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], Sabrina M Corsetti[16], Thomas M Baer[17], Marisa C Eisenberg[16], Karl Falb[16], Yitao Huang[16], Emily T Martin[16], Ella McCauley[16], Robert L Myers[16], Tom Schwarz[16], Daniel Sheldon[1], Graham Casey Gibson[1], Rose Yu[18,19], Liyao Gao[20], Yian Ma[18], Dongxia Wu[18], Xifeng Yan[21], Xiaoyong Jin[21], Yu-Xiang Wang[21], YangQuan Chen[22], Lihong Guo[23], Yanting Zhao[24],Quanquan Gu[25], Jinghui Chen[25], Lingxiao Wang[25], Pan Xu[25], Weitong Zhang[25], Difan Zou[25], Hannah Biegel[26], Joceline Lega[26], Timothy L Snyder[27], Davison D Wilson[27], Steve McConnell[28], Robert Walraven[9], Yunfeng Shi[29], Xuegang Ban[20], Qi-Jun Hong[30,31], Stanley Kong[32], James A Turtle[33], Michal Ben-Nun[33], Pete Riley[33], Steven Riley[34], Ugur Koyluoglu[35], David DesRoches[35], Bruce Hamory[35], Christina Kyriakides[35], Helen Leis[35], John Milliken[35], Michael Moloney[35], James Morgan[35], Gokce Ozcan[35], Chris Schrader[35], Elizabeth Shakhnovich[35], Daniel Siegel[35], Ryan Spatz[35], Chris Stiefeling[35], Barrie Wilkinson[35], Alexander Wong[35], Zhifeng Gao[36], Jiang Bian[36], Wei Cao[36], Juan Lavista Ferres[36], Chaozhuo Li[36], Tie-Yan Liu[36], Xing Xie[36], Shun Zhang[36], Shun Zheng[36], Alessandro Vespignani[37,38], Matteo Chinazzi[37], Jessica T Davis[37], Kunpeng Mu[37], Ana Pastore y Piontti[37], Xinyue Xiong[37], Andrew Zheng[39], Jackie Baek[39], Vivek Farias[40], Andreea Georgescu[39], Retsef Levi[40], Deeksha Sinha[39], Joshua Wilde[39], Nicolas D Penna[41], Leo A Celi[41], Saketh Sundar[42], Sean Cavany[43], Guido España[43], Sean Moore[43], Rachel Oidtman[43,44], Alex Perkins[43], Dave Osthus[45], Lauren Castro[45], Geoffrey Fairchild[45], Isaac Michaud[45], Dean Karlen[46,47], Elizabeth C Lee[48], Juan Dent[48], Kyra H Grantz[48], Joshua Kaminsky[48], Kathryn Kaminsky[9], Lindsay T Keegan[49], Stephen A Lauer[48], Joseph C Lemaitre[50], Justin Lessler[48], Hannah R Meredith[48], Javier Perez-Saez[48], Sam Shah[9], Claire P Smith[48], Shaun A Truelove[48], Josh Wills[9], Matt Kinsey[51], RF Obrecht[51], Katharine Tallaksen[51], John C Burant[9], Lily Wang[8], Lei Gao[8], Zhiling Gu[8], Myungjin Kim[8], Xinyi Li[52], Guannan Wang[53], Yueying Wang[8], Shan Yu[54], Robert C Reiner[20], Ryan Barber[20], Emmanuela Gaikedu[20], Simon Hay[20], Steve Lim[20], Chris Murray[20], David Pigott[20], B Aditya Prakash[55], Bijaya Adhikari[55], Jiaming Cui[55], Alexander Rodríguez[55], Anika Tabassum[55,56], Jiajia Xie[55], Pinar Keskinocak[55], John Asplund[57], Arden Baxter[55], Buse Eylul Oruc[55], Nicoleta Serban[55], Sercan O Arik[58], Mike Dusenberry[58], Arkady Epshteyn[58], Elli Kanal[58], Long T Le[58], Chun-Liang Li[58], Tomas Pfister[58], Dario Sava[58], Rajarishi Sinha[58], Thomas Tsai[59], Nate Yoder, Jinsung Yoon[58], Leyou Zhang[58], Sam Abbott[60], Nikos I Bosse[60], Sebastian Funk[60], Joel Hellewel[60], Sophie R Meakin[60], James D Munday[60], Katherine Sherratt[60], Mingyuan Zhou[11], Rahi Kalantari[11], Teresa K Yamana[61], Sen Pei[61], Jeffrey Shaman[61], Turgay Ayer[55,62], Madeline Adee[63], Jagpreet Chhatwal[63], Ozden O Dalgic[64], Mary A Ladd[63], Benjamin P Linas[65], Peter Mueller[63], Jade Xiao[55], Michael L Li[39], Dimitris Bertsimas[40], Omar Skali Lami[39], Saksham Soni[39], Hamza Tazi Bouardi[39], Yuanjia Wang[61], Qinxia Wang[61], Shanghong Xie[61], Donglin Zeng[66], Alden Green[67], Jacob Bien[14], Addison J Hu[67], Maria Jahja[67], Balasubramanian Narasimhan[68], Samyak Rajanala[68], Aaron Rumack[67], Noah Simon[20], Ryan Tibshirani[67], Rob Tibshirani[68], Valerie Ventura[63], Larry Wasserman[63], Eamon B O'Dea[69], John M Drake[69], Robert Pagano[9], 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, Heidelberg Institute for Theoretical Studies
[5]In-Q-Tel
[6]Karlsruhe Institute of Technology (KIT), Institute for Stochastics
[7]Institute of Mathematical Statistics and Actuarial Science, University of Bern
[8]Iowa State University
[9]No affiliation
[10]Wadhwani Institute of 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]University of Michigan - Ann Arbor
[17]Trinity University, San Antonio
[18]University of California, San Diego
[19]Northeastern University
[20]University of Washington
[21]University of California at Santa Barbara
[22]University of California, Merced
[23]Jilin University
[24]University of Science and Technology of China
[25]University of California, Los Angeles
[26]University of Arizona
[27]Snyder Wilson Consulting, Inc.
[28]Construx
[29]Rensselaer Polytechnic Institute
[30]Brown University
[31]Amazon.com Inc
[32]Manhasset Secondary School
[33]Predictive Science, Inc
[34]Imperial College, London
[35]Oliver Wyman
[36]Microsoft
[37]Laboratory for the Modeling of Biological and Socio-technical Systems, Northeastern University
[38]Institute for Scientific Interchange Foundation
[39]Operations Research Center, Massachusetts Institute of Technology
[40]Sloan School of Management, Massachusetts Institute of Technology
[41]Laboratory for Computational Physiology, Massachusetts Institute of Technology
[42]River Hill High School
[43]University of Notre Dame
[44]University of Chicago
[45]Los Alamos National Laboratory
[46]University of Victoria
[47]TRIUMF
[48]Johns Hopkins Bloomberg School of Public Health
[49]University of Utah

[50]École Polytechnique Fédérale de Lausanne
[51]Johns Hopkins University Applied Physics Lab
[52]Clemson University
[53]College of William & Mary
[54]University of Virginia
[55]Georgia Institute of Technology
[56]Virginia Tech
[57]Metron, Inc
[58]Google Cloud
[59]Harvard University
[60]London School of Hygiene & Tropical Medicine
[61]Columbia University
[62]Emory University Medical School
[63]Massachusetts General Hospital
[64]Value Analytics Labs
[65]Boston University School of Medicine
[66]University of North Carolina Chapel Hill
[67]Carnegie Mellon University
[68]Stanford University
[69]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. In 2020, the COVID-19 Forecast Hub (https://covid19forecasthub.org/) collected, disseminated, and synthesized hundreds of thousands of specific predictions from more than 50 different academic, industry, and independent research groups. This manuscript systematically evaluates 23 models that regularly submitted forecasts of reported weekly incident COVID-19 mortality counts in the US at the state and national level. One of these models was a multi-model ensemble that combined all available forecasts each week. The performance of individual models showed high variability across time, geospatial units, and forecast horizons. Half of the models evaluated showed better accuracy than a naïve baseline model. In combining the forecasts from all teams, the ensemble showed the best overall probabilistic accuracy of any model. Forecast accuracy degraded as models made predictions farther into the future, with probabilistic accuracy at a 20-week horizon more than 5 times worse 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 responses to infectious disease pandemics require federal, state, and local leaders to make timely decisions in order to reduce disease transmission. During the 2020 SARS-CoV-2 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 changes 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 1 or 2 months, forecasts can inform operational decisions about allocation of healthcare supplies (e.g., personal protective equipment, therapeutics, and vaccines), staffing needs, or school closures.[3] Providing prediction uncertainty is critical for such decisions, as it allows policy makers to assess the most likely and plausible worst-case scenarios.[3]

Other modeling approaches that focus on the recent past or distant future also support epidemic decision-making in real-time.[4–6] Some “nowcasting” models estimate current trends of an outbreak (e.g., values of the effective reproduction number) and can provide situational awareness about recent trends and changes in transmission, especially in the context of data that are incomplete from recent weeks.[7–9] Additionally, models that consider counterfactual scenarios provide quantitative information about hypothetical futures that might arise under different actionable options, sometimes at much longer horizons (e.g., months to years).[10–12] These projections typically are not evaluable against actual observed data in the same way that forecasts are due to a lack of “ground truth” data from counterfactual futures. However, projections can provide valuable information about what outcomes might occur under different interventions or treatment regimes, and they can help inform long-term resource planning. Forecasts, therefore, occupy a unique niche in infectious disease modeling, as they provide opportunities to concretely evaluate the accuracy of different 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 in 2020. Publicly accessible forecasts reflected 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 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 an academic research lab at the University of Massachusetts Amherst to create the COVID-19 Forecast Hub

(https://covid19forecasthub.org/).[13] Launched in April 2020, the Forecast Hub collected and archived forecasts. From these, 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).

Ensemble approaches have previously demonstrated superior performance compared with single models in forecasting influenza,[10,14,15] Ebola,[16] and dengue fever outbreaks.[17] Preliminary research suggested that COVID-19 ensemble forecasts were also more accurate and precise than individual models in the early phases of the pandemic.[18,19] As has been seen in research across disciplines, ensemble approaches are able to draw from and incorporate the information from multiple forecasts, each with their own strengths and limitations, to create highly accurate predictions with well-calibrated uncertainty.[20–25] Additionally, synthesizing multiple models removes the risk of over-reliance on any single approach for accuracy or stability. Individual models often rely on a small number of modelers (sometimes only one) and may be subject to unplanned delays. Collaborative efforts with multiple models as inputs are less susceptible to unanticipated interruptions.

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.[26] 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. Models trained using historical data may lack sufficient characterization of all underlying uncertainties.[27]

Hence, 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.

In this analysis, we sought to evaluate the accuracy and precision of individual and ensemble probabilistic forecasts submitted to the Forecast Hub from mid-May through late December 2020, focusing on forecasts of weekly incident deaths. 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.

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.[28] 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 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, but in some cases, these anomalous observations were left in the final dataset (Supplemental Figure 1). In settings where the true dates of deaths were known, these observations could be re-distributed over previous time points, thus adjusting the previous data. In settings where the true dates were not known, the data reflect dramatically inflated observations from the week in which the backlogs were reported. These updates were made available in a public GitHub repository (https://github.com/CSSEGISandData/COVID-19/tree/master/csse_covid_19_data#data-modification-records).

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.[29] 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 include predictions for the national level, for any of the 50 states and for American Samoa, the District of Columbia, Guam, the Northern Mariana Islands, Puerto Rico, and the US Virgin Islands. Incident death forecasts could be submitted with predictions for time periods that were during 1- 20 weeks after the week in which a forecast was submitted.

Weekly 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.[30] As an example of a forecast and the corresponding observation, forecasts submitted during Tuesday, October 6 (day 3 of EW41) and Monday, October 12 (day 2 of EW42) contained a “1-week ahead” forecast of incident deaths that corresponded to the total number of deaths observed in EW42, a 2-week ahead forecast corresponded to the total number of deaths

in week EW43, etc… 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.

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 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 28 weeks beginning in EW20 and ending with EW47, 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 23 models that had at least 19 eligible weeks during this time period (Figure 1c).

Forecast evaluation period

Forecasts were evaluated based on submissions in a continuous 31-week period starting in mid-May and ending in mid-December (EW20 – EW50, Figure 1). Forecasts were scored using CSSE data available as of January 3, 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. During the last 3 weeks of this evaluation period (EW48 – EW50), all 1- through 4-week ahead forecasts could not be evaluated based on recent data, therefore, these weeks were not included in determining eligibility criteria (Supplemental Figure 2).

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 8 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, included territories, and the national level.

Our 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 one time, in New Jersey during EW35.

Occasionally, large retrospectively identified “backlogs” of deaths were reported by CSSE on a single day for a given state. Examination of data before and after revision showed that most revisions to weekly observations were small. For this reason, locations and dates affected by such backlog reporting were not removed from evaluation (Supplemental Figure 1).

Forecast models

We compared 23 models that submitted eligible forecasts for at least 19 of the 28 weeks considered in the 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 or comparison for all models. This baseline model forecasted a predictive median incidence equal to the number of reported deaths in the most recent week, 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.[29] We evaluated more complex ensemble methods, but they did not show consistent improvements in accuracy.[31]

Forecast Submission Timing

Because this was a real-time forecasting project, forecasts were occasionally submitted late and/or resubmitted. Fifty-five of the 598 (9%) forecast submissions we included in the evaluation 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. 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 (i.e., 4 weeks or less into the future) and, in a single analysis, for longer-term projections (i.e., 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 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-format 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.[32]

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 $k = 1, ..., K$ and $w_0 = 1/2$. In our setting, we used data on $K = 11$ interval scores, for $\alpha = 0.02, 0.05, 0.1, 0.2, \ldots, 0.9$.

This particular choice of weights for WIS has been shown to be equivalent to the quantile loss function and to approximate the commonly used continuous ranked probability score (CRPS).[33] 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.[32,33] 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 directly calculated if only a set of quantiles of the predictive distribution are available.

We also used prediction interval coverage, the proportion of times a prediction interval of a certain level covered the true 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 with which to assess the calibration of the forecasts. We compute prediction interval coverage for a set of observations $(y_{1:N})$ and prediction interval bounds with an uncertainty level $1 - \alpha$, $(l_{\alpha,1:N}, u_{\alpha,1:N})$ as

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

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,[34–36] 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.

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

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

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* . The central assumption here is that performing well relative to individual models 1, …, *M* (not including the baseline) is similarly difficult for each week and location so that no model can gain an advantage by focusing on just some of them. We note that the baseline model is not included in these pairwise comparisons, because the difficulty of beating the baseline model can vary considerably over time and space. 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) [13] 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 software language (v 4.0.2).[37]

Results

Summary of models

During the evaluation period, New York, California, New Jersey, Texas, and Pennsylvania had the highest numbers of incident deaths seen in a given week (in descending order, Figure 1A). This timeframe captured a late summer increase in several locations, but missed the first national increase in early May (Figure 1B).

The number of models that submitted forecasts of incident deaths and were screened for eligibility increased from 13 models at the beginning of the evaluation period to as many as 49 in early December (Figure 1C). Not all models submitted every week, and some models submitted forecasts for varying numbers of locations each week (Supplemental Figure 2). Twenty-three models met our inclusion criteria (see Methods), yielding 598 submission files with 191,013 specific predictions for unique combinations of targets and locations. Five of these models submitted forecasts for each of the 31 evaluated weeks and seven models submitted forecasts for all 55 locations.

The submitted forecasts used different data sources and made varying assumptions about future transmission patterns (Table 1). Twenty-two models incorporated data on prior deaths to create forecasts. This included all models except the COVIDhub-ensemble, which did not directly use surveillance data to create forecasts. Additionally, all models other than the ensemble, the baseline model, PSI-Draft, UT-Mobility, and YYG-ParamSearch used case data as inputs to their forecast models. Eight models included data on COVID-19 hospitalizations, seven models incorporated demographic data, and nine models used mobility data. Of the 23 models evaluated, five assumed that social distancing and other behavioral patterns would change over the 4-week prediction period and 19 assumed that social distancing measures would remain unchanged in the forecasted weeks.

Overall model accuracy

Led by the ensemble model, which showed the best average probabilistic accuracy of all models across the evaluation period, half of the evaluated models achieved better accuracy than the baseline in forecasting incident deaths (Table 2). The COVIDhub-ensemble model achieved a relative weighted interval score (relative WIS, θ_m^*) of 0.63, which can be interpreted as it achieving, on average, 37% less probabilistic error than the baseline forecast in the evaluation period. An additional three models achieved a relative WIS of less than 0.75. In total, 11 models had a relative WIS of less than 1, indicating lower probabilistic forecast error than the baseline model, and 11 had a relative WIS of 1 or greater. Values of relative WIS and rankings of models were robust to different sets of models being included or excluded (Supplemental Table 2).

While forecasts from 11 models showed lower average error than the baseline, absolute measures of calibration (empirical coverage rates of prediction intervals) varied among the models (Table 2). When all forecast horizons, weeks, and locations are considered, several models achieved near nominal coverage rates for both the 50% and 95% prediction intervals. Four models achieved coverage rates within 5% for the 50% prediction interval and two other models achieved near nominal coverage for 95% prediction intervals. Eight models had very low coverage rates (less than 50% for the 95% prediction intervals or less than 15% for the 50% prediction intervals).

Models with simple data inputs were some of the most accurate stand-alone models. Of the top five individual models based on relative WIS (YYG-ParamSearch, UMass-MechBayes, OliverWyman-Navigator, CMU-TimeSeries, and GT-DeepCOVID) only two used data beyond the epidemiological 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 consisted of both models with mechanistic components (YYG-ParamSearch, UMass-MechBayes, OliverWyman-Navigator) and purely statistical ones (CMU-TimeSeries and GT-DeepCOVID). Three of the 10 individual models that performed better than the baseline (OliverWyman-Navigator, GT-DeepCOVID, and IHME-SEIR) used data other than epidemiological surveillance data (e.g., demographics or mobility) in their model.

Model accuracy rankings are highly variable

We ranked models based on WIS for each combination of location, target, and time across all 6,486 possible predicted observations (Figure 2). All models showed large variability in skill relative to other models, with each model having observations for which it had the lowest WIS and thereby a standardized rank of 1. 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. Some models 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, see Table 2, Supplemental Figure 3) resulting in either strong scores for being very close to the truth or harsh penalties for being far from the truth. If models were equally accurate, distributions of standardized ranks would be approximately uniform.

Forecast accuracy declines in absolute terms as short-term horizons increase

Averaging across all states and weeks in the evaluation period, forecasts from all models showed lower accuracy and higher variance as the forecast horizon moved from 1 to 4 weeks ahead (Figure 3). At a 1-week horizon, the baseline forecasts had an average WIS of 26.8 and the ensemble had an average WIS of 18.6. Eight models showed lower average WIS than baseline at a 1-week horizon, although three of those models (YYG-ParamSearch, Karlen-pypm, and CMU-TimeSeries) had 9or 10 missing weeks out of the 31 evaluated. The COVIDhub-ensemble model consistently outperformed the COVIDhub-baseline model, with similar error at a 3-week horizon (average WIS 28.5) as the COVIDhub-baseline model had at a 1-week horizon (average WIS 26.8). At a 4-week horizon, the baseline forecasts showed an average WIS of 56.8. Thirteen models showed lower average WIS than the baseline at a 4-week horizon.

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 3).

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 probabilistic error than both the baseline model and the average error of all models across the entire evaluation period. In weeks where the COVIDhub-ensemble forecast showed the worst probabilistic accuracy, other

models also showed lower predictive performance. As an example, the COVIDhub-ensemble 1-week ahead forecast for EW49 (ending December 5) yielded its highest average WIS across all weeks (mean WIS = 43.7), and three out of 17 other models that submitted for the same locations outperformed it. The 4-week ahead COVIDhub-ensemble forecasts were also worse in EW49 than in any other week during the evaluation period (mean WIS = 98.9), and nine out of the 17 models outperformed the ensemble that week at a forecast horizon of 4 weeks.

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 December, Figure 4). In forecasts submitted during the first week of July (EW27), the baseline model had a high 1-week ahead error due to a large number of new deaths reported in the prior week in New Jersey. In general, other models did not show unusual errors in their forecasts originating from these data, suggesting that their approaches (either via hard-coded model robustness or manual adjustments) were robust to changes in reporting.

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). Of the models that submitted for all locations, the ensemble model had the highest fraction of the 55 locations with improved accuracy over baseline, with equivalent or improved performance in all locations. Ensemble forecasts of incident deaths showed the largest relative accuracy improvements in New York (relative WIS = 0.3), New Jersey (relative WIS = 0.3), and Massachusetts (relative WIS = 0.4) and the lowest relative accuracy in Guam (relative WIS = 1.0). Improved relative accuracy over the baseline by a large number of models may be associated with large data revisions (in New York and New Jersey) or outbreaks during the evaluation period that showed a fast rise and fall (as in, e.g., Connecticut and Massachusetts) where the baseline model may not have performed as well (Supplemental Figure 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 in average WIS from all teams submitting showed that 4-week ahead forecasts had roughly twice the error of 1-week ahead forecasts, a relationship that was consistent across all weeks. All longer-term forecasts showed less accuracy than 1- and 4-week ahead forecasts. There was not a clear trend in probabilistic model accuracy between 8- and 20-week horizons, potentially due to the small number of weeks and models evaluated For the 2 teams who made 20-week ahead forecasts for all 50 states, average WIS was 5-6 times higher at a 20-week horizon than it was at a 1-week horizon. No model made forecasts for horizons of 8 or greater that were calibrated at the 95% level, with coverage ranging from 5% to 80% depending on the model and horizon. Average coverage across all models for horizons of 8 or more was always less than 50%, with no clear trend in changing coverage as the horizon increased.

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. Using a rich dataset of outputs from dozens of COVID-19 models, we have quantified the relative probabilistic accuracy of 23 models including an ensemble and a baseline model.

This paper provides a comparative look at the probabilistic accuracy of different modeling approaches during the COVID-19 pandemic in the US during May – December 2020. These evaluations were adjusted for regions and time periods, and three evaluation metrics were analyzed. We treated the relative WIS as our primary evaluation metric to assess the accuracy of the entire predictive distributions submitted. We used 95% PI coverage and 50% PI coverage, which only assess specific features of a predictive distribution, as our secondary evaluation metrics. The results presented in this manuscript will be updated as additional periods are analyzed.

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.[38,39] The Forecast Hub website, on average, received more than 40,000 views per month during May – December 2020.

The number of teams and forecasts contributing to the COVID-19 ensemble forecast model has exceeded forecasting activity for any prior pandemic. Additionally, The open-science orientation of this project ensures that these data generated in real-time can and will be reused by researchers for many years to come.

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.[3,15–17] 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.

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.[10,17,40] 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.
- No model performed best on all three focal metrics (Table 2). The ensemble model and three of the individual models (OliverWyman-Navigator, UMass-MechBayes, and YYG-ParamSearch) performed well on the relative WIS metric and one of the prediction interval coverage metrics. One additional model (LANL-GrowthRate) performed close to the best on 95% PI coverage.
- 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). In part because of this variability, retrospective experiments that cover most of the evaluation period have shown it to be very hard to improve on the median ensemble approach by using "trained" ensembles that estimate weights for component models.[31]
- Forecast accuracy and calibration degraded as the horizon projected farther away from the current observations (Figure 6).

Rigorous evaluation of forecast accuracy faces many challenges in practice. The large variation in forecast errors across targets, submission weeks, and locations (Supplemental Figure 4) makes it difficult to create simple and clean comparisons of models. Additionally, forecast comparison is challenging because teams have submitted forecasts for different lengths of time, different locations, and for different numbers of horizons (Figure 6, Supplemental Figure 2). Some teams have also changed their models over time (Table 1, Supplemental Table 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. Finally, ground truth data are not static. They can be later revised as more data become available (Supplemental Figure 1). Different sources for ground truth data can also have substantial differences that impact model performance.

Short-term forecasts of COVID-19 mortality have informed public health response and risk communication for the pandemic. However, these forecasts are only one component of a comprehensive public health data and modeling system needed to help inform outbreak response. 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.

uhh

Citations

1. Davies, S. E. & Youde, J. R. *The Politics of Surveillance and Response to Disease Outbreaks: The New Frontier for States and Non-state Actors*. (Routledge, 2016).
2. Polonsky, J. A. *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. Lutz, C. S. *et al.* Applying infectious disease forecasting to public health: a path forward using influenza forecasting examples. *BMC Public Health* **19**, 1659 (2019).
4. Bausch, D. G. & Edmunds, J. Real-Time Modeling Should Be Routinely Integrated into Outbreak Response. *Am. J. Trop. Med. Hyg.* **98**, 1214–1215 (2018).
5. Houlihan, C. F. & Whitworth, J. A. G. Outbreak science: recent progress in the detection and response to outbreaks of infectious diseases. *Clin. Med.* **19**, 140 (2019).
6. Heesterbeek, H. *et al.* Modeling infectious disease dynamics in the complex landscape of global health. *Science* **347**, aaa4339 (2015).
7. Osthus, D., Daughton, A. R. & Priedhorsky, R. Even a good influenza forecasting model can benefit from internet-based nowcasts, but those benefits are limited. *PLoS Comput. Biol.* **15**, e1006599 (2019).
8. McGough, S. F., Johansson, M. A., Lipsitch, M. & Menzies, N. A. Nowcasting by Bayesian Smoothing: A flexible, generalizable model for real-time epidemic tracking. *PLoS Comput. Biol.* **16**, e1007735 (2020).
9. Brooks, L. C., Farrow, D. C., Hyun, S., Tibshirani, R. J. & Rosenfeld, R. Nonmechanistic forecasts of seasonal influenza with iterative one-week-ahead distributions. *PLoS Comput. Biol.* **14**, e1006134 (2018).

10. McGowan, C. J. *et al.* Collaborative efforts to forecast seasonal influenza in the United States, 2015–2016. *Sci. Rep.* **9**, 683 (2019).
11. Shea, K. *et al.* Harnessing multiple models for outbreak management. *Science* **368**, 577–579 (2020).
12. Shea, K. *et al.* COVID-19 reopening strategies at the county level in the face of uncertainty: Multiple Models for Outbreak Decision Support. *medRxiv* 2020.11.03.20225409 (2020) doi:10.1101/2020.11.03.20225409.
13. Cramer, E. *et al. COVID-19 Forecast Hub: 4 December 2020 snapshot*. (2020). doi:10.5281/zenodo.4305938.
14. Reich, N. G. *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).
15. Reich, N. G. *et al.* Accuracy of real-time multi-model ensemble forecasts for seasonal influenza in the U.S. *PLoS Comput. Biol.* **15**, e1007486 (2019).
16. Viboud, C. *et al.* The RAPIDD ebola forecasting challenge: Synthesis and lessons learnt. *Epidemics* **22**, 13–21 (2018).
17. Johansson, M. A. *et al.* An open challenge to advance probabilistic forecasting for dengue epidemics. *Proc. Natl. Acad. Sci. U. S. A.* **116**, 24268–24274 (2019).
18. Funk, S. *et al.* Short-term forecasts to inform the response to the Covid-19 epidemic in the UK. *medRxiv* 2020.11.11.20220962 (2020) doi:10.1101/2020.11.11.20220962.
19. Taylor, K. S. & Taylor, J. W. A Comparison of Aggregation Methods for Probabilistic Forecasts of COVID-19 Mortality in the United States. *arXiv:2007.11103 [stat]* (2020).
20. Bates, J. M. & Granger, C. W. J. The Combination of Forecasts. *J. Oper. Res. Soc.* (2017) doi:10.1057/jors.1969.103.

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

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

23. Leutbecher, M. & Palmer, T. N. Ensemble forecasting. *J. Comput. Phys.* **227**, 3515–3539 (2008).

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

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

26. Moran, K. R. *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).

27. Friedman, J. *et al.* Predictive performance of international COVID-19 mortality forecasting models. *medRxiv* 2020.07.13.20151233 (2020) doi:10.1101/2020.07.13.20151233.

28. Dong, E., Du, H. & Gardner, L. An interactive web-based dashboard to track COVID-19 in real time. *Lancet Infect. Dis.* **20**, 533–534 (2020).

29. Ray, E. L. *et al.* Ensemble Forecasts of Coronavirus Disease 2019 (COVID-19) in the U.S. *medRxiv* 2020.08.19.20177493 (2020) doi:10.1101/2020.08.19.20177493.

30. MMWR Weeks. *CDC* https://wwwn.cdc.gov/nndss/document/MMWR_Week_overview.pdf.

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

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

33. Bracher, J., Ray, E. L., Gneiting, T. & Reich, N. G. Evaluating epidemic forecasts in an interval format. *arXiv:2005.12881 [q-bio, stat]* (2020).

34. Soloman, S. R. & Sawilowsky, S. S. Impact of Rank-Based Normalizing Transformations on the Accuracy of Test Scores. *J. Mod. Appl. Stat. Methods* **8**, 448–462 (2009).

35. Wu, S., Crestani, F. & Bi, Y. *Evaluating Score Normalization Methods in Data Fusion*. vol. 4182 (2006).

36. Renda, M. E. & Straccia, U. Web metasearch: rank vs. score based rank aggregation methods. in *Proceedings of the 2003 ACM symposium on Applied computing* 841–846 (Association for Computing Machinery, 2003). doi:10.1145/952532.952698.

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

38. Boice, R. B. J. Where The Latest COVID-19 Models Think We're Headed — And Why They Disagree. *FiveThirtyEight* https://projects.fivethirtyeight.com/covid-forecasts/ (2020).

39. CDC. COVID-19 Forecasts: Deaths. https://www.cdc.gov/coronavirus/2019-ncov/covid-data/forecasting-us.html (2021).

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

41. COVID-19 Findings, Simulations | Shaman Group. https://blogs.cuit.columbia.edu/jls106/publications/covid-19-findings-simulations/.

42. Pei, S. & Shaman, J. *Initial Simulation of SARS-CoV2 Spread and Intervention Effects in the Continental US*. http://medrxiv.org/lookup/doi/10.1101/2020.03.21.20040303 (2020).

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

44. Rodríguez, A. *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 vol. 35 Forthcoming (2021).

45. Wang, L. *et al.* Spatiotemporal Dynamics, Nowcasting and Forecasting of COVID-19 in the United States. *arXiv:2004.14103 [stat]* (2020).
46. Lemaitre, J. C. *et al.* A scenario modeling pipeline for COVID-19 emergency planning. *medRxiv* 2020.06.11.20127894 (2020) doi:10.1101/2020.06.11.20127894.
47. Case studies and reports. https://pypm.github.io/home/.
48. Karlen, D. Characterizing the spread of CoViD-19. *arXiv:2007.07156 [physics, q-bio, stat]* (2020).
49. LANL COVID-19 Cases and Deaths Forecasts. https://covid-19.bsvgateway.org/.
50. Chinazzi, M. *et al.* The effect of travel restrictions on the spread of the 2019 novel coronavirus (COVID-19) outbreak. *Science* **368**, 395–400 (2020).
51. EpiGro & EpiCovDA. https://jocelinelega.github.io/EpiGro/.
52. Gibson, G. C., Reich, N. G. & Sheldon, D. REAL-TIME MECHANISTIC BAYESIAN FORECASTS OF COVID-19 MORTALITY. *medRxiv* 2020.12.22.20248736 (2020) doi:10.1101/2020.12.22.20248736.
53. Rowland, M. A. *et al. COVID-19 infection data encode a dynamic reproduction number in response to policy decisions with secondary wave implications*. https://www.researchsquare.com/article/rs-75665/v1 (2020).
54. COVID-19 Projections Using Machine Learning. https://covid19-projections.com/.
55. O'Dea, E. *e3bo/random-walks*. (2021).
56. Baek, J. *et al.* The Limits to Learning an SIR Process: Granular Forecasting for Covid-19. *arXiv:2006.06373 [cs, stat]* (2020).
57. Wang, R., Maddix, D., Faloutsos, C., Wang, Y. & Yu, R. AutoODE: Bridging Physics-based and Data-driven modeling for COVID-19 Forecasting.

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

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

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.
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, 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
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_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.
NotreDame-mobility and *NotreDame-FRED*: NSF RAPID DEB 2027718
UA-EpiCovDA: NSF RAPID Grant # 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. There were 23 models from 22 teams whose models were evaluated. 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**
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 [29]	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[41–43]	Yes
GT-DeepCOVID	CTP	CTP, HHS, CN	J		G,A	Data-driven approach based on deep learning for forecasting mortality and hospitalizations[44]	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 [45]	No
JHU_IDD-CovidSP[c]	J,UF		J,UF	Cen	Cen	Metapopulation model with commuting, nonpharmaceutical interventions, and stochastic SEIR disease dynamics[46]	No
Karlen-pypm	CTP, HHS	J	J			Finite time difference equations implemented as a general-purpose population modelling framework[47,48]	No
LANL-GrowthRate[d]	J		J			Statistical dynamical growth model accounting for population susceptibility[49]	No
MOBS-GLEAM_COVID	J	HHS	J	Cen	G	Metapopulation, age-structured SLIR model with mobility and nonpharmaceutical interventions[50]	No
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
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.)
RobertWalraven-ESG	J		J			Multiple skewed gaussian mathematical fit	No
PSI-DRAFT			J	Cen		Age-stratified compartmental SEIRX model with time-dependent reproduction number	No
UA-EpiCovDA[e]	CTP		CTP, J			SIR mechanistic model with data assimilation[51]	No

UCLA-SuEIR	J	CTP	J			SEIR model variant considering both untested and unreported cases	Yes
UMass-MechBayes	J		J			Bayesian compartmental model with observations on incident case counts and incident deaths[52]	No
UMich-RidgeTfReg[f]	J		J		G	Ridge regression model using confirmed case and death reports to generate predictions	No
USACE-ERDC_SEIR	J,UF	CTP	J,UF			SEIR model with additional compartments for unreported infections and isolated individuals[53]	No
UT-Mobility			J		SG	Bayesian multilevel negative binomial regression model	No
YYG-ParamSearch			J			SEIR model with a machine learning layer[54]	Yes

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)[28], 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.

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

Table 2: Summary accuracy metrics for all submitted forecasts, 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 (largest number in bold). 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
CMU-TimeSeries	4052	0.69	0.35	0.76	0.78
Covid19Sim-Simulator	5098	0.31	0.09	0.95	0.79
COVIDhub-baseline	**5896**	0.84	**0.48**	1.00	1.00
COVIDhub-ensemble	5296	0.87	**0.47**	**0.63**	0.70
CU-select	4896	0.81	0.40	1.00	1.08
GT-DeepCOVID	4990	0.83	0.37	0.83	0.90
IHME-SEIR	3947	0.68	0.3	0.90	0.95
IowaStateLW-STEM	4125	0.45	0.18	1.05	0.96
JHU_IDD-CovidSP	**5896**	0.80	0.35	1.10	1.23
Karlen-pypm	4096	0.82	0.42	0.86	0.91
LANL-GrowthRate	5096	**0.91**	0.42	0.89	1.03
MOBS-GLEAM_COVID	**5896**	0.69	0.38	1.01	0.99
NotreDame-mobility	4896	0.50	0.24	1.43	1.25
OliverWyman-Navigator	5280	0.85	**0.47**	0.68	0.73
PSI-DRAFT	4492	0.37	0.16	1.75	1.48
RobertWalraven-ESG	4470	0.09	0.04	1.94	1.51
UA-EpiCovDA	4896	0.68	0.36	0.91	0.97
UCLA-SuEIR	5096	0.20	0.07	1.22	0.99
UMass-MechBayes	**5896**	**0.95**	0.58	0.66	0.73
UMich-RidgeTfReg	3435	0.35	0.13	1.51	1.35
USACE-ERDC_SEIR	4646	0.15	0.05	2.49	1.96
UT-Mobility	5297	0.70	0.32	1.40	1.43
YYG-ParamSearch	4196	0.82	**0.46**	0.64	**0.68**

Figure 1: Overview of the evaluation period included in the paper. (A) The reported number of incident weekly COVID-19 deaths by state or territory, per JHU CSSE reports. Locations are sorted by the maximum value of incident deaths in one week. (B) The time-series of weekly incident deaths at the national level overlaid with one example forecast from the COVIDhub-ensemble model. Submitted forecasts provide quantiles that specify prediction intervals at different levels of uncertainty. (C) The number of models submitting forecasts for incident deaths each week. Weeks in which the COVIDhub-ensemble was submitted are shown with a red asterisk.

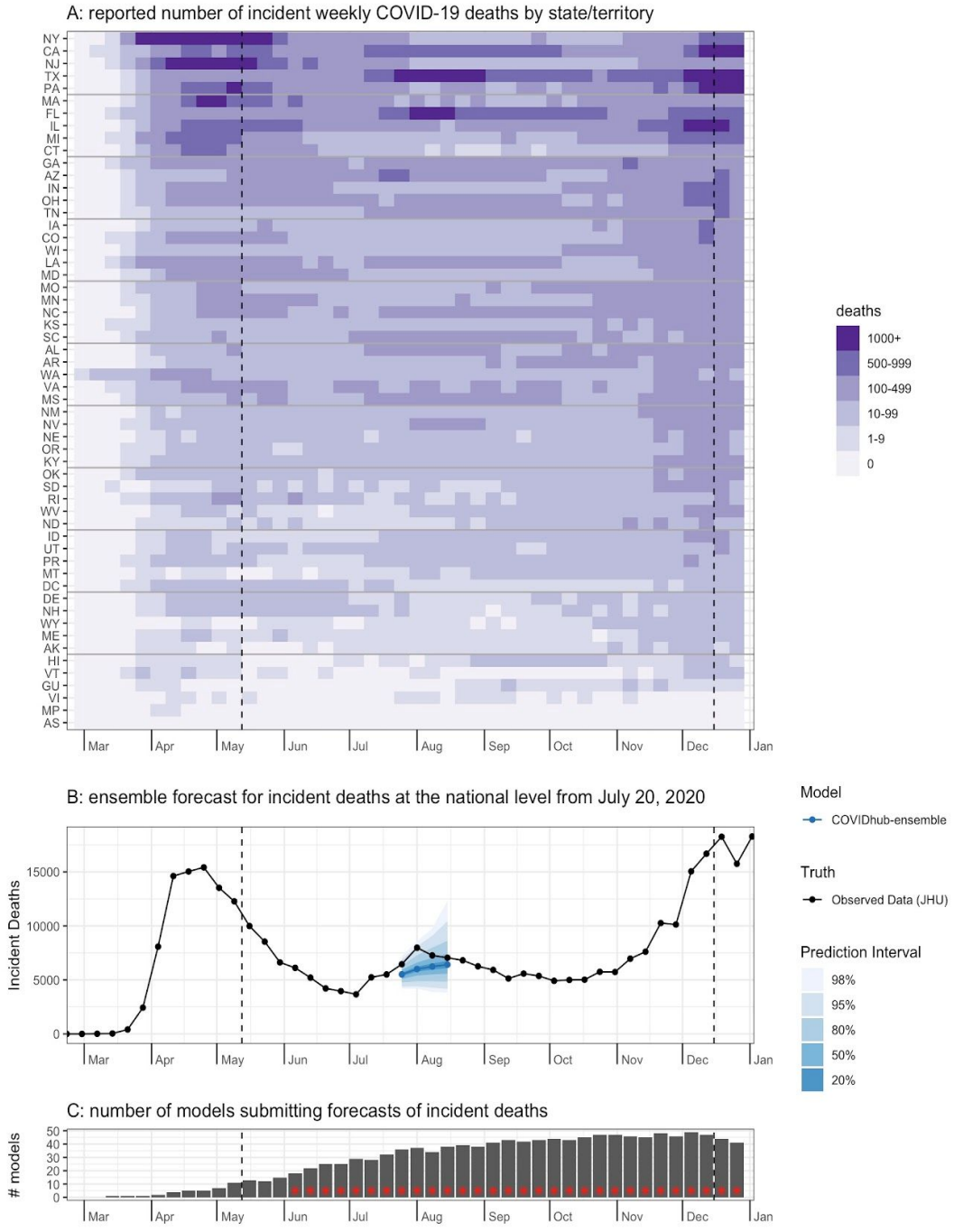


Figure 2: A comparison of each models' distribution of standardized rank of weighted interval scores (WIS) for each location-target-week observation. A standardized rank of 1 indicates that the model had the best WIS for that particular location, target, and week and a value of 0 indicates it had the worst WIS. Intermediate values indicate rankings relative to the best and worst model. 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 models' 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 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. Some models show a bimodal distribution, with one mode in the yellow region and another in the purple region. In these cases, the models frequently made overconfident predictions (Table 2, Supplemental Figure 3) resulting in either strong scores for being very close to the truth or harsh penalties for being far from the truth. 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.

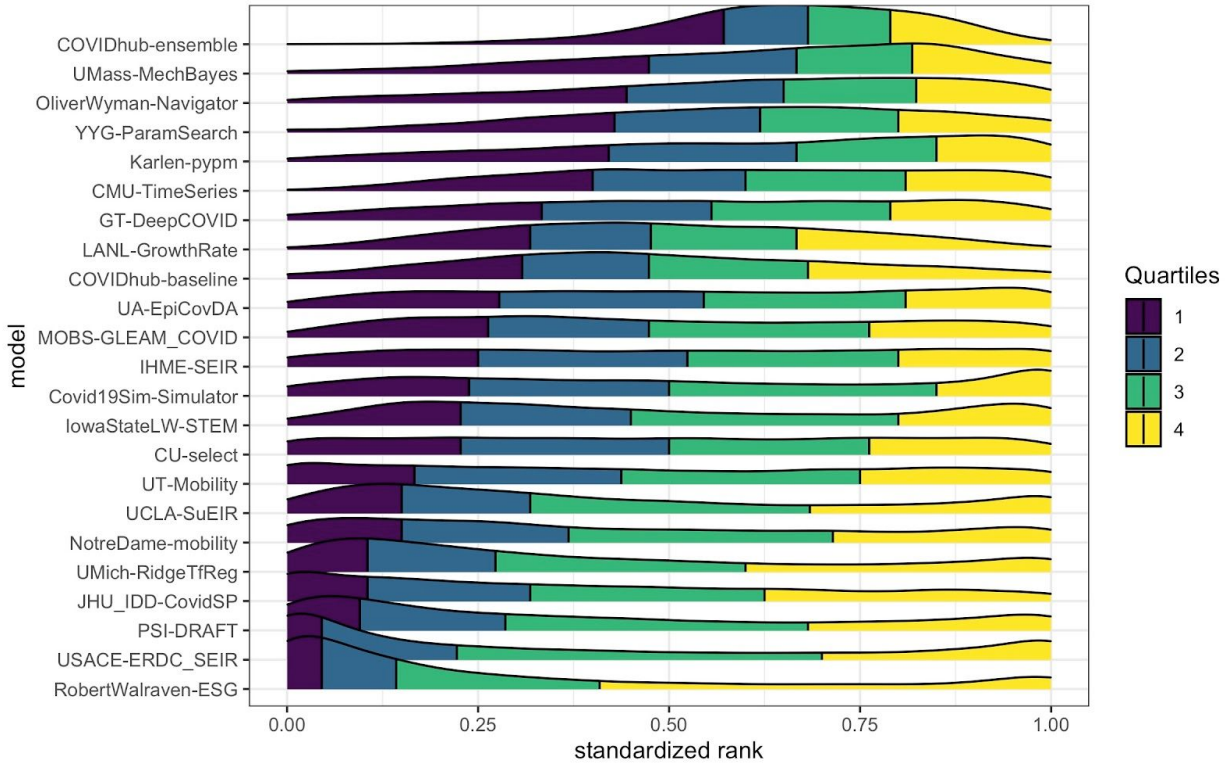


Figure 3: Average weighted interval score (WIS) by evaluation week for each model across all 50 states. The four panels represent each of the 1 through 4 week ahead forecast horizons. Each point represents an average WIS calculated from available states for a particular week. Points colored purple indicate weeks for which the corresponding model did not make forecasts for all 50 states. 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 overall average of the baseline model for each horizon. With longer horizons, there is larger variation across the overall average WIS values (shown by the differences across the orange "x" marks), and larger variation within each model's weekly average WIS (shown by a wider range in average WIS points for each model). For a horizon of 1 week, the baseline model performs approximately as well as most models, with only 5 models outperforming the baseline. At a horizon of 4 weeks, there is more variability in average WIS values and 10 models outperform the baseline model. Models missing more than 4 weeks of forecast submissions are highlighted in red text.

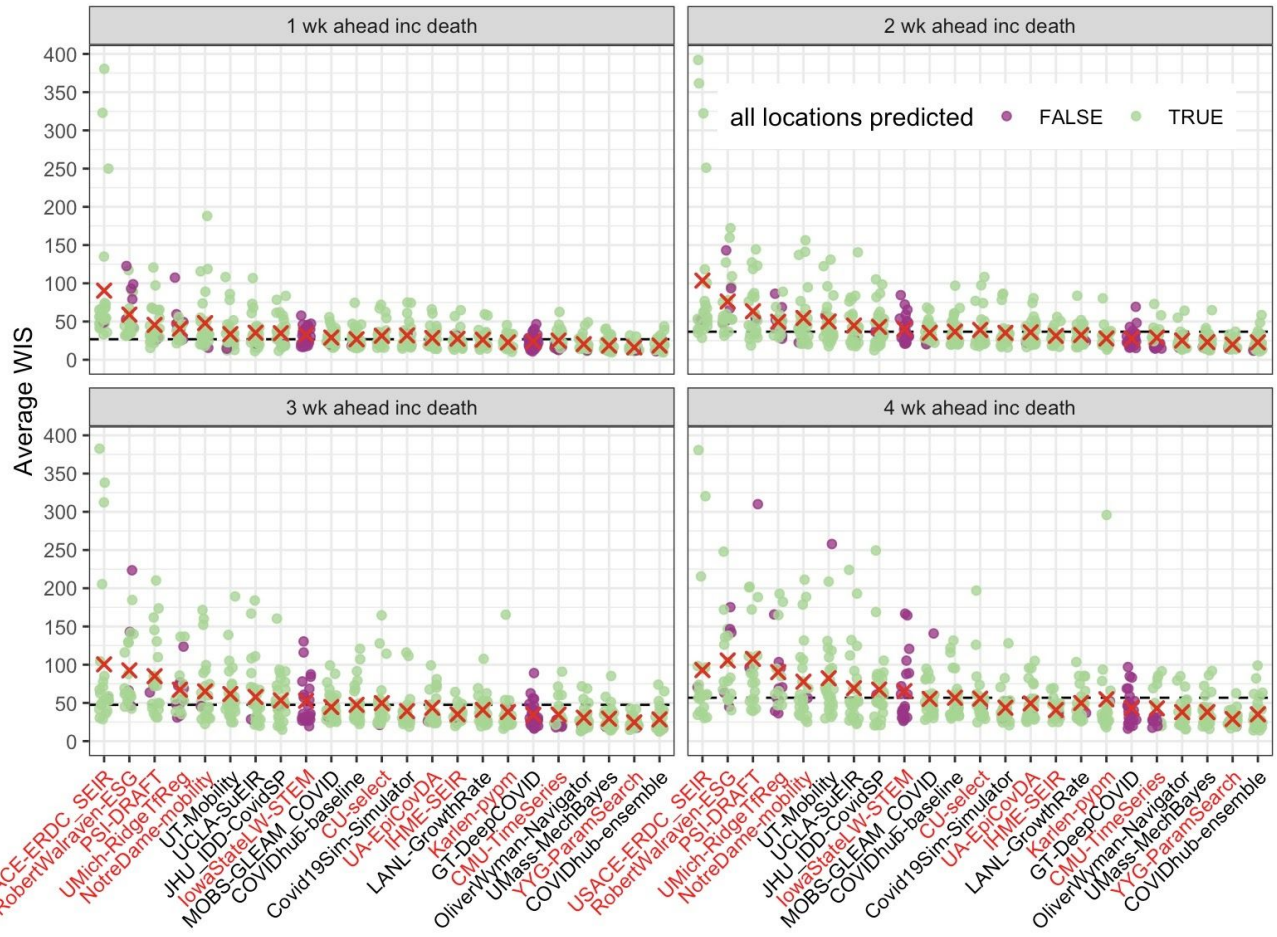


Figure 4: Average weighted interval score (WIS) by the target forecasted week for each model across all 50 states. Panel A shows the observed weekly COVID-19 deaths based on the CSSE reported data as of December 13, 2020. Panel B shows the average 1-week ahead WIS values per model (in grey). For all 21 weeks in which the ensemble model (red) is present, this model has lower WIS values than the baseline model (green) and the average score of all models (blue). Across submission weeks, there is variation in the WIS for each model. The WIS for each model is lowest in weeks where there is stability in the number of incident deaths. In submission weeks where there were large increases in incident deaths such as EW 27 and EW 31, there were also increases in the WIS values for a 1 week ahead horizon. Panel C shows the average 4-week ahead WIS by model. Similar to the 1 week ahead horizon, the ensemble model consistently has a lower WIS than the baseline model and the average across all models. For the 4-week ahead target, the variation in WIS across models is larger.

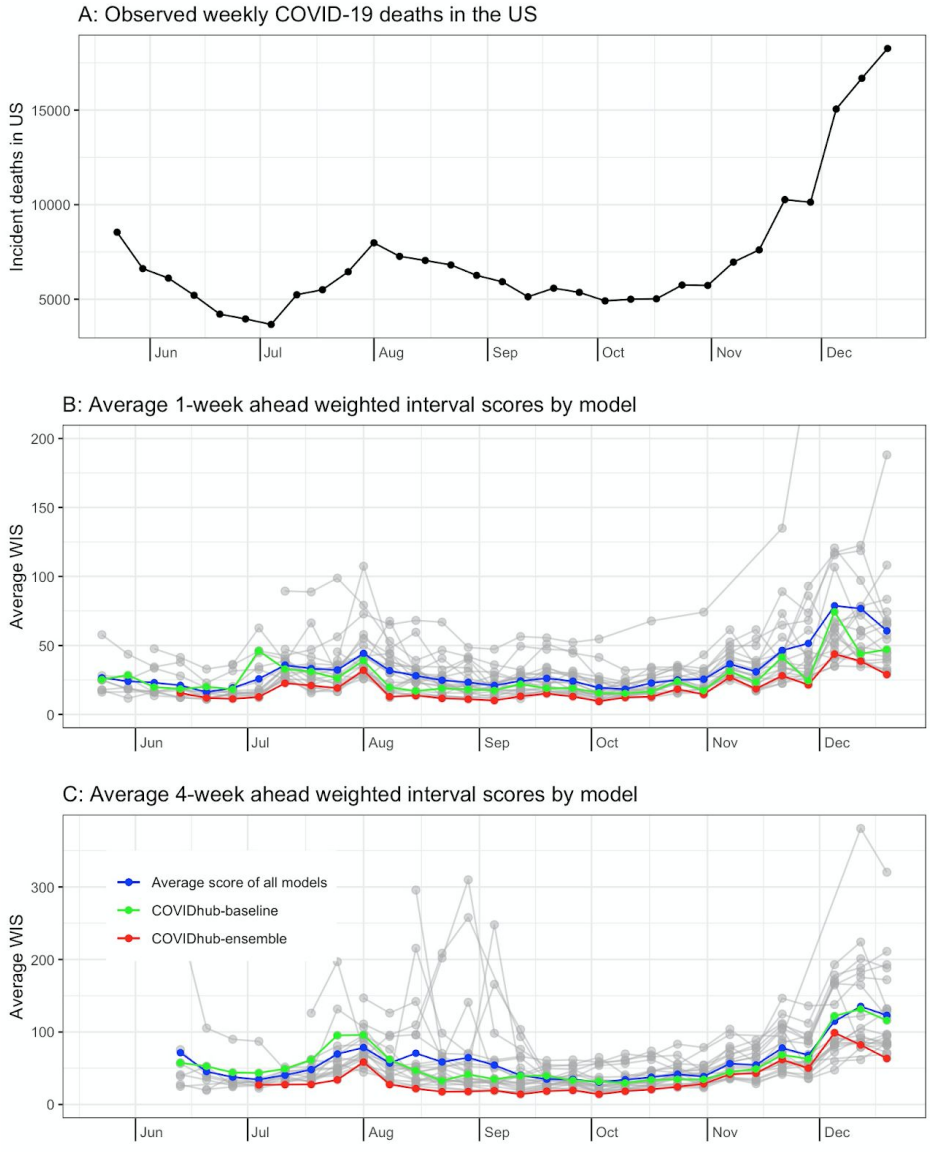


Figure 5: Relative weighted interval score (WIS) by location for each model across all horizons and submission weeks. The value in each box represents the relative WIS calculated from 1- to 4-week ahead targets available for a model at each location. Points are colored based on the relative WIS compared 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. Locations are sorted by maximum value of incident deaths. Teams on the x-axis are listed from their highest to lowest relative WIS values (Table 2). The COVIDhub-ensemble achieved the lowest average WIS overall and performed at least as well as the baseline in every location. Models missing more than 4 weeks of forecast submissions are highlighted in red text.

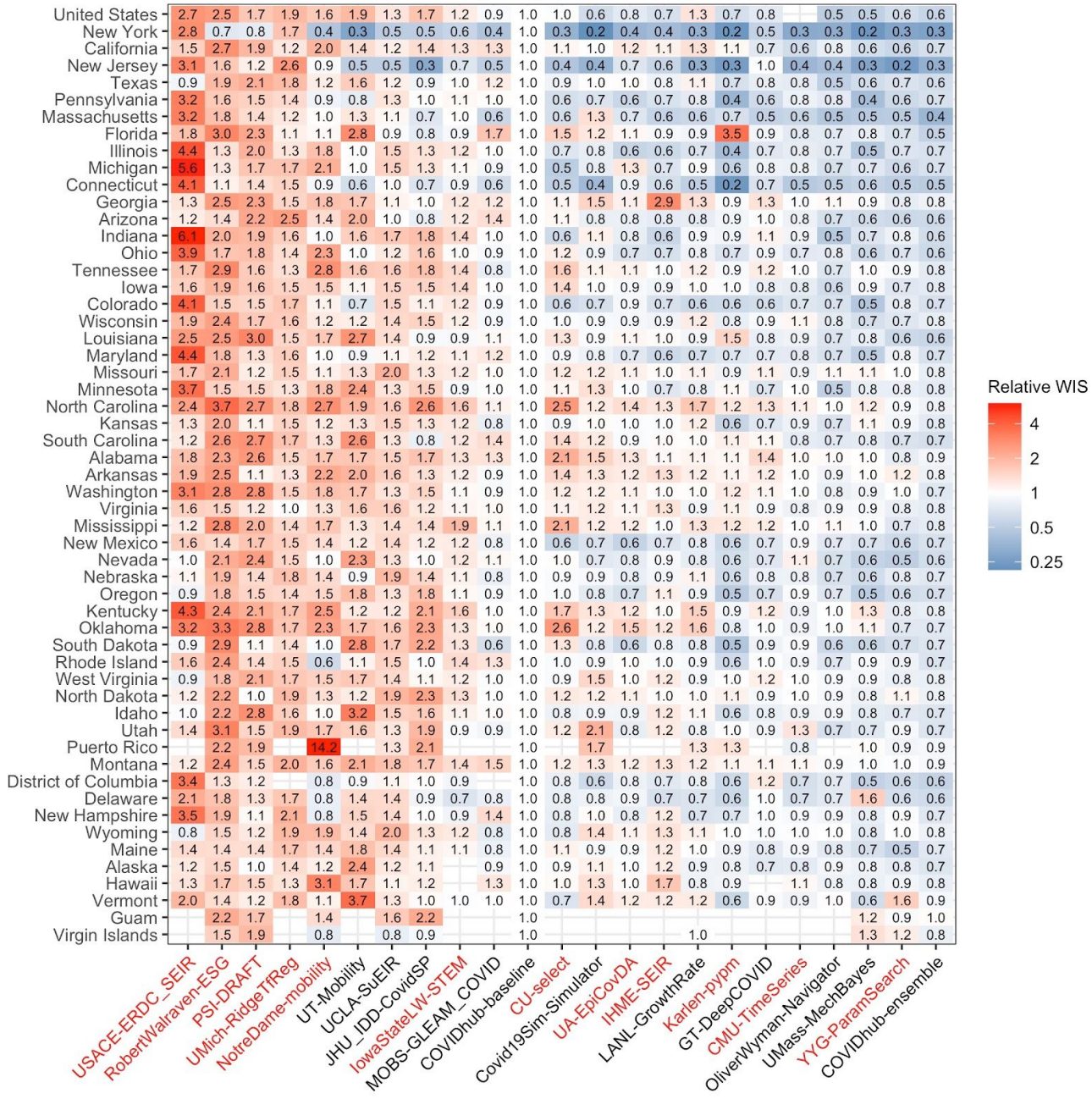
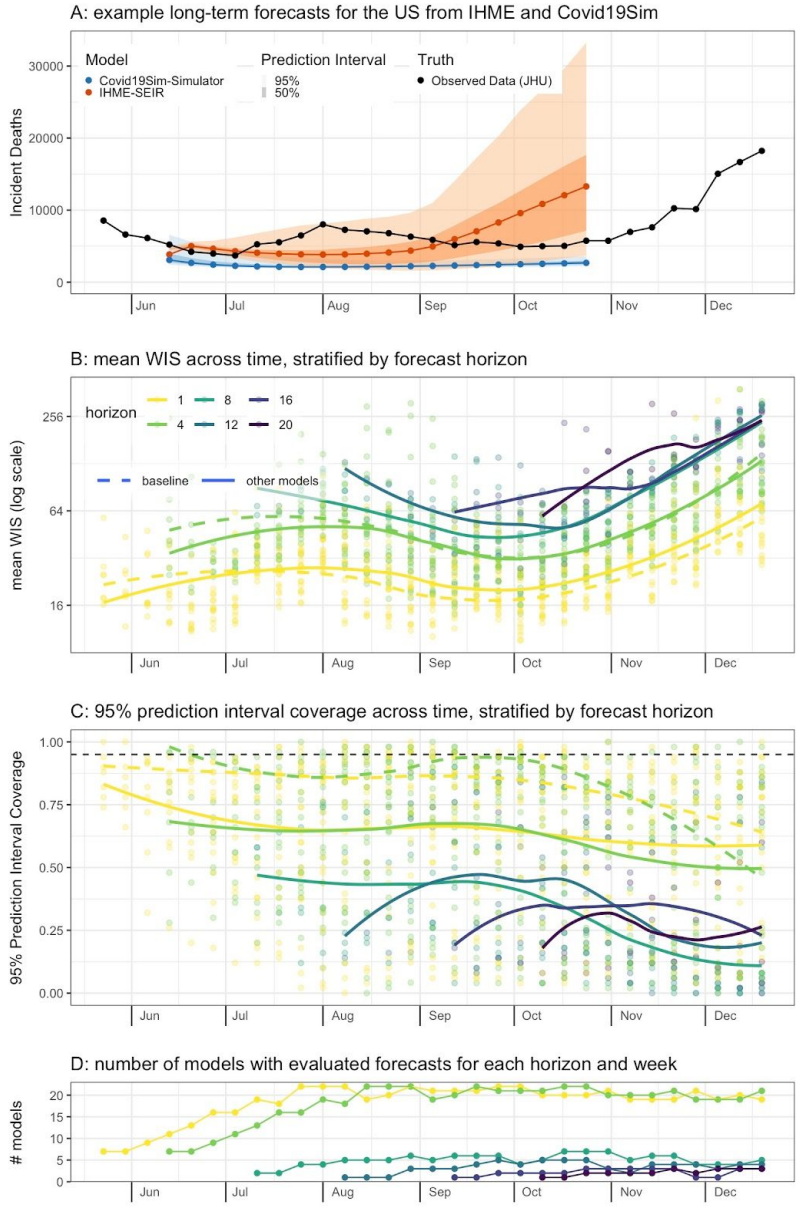


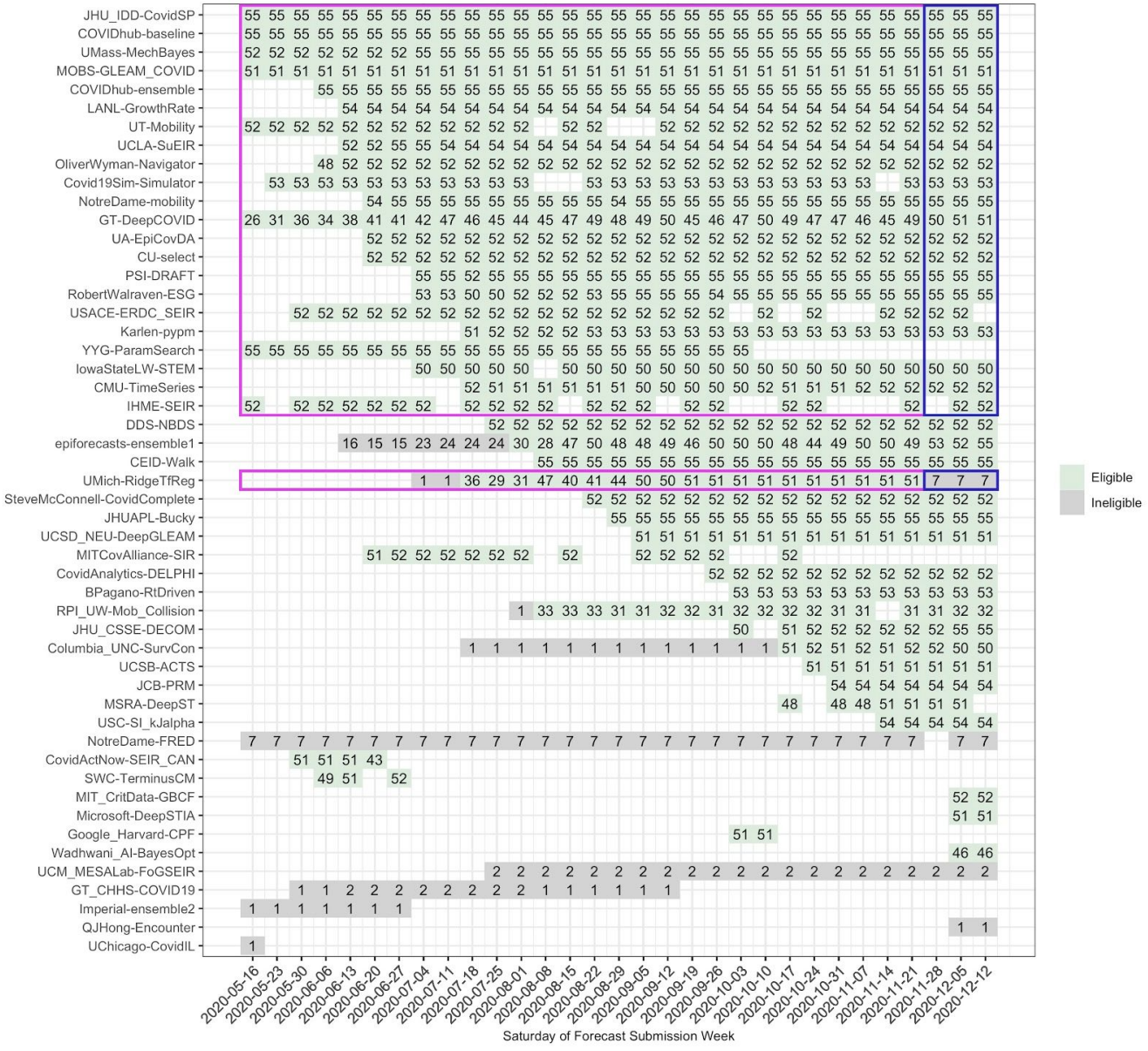
Figure 6: Evaluation of long-range forecast performance. (A) Two 20-week-ahead probabilistic forecasts submitted in early June (EW23). (B) Points show values of mean weighted interval score (WIS) for specific models and target forecast week across all states. The solid line shows the smooth trend in mean WIS across all non-baseline models, and the dashed line shows the trend for the baseline model. Lines are colored by horizon, with darker lines indicating forecasts targeting weeks further in the future. Across all weeks, mean 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 mean WIS as was achieved at a 4-week ahead horizon. (C) 95% prediction interval coverage rates stratified by color as in panel B. Coverage rates for 8- through 20-week ahead horizons were all on average below 50%. The horizontal dashed line shown at 0.95 indicates the expected coverage rate. (D) The number of models evaluated at each horizon in each week.



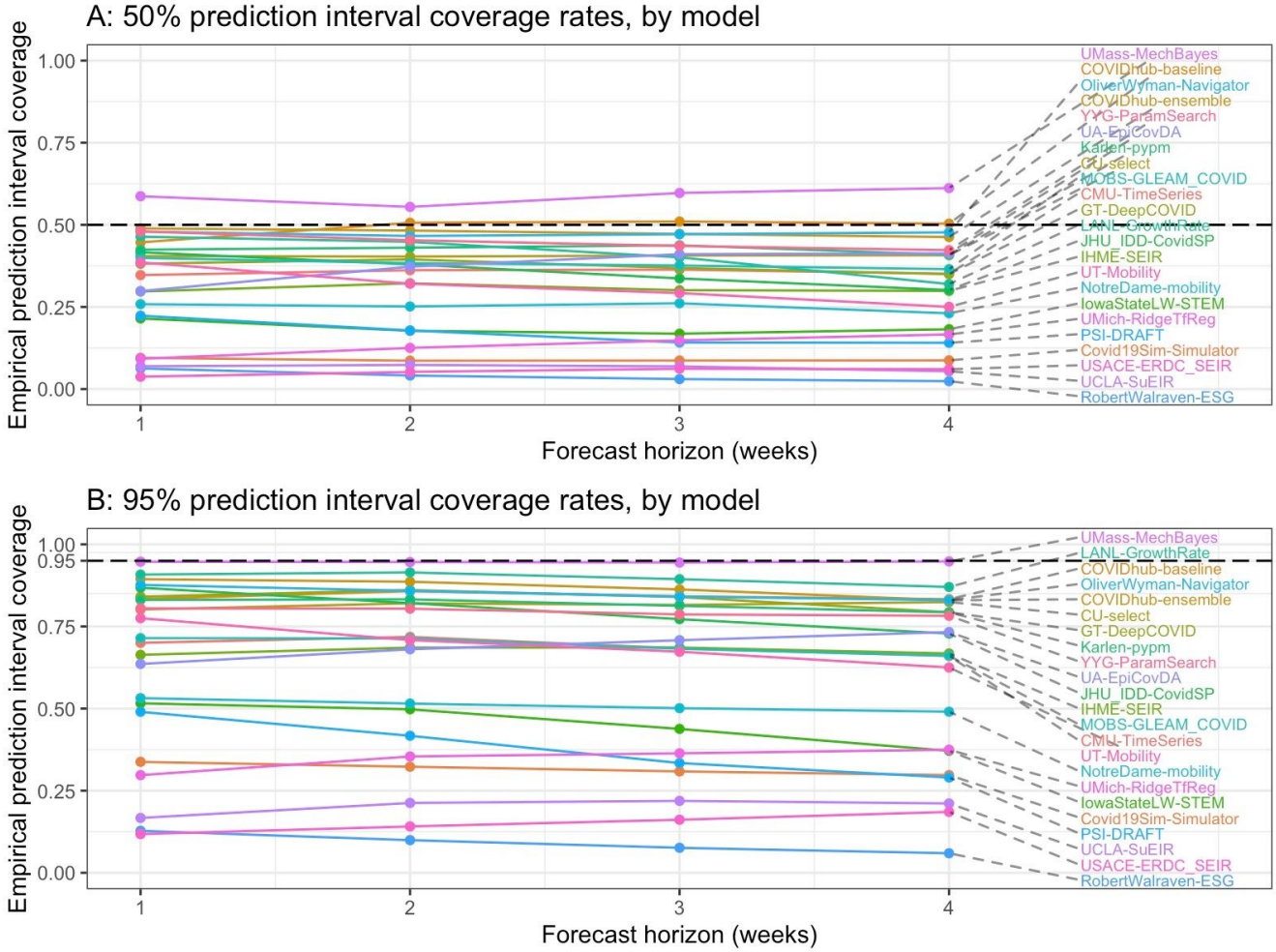
Supplemental Tables and Figures

Supplemental Figure 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 November 30, 2020. 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 occurred in a single week. In other locations, such as New Jersey, a large number of retrospective deaths were initially added to EW 26, 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 EW15 to EW 35. In locations where it was unknown when the deaths occurred, the spikes in data were not revised. This occured in Delaware during EW 26 and EW 35. 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).

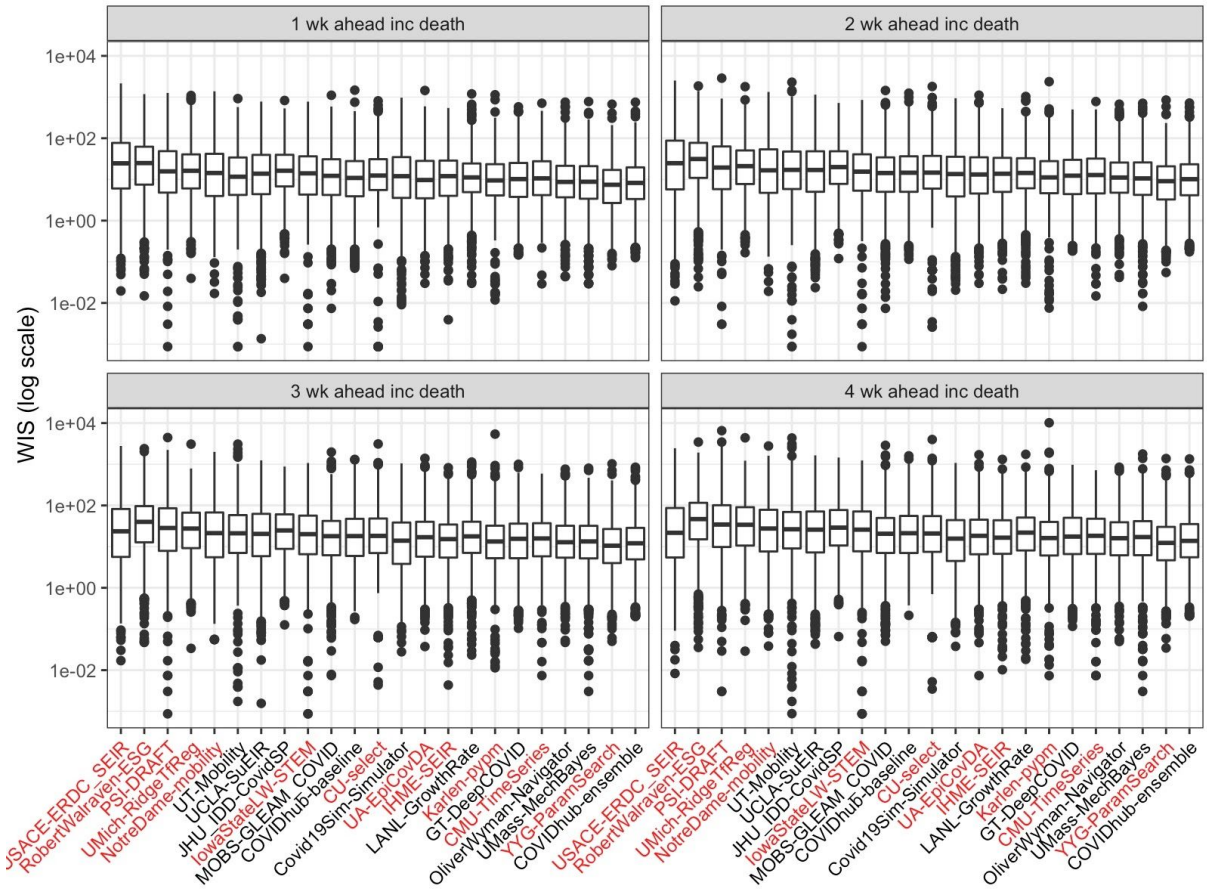
Supplemental Figure 2: Eligibility of weekly submissions from primary models that contributed incident mortality forecasts to the Forecast Hub. Text in each cell indicates the number of locations (out of the 55 evaluated) for which the model submitted incident death forecasts with all 23 quantiles. Cells that were determined to be an eligible submission are highlighted in green. The weeks used to determine overall model eligibility are outlined by a pink border, including submission weeks from 2020-05-16 through 2020-11-21. The few weeks at the end of the eligibility period that were evaluated at 1- through 3-week horizons are highlighted in blue.



Supplemental Figure 3: Summary accuracy metrics for all submitted forecasts for forecast horizons of 1 and 4 weeks, aggregated across location, and week. Forecasts for any available forecasted location (nation, state, or territory) were included in this analysis. The two panels represent the prediction interval (PI) coverage rates of 50% and 95% PIs. If the model is well calibrated, the values in the top panel should be close to 0.50 and the values in the bottom panel should be close to 0.95. The values on the horizontal axis represent the 1- to 4-week horizons forecasted each week. The value on the vertical axis represents the PI coverage.



Supplemental Figure 4: Boxplots of all observed WIS, by model and forecast horizon across all 50 states. The four panels represent each of the 1 through 4 week ahead forecast horizons. The boxplots represent the median and interquartile range of the model's average WIS aggregated by submission week and location. Based on this aggregation, the ensemble model has the lowest WIS and 10 models outperform the baseline for each horizon. Models missing more than 4 weeks of forecast submissions are highlighted in red text.



Supplemental Table 1: Summary of 25 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.**	**Demog.**	**Deaths**	**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
CEID-Walk				J		Random walk model starting from the most recent observation with a dispersion based on the spread of the last 5 observations[55]	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
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
Google_Harvard-CPF	J	CTP	BQ	J	DL	Extended SEIR model with hospitalization compartments and trainable encoders that process static and time-varying covariates to extract information from. rained in an end-to-end way with partial teacher forcing.	Yes (CHC)
GT_CHHS-COVID19	GA DPH, NC DHHS		Cen	GA DPH, NC DHHS	Cen, SG, SL	Agent-based simulation disease spread model assuming heterogeneous population mixing to predict the spread pattern geographically over time.	Yes
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		Cen	J	SG	County-level, empirical machine learning model driven by epidemiological, mobility, demographic, and behavioral data.	No
JHUAPL-Bucky	J	HHS	Cen	J	SG, PIQ	Spatial compartment model using public mobility data. Local parameters.	
MIT_CritData-GBCF	J		Cen	J	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		Cen, CDC, CL, UM	NYT	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.[56]	No
MRSA-DeepST	J			J		Deep spatio-temporal network with knowledge-based SEIR as a regularizer under the assumption of spatio-temporal process in pandemic of different regions.	
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)
QJHong-Encounter	J	CTP		J		SEIR model using encounter density to predict reproductive number	No
RPI_UW-Mob_Collision				J	G	A mobility-informed simplified SIR model motivated by collision theory.	No
SteveMcConnell-CovidComplete	CTP		Cen	J, CTP		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.	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
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[57]	No
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.[58,59]	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

BQ = Bigquery public datasets (https://cloud.google.com/bigquery/public-data), 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)[28] , 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 weighted interval score (rel WIS, θ_m^*) across two 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. For example, CU-select had a relative WIS of 1.00 in Table 2 and a relative WIS of 0.94 when earlier weeks were omitted.

	Table 2 results		**Sensitivity Analysis 1**		**Sensitivity Analysis 2**	
time period evaluated:	EW21-EW51		EW30-EW51		EW30-EW51	
inclusion criteria:	<=9 missing submissions		all models from Table 2		<= 3 missing submissions	
model	**% max obs**	**rel WIS**	**% max obs**	**rel WIS**	**% max obs**	**rel WIS**
COVIDhub-ensemble	89.8	0.63	100.0	0.65	100.0	0.66
YYG-ParamSearch	71.2	0.65	58.5	0.65	-	-
UMass-MechBayes	100.0	0.67	100.0	0.69	100.0	0.69
OliverWyman-Navigator	89.6	0.69	100.0	0.71	100.0	0.71
CMU-TimeSeries	68.7	0.76	98.9	0.80	98.9	0.81
GT-DeepCOVID	84.6	0.83	90.7	0.84	90.7	0.84
Karlen-pypm	69.5	0.86	100.0	0.91	100.0	0.88
IHME-SEIR	66.9	0.90	62.2	1.00	-	-
LANL-GrowthRate	86.4	0.90	100.0	0.95	100.0	0.94
UA-EpiCovDA	83.0	0.92	100.0	0.91	100.0	0.91
Covid19Sim-Simulator	86.5	0.94	85.4	1.04	85.4	1.04
COVIDhub-baseline	100.0	1.00	100.0	1.00	100.0	1.00
CU-select	83.0	1.00	100.0	0.94	100.0	0.92
MOBS-GLEAM_COVID	100.0	1.01	100.0	1.06	100.0	1.04
IowaStateLW-STEM	70.0	1.05	91.3	1.08	91.3	1.09
JHU_IDD-CovidSP	100.0	1.11	100.0	1.10	100.0	1.11
UCLA-SuEIR	86.4	1.21	100.0	1.30	100.0	1.32
UT-Mobility	89.8	1.41	85.4	1.64	85.4	1.60
NotreDame-mobility	83.0	1.42	100.0	1.51	100.0	1.52
UMich-RidgeTfReg	58.3	1.50	83.9	1.57	83.9	1.55
PSI-DRAFT	76.2	1.74	99.9	1.84	99.9	1.81
RobertWalraven-ESG	75.8	1.92	99.4	1.97	99.4	1.93
USACE-ERDC_SEIR	78.8	2.46	79.3	3.01	-	-