medRxiv preprint doi: https://doi.org/10.1101/2022.04.20.22274097; this version posted April 21, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

Accuracy of US CDC COVID-19 Forecasting Models

Aviral Chharia[1 2 3 +], Govind Jeevan[1 2 +], Rajat Aayush Jha[1 2 +], Meng Liu[1 4 +], Jonathan M Berman[1 5], and Christin Glorioso[1 2 6 *]

[1]Academics for the Future of Science, Cambridge, MA 02139, USA
[2]Data Informatics Center for Epidemiology, PathCheck Foundation, Cambridge, MA, USA
[3]Mechanical Engineering Department, Thapar Institute of Engineering and Technology, Patiala, PB 147004, India
[4]Department of Industrial and Manufacturing Engineering, Penn State University, PA 16802, USA
[5]Department of Basic Science, New York Institute of Technology College of Osteopathic Medicine at Arkansas State University, Jonesboro, Ar, USA
[6]Department of Anatomy, University of California, San Francisco, San Francisco, CA 94143, USA
[*]Corresponding Author: gloriosoca@gmail.com
[+]Authors claim equal contribution

ABSTRACT

Accurate predictive modeling of pandemics is essential for optimally distributing resources and setting policy. Dozens of case predictions models have been proposed but their accuracy over time and by model type remains unclear. In this study, we analyze all US CDC COVID-19 forecasting models, by first categorizing them and then calculating their mean absolute percent error, both wave-wise and on the complete timeline. We compare their estimates to government-reported case numbers, one another, as well as two baseline models wherein case counts remain static or follow a simple linear trend. The comparison reveals that more than one-third of models fail to outperform a simple static case baseline and two-thirds fail to outperform a simple linear trend forecast. A wave-by-wave comparison of models revealed that no overall modeling approach was superior to others, including ensemble models, and error in modeling has increased over time during the pandemic. This study raises concerns about hosting these models on official public platforms of health organizations including the US-CDC which risks giving them an official imprimatur and further raising concerns if utilized to formulate policy. By offering a universal evaluation method for pandemic forecasting models, we expect this work to serve as the starting point towards the development of more sophisticated models.

Introduction

The COVID-19 pandemic (1) has resulted in at least 80.3 million confirmed cases and nearly 1 million deaths in the United States (US) alone. Worldwide, cases exceed 500 million, with at least 6 million deaths (2). The pandemic has affected every country and continues to present a major threat to global health. This has caused a critical need to study the transmission of emerging infectious diseases in order to make accurate case forecasts, especially during disease outbreaks. Case prediction models are useful for developing pandemic preventive and control methods, such as suggestions for healthcare infrastructure needs, isolation of infected persons, and contact activity tracking. Accurate models can allow better decision-making about the degree of precautions necessary for a given region at a particular time, which regions to avoid travel to and the degree of risk in various activities like public gatherings. Likewise, models can be used to proactively prepare for severe surges in cases by allocating resources such as oxygen or personnel. Collecting and presenting these models gives public health officials, and organizations such as the United States Centers for Disease Control and Prevention (US-CDC) (3), a mechanism to disseminate these predictions to the public, but risks giving them an official imprimatur, suggesting that these models were either developed by a government agency or endorsed by them.

Since the start of the pandemic, dozens of case prediction models have been designed using a variety of methods. Each of these models depend on data available about cases in the US, derived from a heterogeneous system of reporting, which can vary by county and suffer from regional, temporal delays. For example, some counties may collect data over several days and make it public at once, which creates an illusion of a sudden burst of cases. In counties with less robust testing programs, the lack of data can limit modeling accuracy. These methods are not uniform or standardized between groups that perform data collection, resulting in unpredictable errors. Underlying biases in the data, such as under-reporting, can produce predictable errors in model quality, requiring models to be adjusted to predict future erroneous reporting rather than actual case numbers. Such under-reporting has been identified by serology data (4; 5). Moreover, there is no universally agreed upon system for assessing and comparing the accuracy of case prediction models. Often published models use different methods, which makes direct comparisons difficult. The CDC has taken in data of case prediction models in a standardized way which makes direct

NOTE: This preprint reports new research that has not been certified by peer review and should not be used to guide clinical practice.

medRxiv preprint doi: https://doi.org/10.1101/2022.04.20.22274097; this version posted April 21, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

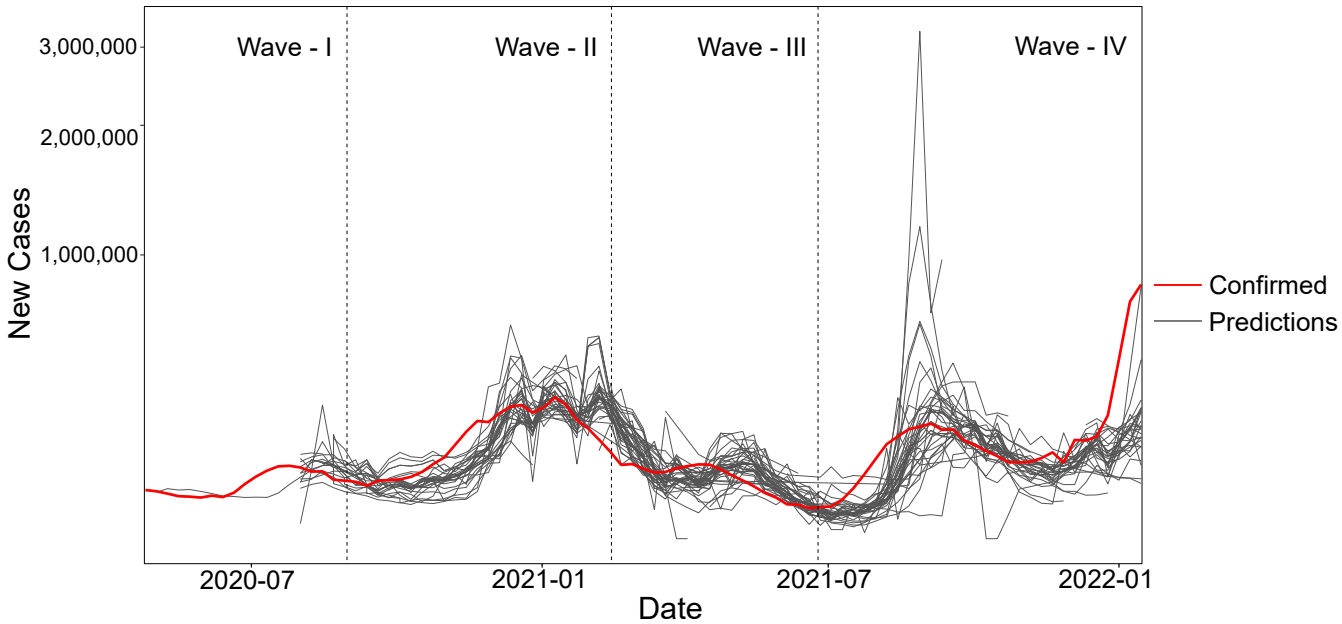

Figure 1. Visual overlay of real case counts and predicted case counts across all waves examined. Actual case counts are shown in red, predicted counts are shown in grey with each trace representing a different CDC COVID-19 forecasting model on the sqrt plot.

comparisons possible (3). In this study, we use Mean Absolute Percent Error (MAPE) compared to the true case numbers to compare models for normalization purposes. First, we consider which models have the most accurately predicted case counts with the least MAPE. Next, we divide these models into five broad sub-types based on approach, i.e., epidemiological (or compartment) models, machine learning approaches, ensemble approaches (which are constructed from the predictions of multiple individual models), hybrid approaches, and other approaches, and compare the overall error of models using these approaches. We also consider which exclusion criteria might produce ensemble models with the greatest accuracy and predictive power.

A few studies (6; 7) have compared COVID-19 case forecasting models. However, the present study is unique in several aspects. First, it is focused on prediction models of US cases and takes into consideration all CDC models that pass the set inclusion criterion. Second, since these models were uploaded in a standardized format they can be compared across several dimensions such as R_0, peak timing error, percent error, and model architecture. Third, through the study, we seek to answer several unaddressed questions relevant to pandemic case modeling. These include- (1) can we establish a metric to uniformly evaluate pandemic forecasting models? (2) what are the top-performing models during the four COVID-19 waves in the US and how do these fare on the complete timeline? (3) are there categories or classes of models that perform significantly better than others? (4) how do model predictions fare with increased forecast horizons? and (5) how do the models compare to two simple baselines?

Results

Dozens of groups have developed models to predict COVID-19 case-counts (refer Figure 1). When these models are overlaid real world data, visually several features stand out. On aggregate, models tend to approach the correct peak during various waves of the pandemic. However, some models undershot, some overshot, and many lagged the leading edge of real-world data by several weeks.

Direct comparison of error based on the difference from real-world data potentially excludes important dimensions of model accuracy. For example, a model that accurately predicts the time course of disease cases (but underestimates cases by 20% at any given point) might have greater utility for making predictions about when precautions are necessary when compared to a model that predicts case numbers with only a 5% error but estimates peak cases two weeks late. To assess the degree of timing error, the peak of each model was compared to the true peak of cases that occurred within the model time window. MAPE i.e., the ratio of the error between the true case count and a model's prediction to the true case count, is a straightforward way of representing the quality of predictive models and comparing between model types. This measure has some advantages over other methods of measuring forecast error. Because it deals with negative residuals by taking an absolute value rather than

medRxiv preprint doi: https://doi.org/10.1101/2022.04.20.22274097; this version posted April 21, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

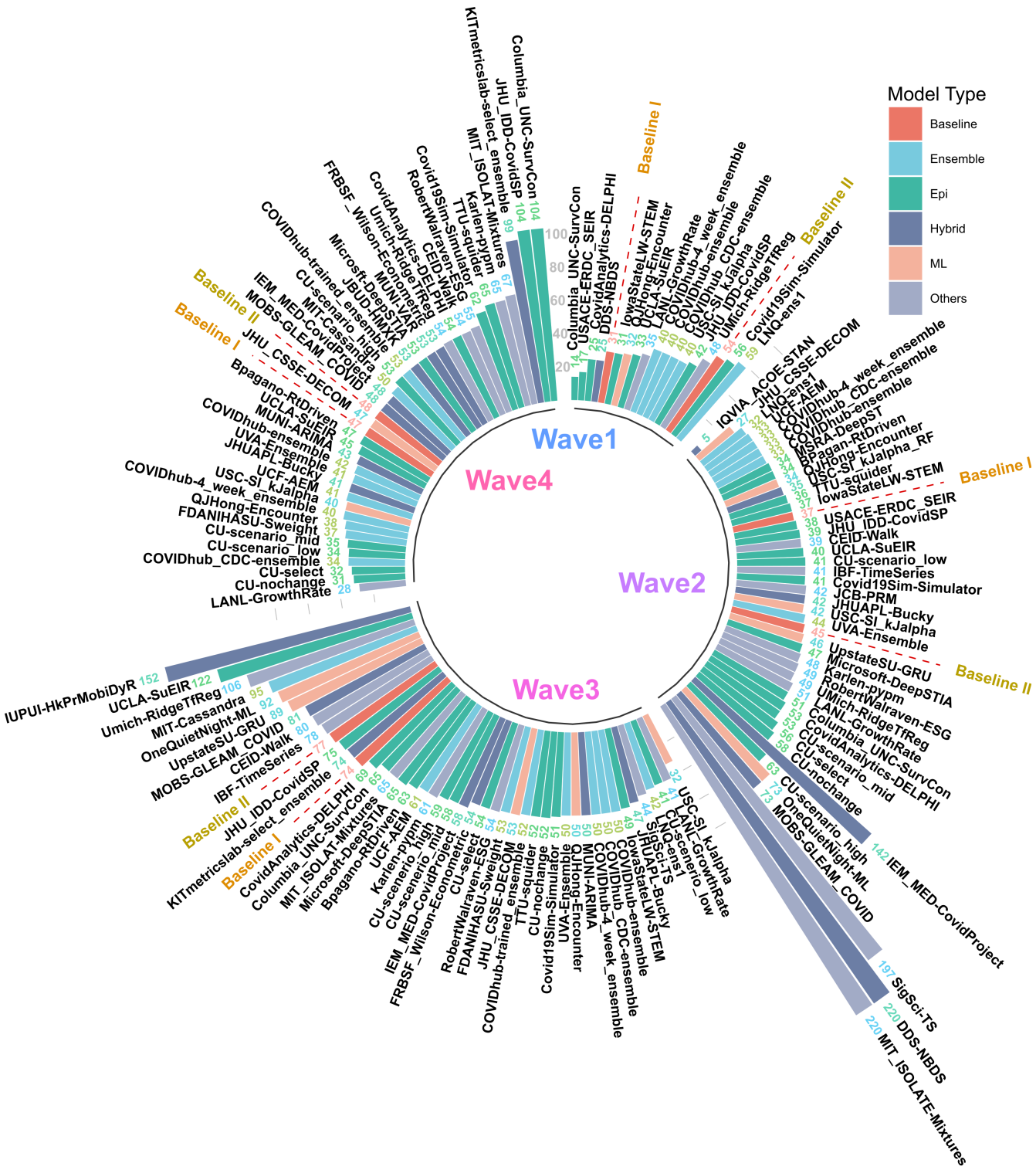

Figure 2. MAPE values of US CDC Case Prediction models in wave-I to IV. Models are sorted in descending order of MAPE. The color scheme represents the model category. Here "Baselines" are represented in red.

medRxiv preprint doi: https://doi.org/10.1101/2022.04.20.22274097; this version posted April 21, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

a squared value, reported errors are proportional. Percent error is also conceptually straightforward to understand. However, MAPE has advantages over absolute error. A model that predicts 201,000 cases when 200,000 cases occur has good accuracy but the same absolute error as a model that predicts 1100 cases when 100 occur. MAPE accounts for this by normalizing population size.

We analysed the MAPE values of all models wave-wise as shown in Figure 2. The MAPE of all models was compared to two "Baseline" models, which represented either an assumption that case counts would remain the same as the previous week (Baseline-I) or a simple linear model following the previous week's case counts (Baseline-II). In the first wave of the pandemic in the US, 'Columbia_UNC-SurvCon' achieved the lowest MAPE = 14%, closely followed by 'USACE-ERDC_SEIR' (MAPE = 17%) and 'CovidAnalytics-DELPHI' (MAPE = 25%) models. Here, only 4 models performed better than both baselines. Three of these were epidemiological models while one was a hybrid model. From the t-test, on MAPE values of models categorized on the basis of model type (refer Figure 4), it can be inferred that during the first wave, hybrid models performed the best and attained the lowest MAPE. This was followed by the epidemiological models and those based on machine learning. On the other hand, ensemble models had the largest MAPE during this wave and none of them surpassed the Baseline-I MAPE, i.e., 31%. During the second wave, 'IQVIA_ACOE-STAN' performed the best with an MAPE score of 5% (see Figure 2). In this wave, a total of 13 models performed better than both baselines, with MAPE ranging from 5 to 37. These included 5 ensemble models, 4 epidemiological models, 2 machine learning models and 2 hybrid models. All ensemble models exceed Baseline-I performance (that had MAPE = 37%), with the exception of 'UVA-Ensemble'. The epidemiological models showed a staggered MAPE distribution. Followed by the hybrid and the models categorized as 'other' model sub-types, these have the lowest average MAPE in wave-II. In contrast to wave-I, ensemble models provide the best forecasts in wave-II (see Figure 4). Here, hybrid models are the worst performing models. During wave-III (see Figure 4), ensemble models performed similarly to wave-I. Baseline models had a relatively elevated high MAPE with Baseline-I and II MAPE scores being 74% and 77% respectively. In wave-III, 'USC-SI_kJalpha' is the best-performed model with MAPE= 32% (see Figure 2). Here, 32 models performed better than both baseline models. These included 12 compartment models, 3 machine learning models, 4 hybrid models, 8 ensemble models, and 5 un-categorized models. In wave-IV of the pandemic, a number of models performed similarly between a MAPE of 28% and baseline of 47% (Figure 2). Ensemble models performed the best whereas epidemiological models had the highest MAPE during this wave. Baseline-I and II MAPE scores were 47% and 48% respectively. In wave-IV, 'LANL-GrowthRate' is the best performed model with MAPE= 28% (Figure 2). In the fourth wave, 17 models performed better than both baseline models. These included 6 compartment models, 7 ensemble models, 1 machine learning model, 2 hybrid models, and 1 uncategorized model.

The MAPE values of all models over the complete timeline was also analysed (see Figure 3). Here, we find 'IQVIA_ACOE-STAN', 'USACE-ERDC_SEIR', 'MSRA-DeepST', and 'USC-SI_kJalpha_RF' achieve the best performance with low MAPE ranging from 5% to 35%. In a comparison of overall performance, ensemble models performed significantly better than all other model types (3). However the performance of ensemble models was not "significantly" better than the baseline models (no change or simple linear model) which performed better than both machine learning and epidemiological models overall (4). MAPE of the top-10 performing models over all waves were examined (refer Figure 6). Though we examined 54 models overall (reported 51 in the full timeline), many of these did not make predictions during the first wave of the pandemic when data was less available and therefore do not appear in the wave-wise analysis. The number of weeks for which each model provides predictions were highly variable, as seen in Figure 7.

The MAPE of models in the US-CDC database increased each week out from the time of prediction. Figure 5 depicts a strictly increasing rise in MAPE with the increase in the forecast horizon. In other words, the accuracy of predictions declined the further out they were made. At one week from the time of prediction, the MAPE of models examined clustered just below 25% MAPE and declined to about 50% MAPE by four weeks. The MAPE in each week was relatively bi-modal, with several models fitting within a roughly normal distribution and others having a higher MAPE. The distribution of the points indicates that the majority of models project similar predictions for smaller forecast horizons, while the predictions for larger horizons are more spread out. The utility of model interpretation would improve by excluding those models that fall more than one standard deviation (σ) from the average MAPE of models.

Discussion

Accurate modeling is critical in pandemics for a variety of reasons. Policy decisions need to be made by political entities which must follow procedures, sometimes requiring weeks for a proposed policy intervention to become law and still longer to be implemented. Likewise, public health entities such as hospitals, nurseries, and health centers need "lead time" to distribute resources such as staffing, beds, ventilators, and oxygen supplies. However, modeling is limited, especially by the availability of data, particularly in early outbreaks (8). Resources such as ventilators are often distributed heterogeneously (9), leading to a risk of unnecessary mortality. Similarly, the complete homogeneous distribution of resources like masks is generally sub-optimal and may also result in deaths (10). Likewise, it is important to develop means of assessing which modeling tools

medRxiv preprint doi: https://doi.org/10.1101/2022.04.20.22274097; this version posted April 21, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

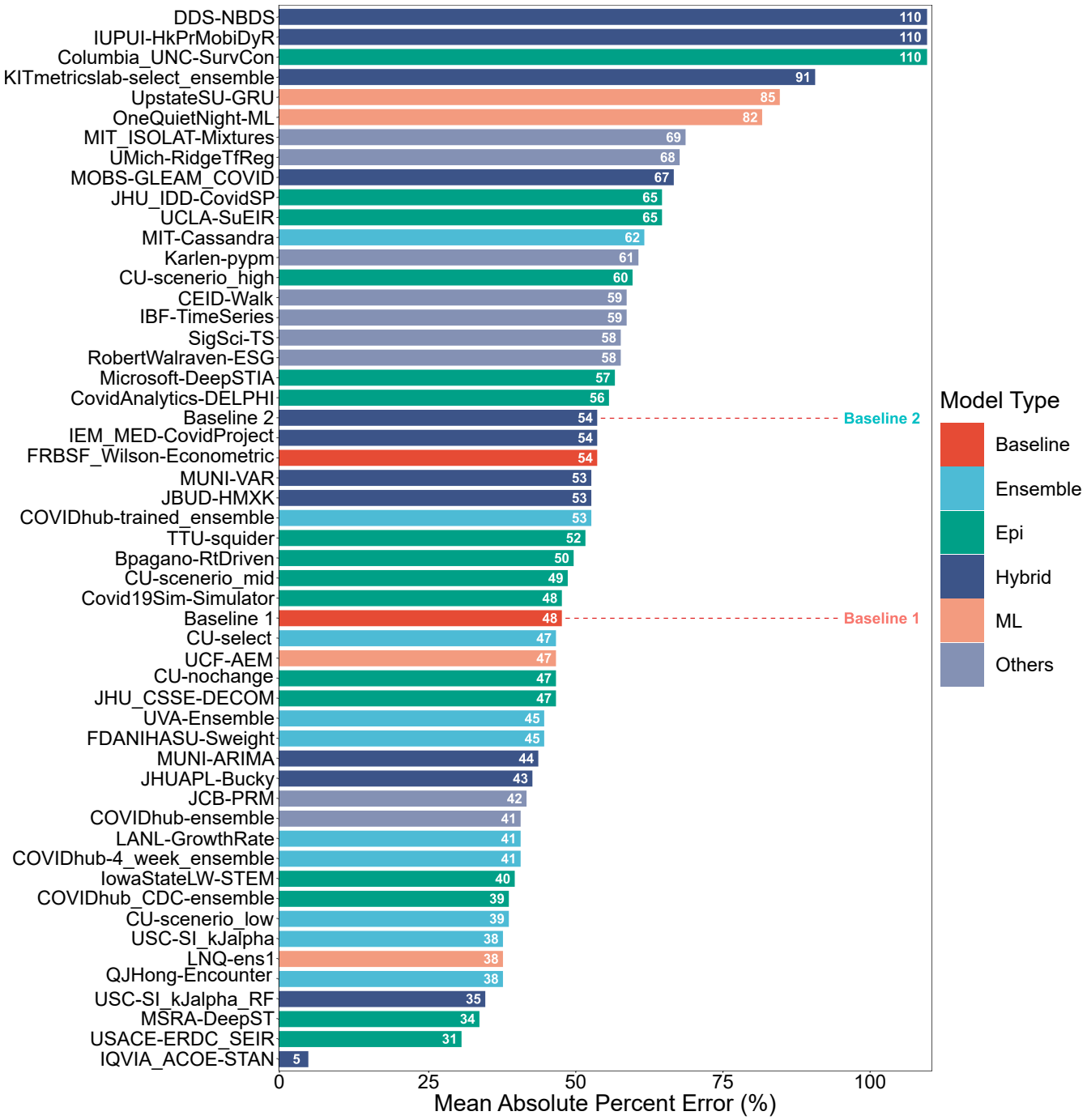

Figure 3. MAPE values of US-CDC case prediction models on the complete timeline, i.e., Wave-I to IV. The y-axis is sorted descending from lowest error to highest. The color scheme represents the model category.

medRxiv preprint doi: https://doi.org/10.1101/2022.04.20.22274097; this version posted April 21, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

are most effective and trustworthy. For example, a case forecasting model that consistently makes predictions that fare worse than assuming that case counts will remain unchanged or that they will follow a simple linear model isn't likely to be useful in situations where modeling is critical. The use of these baselines allows for the exclusion of models that "fail" to predict case counts adequately.

The average MAPE of successful models which we define as lower MAPE than either baseline model varied between methods depending on the wave they were measured in. In the first wave, epidemiological models had a mean MAPE of 31.11, and machine learning models had a mean MAPE of 31.97. In the second wave, these were 44.87 and 44.25 respectively. In the third wave, these were 62.5 and 63.15, and in the fourth wave, these were 91.48 and 43.58 respectively. Therefore, we notice that the mean MAPE of models got worse with each wave. This is because each model type is susceptible to changing real-world conditions, such as the emergence of new variants with the potential to escape prior immunity, or a higher R_0, new masking or lockdown mandates, the spread of conspiracy theories, or the development of vaccines which will decrease the number of individuals susceptible to infection.

In waves-2 and 3, the "best" performing model was a hybrid and machine learning model respectively. In waves 1 and 4, the best model was an other and epidemiological model respectively. In each wave, some examples of each model type were successful, and some were unsuccessful. No model category did significantly better than baseline models except in wave-3 (Figure 4). In wave-3, compartment models were significantly better than baseline models, although ensemble models had the lowest overall MAPE. These results do not suggest that either overall modeling technique is inherently superior for predicting future case counts.

Compartment (or epidemiological) models broadly use several "compartments" which individuals can move between such as "susceptible," "infectious," or "recovered," and use real-world data to arrive at estimates for the transition rate between these compartments. However, the accuracy of a compartment model depends heavily on accurate estimates of the R_0 in a population, a variable that changes over time, especially as new virus variants emerge. For example, the emergence of the Iota variant of SARS-CoV-2 (also known as lineage B.1.526) resulted in an unpredicted increase in case counts (11). Machine learning models train algorithms that would be difficult to develop by conventional means. These models "train" on real-world data sets and then make predictions based on that past data. Machine learning models are sensitive to the datasets they are trained on, and small or incomplete datasets produce unexpected results. Hybrid models make use of both compartment modeling and machine learning tools. On the other hand, ensemble models combine the results of multiple other models hoping that whatever errors exist in the other models will "average out" of the combined model. Ensemble models potentially offer an advantage over individual models in that by averaging the predictions of multiple models, flawed assumptions or errors in individual models may "average out" and result in a more accurate model. However, if multiple models share flawed assumptions or data, then averaging these models may simply compound these errors. An individual model may achieve a lower error than multiple flawed models in these cases.

The "baseline" models had a MAPE of 48 and 54 over the entire course of the pandemic in the US. Most models did not perform better than these. Among those that did, we discuss the five with the best performance (in increasing order of their performance). First, "QJHong-Encounter" is a model by Qijung Hong, an Assistant professor at Arizona State University. This model uses an estimate of encounter density (how many potentially infectious encounters people are likely to have in a day) to predict changes in estimated R (reproduction number), and then uses that to estimate future daily new cases. The model uses machine learning. It had a MAPE of 38 over all waves. Second, "USC-SI_kJalpha_RF" is a hybrid model from the University of Southern California Data Science Lab. This model also uses a kind of hybrid approach, where additional parameters are modeled regionally for how different regions have reduced encounters, and machine learning is used to estimate parameters (12). It had a MAPE of 35 over all waves. Third, "MSRA-DeepST" is a SIR hybrid model from Microsoft Research Lab-Asia that combines elements of SEIR models and machine learning. It had a MAPE of 34. Next, "USACE-ERDC_SEIR" is a compartment model developed by the US Army Engineer Research and Development Center COVID-19 Modeling and Analysis Team. It adds to the classic SEIR model, adding additional compartments for unreported infections and isolated individuals. It used Bayesian estimates of prior probability based on subject matter experts to select initial parameters. It had a MAPE of 31. Lastly, "IQVIA_ACOE_STAN" is a machine learning model from IQVIA-Analytics Center of Excellence and has the highest apparent performance on the overall timeframe. The calculated MAPE for this model was 5. Notably, the MAPE of 5 is much lower than the MAPE of 31 of the next closest model.

Although MAPE is superior to other methods of comparing models, there are still some challenges. For example, "IQVIA_ACOE_STAN" appears to be the lowest model by far, but the only data available covers a relatively short time frame, and does not include any changes in case direction from upswings and downswings. This advantages the model compared to other models which might cover time periods where case numbers peak or new variants emerge. This highlights a potential pitfall of examining this data: the models are not studying a uniform time window.

Data reporting is one of the sources of error affecting model accuracy. There is significant heterogeneity in reporting of COVID-19 cases by state. This is caused by varying state laws, resources made available for testing, the degree of sequencing

medRxiv preprint doi: https://doi.org/10.1101/2022.04.20.22274097; this version posted April 21, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

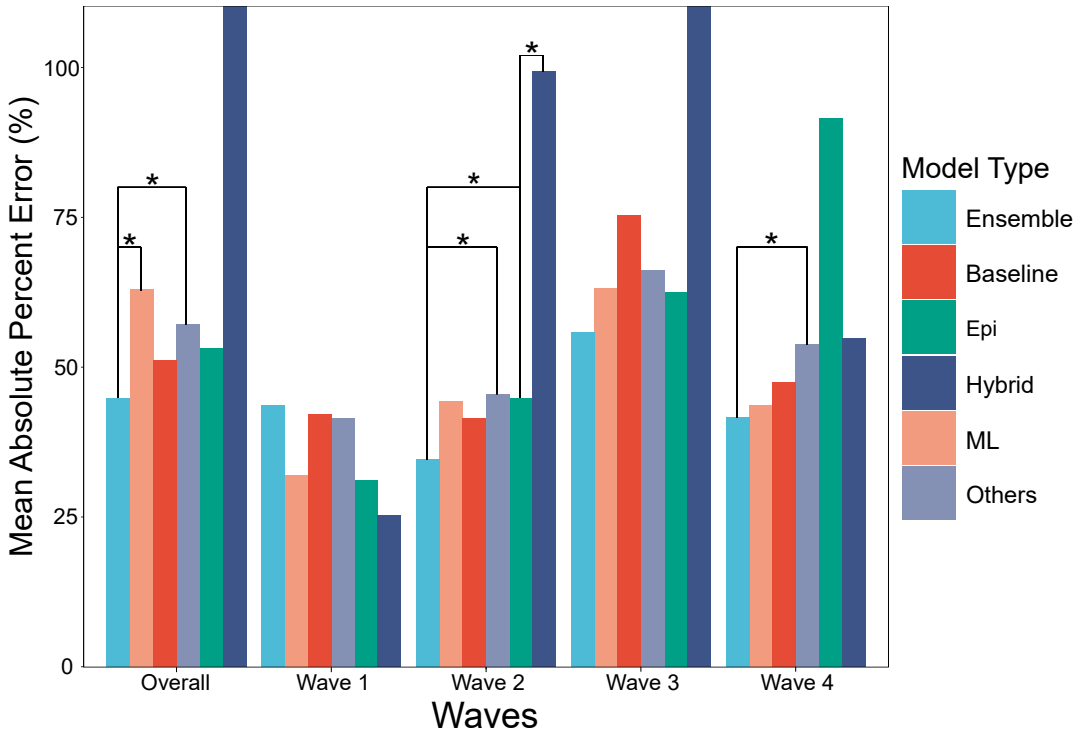

Figure 4. Bar graph showing the paired t-test results. Category-wise error achieved by the models on both overall and wave-wise from wave-1 to wave-4.

being done in each state, and other factors. Additionally, different states have heterogeneity in vaccination rate, population density, implementation of masking and lockdown, and other measures which may affect case count predictions. Therefore, the assumptions underlying different models and the degree to which this heterogeneity is taken into account may result in models having heterogeneous predictive power in different states. Although this same thinking could be extended to the county level, the case count reporting in each county is even more variable and makes comparisons difficult.

To make the comparison between models more even, we used multiple times segments to represent the various waves during the pandemic. Models that have made fewer predictions, particularly avoiding the "regions of interest" such as a fresh wave or a peak, would only be subjected to a less challenging evaluation than the models that covered most of the timeline. Further, the utility of models with only a few predictions within "regions of interest" is also questionable. Considering these aspects, we define 04 time segments corresponding to each of the major waves in the US. Within each of these waves of interest, we consider only models that made a significant number of predictions for the purpose of comparison. Such a compartmentalized comparison is now straightforward, as all models within a time segment can now have a common evaluation metric.

We focus on evaluating the performance of models for their 4-week ahead forecasts. A larger forecast horizon provides a higher real-world utility in terms of policy-making or taking precautionary steps. We believe this to be a more accurate representation of the predictive abilities of models as opposed to smaller windows in which desirable results could be achieved by simply extrapolating the present trend. Therefore, due to the time delays associated with policy decisions and the movement of critical resources and people, the long-term accuracy of models is of critical importance. To take an extreme example, a forecasting model that only predicted one day in advance would have less utility than one that predicted ten days in advance.

The failure of roughly one-third of models in the CDC database to produce results superior to a simple linear model should raise concerns about hosting these models in a public venue. Without strict exclusion criteria, the public may not be aware that the are significant differences in the overall quality of these models. Each model type is subject to inherent weaknesses of the available data. The accuracy of compartment models is heavily dependent on the quality and quantity of reported data and also depends on a variable that might change with the emergence of new variants. Heterogeneous reporting of case counts, variable accuracy between states, and variable early access to testing resulted in limited data sets. Likewise, it seems that since training sets did not exist, machine learning models were unable to predict the Delta variant surge. Robust evidence-based exclusion criteria and performance-based weighting have the potential to improve the overall utility of future model aggregates and ensemble models.

medRxiv preprint doi: https://doi.org/10.1101/2022.04.20.22274097; this version posted April 21, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

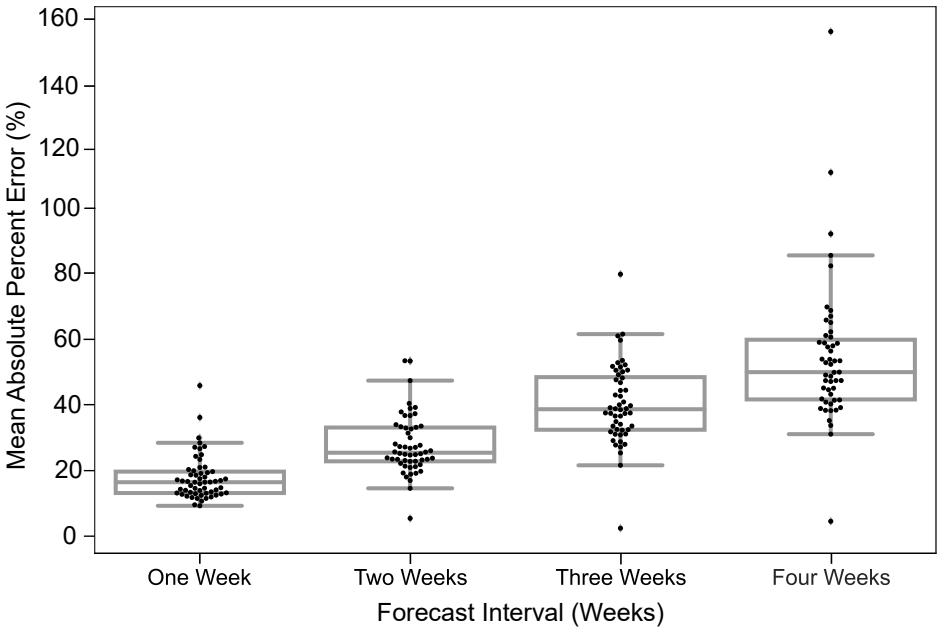

Figure 5. Predictions are most accurate closest to the time of prediction. The MAPE in predictions of all models for different forecast horizons is shown. The dots in each box plot represent the MAPE over all the predictions of a certain model for the corresponding forecast horizon. The y-axis is the MAE between the predicted case count and the reported case count. The x-axis is the forecast horizon.

Because the US-CDC has a primary mission focused on the United States, the models included are focused on United States case counts. However, globally the assumptions necessary to produce an accurate model might differ due to differences in population density, vaccine availability, and even cultural beliefs about health. However, identifying the modeling approaches that work best in the United States provides a strong starting point for global modeling. Some of the differences in modeling will be accounted for by different input data, which can be customized by country or different training sets in the case of machine learning models.

The ultimate measure of forecasting model quality is whether the model makes a prediction that is used fruitfully to make a real-world decision. Staffing decisions for hospitals can require a lead time of 2-4 weeks to prevent over-reliance on temporary workers, or shortages (13). Oxygen has become a scarce resource during the COVID-19 pandemic, and also needs lead time (14). This has had real-world policy consequences as public officials have ordered oxygen imports well after there were needed to prevent shortages (15). Indeed for sufficient time to be available for public officials to enact new policies and for resources to be moved, a time frame of eight weeks is preferable.

The need for accurate predictions weeks in advance is confounded by the declining accuracy of models multiple weeks in advance, especially considering the rise time of new variant waves. During the most recent wave of infections, news media reported on the potential for the "omicron" variant of SARS-CoV-2 to rapidly spread in November of 2021, however in the United States, an exponential rise did was not apparent in case counts until December 14th, when daily new case counts approximated 100k new cases per day, and by January 14th, 2022, new cases exceed 850,000 new cases per day. The time when accurate predictive models are most useful is ahead of rapid rises in cases, something none of the models examined were able to predict, given the rise in SARS-CoV-2 cases that occurred during the pandemic.

Although forecasting models have gained immense attention during the COVID-19 pandemic, many challenges are faced in developing forecasting models that serve the needs of governments and organizations. Since COVID-19 has a long incubation period of up to 14 days, identifying patients beforehand presents its own set of challenges. Patients can infect those in close contact with them throughout the virus incubation period when they may be asymptomatic and unaware that they are infected. In a real-world scenario, it is very difficult to implement a complete lockdown or define a metric for evaluating the effectiveness of a lockdown. Relating the optimal time period of the lockdown, taking into consideration the incubation period of the virus, is a complex task. The lack of clean, structured, and accurate datasets also affects the performance of the case prediction models, making the estimation of patient mortality count and transmission rate significantly harder.

We found that most models are not better than CDC baselines. The benchmark for resource allocation ahead of a wave remains the identification and interpretation of new variants and strains by biologists. We, therefore, propose a reassessment of

medRxiv preprint doi: https://doi.org/10.1101/2022.04.20.22274097; this version posted April 21, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

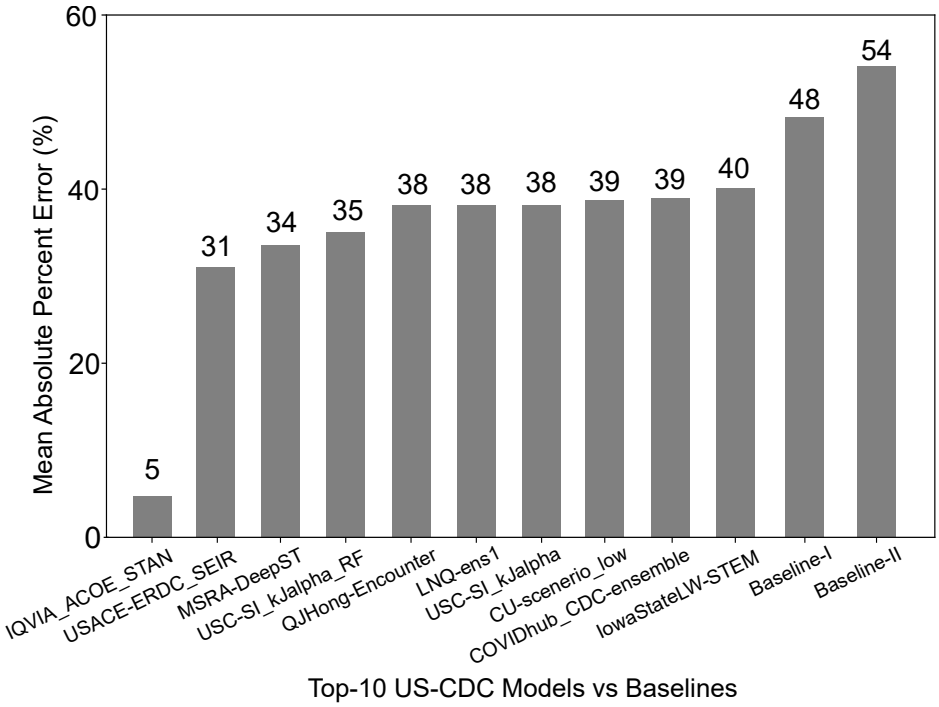

Figure 6. The MAPE of the top-10 best performing models in decreasing order of predictions performance over complete timeline (Wave 1-4 combined), as well as two "Baselines" that either followed a simple linear model based on the previous week's case counts (Baseline-II), or that case counts would remain the same (Baseline-I). Most prediction models did little better than control models.

the role of forecasting models in pandemic modeling. These models as currently implemented can be used to predict the peak and decline of waves that have already been initiated and can provide value to decision-makers looking to allocate resources during an outbreak. Prediction of outbreaks beforehand however still requires "hands-on" identification of cases, sequencing, and data gathering. More robust sources of data on true case numbers, variants, and immunity would be useful to create more accurate models that the public and policy makers can use to make decisions.

Methods

This work compares various US-CDC COVID-19 forecasting models by their quantitative aspects evaluating their performance in strictly numerical terms over various time segments. The US-CDC collects weekly forecasts for COVID cases in four different horizons: 1-week, 2-weeks, 3-weeks and 4-weeks, i.e., each week, the models make a forecast for new COVID cases in each of the four consecutive weeks from the date of the forecast. The forecast horizon is the length of time into the future for which forecasts are to be prepared. In the present study, we focus on evaluating the performance of models for their 4-week ahead forecasts.

The data for the confirmed case counts are taken from the COVID-19 Data Repository (https://github.com/CSSEGISandData/COVID-19/tree/master/csse_covid_19_data/csse_covid_19_daily_reports_us), maintained by the Center for Systems Science and Engineering (CSSE) at the Johns Hopkins University (JHU). The data for the predicted case counts, of all the models, is obtained from the data repository for the COVID-19 Forecast Hub (https://github.com/reichlab/covid19-forecast-hub), which is also the data source for the official US-CDC COVID-19 forecasting page. Both these datasets were preprocessed to remove unwanted data items. For plotting Figure 1, the *Pyplot* module from the *Matplotlib* library in Python was used.

The models are categorized into five different categories- Ensemble, Epidemiological, Hybrid, and Machine Learning. The models which did not broadly fall into these categories were kept in 'others'. These models use very different methods to arrive at predictions. We are comprehensively looking at 51 models. The CDC also uses an ensemble model, and we looked at whether this was better than any individual model. For each model uploaded to the CDC website, MAPE was calculated and reported in this study, and the models were compared wave-wise as shown in various figures. For each model, the model type was noted, as well as the month proposed. The term "wave" here implies a natural pattern of peaks and valleys. As part of their National Forecasts for COVID cases, CDC has reported the results from a total of 54 different models at various instances of

medRxiv preprint doi: https://doi.org/10.1101/2022.04.20.22274097; this version posted April 21, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

Figure 7. Bar plot depicting frequency of 04 week ahead predictions made by models. Here, models were ordered alphabetically on the y-axis. The x-axis represents target dates for which the predictions were made. Dates range from July 2020 to Jan 2022.

medRxiv preprint doi: https://doi.org/10.1101/2022.04.20.22274097; this version posted April 21, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

time during the pandemic. We define the waves, i.e., Wave-I: July, 6th 2020 to August 31st, 2020, Wave-II: September, 1st 2020 to February, 14th 2021, Wave-III: February, 15th 2021 to July, 26th 2021 and Wave-IV: July, 27th 2021 to January, 17th 2022, corresponding to each of the major waves in the US.

The performance of the models was evaluated against two simple baselines. Baseline-I is the 'CovidHub-Baseline' (or CDC's baseline), i.e., the median prediction at all future horizons is the most recent observed incidence. Baseline-II is the linear predictor extrapolation using slope of change in reported active cases between the two weeks preceding date of forecast. These baselines are included in the bar charts (shown in Figure 2 and Figure 3). Within each of the waves of interest, only models that made a significant number of predictions are considered for the purpose of comparison. We only consider models that have made predictions for at least 25% of the target dates covered by the respective time segment for all comparisons in this section. The MAPE was calculated on the four-weeks forecast horizon. Figure 3 illustrates the performance of models across all the waves. Same procedure, as in Figure 2, was followed for plotting Figure 3.

The performance of all the models are compared, wave-wise and on the complete timeline based on the MAPE (or mean absolute percent error). MAPE is defined as the ratio of absolute percentage errors of the predictions. Error refers to the difference between the confirmed case counts and the predicted case counts. Here, n = number of datapoints, A_t is the actual value and P_t denoted the predicted value.

$$MAPE = (1/n)\sum_{t=1}^{n} |(A_t - P_t)/A_t| \tag{1}$$

Figure 4 shows the paired t-test results on the category-wise errors, achieved by the models on overall as well as wave-wise. The mean of the MAPE values is calculated for each category in model-type, for overall and each wave separately. The mean is calculated by adding the MAPE values of all the models in a category and dividing it by the number of models in that category for the corresponding wave. Then, pairwise t-test is performed to determine if there is a significant difference between the means of two groups. We used the *ttest_ind* function of the *statsmodels* module in Python to perform the test. The statistical significance is determined by the p-value given as the output. In null-hypothesis significance testing, the p-value is the probability of obtaining test results at least as extreme as the results actually observed, under the assumption that the null hypothesis is correct. A p-value less than 0.05 (typically ≤ 0.05) is statistically significant. It indicates strong evidence against the null hypothesis, as there is less than 05% probability that null-hypothesis is correct. The null hypothesis is a typical statistical theory which suggests that no statistical relationship and significance exists in a set of given single observed variable, between two sets of observed data and measured phenomena. So, the bar pots having p-value less than 0.05 (statistically significant), are joined using an asterisk.

In Figure 5, box-plots are made representing the MAPE over all the predictions of a certain model for the corresponding forecast horizon. A box-plot is a standardized way of displaying the distribution of data based on a five number summary ("minimum", first quartile (Q1), median, third quartile (Q3), and "maximum"). We used the boxplot function (of *seaborn* library) in python to plot it. *Seaborn* is a Python data visualization library based on *Matplotlib*.

References

1. Zhu, N. *et al.* A novel coronavirus from patients with pneumonia in china, 2019. *New Engl. journal medicine* DOI: https://doi.org/10.1056/NEJMoa2001017 (2020).

2. Dong, E., Du, H. & Gardner, L. An interactive web-based dashboard to track covid-19 in real time. *The Lancet infectious diseases* **20**, 533–534, DOI: https://doi.org/10.1016/S1473-3099(20)30120-1 (2020).

3. CDC. Covid data tracker (2020). https://covid.cdc.gov/covid-data-tracker.

4. Albani, V. V., Loria, J., Massad, E. & Zubelli, J. P. Covid-19 underreporting and its impact on vaccination strategies. *medRxiv* DOI: https://doi.org/10.1186/s12879-021-06780-7 (2021).

5. Stadlbauer, D. *et al.* Repeated cross-sectional sero-monitoring of sars-cov-2 in new york city. *Nature* **590**, 146–150, DOI: https://doi.org/10.1038/s41586-020-2912-6 (2021).

6. Alakus, T. B. & Turkoglu, I. Comparison of deep learning approaches to predict covid-19 infection. *Chaos, Solitons & Fractals* **140**, 110120, DOI: https://doi.org/10.1016/j.chaos.2020.110120 (2020).

7. Fonseca i Casas, P., García i Carrasco, V. & Garcia i Subirana, J. Seird covid-19 formal characterization and model comparison validation. *Appl. Sci.* **10**, 5162, DOI: https://doi.org/10.3390/app10155162 (2020).

medRxiv preprint doi: https://doi.org/10.1101/2022.04.20.22274097; this version posted April 21, 2022. The copyright holder for this preprint **(which was not certified by peer review)** is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

8. Khilji, S. U. S. *et al.* Distribution of selected healthcare resources for influenza pandemic response in cambodia. *Int. journal for equity health* **12**, 1–14, DOI: https://doi.org/10.1186/1475-9276-12-82 (2013).

9. Rudge, J. W. *et al.* Health system resource gaps and associated mortality from pandemic influenza across six asian territories. *PloS one* **7**, e31800, DOI: https://doi.org/10.1371/journal.pone.0031800 (2012).

10. Worby, C. J. & Chang, H.-H. Face mask use in the general population and optimal resource allocation during the covid-19 pandemic. *Nat. communications* **11**, 1–9, DOI: https://doi.org/10.1038/s41467-020-17922-x (2020).

11. Nandakishore, P. *et al.* Deviations in predicted covid-19 cases in the us during early months of 2021 relate to rise in b. 1.526 and its family of variants. *medRxiv* DOI: https://doi.org/10.1101/2021.12.06.21267388 (2021).

12. Srivastava, A., Xu, T. & Prasanna, V. K. Fast and accurate forecasting of covid-19 deaths using the sikjα model, DOI: https://doi.org/10.48550/arXiv.2007.05180 (2020).

13. Drake, R. G. Does longer roster lead-time reduce temporary staff usage? a regression analysis of e-rostering data from 77 hospital units. *J. Adv. Nurs.* **74**, 1831–1838, DOI: https://doi.org/10.1111/jan.13578 (2018).

14. Bonnet, L., Carle, A. & Muret, J. In the light of covid-19 oxygen crisis, why should we optimise our oxygen use? *Anaesthesia, Critical Care & Pain Medicine* DOI: https://dx.doi.org/10.1016%2Fj.accpm.2021.100932 (2021).

15. Bhuyan, A. Experts criticise india's complacency over covid-19. *The Lancet* **397**, 1611–1612, DOI: https://doi.org/10.1016/S0140-6736(21)00993-4 (2021).

16. Khan, Z. S., Bussel, F. V. & Hussain, F. A predictive model for covid-19 spread applied to eight us states, DOI: https://doi.org/10.48550/arXiv.2006.05955 (2020).

17. Lemaitre, J. C. *et al.* A scenario modeling pipeline for covid-19 emergency planning. *Sci. reports* **11**, 1–13, DOI: https://doi.org/10.1038/s41598-021-86811-0 (2021).

18. Wang, L. *et al.* Spatiotemporal dynamics, nowcasting and forecasting of covid-19 in the united states, DOI: https://doi.org/10.48550/arXiv.2004.14103 (2020).

19. Pagano, B. Covid-19 modeling - bob pagano (2020). https://bobpagano.com/covid-19-modeling/.

20. Zou, D. *et al.* Epidemic model guided machine learning for covid-19 forecasts in the united states. *medRxiv* DOI: https://doi.org/10.1101/2020.05.24.20111989 (2020).

21. Chhatwal, J. *et al.* Covid19sim-simulator (2020). https://covid19sim.org/.

22. Li, M. L. *et al.* Overview of delphi model v3 - covidanalytics (2020). https://www.covidanalytics.io/DELPHI_documentation_pdf.

23. Wang, Q., Xie, S., Wang, Y. & Zeng, D. Survival-convolution models for predicting covid-19 cases and assessing effects of mitigation strategies. *Front. Public Heal.* **8**, DOI: https://doi.org/10.3389/fpubh.2020.00325 (2020).

24. Pei, S. & Shaman, J. Initial simulation of sars-cov2 spread and intervention effects in the continental us. *medRxiv* DOI: https://doi.org/10.1101/2020.03.21.20040303 (2020).

25. Mayo, M. L. *et al.* Us army engineer research and development center - usace-erdc_seir (2020). https://github.com/erdc-cv19/seir-model.

26. Gao, Z. *et al.* Microsoft - microsoft-deepstia (2020). https://www.microsoft.com/en-us/ai/ai-for-health.

27. Hong, Q.-J. Qjhong - qjhong-encounter (2020). https://github.com/qjhong/covid19.

28. Jo, A. & Cho, J. Onequietnight - onequietnight-ml (2020). https://github.com/One-Quiet-Night/COVID-19-forecast.

29. Marshall, M., Gardner, L., Drew, C., Burman, E. & Nixon, K. Johns hopkins center for systems science and engineering - jhu_csse-decom (2020). https://systems.jhu.edu/research/public-health/predicting-covid-19-risk/.

30. Zhang-James, Y. *et al.* Suny upstate and su covid-19 prediction team - upstatesu-gru (2021). https://zoltardata.com/model/ylzhang29.github.io/UpstateSU-GRU-Covid.

medRxiv preprint doi: https://doi.org/10.1101/2022.04.20.22274097; this version posted April 21, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

31. Wattanachit, N. & Evan L. Ray, N. R. Covid-19 forecast hub - covidhub-ensemble (2020). https://covid19forecasthub.org/.

32. Wang, D., Summer, T., Zhang, S. & Wang, L. University of central florida - ucf-aem (2020). https://github.com/UCF-AEM/UCF-AEM.

33. Wolfinger, R. & Lander, D. Locknquay - lnq-ens1 (2020). https://www.kaggle.com/c/covid19-global-forecasting-week-5/overview.

34. Ray, E. L. & Tibshirani, R. Covid-19 forecast hub - covidhub-baseline (2020). https://covid19forecasthub.org/.

35. Ray, E. L., Cramer, E., Gerding, A. & Reich, N. Covid-19 forecast hub-covidhub-trained_ensemble (2021). https://covid19forecasthub.org/.

36. Adiga, A. *et al.* University of virginia, biocomplexity covid-19 response team - uva-ensemble (2020). https://biocomplexity.virginia.edu/.

37. Perakis, G. *et al.* Mit-cassandra (2021). https://github.com/oskali/mit_cassandra.

38. Yogurtcu, O. N. *et al.* A quantitative evaluation of covid-19 epidemiological models. *medRxiv* DOI: https://doi.org/10.1101/2021.02.06.21251276 (2021).

39. Kinsey, M. *et al.* Johns hopkins university applied physics lab - jhuapl-bucky (2020). https://docs.buckymodel.com/.

40. Wilson, D. J. Weather, mobility, and covid-19: A panel local projections estimator for understanding and forecasting infectious disease spread, DOI: https://doi.org/10.24148/wp2020-23 (2020).

41. Suchoski, B., Stage, S., Gurung, H. & Baccam, S. Iem med - iem_med-covidproject (2020). https://iem-modeling.com/.

42. Vespignani, A. *et al.* Mobs lab at northeastern - mobs-gleam_covid (2020). https://uploads-ssl.webflow.com/58e6558acc00ee8e4536c1f5/5e8bab44f5baae4c1c2a75d2_GLEAM_web.pdf.

43. Rahi Kalantari, M. Z. Discrete dynamical systems - dds-nbds (2020). https://dds-covid19.github.io/.

44. Jain, C. L. Institute of business forecasting - ibf-timeseries (2020). https://ibf.org/.

45. Walraven, R. Robert walraven - robertwalraven-esg (2020). http://rwalraven.com/COVID19.

46. Corsetti, S. *et al.* The university of michigan - umich-ridgetfreg (2020). https://gitlab.com/sabcorse/covid-19-collaboration.

47. Karlen, D. Characterizing the spread of covid-19 (2020). https://arxiv.org/abs/2007.07156.

48. Osthus, D. *et al.* Los alamos national labs - lanl-growthrate (2020). https://covid-19.bsvgateway.org/.

49. Burant, J. John burant (jcb) - jcb-prm (2020). https://github.com/JohnBurant/COVID19-PRM.

50. Nagraj, V., Hulme-Lowe, C., Guertin, S. L. & Turner, S. D. Focus: Forecasting covid-19 in the united states. *medRxiv* DOI: https://doi.org/10.1101/2021.05.18.21257386 (2021).

51. O'Dea, E. University of georgia center for the ecology of infectious diseases forecasting working group - ceid-walk (2020). https://github.com/e3bo/random-walks.

52. Sarker, A., Jadbabaie, A. & Shah, D. Idss covid-19 collaboration (isolat) at mit - mit_isolat-mixtures (2021). https://idss.mit.edu/vignette/real-time-mixture-based-predictions/.

Author contributions statement

A.C. worked on literature survey, wrote first draft and managed the project. G.J. worked on the code for pulling CDC-model predictions data and calculating MAPE errors. R.J. and M.L. prepared the manuscript figures. J.B. improved the discussion section, provided feedback and contributed to the development of the manuscript. C.G. thought of the idea of the project, supervised the project, helped to draft, review and edit the manuscript. All authors have read the final manuscript.

medRxiv preprint doi: https://doi.org/10.1101/2022.04.20.22274097; this version posted April 21, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

Data and Code Availability

The data used in this study is publicly available from the COVID-19 Data Repository, maintained by the Center for Systems Science and Engineering (CSSE) at Johns Hopkins University (https://github.com/CSSEGISandData/COVID-19/tree/master/csse_covid_19_data/c The data for the predicted case counts, of all the models, is obtained from the data repository for the COVID-19 Forecast Hub (https://github.com/reichlab/covid19-forecast-hub), the data source for the official US-CDC COVID-19 forecasting page. The source code that supports the findings of this research is available from the corresponding author upon request.

Conflicts of Interests

The authors declare no competing interests.

Supplementary Materials

The supplementary section attached with the manuscript contains additional figures and tables that support the main content.

medRxiv preprint doi: https://doi.org/10.1101/2022.04.20.22274097; this version posted April 21, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

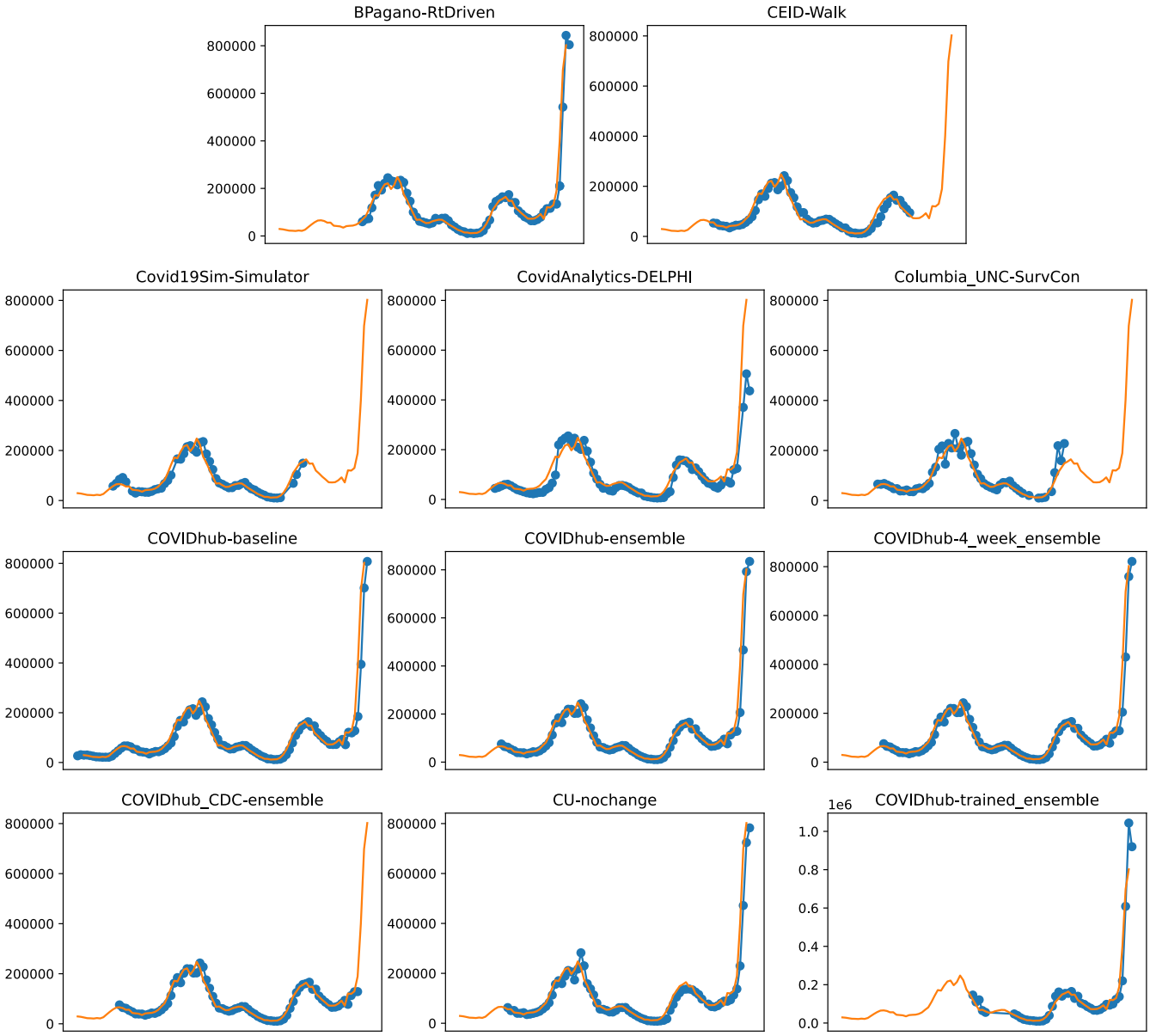

Figure 8. Plots depicting the peak predictions of US-CDC models- 'BPagano-RtDriven', 'CEID-Walk', 'Covid19Sim-Simulator', 'CovidAnalytics-DELPHI', 'Columbia_UNC-SurvCon', 'COVIDhub-baseline', 'COVIDhub-ensemble', 'COVIDhub-4_week_ensemble', 'COVIDhub_CDC-ensemble', 'CU-nochange', 'COVIDhub-trained_ensemble', over complete timeline

medRxiv preprint doi: https://doi.org/10.1101/2022.04.20.22274097; this version posted April 21, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

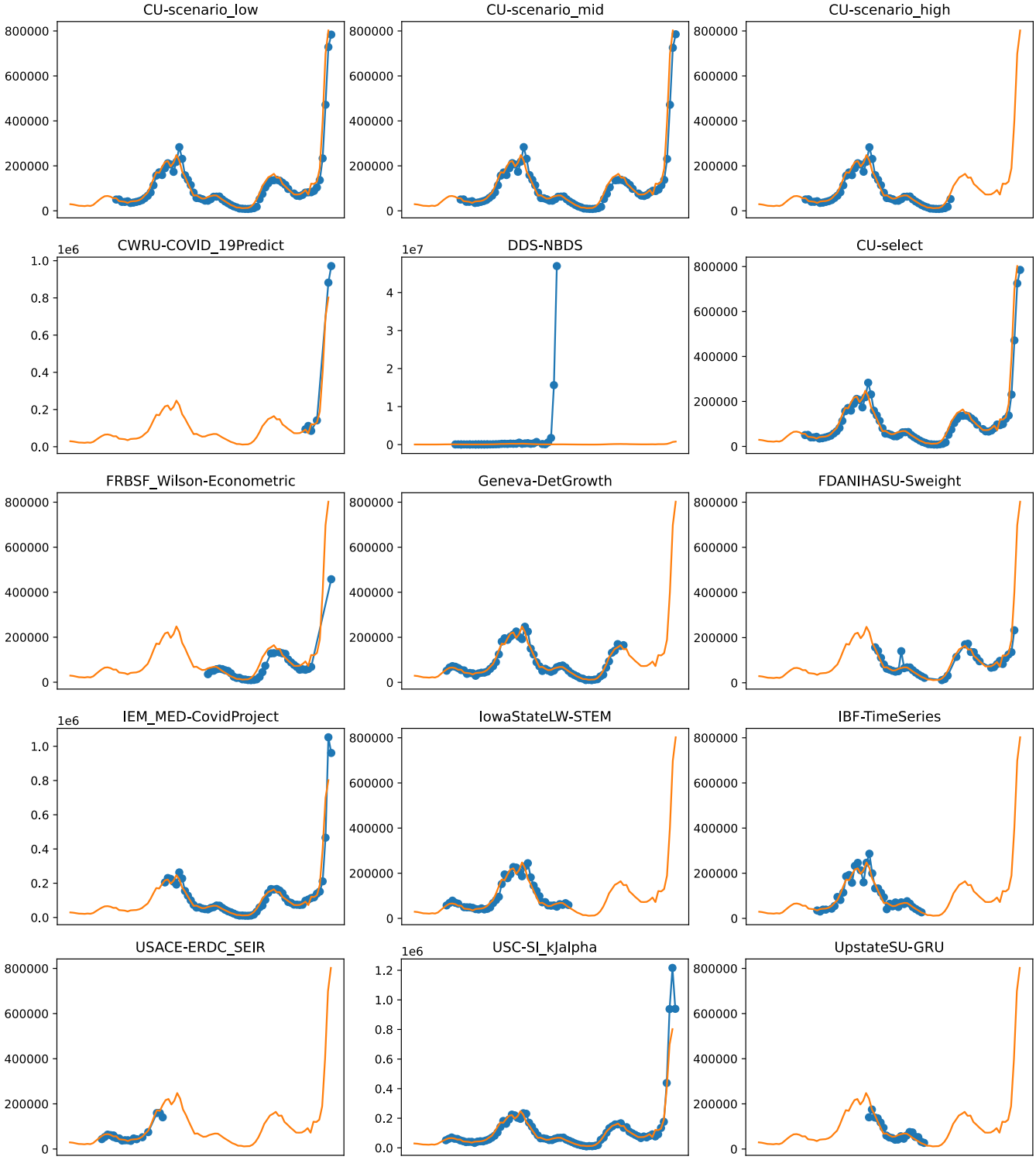

Figure 9. Plots depicting the peak predictions of various US-CDC models - 'CU-scenerio_low', 'CU-scenario_mid', 'CU-scenerio_high', 'CWRU-COVID_19Predict', 'DDS-NBDS', 'CU-select', 'FRBSF_Wilson-Econometric', 'Geneva-DetGrowth', 'FDANIHASU-Sweight', 'IEM_MED-CovidProject', 'IowaStateLW-STEM', 'IBF-TimeSeries', 'USACE-ERDC_SEIR', 'USC-SI_KJalpha', 'UpstateSU-GRU', over complete timeline

medRxiv preprint doi: https://doi.org/10.1101/2022.04.20.22274097; this version posted April 21, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

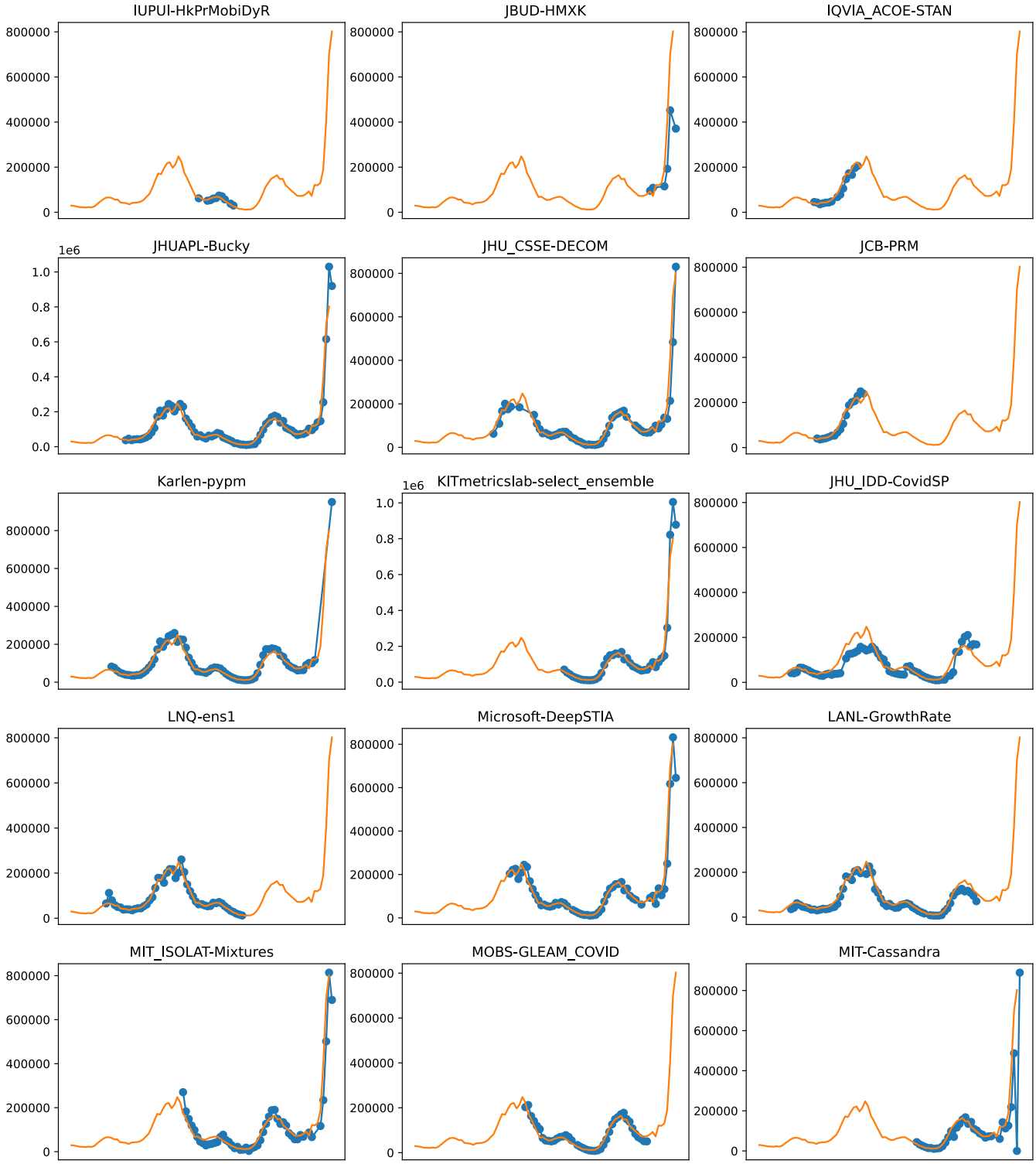

Figure 10. Plots depicting the peak predictions of various US-CDC models - 'IUPUI-HkPrMobiDyR', 'JBUD-HMXK', 'IQVIA_ACOE-STAN', 'JHUAPL-Bucky', 'JHU_CSSE-DECOM', 'JCB-PRM', 'Karlen-pypm', 'KITmetricslab-select_ensemble', 'JHU_IDD-CovidSP', 'LNQ-ens1', 'Microsoft-DeepSTIA', 'LANL-GrowthRate','MIT_ISOLAT-Mixtures', 'MOBS-GLEAM_COVID', 'MIT-Cassandra', over complete timeline

medRxiv preprint doi: https://doi.org/10.1101/2022.04.20.22274097; this version posted April 21, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

Figure 11. Plots depicting the peak predictions of various US-CDC models - 'MUNI-ARIMA', 'MUNI-VAR', 'MSRA-DeepST', 'OneQuietNight-ML', 'prolix-euclidean', 'OliverWyman-Navigator', 'RobertWalraven-ESG', 'SDSC_ISG-TrendModel', 'QJHong-Encounter', 'TTU-squider', 'UCF-AEM', 'SigSci-TS', 'UCLA-SuEIR', 'UMich-RidgeTfReg' and 'UChicagoCHATTOPADHYAY-UnIT' over complete timeline

medRxiv preprint doi: https://doi.org/10.1101/2022.04.20.22274097; this version posted April 21, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

Table S1. Various US-CDC COVID-19 case prediction models- their description, features employed for case prediction, the method used, assumptions made related to public health interventions.

No.	Model	Features employed for Case Predictions	Proposed by	Method Used	Assumptions
Epidemiological/ Compartmental Models					
01.	TTU-Squider (16)	Takes into account power-law incident rate, separate compartments for silent spreaders, quarantine/hospital isolation of confirmed infected individuals, restrictions on social contact, possible loss of immunity for recovered individuals.	Hussain Lab, Texas Tech University	SIR	Assumes that effects of interventions are reflected in observed data and will continue going forward.
02.	JHU-IDD (17)	Accounts for uncertainty in epidemiological parameters including R0, spread of more transmissible variants, infectious period, time delays to health outcomes and effectiveness of state-wide intervention policies.	John Hopkins ID Dynamics Working Group	Metapopulation SEIR	Assumes that current interventions will not change during the period forecasted.
03.	IowaStateLW-STEM (18)	A non-parametric space-time disease transmission model for epidemic data to study the spatial-temporal pattern of COVID-19.	Iowa State - Lily Wang's Research Group	Non-parametric spatiotemporal model	No specific assumptions.
04.	BPagano-RtDriven (19)	The effective transmission ratio, Rt, drives the model's projections. To forecast how Rt will change with time, the model analyzes Rt change data through the pandemic and applies a model of that characteristic behavior to forecast infections.	BPagano	SIR	Assumes that effects of interventions are reflected in observed data and will continue going forward.
05.	UCLA-SuEIR (20)	An updated variant of the SEIR Model that takes into consideration the effects of reopenings. It assumes a transition from a virtual 'Quarantined' group to the 'Susceptible' group at a specific rate for the states that have reopened/ partially reopened. It's most notable feature is that it can infer the untested cases as well as unreported cases.	UCLA Statistical Machine Learning Lab.	Modified SEIR	Assumes contact rates will increase as states reopen and calculates the increase in contact rates for each state.
06.	COVID19Sim-Simulator (21)	Uses a validated compartment model defined using SEIR with continuous-time progression to simulate the trajectory of COVID-19 at the state level.	COVID-19 Simulator	SEIR	Based on assumptions about how in the future, the levels of social distancing may evolve.
07.	CovidAnalytics-DELPHI (22)	Introduces new states to accommodate for cases that remained unnoticed, as well as an explicit death state. A nonlinear curve that reflects the government reaction is used to adjust the infection rate. Also, a meta-analysis of 150 factors is used to determine key illness parameters, while epidemiological parameters are fitted to historical death counts and identified cases.	CovidAnalytics at MIT	Augmentation of the SEIR model	–
08.	Columbia_UNC-SurvCon (23)	Takes into consideration transmission throughout the pre-symptomatic incubation phase, employing a time-varying effective R0 to capture the temporal trend of transmission and change in response to a public health intervention, and uses permutation to quantify uncertainty.	Columbia_UNC	–	–
09.	CU-select (24)	Produces several different intervention scenarios, each assuming various interventions and rates of compliance are implemented in the future. This submission selects the weekly scenario believed to be most plausible given current observations and planned intervention policies.	Columbia University	Metapopulation county-level SEIR	–
10.	CU-nochange (24)	Assumes that current contact rates will remain unchanged in the future.	Columbia University	Metapopulation county-level SEIR	–

medRxiv preprint doi: https://doi.org/10.1101/2022.04.20.22274097; this version posted April 21, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

continued table ...

11.	CU-scenario_low (24)	This projection assumes relatively low transmission.	Columbia University	Metapopulation county-level SEIR	–
12.	CU-scenerio_mid (24)	This projection assumes relatively moderate transmission.	Columbia University	Metapopulation county-level SEIR	–
13.	CU-scenerio_high (24)	This projection assumes relatively high transmission.	Columbia University	Metapopulation county-level SEIR	–
14.	USACE-ERDC_SEIR (25)	Bayesian Inference is used to calculate model parameters from observations of the total number of cases. Information from subject matter experts is used to develop a prior probability distribution over the model parameters. The accumulated observations and subject matter knowledge are then coupled with a statistical model of the model-data mismatch to generate a posterior probability distribution across the model parameters. To make forecasts, the parameters that maximize the posterior probability density are utilized.	US Army Engineer Research and Development Center	Process-based mathematical model similar to classic SEIR with additional compartments for unreported infections/ isolated individuals.	Assumes that current interventions will not change during the forecast period. Further assumptions similar to compartmental models (i) modeled populations are large enough that fluctuations in the disease states grow slower than average (ii) recovered individuals are neither infectious nor become susceptible to further infection.
15.	Microsoft-DeepSTIA (26)	Deep Spatio-temporal network with intervention under the assumption of Spatio-temporal process in the pandemic of different regions.	Microsoft	SEIR model on spatiotemporal network	Assumes that current interventions will not change during the period forecasted.
Machine Learning models					
16.	QJHong-Encounter (27)	Uses (1) Reproductive Number (R) and Encounter Density (D) relation in the past as a training set, (2) future D as input, and (3) ML/regression, the model predicts future R, and ultimately future Daily New Cases.	QJHong	ML	Assumes that current interventions will not change during the forecasted period.
17.	OneQuietNight-ML (28)	Uses high-level features of daily case reports and movement trends data to make predictions about future Covid-19 cases.	OneQuietNight	ML	Assumes that current interventions will not change during the forecasted period.
18.	JHU_CSSE-DECOM (29)	County-level, empirical ML model driven by epidemiological, mobility, demographic, and behavioral data.	JHU Center for Systems Science and Engineering	ML	Assumes that current interventions will not change during period forecasted.
19.	UpstateSU-GRU (30)	A Sequence-to-sequence learning framework which is a feed-forward recurrent neural network. The Seq2Seq algorithm trains a model to convert sequences from the input to sequences in the output. The model takes 'm' days data and demographic and health risk indices as input for case predictions. Inputs daily smoothed incident cases and deaths count google mobility index and daily reproduction number. It also aggregates the county demographic and health risk indices to model the baseline risk score.	SUNY Upstate and SU COVID-19 Prediction Team	County-level forecast using RNN seq2seq model with gated recurrent units.	Assumes that current interventions will not change during period forecasted.

medRxiv preprint doi: https://doi.org/10.1101/2022.04.20.22274097; this version posted April 21, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

continued table ...

Ensemble models					
20.	USC-SI_kJalpha (12)	Considers the influence of parameters learnt via rapid linear regressions, with a focus on reducing hardware requirements and making quicker predictions without compromising performance. A logistic regression model is used to fit the number of samples for each variation on a given day. The model takes into account changing patterns by focusing more on data that has recently been examined. It makes use of a Random Forest to reflect empirical errors in quantile projections, accounting for future trend changes.	University of Southern California	SIR	Assumes that current interventions will not change during period forecasted.
21.	COVIDhub-ensemble (31)	An ensemble, or model average, of submitted forecasts to the COVID-19 Forecast Hub.	COVID-19 Forecast Hub	–	–
22.	UCF-AEM (32)	Combines a traditional SEIR model with mixture modeling and uses ensemble neural networks to extract information from a complicated mixture modeling system.	University of Central Florida	SEIR model informed with ensemble neural networks.	Assumes that current interventions will not change during the forecasted period.
23.	LNQ-ens1 (33)	Uses an ensemble of three models; two fit with LightGBM, and the third being a neural net. Ensemble weights are chosen each week manually based on performance in the previous week.	LockNQuay	Ensemble of three different models.	Assumes that intervention effects are reflected in observable data and will continue in the future.
24.	COVIDhub-baseline (34)	Baseline model for predictions. The most recent observed incidence is the median projection for all future horizons. From one week to the next, the slope of the predicted medians for cumulative values will be constant and equal to the previously observed slope. The model looks at how much incidence has varied from week to week in the past to generate a distribution around the median, and it allows for the possibility that similar fluctuations will occur again in the future.	COVID-19 Forecast Hub	–	–
25.	COVIDhub-trained_ensemble (35)	A weighted ensemble combination of all component model forecasts.	COVID-19 Forecast Hub	Ensemble	–
26.	UVA-Ensemble (36)	Combines models using Bayesian model averaging. The auto-regressive method with features including mobility, other county case counts time-series, an LSTM model with mobility data as an additional predictor, and PatchSim, an SEIR variant with the interaction between counties modeled using commuter data and calibrated on new confirmed cases.	University of Virginia, Biocomplexity COVID-19 Response Team	Ensemble of three different models.	Impact of interventions is represented in observed data in two of three models, while the third assumes that interventions will change in the future.
27.	Caltech CS156	Based on the Ensemble of 14 different ML models including- a) Feedforward Neural Network, b) Quantile Neural Network, c) LSTM, d) Conditional LSTM, e) Encoder-Decoder Conditional LSTM, f) Autoregressive, g) Sessional Autoregressive, h) Decision Tree, i) Gradient- Boosted Decision Tree, j) K-NN, k) Gaussian Process, l) Bayesian epidemiological, m) Two-group epidemiological, n) Curve-fitting Model.	California Institute of Technology	Ensemble of fourteen different models	–
28.	MIT-Cassandra (37)	Based on the ensemble of predictions from four models, including 1) MDP feature representation, 2) KNN time-series, 3) Bi-LSTM time-series, 4) C-SEIRD epidemiological.	MIT Cassandra Ensemble of four different models	Assumes that current interventions will remain in place indefinitely.	Others/ Not Defined
29.	FDANIHASU-Sweight (38)	An ensemble of submitted forecasts to the COVID-19 Forecast Hub. The ensembles are formed by weighting the individual model forecasts with their past performances	FDANIHASU	Ensemble	–

medRxiv preprint doi: https://doi.org/10.1101/2022.04.20.22274097; this version posted April 21, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

continued table ...

Hybrid models				
30. JHUAPL-Bucky (39)	Uses public mobility data to build a Spatial compartment model.	JHUApplied Physics Lab	Spatial compartment model	–
31. FRBSF_Wilson-Econometric (40)	An econometric model that connects the current transmission rate with the fraction of the population that is vulnerable to the shift in new infections from now until a future horizon. The current transmission rate is assumed to be caused by people's mobility and the weather. Mobility, weather, and acquired natural immunity are significant components of the model; there are two additional important parts of the econometric model. The first component includes infection growth lag, which implies that infection growth lag predicts future infection growth. The second component is country-specific intercepts (fixed effects), which allow each county to have a distinct mean level of infection increase regardless of the other model features.	Federal Reserve Bank of San Francisco/Wilson	SIR-derived econometric county panel data model with transmission rate assumed to be a function of weather and mobility	Assumes that intervention effects are reflected in observable data and will continue in the future.
32. IEM_MED-CovidProject (41)	Uses an AI model to fit data from various sources and project new cases of COVID-19. Assumes that the R-value (average number of secondary infections) changes quite rapidly over time due to changes in human behavior and uses a sliding window that fits the data and finds the best R values for each window.	IEM MED	SEIR model with ML	Assumes that current interventions will not change during the forecasted period.
33. MOBS-GLEAM_COVID (42)	A metapopulation method is used. The world is divided into geographical subpopulations, and human mobility between subpopulations is depicted on a network. This data layer on mobility identifies the number of persons travelling from between sub-populations. The mobility network is made up of many mobility processes, ranging from short-distance commuting to intercontinental travel. Superimposed on the globe population and mobility layers is an agent-based epidemic model that describes the infection and population dynamics.	MOBS Lab at Northeastern	Metapopulation age-structured SLIR	Assumes that social distancing policies in place at the date of calibration are extended for the future weeks.
34. DDS-NBDS (43)	Jointly modeling daily deaths and cases using a negative binomial distribution based non-parametric Bayesian generalized linear dynamical system (NBDS).	Team DDS	Bayesian hierarchical model	Assumes that intervention effects are reflected in observable data and will continue in the future.
Other models				
35. IBF-TimeSeries (44)	Combines mechanistic disease transmission model with a curve-fitting approach.	Institute of Business Forecasting	Modified Time Series Model	Do not make any specific assumptions.
36. RobertWalraven (45)	Uses a skewed Gaussian distribution with four empirical parameters: height, position, left growth rate, and right decay rate. The model makes no epidemiological assumptions and has no epidemiological parameters.	Robert Walraven	Skewed Gaussian distribution	Assumes that current interventions will not change during the forecasted period.
37. UMich-RidgeTfReg (46)	This model is based on ridge regression (penalized Ordinary Least Squares regression) to make predictions without relying on external assumptions. The model uses Finite Impulse Response filtering to forecast confirmed cases each day as a function of prior day numbers.	The University of Michigan	Ridge regression	Assumes that current interventions will not change during the forecasted period.
38. Karlen-pypm (47)	Uses Discrete-time difference equations with long periods of the constant transmission rate.	Karlen Working Group	Discrete-time difference equations	Assumes that intervention effects are reflected in observable data and will continue in the future.

medRxiv preprint doi: https://doi.org/10.1101/2022.04.20.22274097; this version posted April 21, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

continued table ...

39.	LANL-GrowthRate (48)	Two processes are represented by the model. The first step is to create a statistical model that depicts how the number of COVID-19 infections varies over time. The second model correlates the number of infections with the reported data. The underlying numbers of susceptible and infected cases in the population at the preceding time step, scaled by the size of the state's initial susceptible population, are used to map the rise of new instances. These two components' weights are dynamically adjusted to the observed data.	Los Alamos National Labs	Statistical dynamical growth model (accounts for population susceptibility)	Assumes that interventions implemented on the first day of the forecast will continue over the following four weeks.
40.	JCB-PRM (49)	Built on observations of macro-level societal and political responses to COVID19 characterized only in terms of infections and deaths. Assumes that although individuals and policymakers have responded to the epidemic with a wide variety of behavioral modifications and policy actions, the actual net impact of the measures taken, though not identical across time or geography, is predictable. The model identifies 'acceptability' ranges from observation of the epidemic up to the current time.	John Burant (JCB)	Phenomenological statistical model	The incidence of COVID-19 in the population determines the strength and impact of control measures in the future.
41.	SigSci-TS (50)	Time series forecasting using ARIMA for case forecasts and lagged cases for death forecasts.	Signature Science FOCUS	Autoregressive time-series model	Assumes that current interventions will not change during the forecasted period.
42.	CEID-Walk (51)	The model is based on a random walk with no drift. The variance in step size of random walk is estimated using the last few observations of a target time series.	University of Georgia Center for the Ecology of Infectious Diseases Forecasting Working Group	Statistical random walk model	Assumes that social distancing policies in place at the date of calibration are extended for the future weeks.
43.	MIT_ISOLAT-Mixtures (52)	A non-mechanistic, non-parametric forecasting model that forecasts time series as a sum of bell curves. The confidence intervals are calculated by applying a multiplicative log-Gaussian perturbation to the observed time series.	IDSS COVID-19 Collaboration (Isolat) at MIT Mixture model	–	Assumes that current interventions will remain in place indefinitely.