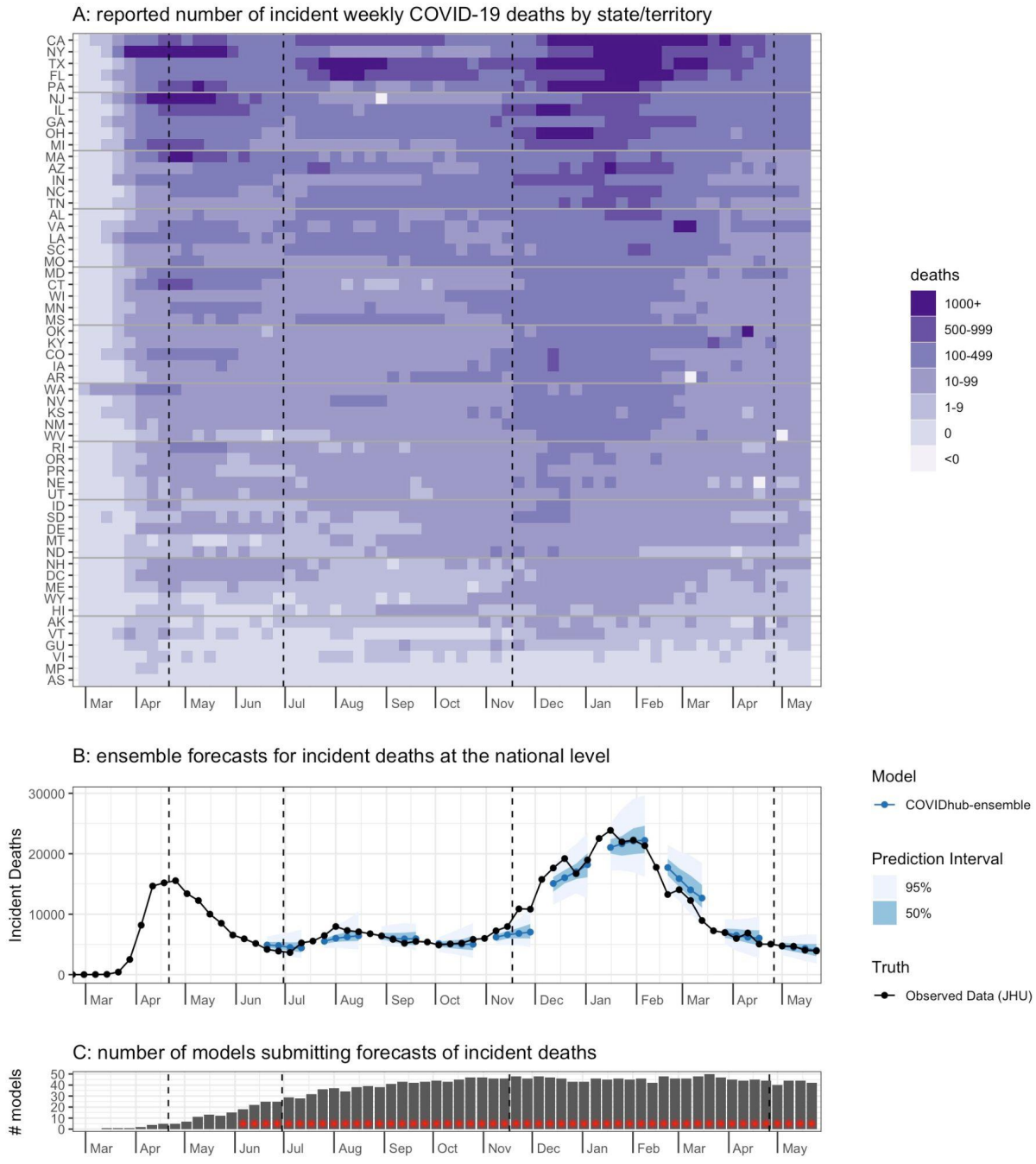


Figure 2: A comparison of each model's distribution of standardized rank of weighted interval scores (WIS) for each location-target-week observation. A standardized rank of 1 indicates that the model had the best WIS for that particular location, target, and week and a value of 0 indicates it had the worst WIS. The density plots show smoothly 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. The COVIDhub-ensemble was the only model that ranked in the top half of all models (standardized rank > 0.5) for over 75% of the observations it forecasted. Observations in this figure included predictions for the national level, all 50 states, and 5 US territories.

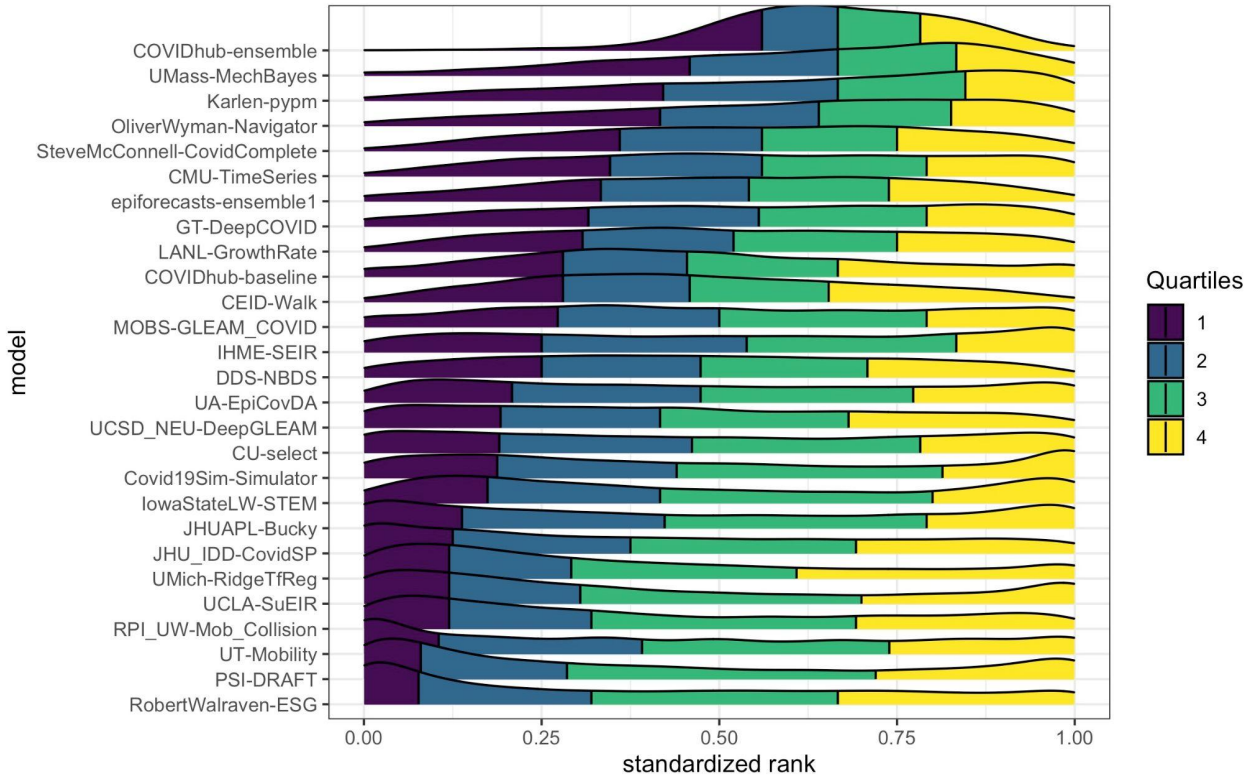


Figure 3: Boxplot distributions of average weighted interval score (WIS, y-axis on log scale) for each week and model across all 50 states. The two panels represent 1 and 4 week ahead forecast horizons. The boxplots summarize the distribution of average WIS values for each week, averaging across all available locations for each model. The "x" marks indicate the average WIS for each model. Models are ordered along the x-axis by their relative WIS (Table 2). The horizontal dashed lines indicate the average baseline WIS for each horizon.

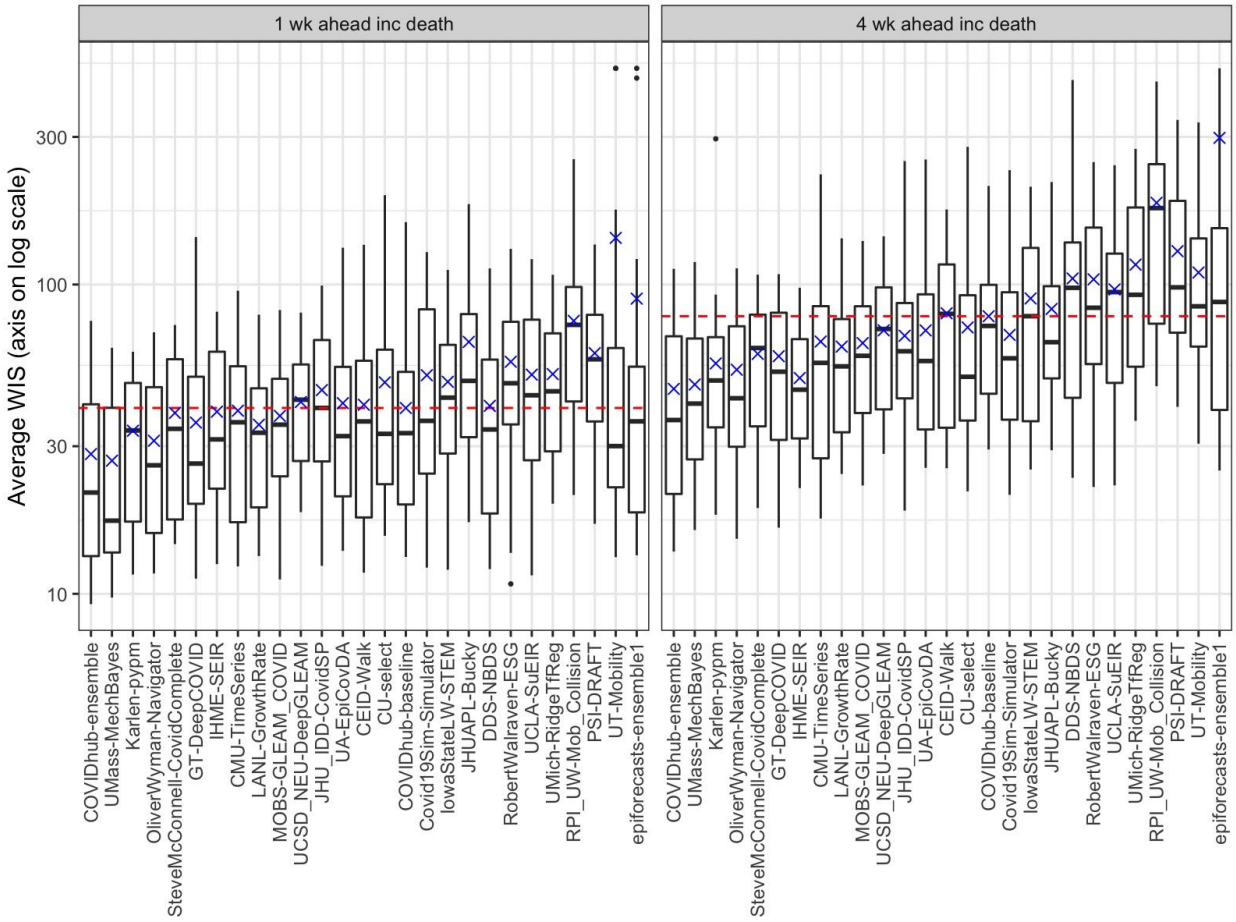


Figure 4: Average WIS by the target forecasted week for each model across all 50 states. Panel A shows the observed weekly COVID-19 deaths based on the CSSE reported data as of May 25, 2021. Panel B shows the average 1-week ahead WIS values per model (in grey). For all 21 weeks in which the ensemble 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). Across submission weeks, there is variation in the WIS for each model. The WIS for each model is lowest in weeks where there are stable and low numbers of incident deaths. The y-axes are truncated in panels B and C for readability of the majority of the data.

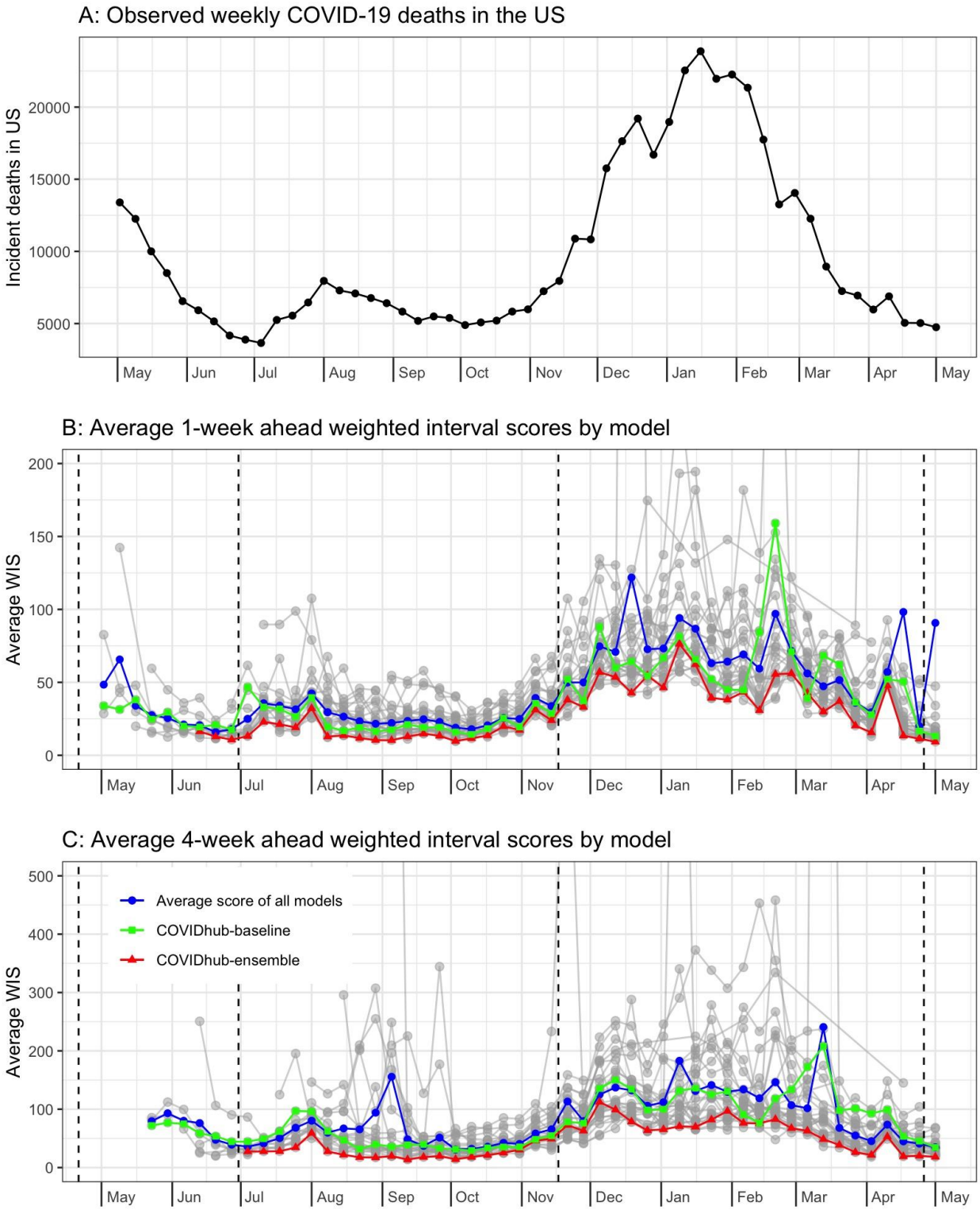


Figure 5: Relative WIS by location for each model across all horizons and submission weeks. The value in each box represents the relative WIS calculated from 1- to 4-week ahead targets available for a model at each location. Boxes are colored based on the relative WIS compared to the baseline model (θ^*_m, see Methods). Blue boxes represent teams that outperformed the baseline and red boxes represent teams that performed worse than the baseline, with darker hues representing performance further away from the baseline. Locations are sorted by cumulative deaths as of the end of the evaluation period (May 1, 2021). Teams are listed on the horizontal axis in order from the lowest to highest relative WIS values (Table 2). The COVIDhub-ensemble achieved the lowest relative WIS overall and performed at least as well as the baseline in every location.

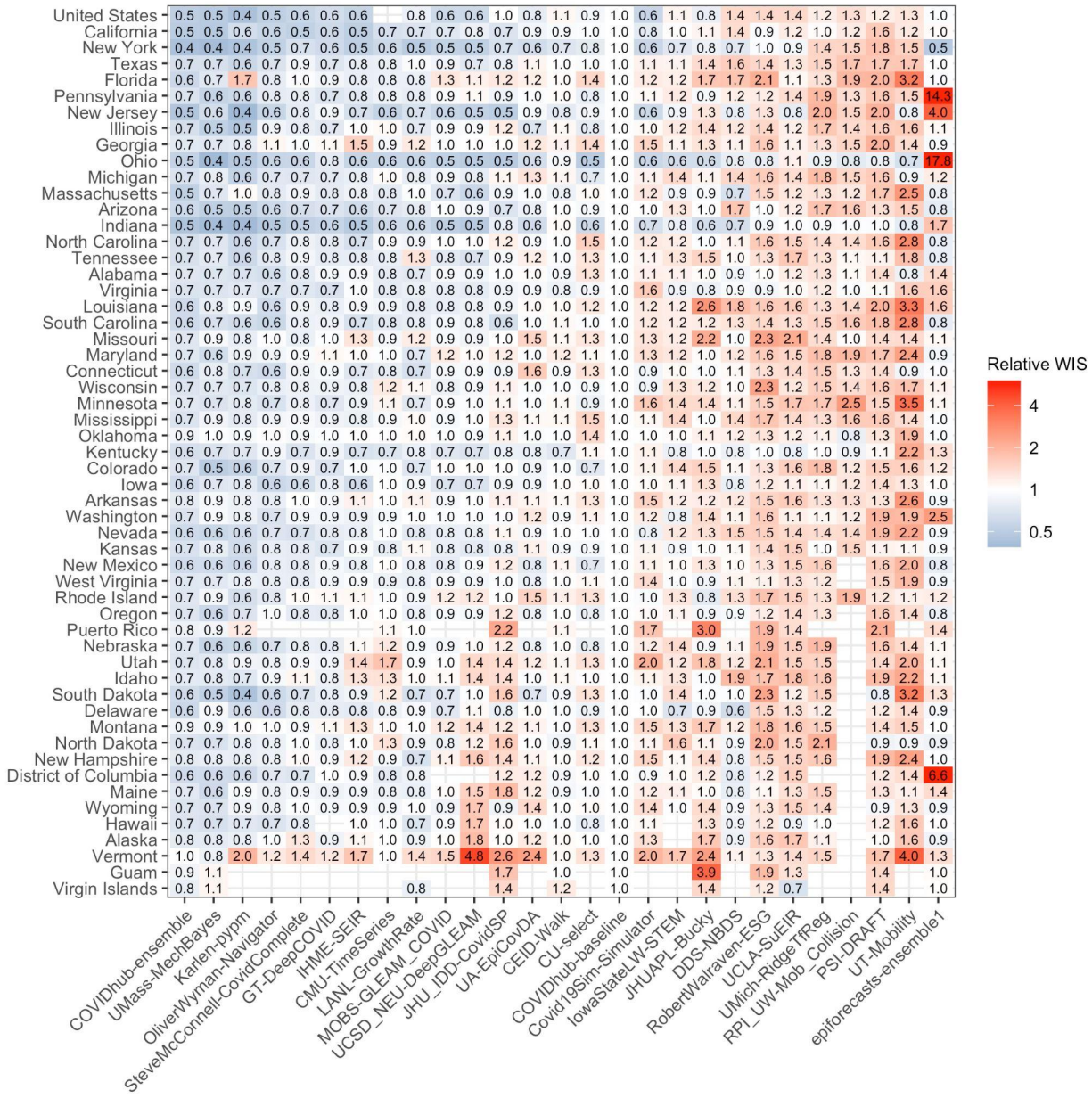
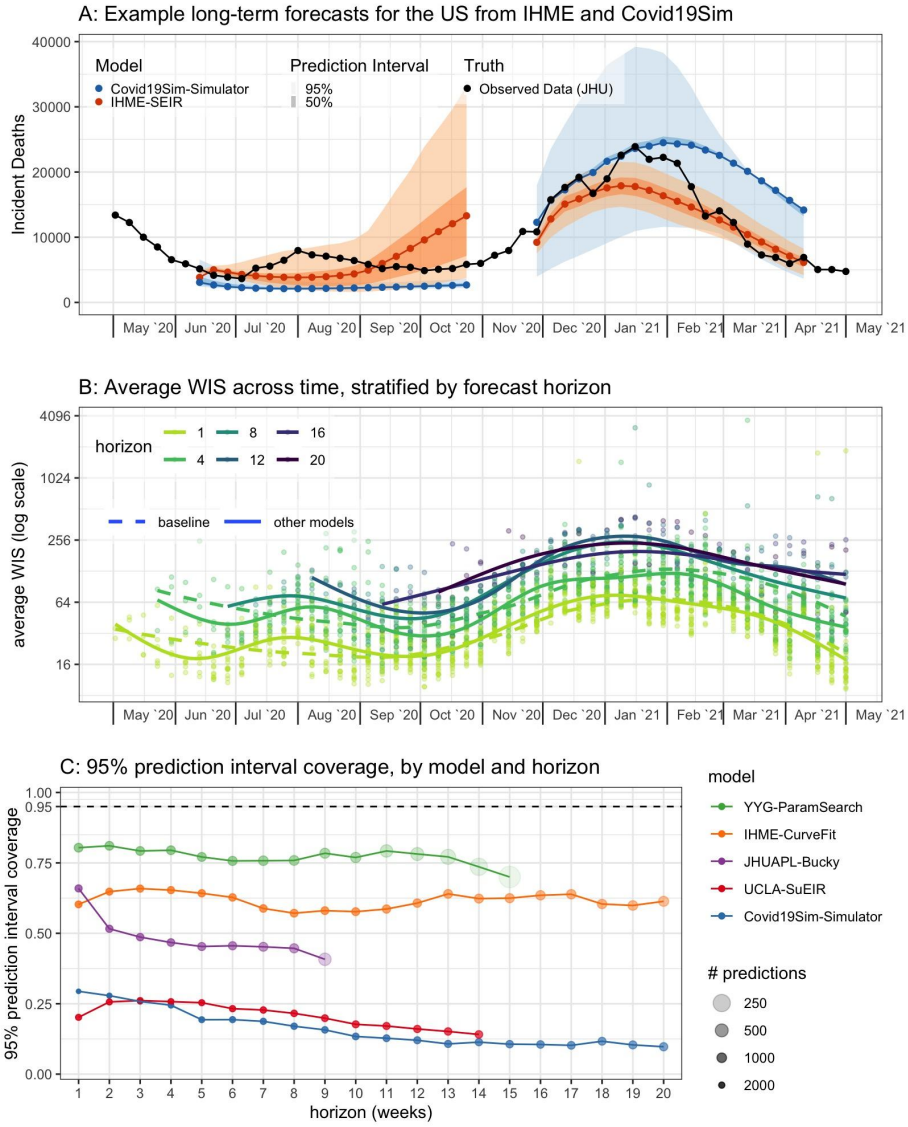


Figure 6: Evaluation of long-range forecast performance. (A) Example 20-week-ahead probabilistic forecasts submitted in early June and late November 2020. (B) Points show values of average WIS for specific models and target forecast week across all states. The solid line shows the smooth trend in average WIS across all non-baseline models, and the dashed line shows the trend for the baseline model (horizons 1 and 4 only). Lines are colored by horizon, with darker lines indicating forecasts targeting weeks further in the future. Across all weeks, average WIS tends to be about twice as high for 4-week ahead as it is for 1-week ahead forecasts. For later weeks, when forecasts at all horizons are able to be evaluated, forecasts for horizons above 8 weeks tend to have about double the average WIS as was achieved at a 4-week ahead horizon. (C) 95% prediction interval coverage rates across horizons for a subset of five models that consistently submitted for more than 8 weekly horizons. Coverage rates for 8- through 20-week ahead horizons were all below the nominal 95%. The horizontal dashed line shown at 0.95 indicates the expected coverage rate. The size of points indicates the number of predictions the coverage rates are based on: smaller points indicate more observations and therefore less variance in the estimated coverage rate.

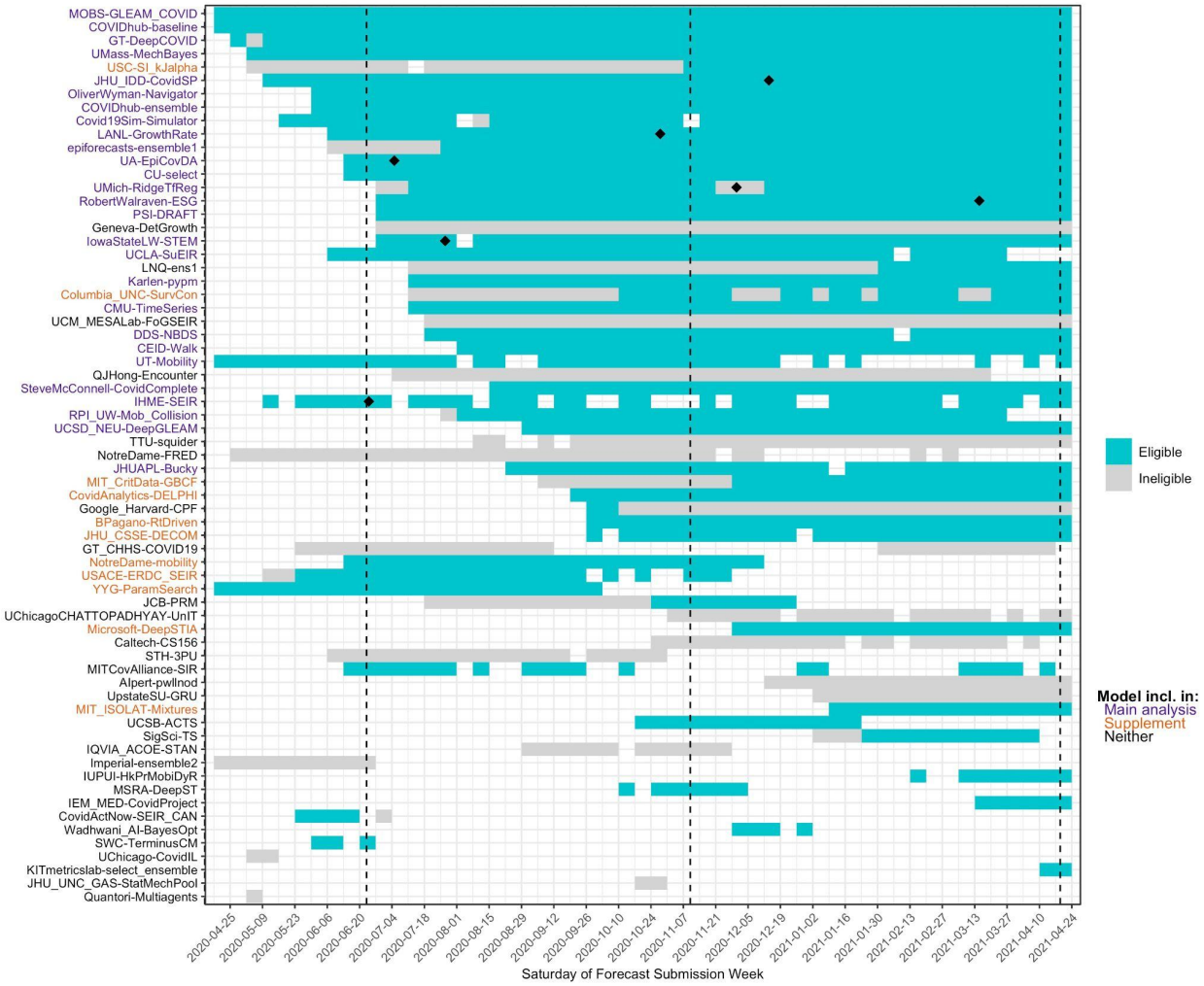


Supplemental Data

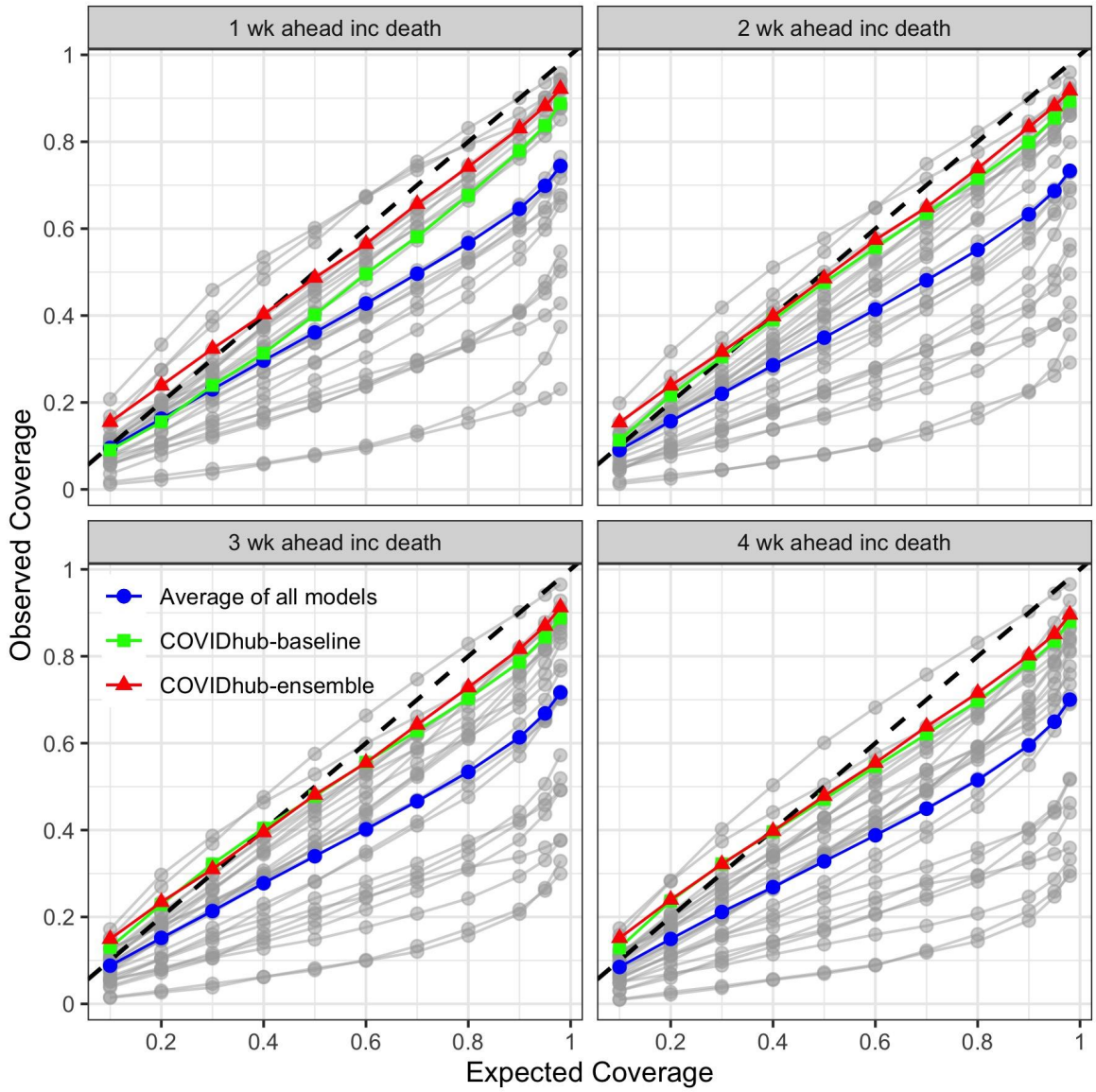
Supplemental File 1: Observed incident deaths in US states and territories over multiple revision dates. Weekly incident deaths are shown for each Monday from June 1, 2020 through May 24, 2021. The latest data revision is shown in pink. If there have been no data revisions in a location, there is a single pink line. If there have been data revisions, additional lines will show on the graph, indicating the last week prior to the data revision. Out of all states and territories evaluated, twelve report data revisions. In some instances, the change is minor, such as in Washington DC and South Carolina where there is a back-distribution of fewer than 5 deaths occurring in a single week. In other locations, larger revisions occurred; for example, in New Jersey a large number of retrospective deaths were initially added to EW 26 of 2020, then later back-distributed over a series of weeks in which the deaths actually occurred. Similarly, Rhode Island had a large number of delayed deaths that were backfilled leading to a discrepancy between the reported incident cases over revision dates from EW 15 to EW 35 of 2020. In locations where it was unknown when the deaths occurred, the spikes in data were not revised. This occurred in Delaware during EW 26 and EW 35 of 2020. Additional information on the causes of the anomalous data reporting and modification dates can be found in the CSSE GitHub repository (https://github.com/CSSEGISandData/COVID-19/tree/master/csse_covid_19_data).

https://github.com/reichlab/covid19-forecast-evals/blob/main/figures/data_revisions.pdf

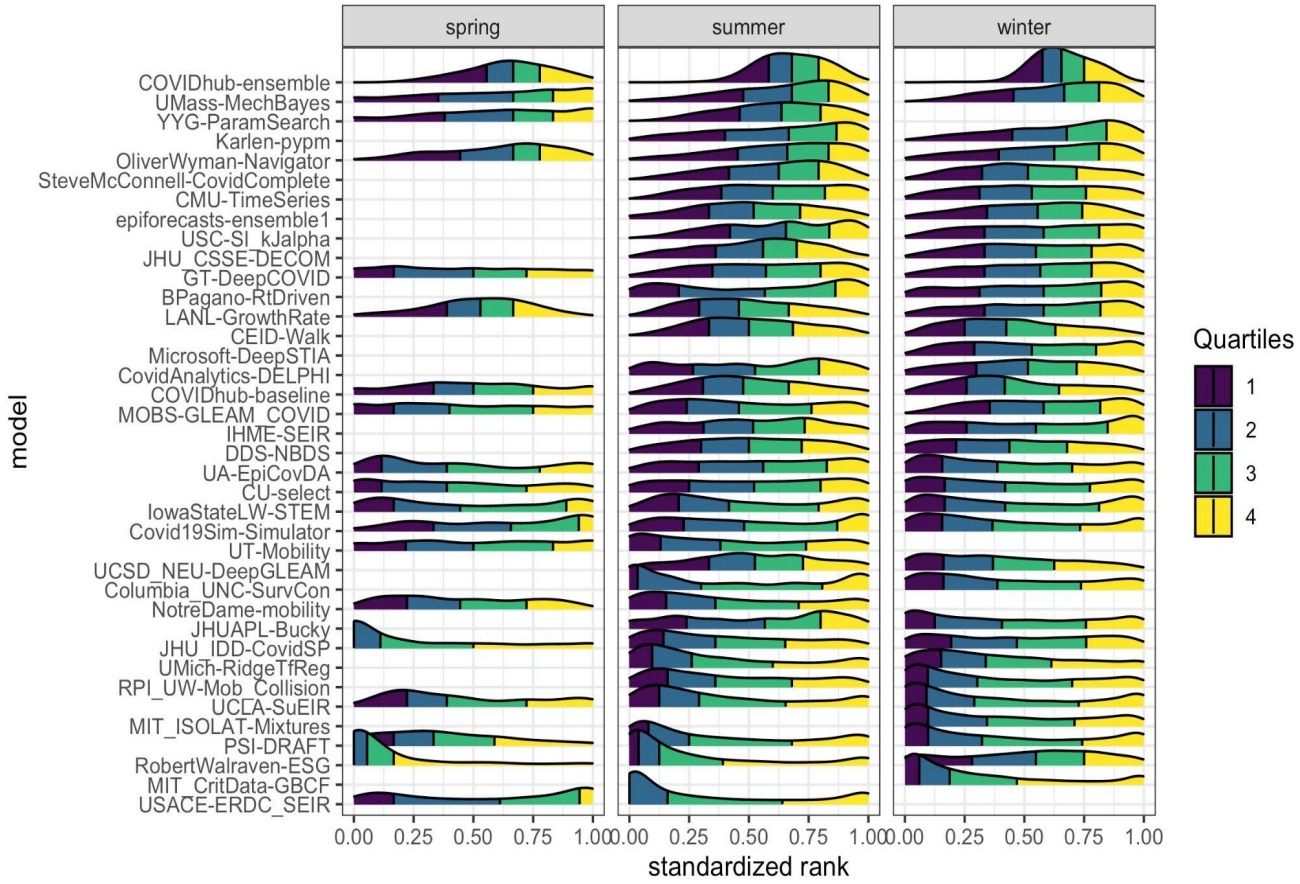
Supplemental Figure 1: Models contributed incident mortality forecasts to the COVID-19 Forecast Hub and were evaluated for eligibility for the analysis in this manuscript. Each cell represents the weekly submission from a particular model (row) in a particular week (column). Forecasts that were determined to be an eligible submission, based on forecasting for at least 25 locations and all of the 1 - 4 week horizons and submitting all quantiles, are highlighted in light blue. Submissions that are not eligible are shown in grey. Model names in purple indicate the teams included in the overall evaluation in the main text. Model names in orange indicate teams that were only included in a phase-specific evaluation included in the supplemental information. Models in black were not evaluated individually at any point. Vertical dashed lines demarcate the three "phases" evaluated separately in the supplement. Black diamonds indicate the timepoints at which models were altered (Table 1 footnotes).



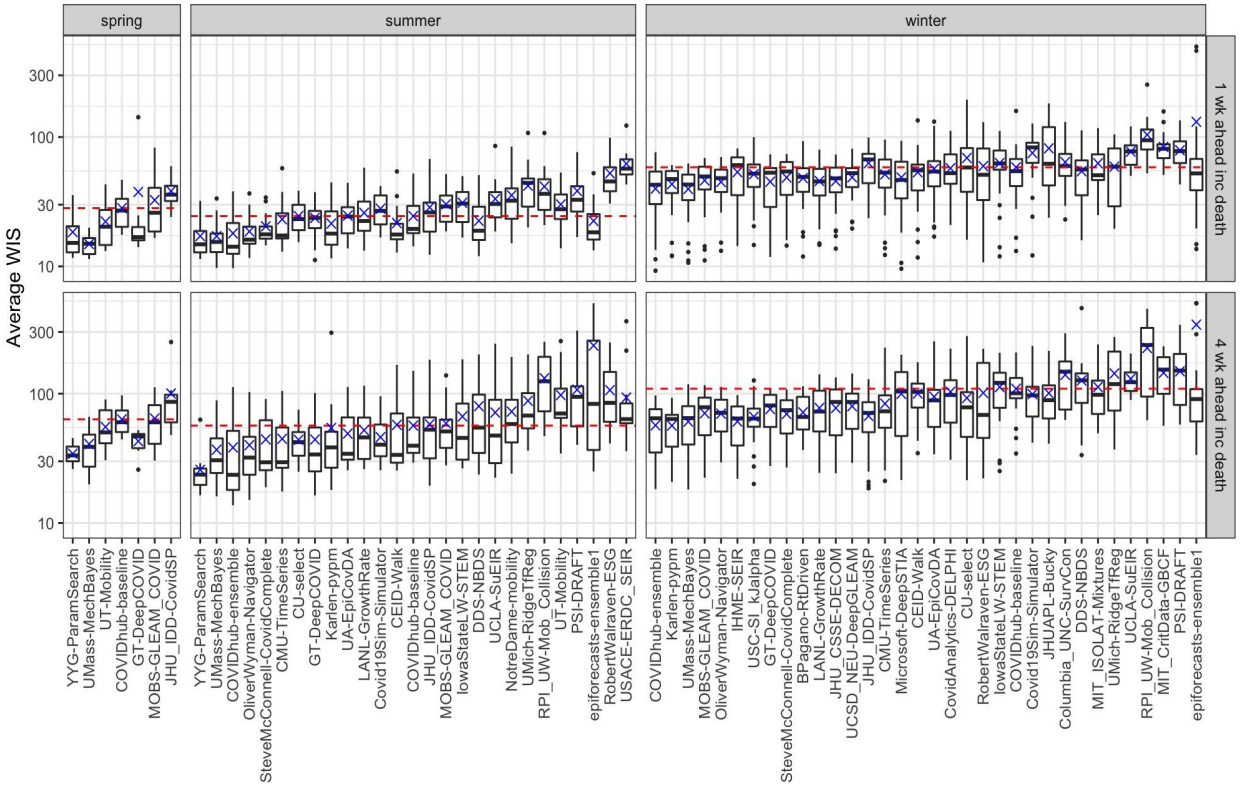
Supplemental Figure 2: Prediction interval coverage for all submitted forecasts with horizons of 1 through 4 weeks, aggregated across submission date, location, and week. Forecasts for any available forecasted location (nation, state, or territory) were included in this analysis. Points indicate PI coverage rates at nominal PI levels of 10%, 20%, 30%, ... 90%, 95%, and 98%. If a model is well calibrated across all PIs, the values should be close to the dashed black line, representing the expected PI coverage. As seen in each panel, few models (grey) have an observed coverage rate at or above the expected coverage rate. When averaged across all models (blue circle) the PI coverage falls below the expected coverage at every level at every horizon. The ensemble (red triangle) is better calibrated than the baseline model (green square) and the model average across nearly all PI levels for 1-week ahead and more than half of the levels for 4-week ahead horizons.



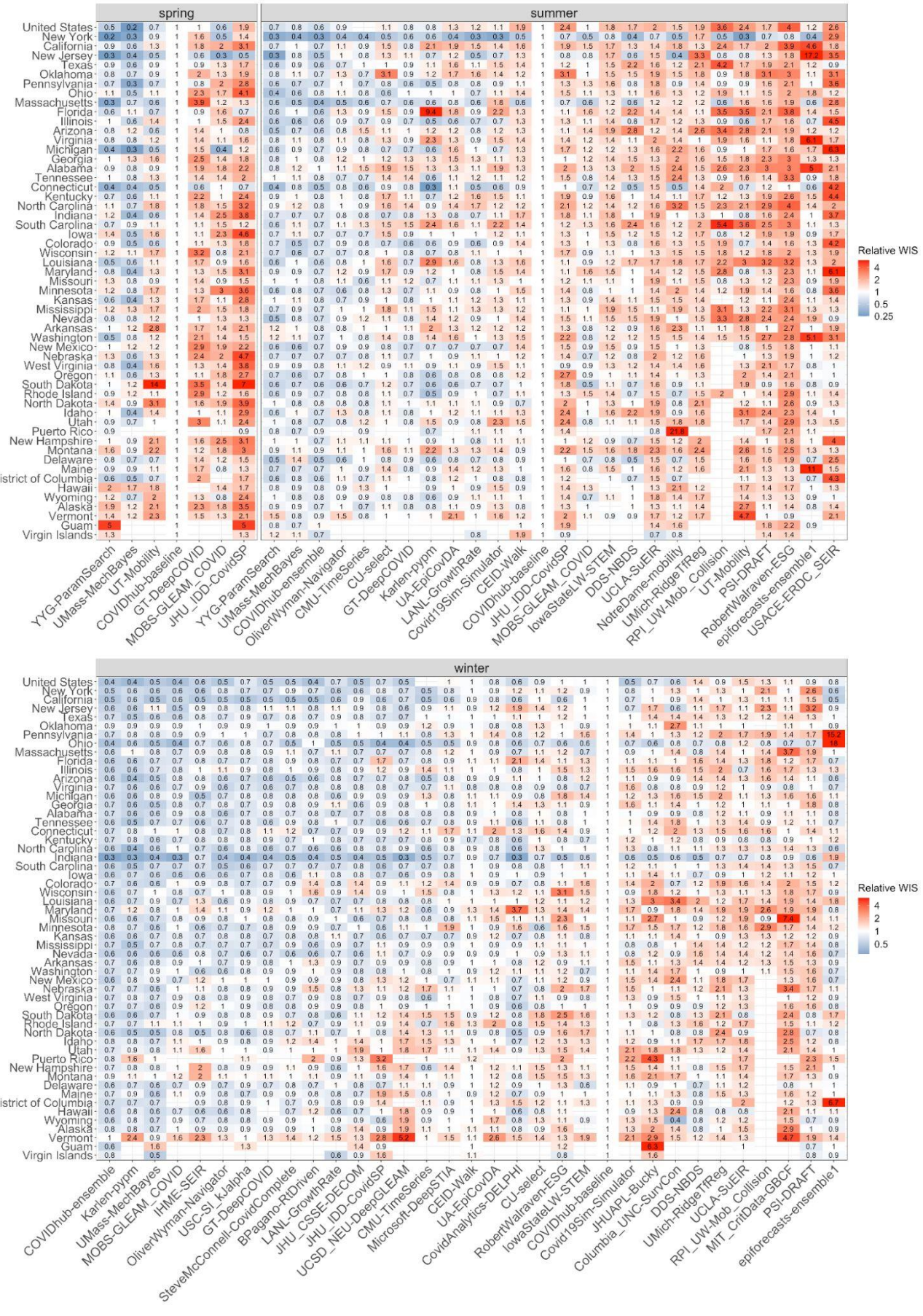
Supplemental Figure 3: A comparison of each model's distribution of standardized rank of weighted interval scores (WIS) for each location-target-week observation stratified by three phases of the pandemic. A standardized rank of 1 indicates that the model had the best WIS for that particular location, target, and week and a value of 0 indicates it had the worst WIS. The density plots show smoothly 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 region containing the bottom quarter of the distribution. The models are ordered by the overall first quartile of the distribution with models that rarely had a low rank near the top. Observations in this figure included predictions for the national level, all 50 states, and 5 US territories. If models were equally accurate, all distributions would be approximately uniform.



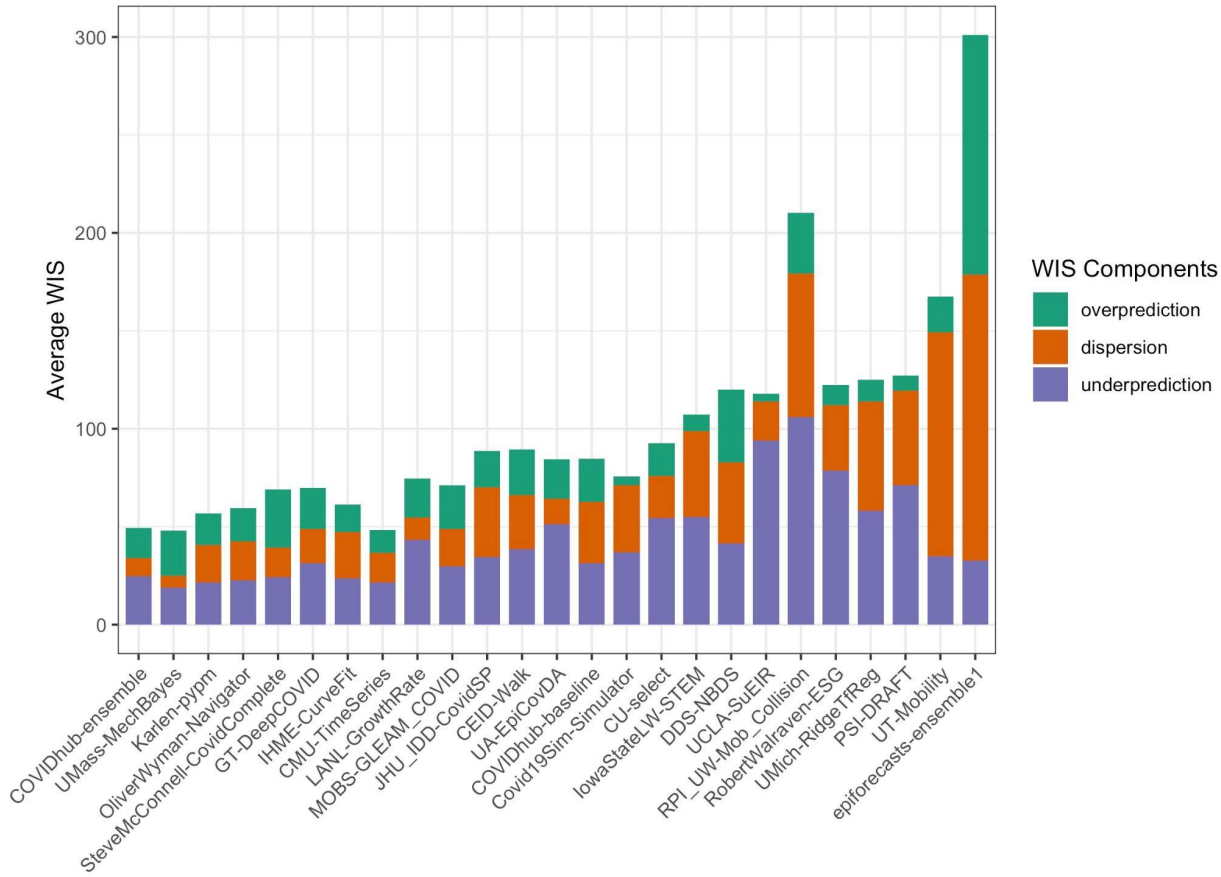
Supplemental Figure 4: Boxplots of average WIS (shown on log scale), by model and forecast horizon and phase. The boxplots represent the median and interquartile range of the model's weekly average WIS aggregated across locations. The baseline median is shown with a dotted red line, and a team's average WIS is shown with a blue "X". The one week ahead forecasts (top row) are more accurate for every model than the 4 week ahead forecasts (bottom row). Models are ordered along the x-axis by their relative WIS within each phase. Based on this aggregation, the ensemble had the lowest median WIS during the summer and winter phases, but other models had lower average WIS values.



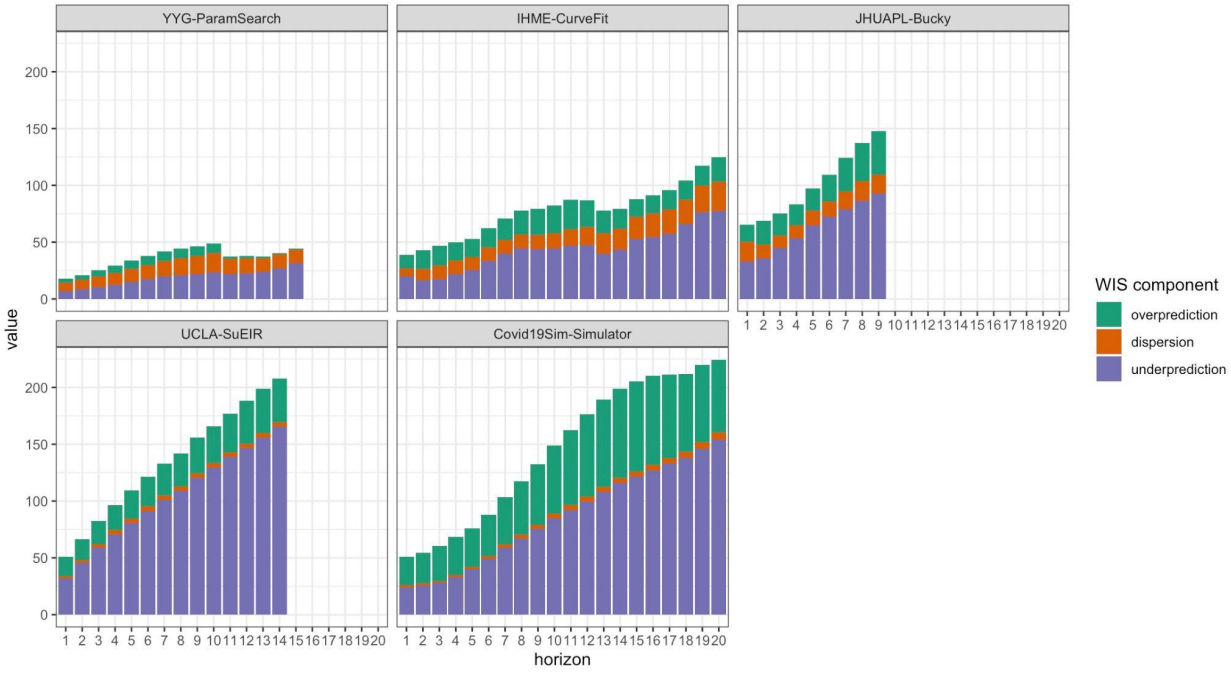
Supplemental Figure 5: Relative WIS by location for each model across all horizons stratified by pandemic phase. The value in each box represents the relative WIS calculated from 1- to 4-week ahead targets available for a model at each location. Points are colored based on the relative WIS compared to the baseline model (θ_m^*, see Methods). Blue boxes represent teams that outperformed the baseline and red boxes represent teams that performed worse than the baseline, with darker hues representing performance further away from the baseline. Teams on the x-axis are listed in order from the highest to lowest relative WIS values within each phase. YYG-ParamSearch achieved the lowest average WIS in the spring and summer and the COVIDhub-ensemble achieved the lowest average WIS in the winter. The COVIDhub-ensemble performed at least as well as the baseline in every location.



Supplemental Figure 6: Decomposition of average WIS scores into underprediction, overprediction, and dispersion, aggregated over locations, horizons, and submission weeks. The sum value of these average metrics add up to the average weighted interval score. Models on the x-axis are ordered by relative WIS values (Table 2). The WIS and relative WIS do not follow the same ordering because the WIS values shown below are not adjusted for prediction difficulty across submission weeks and locations.



Supplemental Figure 7: Average WIS components by horizon. For models that had over 100 forecasts for horizons greater than 8 weeks, the average WIS components of dispersion, underprediction and overprediction are shown. Average WIS in general increases with horizon. Underprediction tends to increase proportionally more than other components.



Supplemental Table 1: Summary of 26 models that contributed to the ensemble forecast but were not individually evaluated due to not having enough eligible submissions during the evaluation period.

	Data Sources Included					Model Information	
Team-Model	**Cases**	**Hosp.**	**Deaths**	**Demog.**	**Mob.**	**Description**	**Assumes social distancing measures change in the future (data source)**
BPagano-RtDriven	J		J			Death-based SIR model that uses the change history of the Covid-19 effective transmission rate to forecast deaths and cases.	No
CovidActNow-SEIR_CAN	NYT		NYT			SEIR model	No
Columbia_UNC-SurvCon	J		J			Survival-convolution model with piecewise transmission rates that incorporates latent incubation period and provides time-varying effective reproductive number.	No
COVIDAnalytics-DELPHI	J		J			SEIR model augmented with underdetection and interventions.	Yes
Google_Harvard-CPF	J	CTP	J	BQ	DL	Extended SEIR model with hospitalization compartments and trainable encoders that process static and time-varying covariates to extract information from. Trained in an end-to-end way with partial teacher forcing.	Yes (CHC)
GT_CHHS-COVID19	GA DPH, NC DHHS		GA DPH, NC DHHS	Cen	Cen, SG, SL	Agent-based simulation disease spread model assuming heterogeneous population mixing to predict the spread pattern geographically over time (49, 50)	Yes
IEM_MED-CovidProject	J		J			SEIR model projections using MCMC to find best parameters to fit actual data.	No
JCB-PRM	J		J			Deterministic model built on observations of macro-level societal and political responses to COVID measured only in terms of infections and deaths.	Yes
JHU_CSSE-DECOM	J		J	Cen	SG	County-level, empirical machine learning model driven by epidemiological, mobility, demographic, and behavioral data.	No

LNQ-ens1				J		County-level ensemble of boosted tree and neural net models	No
MIT_CritData-GBCF	J		J	Cen	PIQ	Gradient boosted regressor with hyperparameter optimization that uses prior COVID-19 cases and deaths as well as static and time-varying county-level covariates. Forecasts at county-level and aggregates to state and national level.	No
MITCovAlliance-SIR	NYT		NYT	Cen, CDC, CL, UM	SG	SIR model trained on public health regions. SIR parameters are functions of static demographic and time-varying mobility features. A two-stage approach that first learns the magnitude of peak infections (51)	No
MIT_ISOLAT-Mixtures	J			J		Non-mechanistic, non-parametric model based on representing time series as a sum of bell curves.	No
Microsoft-DeepSTIA	J	CTP	J		G	A hierarchical spatial-temporal forecasting model that not only follows the time-series trends but also takes into consideration the spatial correlations among different administrative regions.	Yes
MSRA-DeepST	J	CTP	J		G	Deep spatio-temporal network with knowledge-based SEIR as a regularizer under the assumption of spatio-temporal process in pandemic of different regions.	Yes
NotreDame-FRED	NYT		NYT			Agent-based model developed for influenza with parameters modified to represent the natural history of COVID-19.	Yes (IHME COVID-19 health service utilization forecasting Team)
NotreDame-mobility	CTP		J		G,A	Ensemble of nine models that are identical except that they are driven by different mobility indices from Apple and Google. Underlying deterministic, SEIR-like model.	No
USACE-ERDC_SEIR	J,UF	CTP	J,UF			SEIR model with additional compartments for unreported infections and isolated individuals (52)	No
QJHong-Encounter	J	CTP	J			SEIR model using encounter density to predict reproductive number	No
SigSci-TS	J		J			Time series forecasting using ARIMA for case forecasts and lagged cases for death forecasts.	No

SWC-TerminusCM	CTP	CTP	CTP			Mechanistic compartmental model using disease parameter estimates from literature and Bayesian inference.	Yes
UCM_MESALab-FoGSEIR	J		J		G	Modification of integer order SEIR model considering fractional integrals. Considers the age structure and reopening intervention to minimize infections and deaths.	Yes
UCSB-ACTS	J	CTP	J			Data-driven machine learning model that makes predictions by referring to other regions with similar growth patterns and assuming similar development will take place in the current region.	No
UpstateSU-GRU	J		Cen, KFF, BRFSS	J	G	recurrent neural network seq2seq model with the Gated recurrent units (53)	Yes
USC-SIkJalpha	J	HHS	J			Models temporally varying infection, death, and hospitalization rates. Learning is performed by reducing the problem to multiple simple linear regression problems. True susceptible population is identified based on reported cases, whenever mathematically possible (54, 55)	No
Wadhwani_AI-BayesOpt	J		J			Model-agnostic Bayesian optimization ("BayesOpt") approach for learning the parameters of an SEIR-like compartmental model from observed data.	No
YYG-ParamSearch			J	Cen		SEIR model with a machine learning layer (56)	Yes

BQ = Bigquery public datasets (https://cloud.google.com/bigquery/public-data),BRFSS = Behavioral Risk Factor Surveillance System (https://www.cdc.gov/brfss/data_documentation/index.htm), Cen = US Cen (https://www.census.gov/), CHC = COVID Healthcare Coalition https://c19hcc.org/resources/npi-dashboard/), CL = Claritas (https://www.claritascreative.com/covid19), CN = Coronavirus Disease 2019 (COVID-19)-Associated Hospitalization Surveillance Network (COVID-NET) (https://www.cdc.gov/coronavirus/2019-ncov/covid-data/covid-net/purpose-methods.html), CTP = COVID Tracking Project (https://covidtracking.com/), DL= Descartes Labs (https://github.com/descarteslabs/DL-COVID-19), G = Google mobility (https://www.google.com/covid19/mobility/), GA DPH = Georgia Department of Public Health (https://dph.georgia.gov/covid-19-daily-status-report), HHS = Health and human services hospitalizations (https://protect-public.hhs.gov/pages/covid19-module), J = JHU CSSE (https://github.com/CSSEGISandData/COVID-19)(24) , KFF = regional health index from 2019 Kaiser Family Foundation Survey (https://www.kff.org/report-section/ehbs-2019-summary-of-findings/), MMODS = Multi-modal outbreak decision support scenarios (https://midasnetwork.us/mmods/), NC DHHS = NC Department of Health and Human Services (https://covid19.ncdhhs.gov/dashboard), NYT = New York Times (https://github.com/nytimes/covid-19-data), PIQ = Place IQ (https://github.com/COVIDExposureIndices/COVIDExposureIndices), Rt = time-varying reproductive number, SEIR = Susceptible-Exposed-Infectious-Recovered compartmental model, SG = SafeGraph mobility (https://www.safegraph.com/), SIR = Susceptible-Infectious-Recovered compartmental model, SL = StreetLight

(https://www.streetlightdata.com/), UM = University of Michigan Health and Retirement Study (https://hrs.isr.umich.edu/data-products)

Supplemental Table 2: Sensitivity analysis of relative WIS calculations. We computed the relative WIS (rel WIS, θ_m^*) across three different time periods and using two different inclusion criteria, to assess the robustness of the original analysis shown in Table 2. The results show that the values of relative WIS and the ordering of models according to this metric were not strongly sensitive to whether models with smaller numbers of available forecasts were included in the computation of relative WIS. (The "% max obs" column shows the percentage of the maximum possible scores that a given model made.) Some models showed differences in relative WIS when different weeks were included, which is to be expected if models performed better during different phases of the pandemic.

		Table 2	Sensitivity 1	Sensitivity 2
time period evaluated		EW18-2020 - EW17-2021	EW18-2020 - EW17-2021	EW18-2020 - EW17-2021
Inclusion criteria		>= 32 weeks submitted (60%)	>= 37 weeks submitted (70%)	>= 47 weeks (89%)
model	% Max Obs	rel WIS	rel WIS	rel WIS
CEID-Walk	69.38	0.95	-	-
CMU-TimeSeries	76.23	0.79	0.79	-
Covid19Sim-Simulator	86.41	1.01	1.00	-
COVIDhub-baseline	100.00	1.00	1.00	1.00
COVIDhub-ensemble	88.33	0.61	0.61	0.62
CU-select	84.44	0.96	0.96	-
DDS-NBDS	72.78	1.16	1.17	-
epiforecasts-ensemble1	68.34	3.88	3.81	-
GT-DeepCOVID	84.44	0.77	0.78	0.77
IHME-CurveFit	67.45	0.77	0.77	-
IowaStateLW-STEM	75.47	1.03	1.03	-
JHU_IDD-CovidSP	93.57	0.88	0.89	0.92
JHUAPL-Bucky	63.09	1.10	-	-
Karlen-pypm	76.66	0.66	0.67	-
LANL-GrowthRate	86.39	0.80	0.80	-
MOBS-GLEAM_COVID	99.92	0.81	0.82	0.83
OliverWyman-Navigator	88.10	0.70	0.70	0.71
PSI-DRAFT	80.44	1.47	1.49	-
RobertWalraven-ESG	80.26	1.26	1.28	-
RPI_UW-Mob_Collision	40.42	1.34	1.36	-
SteveMcConnell-CovidComplete	66.97	0.74	-	-
UA-EpiCovDA	84.44	0.93	0.93	-
UCLA-SuEIR	79.65	1.28	1.28	-
UCSD_NEU-DeepGLEAM	63.09	0.81	-	-

UMass-MechBayes	96.11	0.61	0.62	0.62
UMich-RidgeTfReg	66.35	1.30	1.31	-
UT-Mobility	69.46	2.66	2.57	-

Supplemental Table 3: Sensitivity analysis examining the impact of excluding data anomalies (outlying observations, or forecasts made from revised data) on the calculations of relative WIS, relative MAE and prediction interval coverage. In general, the metrics do not show large differences based on including or not these anomalous observations in the evaluation.

	Full analysis (Table 2)					Sensitivity analysis (no anomalies)				
		PI Cov					PI Cov			
model	N	95%	50%	relWIS	relMAE	N	95%	50%	relWIS	relMAE
CEID-Walk	7135	0.81	0.46	0.95	1	6608	0.82	0.47	0.95	1.01
CMU-TimeSeries	7840	0.72	0.39	0.79	0.8	7298	0.73	0.4	0.78	0.79
Covid19Sim-Simulator	8886	0.27	0.08	1.01	0.81	8326	0.28	0.08	1.03	0.82
COVIDhub-baseline	10284	0.84	0.44	1	1	9706	0.85	0.46	1	1
COVIDhub-ensemble	9084	0.87	0.47	0.61	0.66	8518	0.89	0.49	0.58	0.64
CU-select	8684	0.68	0.34	0.96	0.93	8127	0.7	0.34	0.99	0.96
DDS-NBDS	7485	0.84	0.4	1.16	1.52	6958	0.85	0.41	1.16	1.51
epiforecasts-ensemble1	7028	0.86	0.45	3.88	3.17	6505	0.87	0.45	1.55	0.93
GT-DeepCOVID	8684	0.82	0.37	0.77	0.85	8146	0.83	0.38	0.74	0.8
IHME-CurveFit	6937	0.64	0.27	0.77	0.8	6483	0.65	0.28	0.76	0.8
IowaStateLW-STEM	7761	0.44	0.18	1.03	0.92	7223	0.45	0.18	1.07	0.94
JHU_IDD-CovidSP	9623	0.8	0.36	0.88	0.99	9070	0.81	0.36	0.92	1.03
JHUAPL-Bucky	6488	0.53	0.24	1.1	1.09	5968	0.53	0.24	1.17	1.14
Karlen-pypm	7884	0.84	0.44	0.66	0.72	7342	0.85	0.46	0.65	0.71
LANL-GrowthRate	8884	0.89	0.4	0.8	0.89	8322	0.9	0.4	0.79	0.89
MOBS-GLEAM_COVID	10276	0.67	0.35	0.81	0.8	9698	0.68	0.36	0.81	0.79
OliverWyman-Navigator	9060	0.83	0.44	0.7	0.74	8498	0.84	0.46	0.68	0.72
PSI-DRAFT	8272	0.34	0.14	1.47	1.24	7721	0.35	0.15	1.55	1.28
RobertWalraven-ESG	8254	0.37	0.21	1.26	1.03	7705	0.38	0.22	1.28	1.04
RPI_UW-Mob_Collision	4157	0.55	0.22	1.34	1.29	3815	0.56	0.22	1.41	1.34
SteveMcConnell-CovidComplete	6887	0.79	0.49	0.74	0.78	6353	0.8	0.5	0.72	0.76
UA-EpiCovDA	8684	0.64	0.33	0.93	0.88	8127	0.65	0.34	0.95	0.89
UCLA-SuEIR	8191	0.24	0.08	1.28	1.08	7652	0.24	0.07	1.28	1.04

UCSD_NEU-DeepGLEAM	6488	0.87	0.55	0.81	0.79	5956	0.89	0.57	0.81	0.78
UMass-MechBayes	9884	0.94	0.56	0.61	0.66	9314	0.95	0.57	0.61	0.65
UMich-RidgeTfReg	6823	0.45	0.23	1.3	1.13	6518	0.47	0.24	1.36	1.17
UT-Mobility	7143	0.67	0.3	2.66	2.45	6718	0.68	0.3	2.85	2.5