Predictive performance of international COVID-19 mortality forecasting models

Joseph Friedman*, Patrick Liu*, Emmanuela Gakidou[†], and the IHME COVID19 Model Comparison Team

Abstract

Forecasting models have provided timely and critical information about the course of the COVID-19 pandemic, predicting both the timing of peak mortality, and the total magnitude of mortality, which can guide health system response and resource allocation[1–4]. Out-of-sample predictive validation—checking how well past versions of forecasting models predict subsequently observed trends—provides insight into future model performance[5]. As data and models are updated regularly, a publicly available, transparent, and reproducible framework is needed to evaluate them in an ongoing manner. We reviewed 384 published and unpublished COVID-19 forecasting models, and evaluated seven models for which publicly available, multi-country, and date-versioned mortality estimates could be downloaded[6–10]. These included those modeled by: DELPHI-MIT (Delphi), Youyang Gu (YYG), the Los Alamos National Laboratory (LANL), Imperial College London (Imperial), and three models produced by the Institute for Health Metrics and Evaluation (IHME), a curve fit model (IHME-CF), a hybrid curve fit and epidemiological compartment model (IHME-CF SEIR), and a hybrid mortality spline and epidemiological compartment model (IHME – MS SEIR). Collectively models covered 169 countries, as well as the 50 states of the United States, and Washington, D.C., and accounted for >99% of all reported COVID-19 deaths on July 20th, 2020. As expected, errors in mortality predictions increased with a larger number of weeks of extrapolation. For the most recent models, released in June, at six weeks of forecasting the best performing model was the IHME-MS SEIR model, with a cumulative MAPE of 10.2%, followed by YYG (11.3%) and LANL (12.6%). Looking across models, errors in cumulative mortality predictions were highest in sub-Saharan Africa and lowest in high-income countries, reflecting differences in data availability and prediction difficulty in earlier *vs*. later stages of the epidemic. For peak timing prediction, among models released in April, median absolute error values at six weeks ranged from 20 days for the Imperial model to 35 days for the YYG model. In sum, we provide a publicly available dataset and evaluation framework for assessing the predictive validity of COVID-19 mortality forecasts. We find substantial variation in predictive performance between models, and note large differences in average predictive validity between regions, highlighting priority areas for further study in sub-Saharan Africa and other emerging-epidemic contexts.

[†]Correspondence to: Emmanuela Gakidou (gakidou@uw.edu).

*These authors contributed equally to the analysis, and are listed in alphabetical order.

Main

As the COVID-19 pandemic has spread across the globe, policymakers and hospital systems have relied on forecasting models predicting mortality to drive real-time policy change and resource allocation. These models have proven to be controversial—iteratively championed for providing essential information, and criticized for providing incorrect predictions in the midst of an evolving, high-stakes public health situation[11–15]. Nevertheless, many assessments of COVID-19 forecasting model performance have relied on a small selection of examples—such as a single iteration of a forecasting model that is updated daily, or estimates for a single country from a global model—to assess predictive validity. A comprehensive approach, including all estimated timeseries and all iterations of each model, is needed to provide additional perspective on overall model performance.

Although more systematic comparisons have been conducted for models describing the epidemic in the United States[7,16–18], similar analyses have not been undertaken elsewhere, despite the global impact of COVID-19. A critical review of each model and its out-of-sample predictive performance is therefore urgently needed, to provide insight into which models may be most helpful in planning for upcoming phases of the pandemic. This may be of particular importance as forecasting models increasingly seek to leverage observed trends in places with established epidemics, such as the United States or Europe, to predict future trends in places with emerging epidemics.

A global framework for COVID-19 mortality forecast comparisons

In order to conduct a systematic comparison of the out-of-sample predictive validity of international COVID-19 forecasting models, a number of issues must be addressed. Looking across models, a high degree of heterogeneity can be observed in numerous dimensions, including sources of input data, frequency of public releases of model estimates, geographies included in the results, and how far into the future predictions are made available for. Differences in each of these areas must be taken into account to provide a fair and relevant comparison.

<u>Input data</u>: A number of sources of input data—describing observed epidemiological trends in COVID-19—exist, and they often do not agree for a given country and time point[19–21]. We chose to use mortality data collected by the Johns Hopkins University Coronavirus Resource Center as the in-sample data against which forecasts were validated at the national level, and data from the New York Times for state-level data for the United States[20]. However, we adjusted for differences in model input data using intercept shifts, whereby all models where shifted to perfectly match the in-sample data for the date in which the model was released (see methods). This avoids unfairly penalizing models using a different input data set. It is important to note, however, that some known issues do exist with each of these data sources, which have led some groups evaluated here to use a mix of these data and country-specific data that are made available through ministries of health and national and local statistical offices.

<u>Frequency of public releases of model estimates</u>: Most forecasting models are updated regularly, but at different intervals, and on different days. Specific days of the week have been associated with a greater number of reported daily deaths. Therefore, previous model comparison efforts in the United States—such as those conducted by the US Centers for Disease Control and Prevention—have required modelers to produce estimates using input data cutoffs from a specific day of the week[22]. For the sake of including all publicly available modeled estimates, we took a more inclusive approach, considering each publicly

released iteration of each model. To minimize the effect of day-to-day fluctuations in death reporting, we focus on errors in cumulative and weekly total mortality, which are less sensitive to daily variation.

Geographies and time periods included in the results: Each model produces estimates for a different set of national and subnational locations, and extrapolates a variable amount of time from the present. Each model was also first released on a different date, and therefore reflects a different window of the pandemic. Here, we also took an inclusive approach, and included estimates from all possible locations and time periods. One exception, however, was that data from Ecuador and Peru were excluded from predictive validity analyses, given large discrepancies in trend and level between official COVID-19 statistics and excess mortality analyses from those two locations, suggesting poor data quality[23]. To increase comparability, we stratified summary error statistics by region, weeks of extrapolation, and month of estimation, and masked summaries reflecting a small number of locations or time points. Estimates were included at the national level for all countries, except the United States, where they were included at the admin-1 (state) level, as they were available for most models. In order to be considered for inclusion, models were required to forecast at least four weeks into the future.

Outcomes: Finally, each model also includes different estimated quantities, including daily and cumulative mortality, number of observed or true underlying cases, and various dimensions of hospital resource utilization. We focus here on mortality, as it was the most widely reported outcome, and it also has a high degree of societal, epidemiological and public health importance.

We reviewed 383 published and unpublished COVID-19 forecasting models (see supplemental file). We excluded models from consideration that did not 1) produce estimates for at least five different countries, 2) did not extrapolate at least four weeks out from the time of estimation, 3) did not estimate mortality, 4) did not provide downloadable, publicly available results, or 5) did not provide date-versioned sets of previously estimated forecasts, which are required to calculate subsequent out-of-sample predictive validity. Seven models which fit all inclusion criteria were evaluated (Table 1). These included those modeled by: DELPHI-MIT (Delphi)[10], Youyang Gu (YYG)[7], the Los Alamos National Laboratory (LANL)[9], Imperial College London (Imperial)[8], and three models produced by the Institute for Health Metrics and Evaluation (IHME)[6]. Beginning March 25th, IHME initially produced COVID forecasts using a statistical curve fit model (IHME-CF), which was used through April 29th for publicly released forecasts[1]. On May 4th, IHME switched to using a hybrid model, drawing on a statistical curve fit first stage, followed a second-stage epidemiolocal model with susceptible, exposed, infectious, recovered compartments (SEIR)[24]. This model—referred to herein as the IHME-CF SEIR model—was used through May 26th. On May 29th, the curve fit stage was replaced by a spline fit to the relationship between log cumulative deaths and log cumulative cases, while the second stage SEIR model remained the same[25]. This model, referred to as the IHME-MS SEIR model, was still in use at the time of this publication. The three IHME models rely upon fundamentally different assumptions and core methodologies, and therefore are considered separately. They were also released during different windows of the pandemic, and are therefore compared in subsequent sections to models released during similar time periods.

In some cases, numerous scenarios were produced by modelling groups, to describe the potential effects of interventions, or future trajectories under different assumptions. In each case the baseline or status quo scenario was selected to evaluate model performance as that represents the modelers' best estimate about the most probable course of the pandemic.

Assessing the magnitude of deaths

The total magnitude of COVID-19 deaths is a key measure for monitoring the progression of the pandemic. It represents the most commonly produced outcome of COVID-19 forecasting models, and perhaps the most widely debated measure of performance. The evaluation framework we developed for assessing how well models predicted the total number of cumulative deaths is shown in Figure 1 for an example country—the United States—and similar figures for all locations included in the study can be found in the supplement. Out-of-sample errors in predicted total cumulative number of deaths were calculated for each model globally by number of weeks of model extrapolation, month of estimation and each super-region used in the Global Burden of Disease Study[26] (Figure 2). In the main text we focus on errors in total cumulative deaths—as opposed to other metrics such as weekly or daily deaths—as it has been most commonly discussed measure, to-date, in academic and popular press critiques of COVID-19 forecasting models. Nevertheless, alternate measures are presented in the extended data figures and online framework. Errors were assessed for systematic upward or downward bias (Extended Data Figure 1), and errors for weekly, rather than cumulative deaths, were also assessed (Extended Data Figure 2). In calculating summary statistics, we used percent errors to control for the large differences in the scale of the epidemic between locations. We calculated medians, rather than means, due to a small number of large magnitude outliers present in a few time-series. Errors from all models were pooled to calculate overall summary statistics, in order to comment on overarching trends by geography and time.

As expected, errors in cumulative mortality predictions increased with a larger number of weeks of extrapolation (Figure 2). For example, pooling errors across models released in May, the median absolute percent error (MAPE) rose from 3.8% at one week of extrapolation to 14.8% at 6 weeks, and similar patterns were observed across the study period. Decreases in predictive ability with extrapolation were similarly noted for errors in weekly deaths (Extended Data Figure 2). Model performance in predicting total cumulative deaths also improved over time, which is to be expected, given the reduced difficulty in predicting the overall scope of an epidemic which has stabilized. Pooling across models, the MAPE at four weeks of extrapolation was 10.2% in June, as compared to 13.4% in May, 25.2% in April, and 94.9% in March. No corresponding improvement over time, however, was seen when examining weekly errors, with a MAPE at 4 weeks of 66.7%, 62.3%, and 65.2% for June, May, and April, respectively. We can therefore conclude that although modelling the cumulative death toll of COVID-19 has become progressively easier over the course of the pandemic, models have made less progress in improving predictions of the level of upcoming deaths occurring each week, an objectively more difficult task.

For models released in June, at four weeks of forecasting, the best performing model was the IHME-MS SEIR model, with a cumulative MAPE of 10.2%, followed by YYG (11.3%) and LANL (12.6%). For models released in May, assessed at six weeks of extrapolation, the IHME-CF SEIR model had a MAPE for cumulative deaths of 11.6%, followed by YYG (13.1%), and LANL (14.0%). For models released in April, at six weeks of extrapolation, the best performance was observed for the LANL model with a MAPE of 26.5%, followed by YYG at 26.9%, and the Delphi model at 29.3%.

In April, most forecasting models were largely unbiased in their short-term forecasts, with median percent error (MPE) values at one week of extrapolation ranging from -5.5% for LANL to +10.3% for Imperial, with a pooled MPE of +0.6%. By six weeks of extrapolation, however, models produced in April proved to be overly optimistic, with a pooled MPE of -10.0%. In contrast, by June, models had become

relatively unbiased in longer-term forecasts, with a pooled MPE at 4 weeks of extrapolation of -0.2%. A notable exception was the Imperial model, which had an MPE of +2,195.8% at six weeks for models released in April, and an MPE of +95.6% at four weeks for models released in June (see the Supplement for a location-specific visualization of all time-series and errors).

Looking across models, errors in cumulative mortality predictions were generally highest in sub-Saharan Africa, with a pooled-error MAPE of 41.8% at four weeks for models released in June, and 37.5% at six weeks for models released in May, compared to 4.6% and 11.1% respectively for countries in the high-income region, which generally had the lowest MAPE values. These systematic regional differences likely reflect both differences in data availability and a different level of prediction difficulty in earlier *vs*. later stages of each country's epidemic. Similar disparities were also seen between regions when examining errors in weekly deaths, and so regional gaps are unlikely to reflect the stage of the epidemic alone.

Assessing the timing of peak mortality

We also assessed how well each model predicted the timing of peak daily deaths—an additional aspect of COVID-19 epidemiology with acute relevance for resource planning. Peak timing may be better predicted by different models than those best at forecasting the magnitude of mortality, and therefore deserves separate consideration as an outcome of predictive performance. In order to assess peak timing predictive performance, we first calculated the observed peak of daily deaths in each location—a task complicated by the highly volatile nature of reported daily deaths values. We therefore smoothed each timeseries of daily deaths, and subsequently calculated the date of the peak observed in each location, as well as the predicted peak for each iteration of each forecasting model (see the methods section). We used various "smoothers" to accomplish this task (see Extended Data Figure 5 and methods section), but present results calculated using a LOESS smoother in the main text (Figures 3 and 4), as it was found to be the most robust to daily fluctuations. Results shown here reflect only those locations for which the peak of the epidemic had passed at the time of publication, and for which at least one set of model results was available seven days or more ahead of the peak date. We stratified predictive validity statistics by the number of weeks in advance of the observed peak that the model was released, as well as the month in which the model was released. There was insufficient geographic variation to stratify results by regional groupings, although that remains an important topic for further study, which will become feasible as the pandemic peaks in a greater number of countries globally.

As expected, forecasting models tended to improve at predicting peak timing as they got closer to the observed peak date in each location. For example, among estimates that were released in April, one week (7 to 13 days) ahead of an observed peak, models showed a median absolute error (MAE) of seven days, which increased to 29 days for models released six weeks in advance (Figure 4).

Unlike the predictive performance observed for the magnitude of mortality, models were not observed to improve greatly in their ability to predict peak timing out-of-sample over time. The only model available in March, the IHME-CF model, had an MAE of 17 days for models released in that month, at six weeks in advance of a peak. Pooling across models, estimates had an MAE of 29 days at six weeks in April, and 34 days at six weeks in May. Among models released in April, MAE values at six weeks ranged from 20 days for the Imperial model to 35 days for the YYG model. For models from May, MAE values at six weeks ranged from 18 for the Imperial model, to 37 for the YYG model. For models released in June, MAE values at 4 weeks ranged from 9 for the Delphi model to 52 for the YYG model. For models

released in April at two weeks of forecasting, the Delphi model had the best performance with an MAE of nine days, followed by IHME-CF (12), YYG (13), and LANL (18).

Models generally were biased towards premature peak prediction, reflected in negative median error (ME) values for most models and estimation months, regardless of lead time. For example, pooling across models, a ME of -17 days, -29 days, and -27 days was observed at six weeks for March, April, and May respectively. Peak timing assessment was only available up to five weeks in advance for models released during the month of June, however a distinct trend emerged, where positive median errors were noted. This seems to reflect a number of locations where true peaks where observed in June, although models predicted peaks would not be seen until later in the summer or fall.

Selecting models to monitor emerging challenges in COVID-19

In this analysis we provide a publicly available dataset and evaluation framework for assessing the predictive validity of COVID-19 mortality forecasts. For many countries, especially those well past the peak of their epidemic, forecasts have largely converged. Nevertheless, we find substantial variation in historical predictive performance between models. This variation may be helpful in assessing probable future performance, especially in emerging epidemics in Latin America, sub-Saharan Africa, and North Africa and the Middle East, which have not yet peaked. Although model performance has generally improved over time, considerable challenges remain. We note large differences in average performance between regions, highlighting priority areas for further study in sub-Saharan Africa and other emerging-epidemic contexts.

We also note that the vast majority of models did not provide sufficient information to be included in this framework, given that publicly available and date-version forecasts were not made available. We would encourage all research groups forecasting COVID-19 mortality to consider providing historical versions of their models in a public platform for all locations, to facilitate ongoing model comparisons. This will improve reproducibility, the speed of development of the modelling science and the ability of policy makers to discriminate between the burgeoning number of models[27].

Our analysis does have limitations that should be considered. Death reporting mechanisms have evolved over time, and these shifts represent a source of error that is difficult to control for, and will affect all models. Additionally, each modeling group uses a different set of input data. In this analysis we have selected data from the Johns Hopkins Coronavirus Resource Center as the in-sample standard, because they are publicly available, and released daily, and are widely used by modelers. Nevertheless, they do have known issues, which has led other groups, including IHME, to include additional sources of data for specific locations. Updating this framework to include a wider array of in-sample data is an important area for future study.

When taking an inclusive approach to including forecasts from various modelling groups, including estimates from a wide range of time periods and geographies, extra care must be taken to ensure comparability between models. We use various techniques to construct fair companions, such as stratifying by region, month of estimation, and weeks of forecasting, and masking summary statistics representing a small number of values. Nevertheless, other researchers may prefer distinct methods of maximizing comparability over a complex and patchy estimate space. Furthermore, the domains we assess—magnitude of total mortality and peak timing—are not an exhaustive list of all possible

dimensions of model performance. By providing an open-access framework to compile forecasts and calculate errors, we hope that other researchers can build on our results to provide additional analyses.

Ultimately, policymakers would benefit from considering a multitude of forecasting models as they consider resource planning decisions related to the response to the COVID-19 pandemic. We provide a publicly available framework and codebase, which will be updated in an ongoing fashion, to continue to monitor model predictions in a timely fashion, and contextualize them with prior predictive performance. It is our hope that this spurs conversation and cooperation amongst researchers, which might lead to more accurate predictions, and ultimately aid in the collective response to COVID-19.

Tables and Figures

Model	Data Access	Model Type	Geographies	Range
IHME - CurveFit	http://www.healthdata.org/covid/data-downloads	Statistical Curve Fit	34 Countries*	August 4th
IHME - CF SEIR	http://www.healthdata.org/covid/data-downloads	Curve fit + SEIR	52 Countries*	August 4th
IHME – MS SEIR	http://www.healthdata.org/covid/data-downloads	Mortality Spline + SEIR	156 Countries*	November 1st
Youyang Gu	https://github.com/youyanggu/covid19_projections	SEIR	73 Countries*	November 1st
MIT - DELPHI	https://github.com/COVIDAnalytics/DELPHI	SEIR	154 Countries*	October 15th
Imperial-LMIC	https://github.com/mrc-ide/global-lmic-reports	SEIR	156 Countries	October 15th
LANL-Growthrate	https://covid-19.bsvgateway.org/	Dynamic Growth	131 Countries*	September 1st

Table 1. Models Included in the Study
All seven models included in the study are shown. The full list of models assessed for inclusion is shown in the supplemental review file.
*Includes state-level estimates for the United States

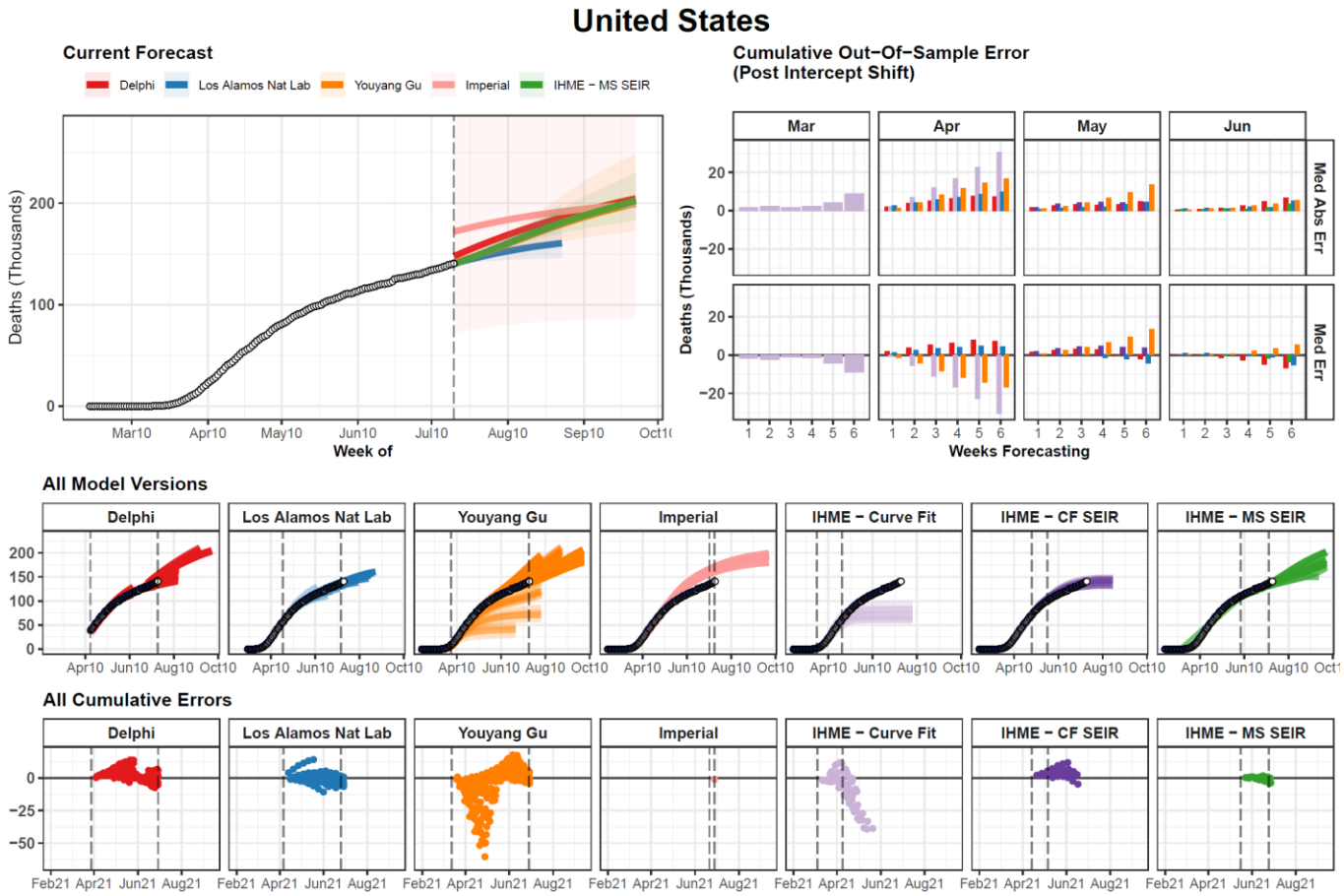


Figure 1. Cumulative Mortality Forecasts and Prediction Errors by Model – Example for United States
The most recent version of each model is shown on the top left. The middle row shows all iterations of each model as separate lines, with the intensity of color indicating model date (darker models are more recent). The vertical dashed lines indicate the first and last model release date for each model. The bottom row shows all errors calculated at weekly intervals from one to six weeks. The top right panel summarizes all observed errors, using median error and median absolute error, by weeks of forecasting, and month of model estimation. Errors incorporate an intercept shift to account for differences in each model's input data. Values are shown for the United States, and similar graphs for all other locations are available in the supplement.

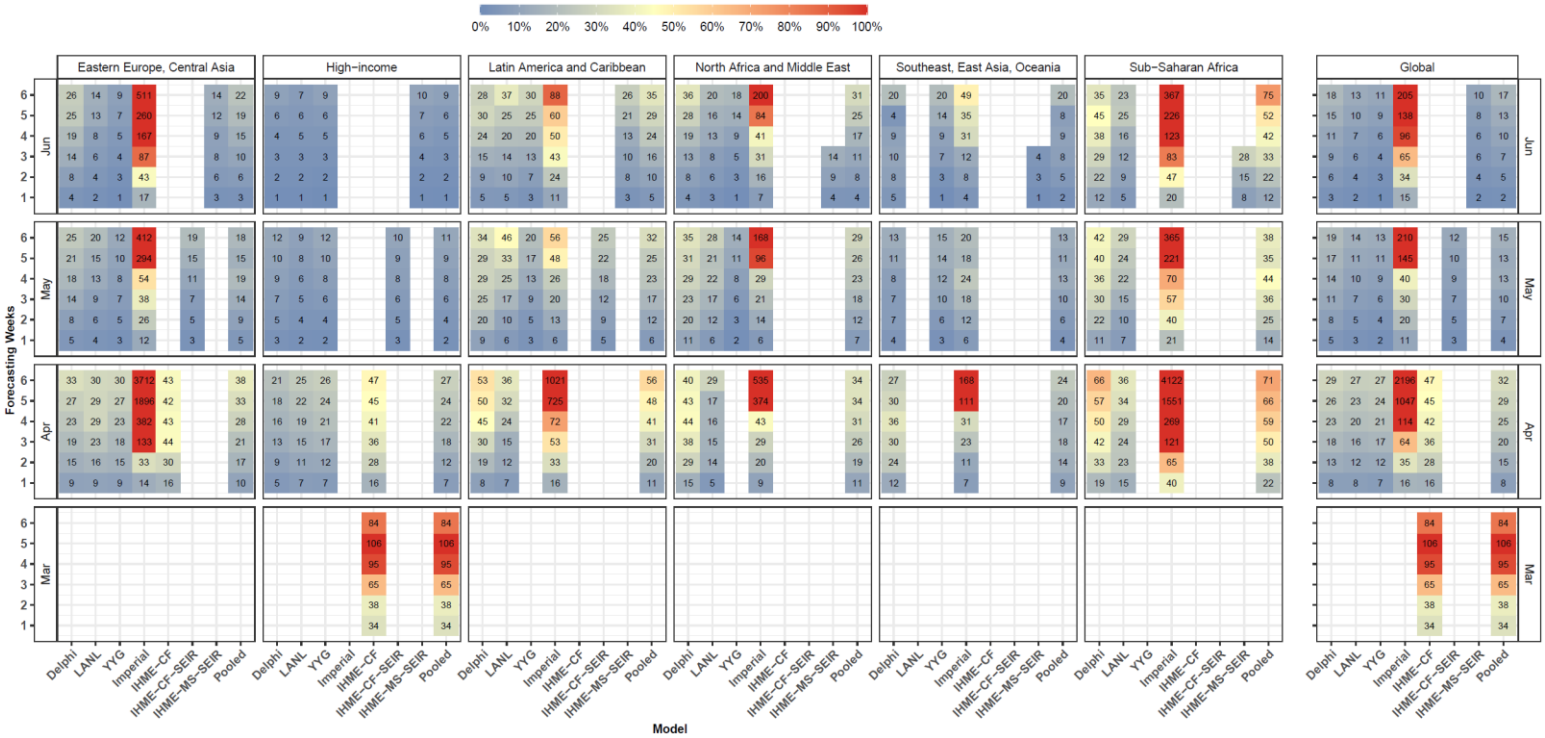


Figure 2. Median Absolute Percent Error for Total Cumulative Errors by Super Region, Month of Estimation, and Weeks of Extrapolation
Median absolute percent error values were calculated across all observed errors at weekly intervals, for each model, by month of estimation, weeks of forecasting, and super regional grouping used in the Global Burden of Disease Study. Values that represent fewer than five locations are masked due to small sample size. Pooled summary statistics reflect values calculated across all errors from all models, in order to comment on aggregate trends by time or geography.

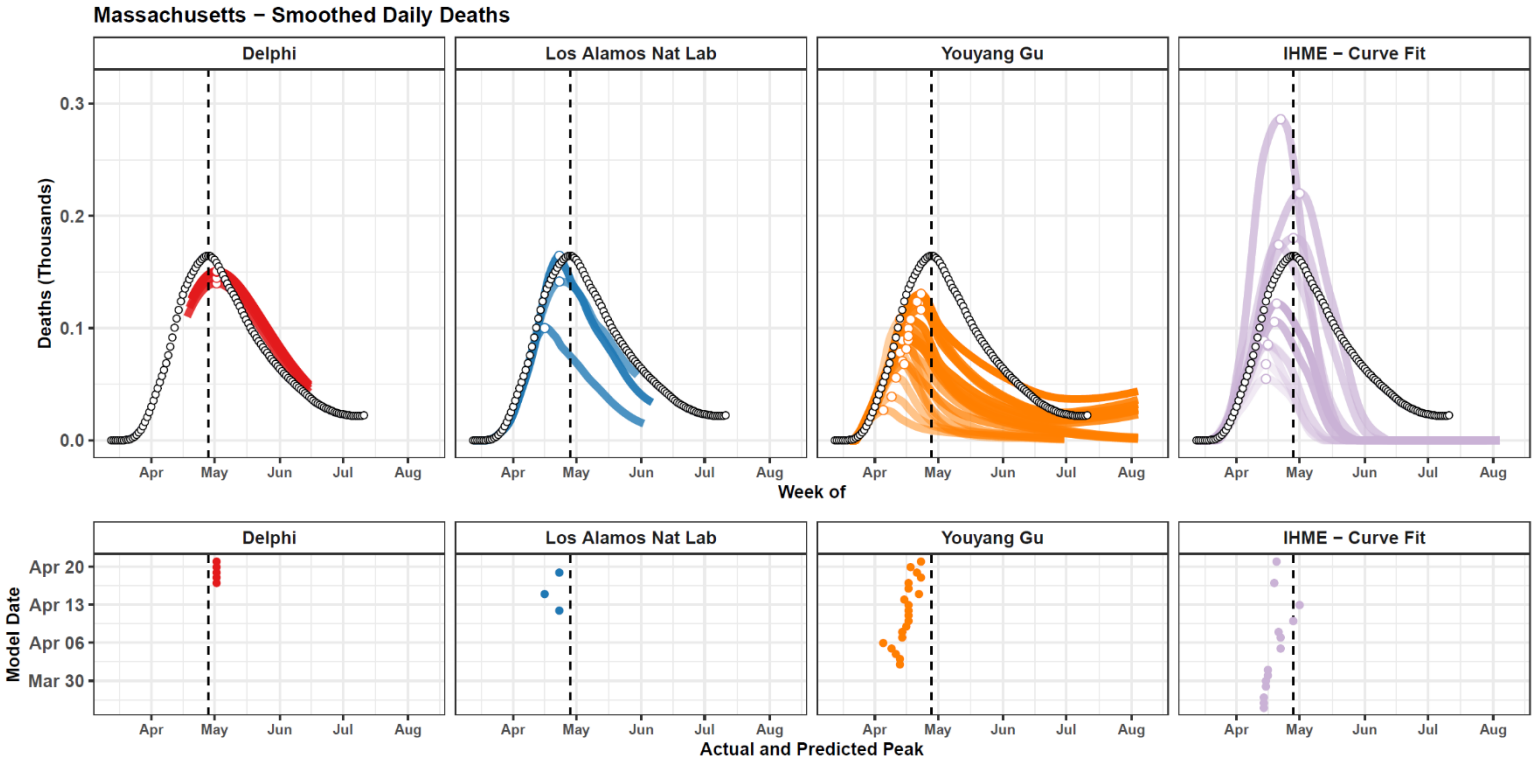


Figure 3. Observed vs Predicted Peak in Daily Deaths– Example for Massachusetts

Observed daily deaths, smoothed using a loess smoother, are shown as black-outlined dots (top). The observed peak in daily deaths is shown with a vertical black line (bottom). Each model version that was released at least one week prior to the observed peak is plotted (top) and its estimated peak is shown with a point (top and bottom). Estimated peaks are shown in the bottom panel with respect to their predicted peak date (x-axis) and model date (y-axis). Values are shown for the Massachusetts, and similar graphs for all other locations are available in the supplement.

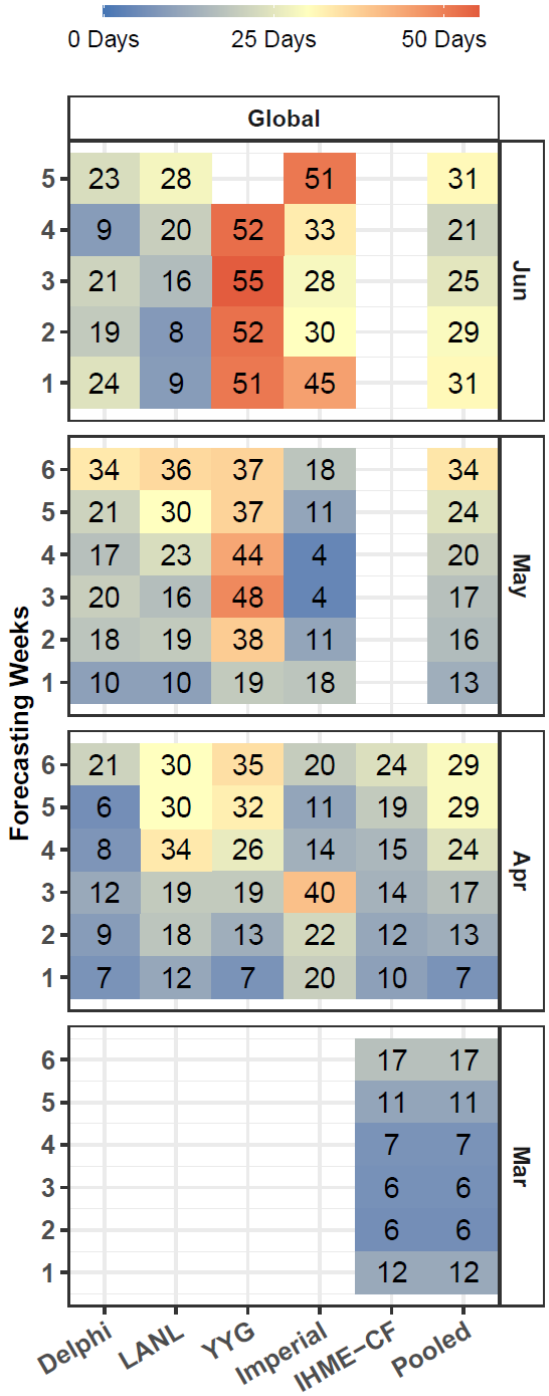


Figure 4. Predictive Validity of Peak Timing

Median absolute error in days is shown by model, number of weeks of forecasting, and month of estimation. Models that are not available for at least 20 peak timing predictions are not shown. Errors only reflect models released at least seven days before the observed peak in daily mortality. One week of forecasting refers to errors occurring from seven to 13 days in advance of the observed peak, while two weeks refers to those occurring from 14 to 20 days prior, and so on, up to six weeks, which refers to 42-48 days prior.

References

1. Team, I. C.-19 health service utilization forecasting & Murray, C. J. Forecasting COVID-19 impact on hospital bed-days, ICU-days, ventilator-days and deaths by US state in the next 4 months. *medRxiv* 2020.03.27.20043752 (2020) doi:10.1101/2020.03.27.20043752.
2. Lu, F. S., Nguyen, A. T., Link, N. B., Lipsitch, M. & Santillana, M. Estimating the Early Outbreak Cumulative Incidence of COVID-19 in the United States: Three Complementary Approaches. *medRxiv* 2020.04.18.20070821 (2020) doi:10.1101/2020.04.18.20070821.
3. Weinberger, D. *et al.* Estimating the early death toll of COVID-19 in the United States. *medRxiv* 2020.04.15.20066431 (2020) doi:10.1101/2020.04.15.20066431.
4. Epidemic Model Guided Machine Learning for COVID-19 Forecasts in the United States | medRxiv. https://www.medrxiv.org/content/10.1101/2020.05.24.20111989v1.
5. Tashman, L. J. Out-of-sample tests of forecasting accuracy: an analysis and review. *International Journal of Forecasting* **16**, 437–450 (2000).
6. COVID-19 estimation updates. *Institute for Health Metrics and Evaluation* http://www.healthdata.org/covid/updates (2020).
7. Gu, Y. COVID-19 Projections Using Machine Learning. https://covid19-projections.com/.
8. Imperial College COVID-19 LMIC Reports. https://mrc-ide.github.io/global-lmic-reports/.
9. Los Alamos Natinoal Laboratory COVID-19 Confirmed and Forecasted Case Data. https://covid-19.bsvgateway.org/.
10. MIT DELPHI Epidemiological Case Predictions COVIDAnalytics. https://www.covidanalytics.io/projections.
11. Holmdahl, I. & Buckee, C. Wrong but Useful — What Covid-19 Epidemiologic Models Can and Cannot Tell Us. *New England Journal of Medicine* **0**, null (2020).
12. Ioannidis, J. P. A. Coronavirus disease 2019: The harms of exaggerated information and non-evidence-based measures. *European Journal of Clinical Investigation* **50**, e13222 (2020).
13. Bui, Q., Katz, J., Parlapiano, A. & Sanger-Katz, M. What 5 Coronavirus Models Say the Next Month Will Look Like. *The New York Times* (2020).
14. Bui, Q., Katz, J., Parlapiano, A. & Sanger-Katz, M. Coronavirus Models Are Nearing Consensus, but Reopening Could Throw Them Off Again. *The New York Times* (2020).
15. Caution Warranted: Using the Institute for Health Metrics and Evaluation Model for Predicting the Course of the COVID-19 Pandemic. *Annals of Internal Medicine*.
16. Reich Lab COVID-19 Forecast Hub. https://reichlab.io/covid19-forecast-hub/.
17. Project Score Data: COVID-19 Forecasts - Zoltar. https://zoltardata.com/project/44/score_data.
18. UCLAML Combating COVID-19. http://covid19.uclaml.org/compare.
19. Coronavirus Pandemic (COVID-19) - Statistics and Research - Our World in Data. https://ourworldindata.org/coronavirus.
20. *nytimes/covid-19-data*. (The New York Times, 2020).
21. COVID-19 Map. *Johns Hopkins Coronavirus Resource Center* https://coronavirus.jhu.edu/map.html.
22. CDC. Coronavirus Disease 2019 (COVID-19). *Centers for Disease Control and Prevention* http://www.cdc.gov/coronavirus/2019-ncov/covid-data/forecasting-us.html (2020).

23. A greater tragedy than we know: Excess mortality rates suggest that COVID-19 death toll is vastly underestimated in LAC. *UNDP* https://www.latinamerica.undp.org/content/rblac/en/home/presscenter/director-s-graph-for-thought/a-greater-tragedy-than-we-know--excess-mortality-rates-suggest-t.html.
24. IHME COVID-19 Estimation Update: May 4th, 2020. http://www.healthdata.org/sites/default/files/files/Projects/COVID/Estimation_update_050420.pdf.
25. IHME COVID-19 Estimation Update: May 29th, 2020. http://www.healthdata.org/sites/default/files/files/Projects/COVID/Estimation_update_05.30.2020.pdf.
26. Foreman, K. J. *et al.* Forecasting life expectancy, years of life lost, and all-cause and cause-specific mortality for 250 causes of death: reference and alternative scenarios for 2016–40 for 195 countries and territories. *The Lancet* (2018) doi:10.1016/S0140-6736(18)31694-5.
27. Rivers, C. & George, D. How to Forecast Outbreaks and Pandemics. (2020).
28. Dicker, D. *et al.* Global, regional, and national age-sex-specific mortality and life expectancy, 1950–2017: a systematic analysis for the Global Burden of Disease Study 2017. *The Lancet* **392**, 1684–1735 (2018).

Methods

In order to construct a framework to compare the out-of-sample predictive performance of COVID-19 mortality forecasting models, we conducted a comprehensive review of published and unpublished papers and models related to COVID forecasting. We identified models with date-versioned, publicly available mortality estimates for multiple countries, and create a code base to automatically download and process them into a common framework, as models update daily. We calculated errors for both the magnitude of mortality, and the timing of peak daily mortality. Finally, we calculated summary statistics describing the errors over a number of dimensions, including region, weeks of extrapolation, and month of estimation.

Comprehensive Review and Data Compilation

Conducting a comprehensive review of COVID-19 mortality forecasting is complicated by the unpublished nature of many models. In general, the pace of COVID-19 research has been rapid, to provide up-to-date evidence for an evolving situation. We therefore draw on both traditional and non-traditional sources to find models. We used PubMed to identify models published in journal articles, medRxiv to identify models published in preprint articles, and a collection of models curated by the Reich Lab and the US Centers for Disease Control describing forecasts for the United States, many of which also produce estimates internationally[16]. In PubMed and medRxiv we selected any articles with the terms "COVID", or "Coronavirus" present, as well as one of either term "Project" or "Forecast". At the time of publication of this article, n=686 articles had been screened. The systematic review framework was created to automatically download new citations for review, and subsequently be updated in a prospective manner, to stay current on future forecasting model efforts.

Each article or model was screened for 5 inclusion criteria:

1) Including a forecasting component (n=383),
2) and making projections for at least five locations (n=59),
3) and making projections for COVID-19 mortality (n=39),
4) and making projections at least four weeks into the future (n=35),
5) and providing publicly available, date-versioned estimates (n=7).

The seven models which fit these criteria are described in the main text (Table 1). All screened articles, including their inclusion and exclusion criteria, are described in the review supplemental file. The code used to compile candidate models and articles, and conduct the systematic review, is presented along with the remainder of the codebase.

For the seven models that were determined to meet all inclusion criteria, a codebase was developed to automatically download each date-versioned set of estimates as they became available. The model date (or estimation date) was assigned as the date on which the estimated became publicly available.

Locations were mapped onto the location hierarchy used by the Global Burden of Disease Study (GBD)[28] that categorizes all countries into regions, and regions into super-regions. When estimates were available at multiple geographic levels, the admin-0 (national) level results were used in all cases, except the United States, where both admin 0 and admin-1 (state) level results were used.

For models that provided only daily deaths, cumulative deaths were calculated as a rolling sum. Similarly, for models that provided only cumulative deaths, daily deaths were calculating by taking the daily first difference of cumulative deaths.

We chose to use mortality data collected by Johns Hopkins University Coronavirus Resource Center as the in-sample data against which forecasts were validated at the national level, and data from the New York Times for state-level data for the United States[20]. Data from both providers were mapped onto the GBD location hierarchy. For all analyses, the most recent set of input data were used, reflecting any potential revisions of historical trends.

Mortality Magnitude Predictive Validity

Before out-of-sample errors could be assessed for COVID-19 mortality, differences in input data sources between models were investigated and controlled. Estimates for the same locations, from different models, can differ greatly in magnitude of estimated mortality when they use input data sources that use different methodologies. To create a fair comparison, before errors were calculated, each model-, model-date-, and location-specific timeseries was shifted to match the "true" in-sample data for that model's date of release. This was accomplished by calculating the timeseries-specific difference on the model date, and apply it to the entire timeseries as a fixed intercept shift. Subsequent forecasting errors were calculated using the resulting shifted time-series.

Out-of-sample errors were calculated for each timeseries at weekly intervals, beginning at one week, through six weeks of extrapolation. Summary statistics were first calculated across model-runs for each location, for use in country-specific graphics (Figure 1, for example). Summary statistics included the median absolute error, a measure of accuracy, and median error, a measure of bias. These were calculated separately by model, and by weeks of extrapolation.

Subsequently, errors were summarized between countries. Summary statistics included the median absolute percent error, a measure of accuracy, and median percent error, a measure of bias. Relative error statistics were used for all inter-country comparisons, to account for the substantial differences in magnitude of deaths between locations. Summary statistics were calculated in a stratified manner by regional groupings from the GBD[28], as well as weeks of extrapolation, and month of estimation. Pooled summary statistics were also calculated across models, to provide context about commonalities in trends in predictive performance over time and geographies.

Peak Timing

In order to calculate out-of-sample predictive validity statistics on how well each model predicted the timing of peak daily deaths, we smoothed daily death data, which are highly stochastic, applied an algorithm to detect peaks in both observed data and forecasted model estimates, and calculated errors in the difference in number of days between the observed and estimated peaks.

First, observed daily death data were smoothed to provide stable time-series that could be used for local maxima detection. We used various smoothers to accomplish this task, including a LOESS smoother with a span of 0.33, run separately for each location-specific timeseries, a 7-day rolling average, and a 3-day rolling average applied tenfold to the same timeseries. We chose to present results calculated using the LOESS smoother in the main text, as it was found to be the most robust method to daily stochasticity that could introduce false peaks. Although most models produced smooth timeseries of daily deaths,

some also demonstrated stochasticity, and so all forecasted daily death timeseries were also smoothed with a LOESS smoother.

Peaks in smoothed, observed daily deaths were calculated according to the following algorithm. A peak was defined as:

1) a local maximum *p* in the timeseries at time *t,*
2) where no other point exists in the next 21 days (*t* through *t*+21*)* that exceeds the *p* by more than 20%,
3) *t* does not fall within the last seven days of the timeseries,
4) where *p* represents at least 5 deaths per day,
5) and if multiple such points *p* exist that meet the above criteria, then the first value will be selected.

Peaks in forecasted trends were also identified with the same algorithm. For a time-series in which no peak was identified using the above algorithm, for a location which did have a peak in observed data, the global maximum value was used. This captured errors among models that failed to ever predict a peak, despite a true peak being observed. Errors for locations with a true peak in observed data, for the model runs in which the model date was at least seven days prior to the true detected peak. Errors were defined as the difference between the date of the true peak and the estimated peak from each forecasting model, in days. Summary statistics included the median absolute error in days, as a measure of accuracy, and the median error in days as a measure of bias. Errors were stratified by model, and weeks of extrapolation, which was defined as:

$$Weeks\ of\ extrapolation\ =\ floor((peak\ date - model_release_date)/7)$$

Summary statistics were masked for models that were not released in time to produce peak timing estimates for at least 20 total locations. Due to limited regional coverage it was not possible to stratify results by geography. This will likely become feasible as more locations pass their peak of daily mortality.

Code Availability

All code used for these analyses is publicly available at:
https://github.com/pyliu47/covidcompare.

Data Availability

All data used for these analyses are publicly available at:
https://github.com/pyliu47/covidcompare.

Acknowledgements

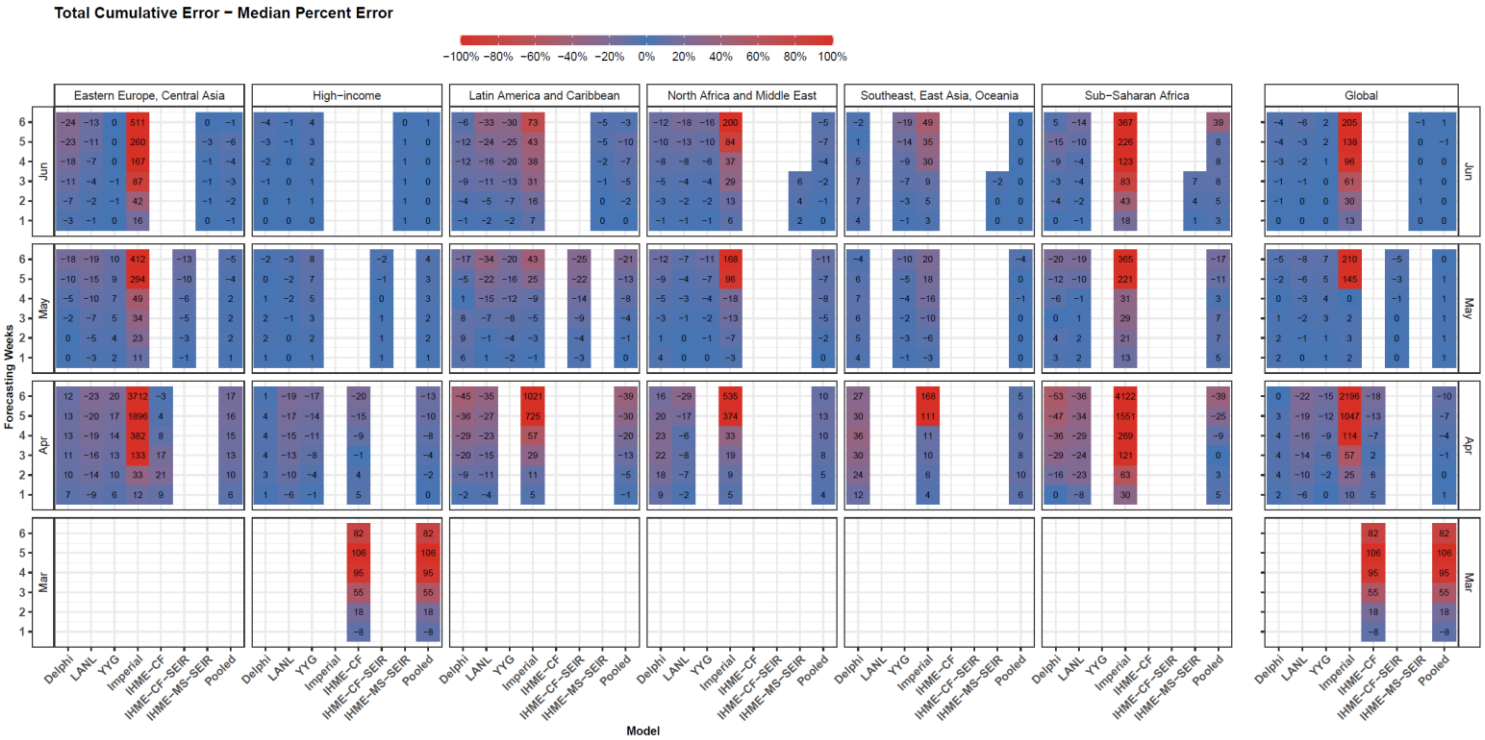
This work was primarily supported by the Bill & Melinda Gates Foundation. J.F. received support from the UCLA Medical Scientist Training program (NIH NIGMS training grant GM008042).

Competing Interests

The authors declare no competing interests.

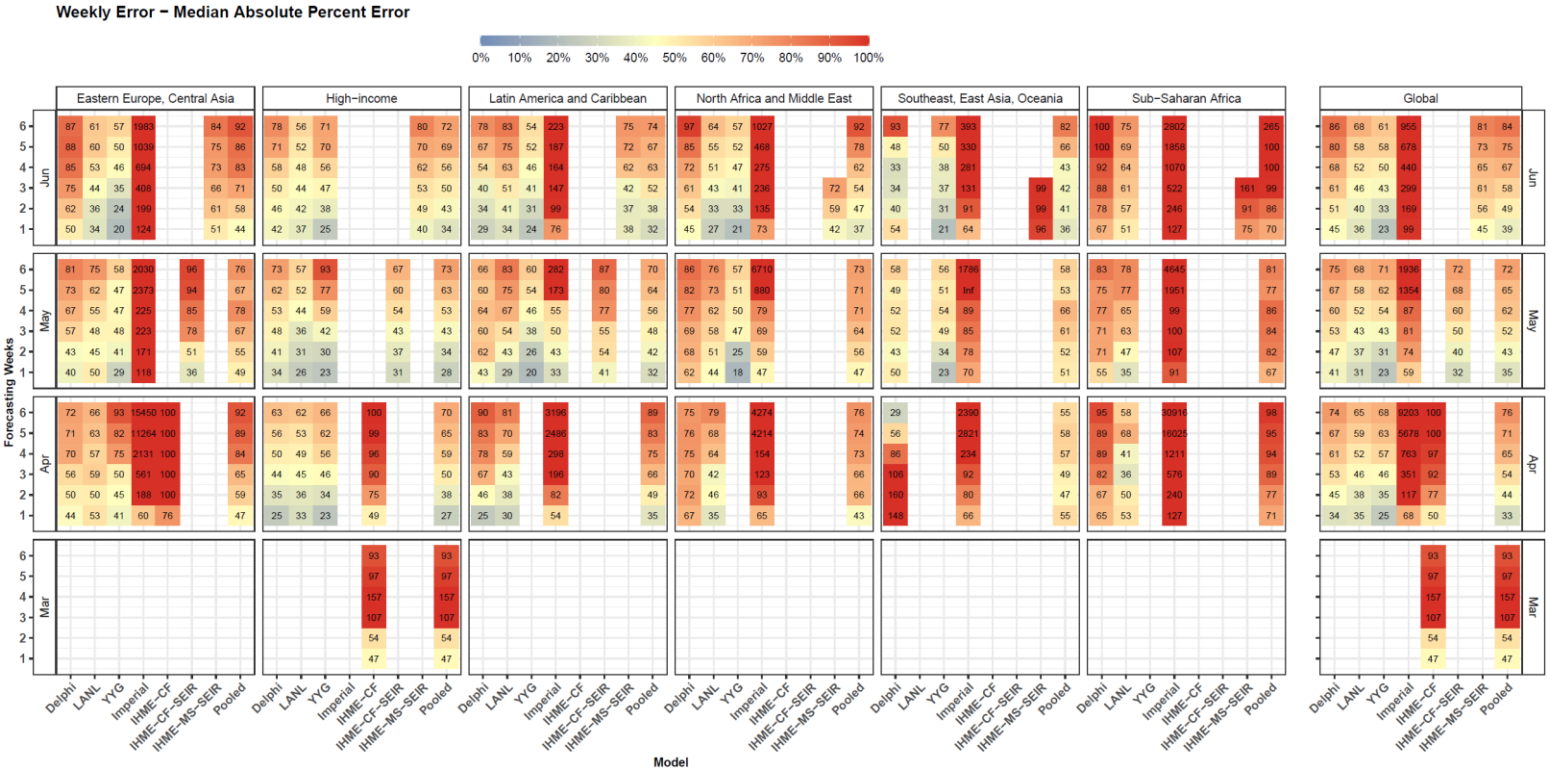
Correspondence and requests for materials should be addressed to E.G.

Extended Data Figures



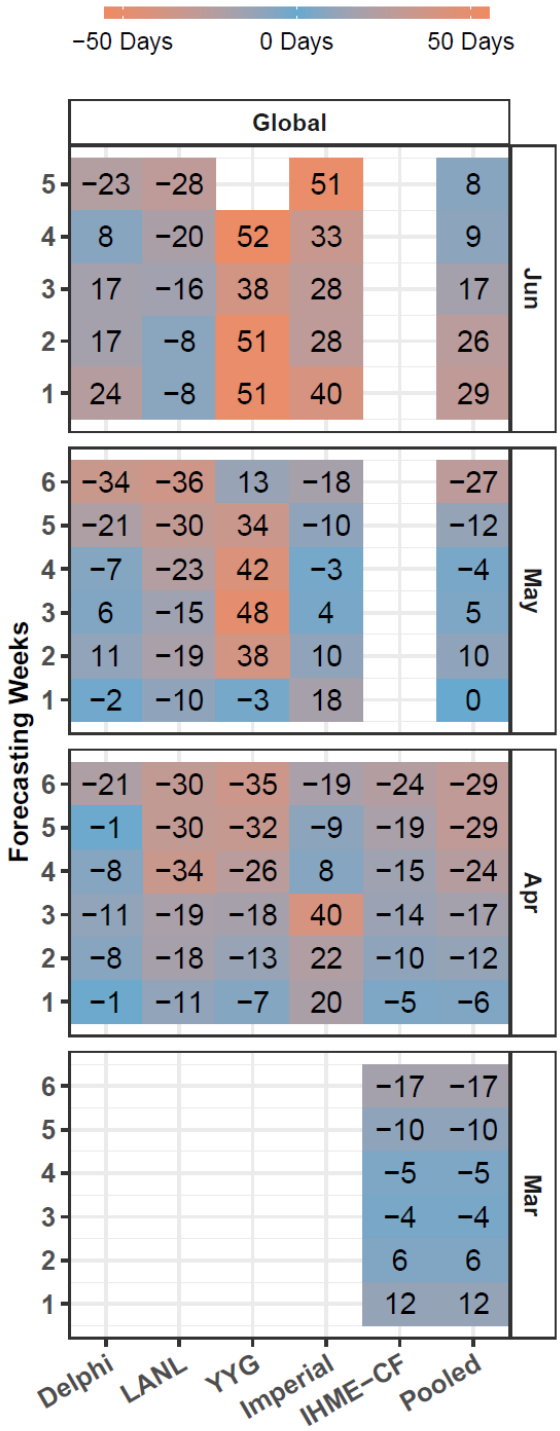
Extended Data Figure 1. Bias in Cumulative Mortality Predictions

Median percent error values, a measure of bias, were calculated across all observed errors at weekly intervals, for each model, by month of estimation, weeks of forecasting, and super regional grouping used in the Global Burden of Disease Study. Values that represent fewer than five locations are masked due to small sample size.

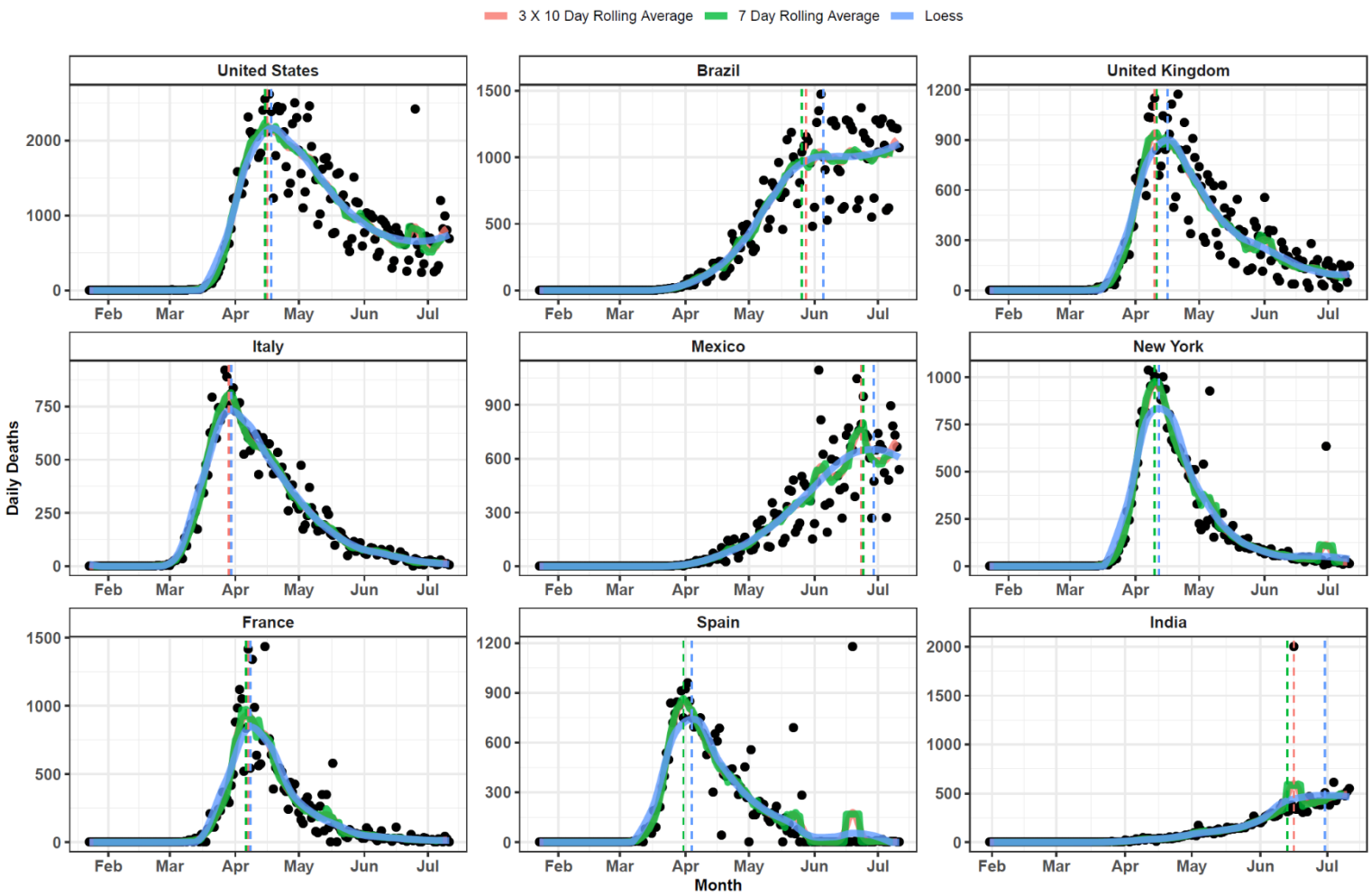


Extended Data Figure 2. Predictive Validity of Weekly Mortality - Median Absolute Percent Error

Median absolute percent error values were calculated for weekly mortality rates across all observed errors at weekly intervals, for each model, by month of estimation, weeks of forecasting, and super regional grouping used in the Global Burden of Disease Study. Values that represent fewer than 5 locations are masked due to small sample size.

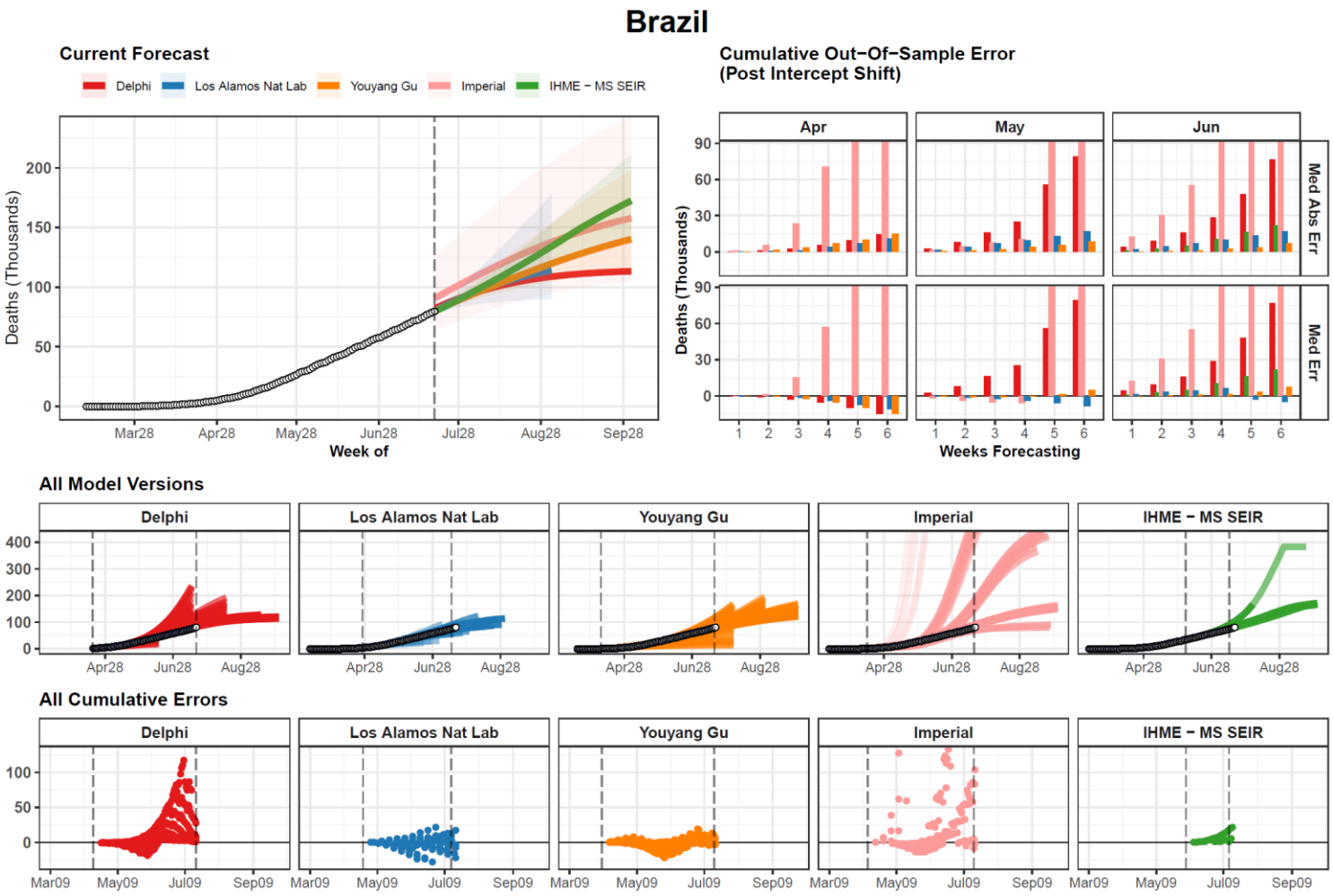


Extended Data Figure 3. Bias in Peak Timing by Month of Estimation
Median error in days are shown by model, weeks of forecasting, and estimation month. Values that represent fewer than ten errors for a given model and forecasting-week are not shown. Errors only reflect models released at least seven days before the observed peak in daily mortality.

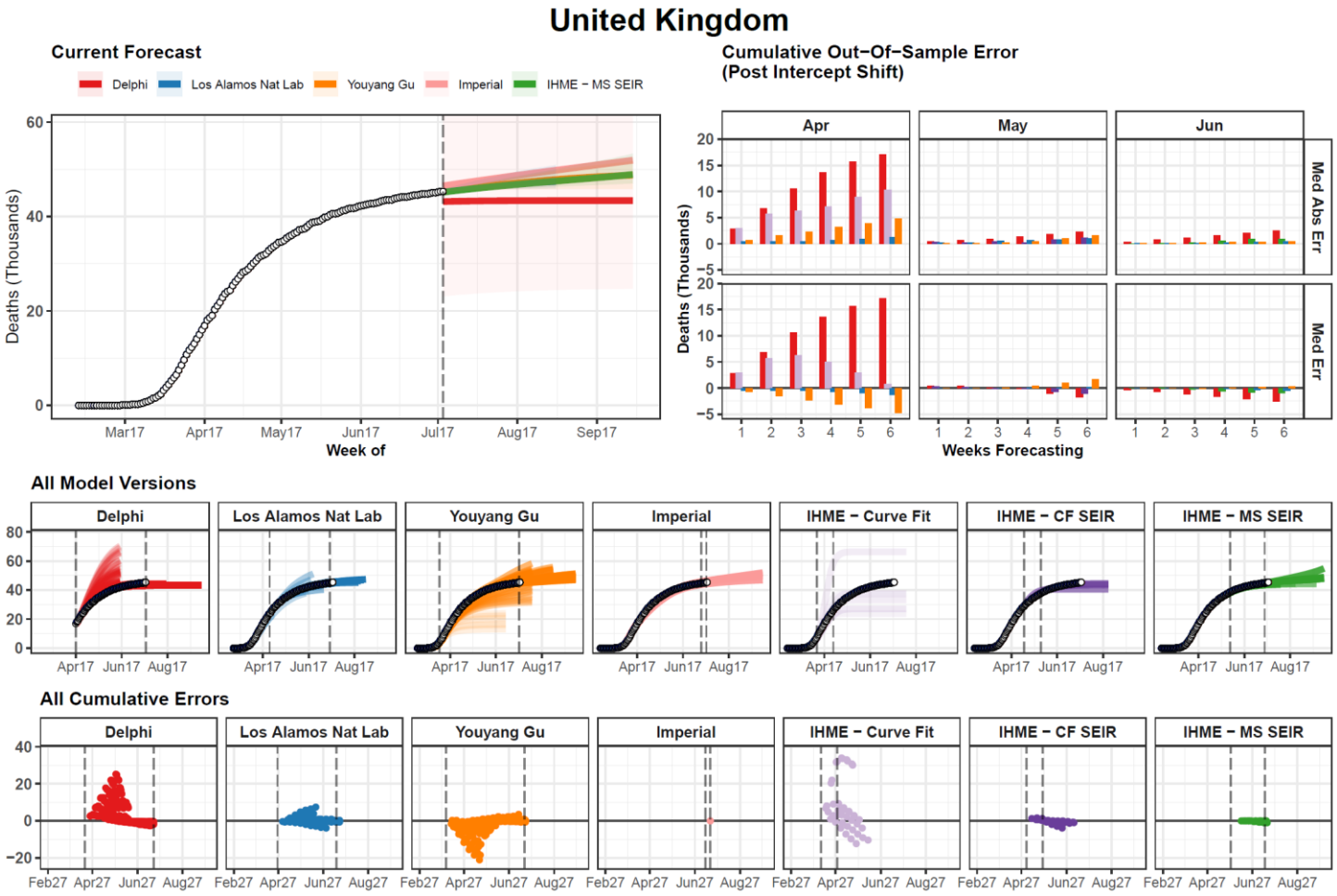


Extended Data Figure 4. Smoothing Method – Example for nine Countries

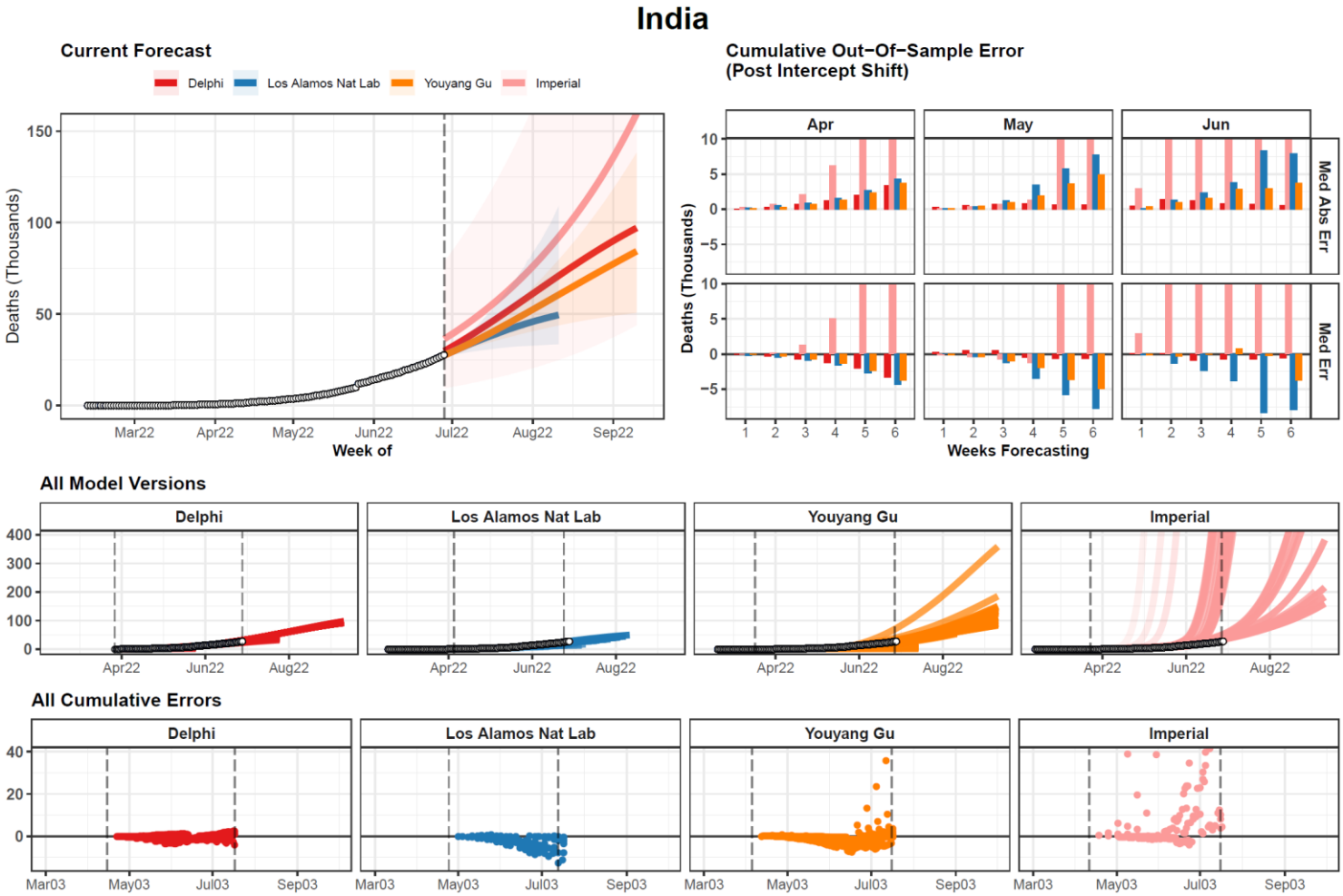
Daily deaths are shown for nine locations, as well as three methods used to smooth them prior to peak date calculation. Calculated peaks from each method are shown with dashed vertical lines. Smoothing method is shown by color. Graphs for all locations are shown in the supplement.



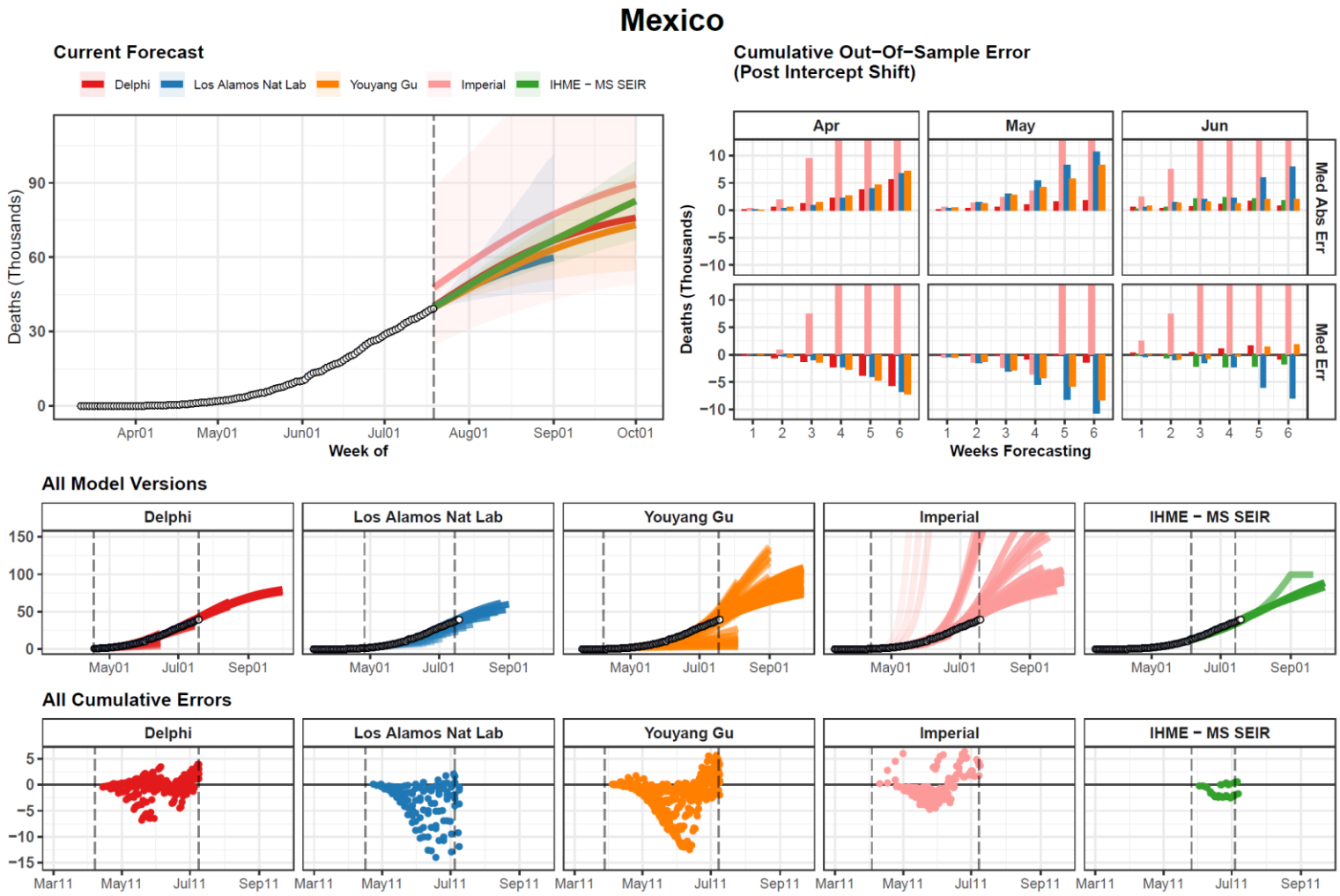
Extended Data Figure 5. Cumulative Mortality Forecasts and Prediction Errors by Model – Example for Brazil



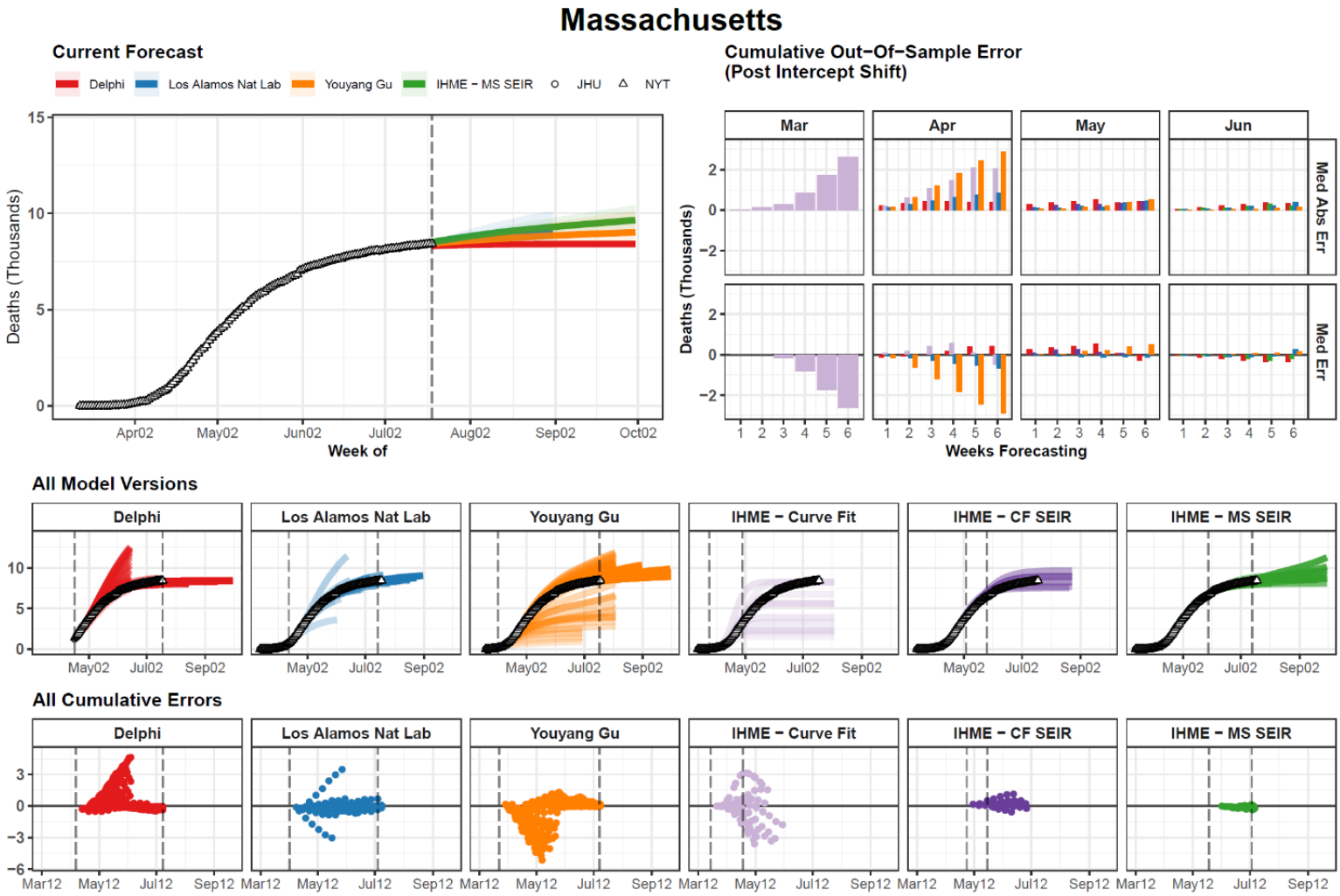
Extended Data Figure 6. Cumulative Mortality Forecasts and Prediction Errors by Model – Example for United Kingdom



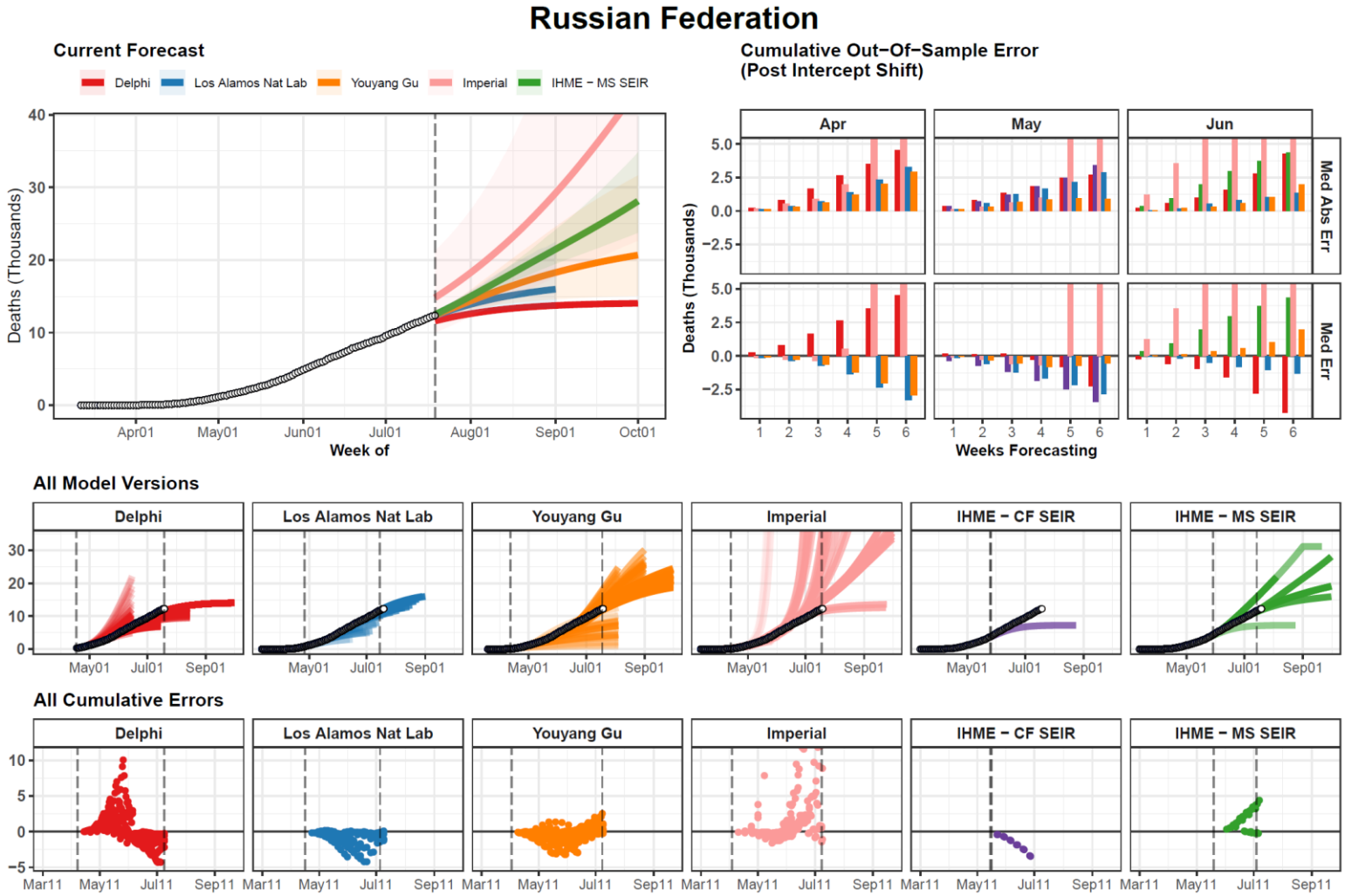
Extended Data Figure 7. Cumulative Mortality Forecasts and Prediction Errors by Model – Example for India



Extended Data Figure 8. Cumulative Mortality Forecasts and Prediction Errors by Model – Example for Mexico



Extended Data Figure 9. Cumulative Mortality Forecasts and Prediction Errors by Model – Example for Massachusetts



Extended Data Figure 10. Cumulative Mortality Forecasts and Prediction Errors by Model – Example for Russian Federation