medRxiv preprint doi: https://doi.org/10.1101/2020.07.13.20151233; this version posted August 26, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.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 4.0 International license .

Predictive performance of international COVID-19 mortality forecasting models

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

Abstract

Background: Forecasts and alternative scenarios of the COVID-19 pandemic have been critical inputs into a range of important decisions by healthcare providers, local and national government agencies and international organizations and actors. Hundreds of COVID-19 models have been released. Decision-makers need information about the predictive performance of these models to help select which ones should be used to guide decision-making.

Methods: We identified 383 published or publicly released COVID-19 forecasting models. Only seven models met the inclusion criteria of: estimating for five or more countries, providing regular updates, forecasting at least 4 weeks from the model release date, estimating mortality, and providing date-versioned sets of previously estimated forecasts. These models included those produced by: a team at MIT (Delphi), Youyang Gu (YYG), the Los Alamos National Laboratory (LANL), Imperial College London (Imperial) the USC Data Science Lab (SIKJalpha), and three models produced by the Institute for Health Metrics and Evaluation (IHME). For each of these models, we examined the median absolute percent error—compared to subsequently observed trends—for weekly and cumulative death forecasts. Errors were stratified by weeks of extrapolation, world region, and month of model estimation. For locations with epidemics showing a clear peak, each model's accuracy was also evaluated in predicting the timing of peak daily mortality.

Results: Across models, the median absolute percent error (MAPE) on cumulative deaths for models released in June rose with increased weeks of extrapolation, from 2.3% at one week to 32.6% at ten weeks. Globally, ten-week MAPE values were lowest for IHME-MS-SEIR (20.3%) and YYG (22.1). Across models, MAPE at six weeks were the highest in Sub-Saharan Africa (55.6%), and the lowest in high-income countries (7.7%). Median absolute errors (MAE) for peak timing also rose with increased forecasting weeks, from 14 days at one week to 30 days at eight weeks. Peak timing MAE at eight weeks ranged from 24 days for the IHME Curve Fit model, to 48 days for LANL.

Interpretation: Five of the models, from IHME, YYG, Delphi, SIKJalpha and LANL, had less than 20% MAPE at six weeks. Despite the complexities of modelling human behavioural responses and government interventions related to COVID-19, predictions among these better-performing models were surprisingly accurate. Forecasts and alternative scenarios can be a useful input to decision-makers, although users should be aware of increasing errors with a greater amount of extrapolation time, and corresponding steadily widening uncertainty intervals further in the future. The framework and publicly available codebase presented can be routinely used to evaluate the performance of all publicly released models meeting inclusion criteria in the future, and compare current model predictions.

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

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

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/10.1101/2020.07.13.20151233; this version posted August 26, 2020. 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 4.0 International license .

Background

Forecasts and alternative scenarios of COVID-19 have been critical inputs into a range of important decisions by healthcare providers, local and national government agencies and international organizations and actors[1–4]. For example, hospitals need to prepare for potential surges in the demand for hospital beds, ICU beds and ventilators[1]. National critical response agencies such as the US Federal Emergency Management Agency have scarce resources including ventilators that can be moved to locations in need with sufficient notice[5,6]. Longer range forecasts are important for decisions such as the potential to open schools, universities and workplaces, and under what circumstances[7]. Much longer-range forecasts—six months to a year—are important for a wide range of policy choices, where efforts to reduce disease transmission must be balanced against economic outcomes such as unemployment and poverty[8]. Furthermore, vaccine and new therapeutic trialists need to select locations that will have sufficient transmission to test new products in the time frame when phase three clinical trials are ready to be launched. Nevertheless, hundreds of forecasting models have been published and/or publicly released, and it is often not immediately clear which models have had the best performance, or are most appropriate for predicting a given aspect of the pandemic.

Existing COVID-19 forecasting models differ substantially in methodology, assumptions, range of predictions, and quantities estimated. Furthermore, mortality forecasts for the same location have often differed substantially, in many cases by more than an order of magnitude, even within a six-week forecasting window. The challenge for decision-makers seeking input from models to guide decisions, which can impact many thousands of lives, is therefore not the availability of forecasts, but guidance on which forecasts are likely to be most accurate. Out-of-sample predictive validation—checking how well past versions of forecasting models predict subsequently observed trends—provides insight into future model performance[9]. Although some comparisons have been conducted for models describing the epidemic in the United States[10–13], to our knowledge similar analyses have not been undertaken for models covering multiple countries, despite the growing global impact of COVID-19.

This paper introduces a publicly available dataset and evaluation framework for assessing the predictive validity of COVID-19 mortality forecasts. The framework and associated open-access software can be routinely used to track model performance. This will, overtime, serve as a reference for decision-makers on historical model performance, and provide insight into which models should be considered for critical decisions in the future.

Methods

Systematic Review

386 published and unpublished COVID-19 forecasting models were reviewed (see appendix). Models were excluded from consideration if they 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. Eight models which fit all inclusion criteria were evaluated (Table 1). These included those modelled by: DELPHI-MIT (Delphi)[14,15], Youyang Gu (YYG)[10], the Los Alamos National Laboratory (LANL)[16], Imperial College London (Imperial)[17],the SIKJ-Alpha model from the USC Data

medRxiv preprint doi: https://doi.org/10.1101/2020.07.13.20151233; this version posted August 26, 2020. 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 4.0 International license .

Science Lab (SIKJalpha)[18], and three models produced by the Institute for Health Metrics and Evaluation (IHME)[19]. 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)[20]. 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[21]. 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 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.

Model Comparison Framework

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, in order to provide a fair and relevant comparison.

Input data: 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[22–24]. 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[23,24]. Locations were excluded from the evaluation (including Ecuador and Peru) where models used alternative data sources, such as excess mortality, in settings with known marked under-registration of COVID-19 deaths and cases[25,26]. 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 supplemental methods).

Frequency of public releases of model estimates: 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 cut-offs from a specific day of the week[27]. For the sake of including all publicly available modelled 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.

medRxiv preprint doi: https://doi.org/10.1101/2020.07.13.20151233; this version posted August 26, 2020. 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 4.0 International license .

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. To increase comparability, summary error statistics were stratified by super-region used in the Global Burden of Disease Study[28], weeks of extrapolation, and month of estimation, and we 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 also 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. The focus of this analysis is on mortality, as it was the most widely reported outcome, and it also has a high degree of societal, epidemiological and public health importance.

Comparison of Cumulative Mortality Forecasts

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 main quantity that is considered is 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 appendix. Errors were assessed for systematic upward or downward bias, and errors for weekly, rather than cumulative deaths, were also assessed. In calculating summary statistics, percent errors were used to control for the large differences in the scale of the epidemic between locations. Medians, rather than means, are calculated 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. Results are presented for June in the main text—the most recent month allowing for assessment of errors at ten weeks of forecasting—and errors for all months are shown in the appendix.

Comparison of Peak Daily Mortality Forecasts

Each model was also assessed on how well it 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, the observed peak of daily deaths in each location was estimated first—a task complicated by the highly volatile nature of reported daily deaths values. Each timeseries of daily deaths was smoothed, and the date of the peak observed in each location, as well as the predicted peak for each iteration of each forecasting model was calculated (see supplemental methods). A LOESS smoother was used, 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. Predictive validity statistics were stratified by the number of weeks in advance of the observed peak that

medRxiv preprint doi: https://doi.org/10.1101/2020.07.13.20151233; this version posted August 26, 2020. 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 4.0 International license .

the model was released, as well as the month in which the model was released. Results shown in the main text were pooled across months, as there was little evidence of dramatic differences over time (see appendix). 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.

Results

The evaluation framework developed here 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 appendix. When looking across iterations of forecasts, a wide range of variation can be observed for nearly all of the models. Nevertheless, in many locations, models now have largely reached consensus. Figure 1, and similar figures in the appendix, also highlight the direction of error for each model in each location. Systematic assessments of bias are shown in Figure 2, and Supplemental Figure 2. The Delphi and LANL models from June underestimated mortality, with median percent errors of -6.0% and -6.9% at 6 weeks respectively, while Imperial has tended to vastly overestimated (+227.4%), and the YYG and IHME-MS-SEIR models have been largely unbiased (0.0% and -0.3% respectively).

Overall model performance is shown for cumulative deaths by week in Figure 3. As one might expect, median absolute percent error (MAPE) tends to increase by number of weeks of extrapolation. Across models released in June the MAPE rose from 2.3% at one week to 32.6% at ten weeks. Decreases in predictive ability with greater periods of extrapolation were similarly noted for errors in weekly deaths (Supplemental Figure 3). At the global level, MAPE at six weeks was less than 20% for YYG (9.9%), LANL (12.6%), IHME-MS-SEIR (13.4%), SIKJalpha (16.9%) and Delphi (17.7%). The Imperial model had considerably larger errors, reaching 20-fold higher than other models by 6 weeks. This appears to be largely driven by the aforementioned tendency to overestimate mortality. At ten weeks, MAPE values were lowest for the IHME-MS-SEIR model (20.3%) and YYG (22.1%), while the SIKJalpha model showed intermediate performance (33.1%) and the Imperial model had a substantially elevated MAPE (548.9%).

Figure 3 also shows that model performance varies substantially by region. The lowest errors across models were observed among high-income countries and those in Southeast, East Asia and Oceania, with 6-week MAPE values of 7.7% and 19.6% respectively. In contrast, the largest errors have been seen in sub-Saharan Africa, with a 6-week MAPE of 55.6%, South Asia, with a MAPE of 36.8%, and Latin America and the Caribbean, with a MAPE of 32.3%.

The evaluation framework for exploring the ability of models to predict the timing of peak mortality accurately—a matter of paramount importance for health service planning—is shown in Figure 4 for an example location, Massachusetts. Similar figures for all locations are shown in the appendix. Median absolute errors (MAE) for peak timing also rose with increased forecasting weeks, from 14 days at one week to 30 days at eight weeks. MAE at eight weeks ranged from 24 days for the IHME Curve Fit model to 48 days for the LANL model, with an overall error across models of 30 days. Models were generally biased towards predicting peak mortality too early (Supplemental Figure 4).

medRxiv preprint doi: https://doi.org/10.1101/2020.07.13.20151233; this version posted August 26, 2020. 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 4.0 International license .

Discussion

Eight COVID-19 models were identified that covered more than five countries, were regularly updated, publicly released and provide archived results for past forecasts. Taken together at ten weeks, the models released in June had a median average percent error of 32.6% percent. Errors tend to increase with longer forecasts, rising from 2.2% at one week to 16.5% at 6 weeks. At ten weeks of extrapolation, the best predictive performance was observed for the IHME-MS-SEIR models, with a MAPE of 20.3%, as well as the YYG model, with a MAPE of 22.1%. The projections provided by Imperial had considerably higher error (548.9%) and the SIKJalpha model had an intermediate 33.1% for the same period.

A forecast of the trajectory of the COVID-19 epidemic for a given location depends on three sets of factors: 1) attributes of the virus itself, and characteristics of the location, such as population density and the use of public transport; 2) individual behavioural responses to the pandemic such as avoiding contact with others or wearing a mask; and 3) the actions of governments, such as the imposition of a range of social distancing mandates. Given the complexity of forecasting human and governmental behaviours, especially in the context of a new pandemic, performance of most of the models evaluated here was encouraging. Nevertheless, errors were observed to grow with greater extrapolation time, indicating that governments and planners should recognize the wide uncertainty that comes with longer range forecasts, and plan accordingly. Hospital administrators may want to plan for the higher end of the forecast range, while government policymakers may elect to use the mean forecast, depending on their risk tolerance.

The vast majority of COVID-19 forecasting 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 for modelling science, and the ability of policy makers to discriminate between a burgeoning number of models[29].Many of the models featured in this analysis were generally unbiased, or tended to underestimate future mortality, while other models, such as the Imperial model, as well as many other published models that did not meet our inclusion criteria, tend to substantially overestimate transmission, even within the first 4 weeks of a forecast. This tendency towards over-estimation among SEIR and other transmission-based models is easy to understand given the potential for the rapid doubling of transmission. Nevertheless, sustained exponential growth in transmission is not often observed, likely due to the behavioural responses of individuals and governments; both react to worsening circumstances in their communities, modifying behaviours and imposing mandates to restrict activities. This endogenous behavioural response is commonly included in economic analyses, however, it has not been routinely featured in transmission dynamics modelling of COVID-19. More explicit modelling of the endogenous response of individuals and governments may improve future model performance for a range of models.

Modelling groups are increasingly providing both reference forecasts, describing likely future trends, and alternative scenarios describing the potential effects of policy choices, such as school openings, timing of mandate reimposition, or planning for hospital surges. For these scenarios, the error in the reference forecast—which we describe in this manuscript—is actually less important than the error in the effect implied by the difference between the reference forecast and policy scenario. Unfortunately, evaluating the accuracy of these counterfactual scenarios is a very difficult task. The validity of such

medRxiv preprint doi: https://doi.org/10.1101/2020.07.13.20151233; this version posted August 26, 2020. 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 4.0 International license .

claims depends on the supporting evidence for the assumptions about a policy's impact on transmission. The best option for decision-makers is likely to examine the impact of these policies as portrayed by a range of modelling groups, especially those that have historically had reasonable predictive performance in their reference forecasts.

Given that five very different models demonstrated six-week errors for cumulative deaths below 20%, it would likely be worthwhile to construct an ensemble of these models, and evaluate the performance the ensemble compared to each component. Although from a logistical standpoint, creating an ensemble of the forecasts would be relatively straightforward, it would be more challenging to integrate such a model pool with scenarios assessing policy options, given that the models have highly different underlying structures. Nevertheless, the inclusion of the models shown here, and future models meeting criteria into an ensemble framework, is an important area for future research.

This analysis of the performance of publicly released COVID-19 forecasting models has some limitations. First, we have focused only on forecasts of deaths, as they are available for all models included here. However, hospital resource use is also of critical importance, and deserves future consideration. Nevertheless, this will be complicated by the heterogeneity in hospital data reporting; many jurisdictions report hospital census counts, others report hospital admissions, and still others do not release hospital data on a regular basis. Without a standardized source for these data, assessment of performance can only be undertaken in an ad hoc way. Second, many performance metrics exist which could have been computed for this analysis. We have focused on reporting median absolute percent error, as the metric is quite stable, and provides an easily interpreted number that can be communicated to a wide audience. However, relative error is an exacting standard. For example, a forecast of three deaths in a location that observed only one may represent a 200% error, yet it would be of little policy or planning significance. On the other hand, focusing on absolute error would create an assessment dominated by a limited number of locations with large epidemics. Future assessment could consider different metrics that may offer new insights, although the relative rank of performance by model is likely to be similar.

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 assessed here —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, other researchers can build on the results presented here 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. This study provides 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. As epidemics begin to take off in settings such as sub-Saharan Africa, South Asia, or parts of Latin America,

medRxiv preprint doi: https://doi.org/10.1101/2020.07.13.20151233; this version posted August 26, 2020. 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 4.0 International license .

regularly updating models, and assessing their predictive validity, will be important in order to provide stakeholders with the best possible tools for COVID-19 decision-making.

References

1 Team IC-19 health service utilization forecasting, Murray CJ. Forecasting COVID-19 impact on hospital bed-days, ICU-days, ventilator-days and deaths by US state in the next 4 months. *medRxiv* 2020; : 2020.03.27.20043752.

2 Lu FS, Nguyen AT, Link NB, Lipsitch M, Santillana M. Estimating the Early Outbreak Cumulative Incidence of COVID-19 in the United States: Three Complementary Approaches. *medRxiv* 2020; : 2020.04.18.20070821.

3 Weinberger D, Cohen T, Crawford F, *et al.* Estimating the early death toll of COVID-19 in the United States. *medRxiv* 2020; : 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 (accessed June 23, 2020).

5 Critical Supply Shortages — The Need for Ventilators and Personal Protective Equipment during the Covid-19 Pandemic | NEJM. *New England Journal of Medicine* http://www.nejm.org/doi/full/10.1056/NEJMp2006141 (accessed July 26, 2020).

6 FEMA Administrator March 27, 2020, letter to Emergency Managers Requesting Action on Critical Steps | FEMA.gov. https://www.fema.gov/news-release/2020/03/27/fema-administrator-march-27-2020-letter-emergency-managers-requesting-action (accessed July 26, 2020).

7 Viner RM, Russell SJ, Croker H, *et al.* School closure and management practices during coronavirus outbreaks including COVID-19: a rapid systematic review. *The Lancet Child & Adolescent Health* 2020; **4**: 397–404.

8 Atkeson A. What Will Be the Economic Impact of COVID-19 in the US? Rough Estimates of Disease Scenarios. National Bureau of Economic Research, 2020 DOI:10.3386/w26867.

9 Tashman LJ. Out-of-sample tests of forecasting accuracy: an analysis and review. *International Journal of Forecasting* 2000; **16**: 437–450.

10 Gu Y. COVID-19 Projections Using Machine Learning. https://covid19-projections.com/ (accessed June 23, 2020).

11 Reich Lab COVID-19 Forecast Hub. https://reichlab.io/covid19-forecast-hub/ (accessed June 23, 2020).

12 Project Score Data: COVID-19 Forecasts - Zoltar. https://zoltardata.com/project/44/score_data (accessed June 23, 2020).

13 UCLAML Combating COVID-19. http://covid19.uclaml.org/compare (accessed June 23, 2020).

14 MIT DELPHI Epidemiological Case Predictions COVIDAnalytics. https://www.covidanalytics.io/projections (accessed June 23, 2020).

15 Li ML, Bouardi HT, Lami OS, Trikalinos TA, Trichakis NK, Bertsimas D. Forecasting COVID-19 and Analyzing the Effect of Government Interventions. *medRxiv* 2020; : 2020.06.23.20138693.

16 Los Alamos Natinoal Laboratory COVID-19 Confirmed and Forecasted Case Data. https://covid-19.bsvgateway.org/ (accessed June 23, 2020).

medRxiv preprint doi: https://doi.org/10.1101/2020.07.13.20151233; this version posted August 26, 2020. 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 4.0 International license .

17 Imperial College COVID-19 LMIC Reports. https://mrc-ide.github.io/global-lmic-reports/ (accessed June 23, 2020).

18 Srivastava A, Xu T, Prasanna VK. Fast and Accurate Forecasting of COVID-19 Deaths Using the SIkJα Model. *arXiv:200705180 [physics, q-bio]* 2020; published online July 12. http://arxiv.org/abs/2007.05180 (accessed Aug 23, 2020).

19 COVID-19 estimation updates. Institute for Health Metrics and Evaluation. 2020; published online March 24. http://www.healthdata.org/covid/updates (accessed June 23, 2020).

20 IHME COVID-19 Estimation Update: May 4th, 2020. http://www.healthdata.org/sites/default/files/files/Projects/COVID/Estimation_update_050420.pdf (accessed July 6, 2020).

21 IHME COVID-19 Estimation Update: May 29th, 2020. http://www.healthdata.org/sites/default/files/files/Projects/COVID/Estimation_update_05.30.2020.pdf (accessed July 6, 2020).

22 Coronavirus Pandemic (COVID-19) - Statistics and Research - Our World in Data. https://ourworldindata.org/coronavirus (accessed June 28, 2020).

23 nytimes/covid-19-data. The New York Times, 2020 https://github.com/nytimes/covid-19-data (accessed June 28, 2020).

24 COVID-19 Map. Johns Hopkins Coronavirus Resource Center. https://coronavirus.jhu.edu/map.html (accessed June 23, 2020).

25 Covid-19 data - Tracking covid-19 excess deaths across countries | Graphic detail | The Economist. https://www.economist.com/graphic-detail/2020/07/15/tracking-covid-19-excess-deaths-across-countries (accessed July 26, 2020).

26 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 (accessed July 20, 2020).

27 CDC. Coronavirus Disease 2019 (COVID-19). Centers for Disease Control and Prevention. 2020; published online Feb 11. http://www.cdc.gov/coronavirus/2019-ncov/covid-data/forecasting-us.html (accessed June 23, 2020).

28 Dicker D, Nguyen G, Abate 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* 2018; **392**: 1684–735.

29 Rivers C, George D. How to Forecast Outbreaks and Pandemics. 2020; published online July 5. https://www.foreignaffairs.com/articles/united-states/2020-06-29/how-forecast-outbreaks-and-pandemics (accessed July 8, 2020).

medRxiv preprint doi: https://doi.org/10.1101/2020.07.13.20151233; this version posted August 26, 2020. 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 4.0 International license .

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	159 Countries*	December 1st
Youyang Gu	https://github.com/youyanggu/covid19_projections	SEIR	73 Countries*	November 1st
MIT - DELPHI	https://github.com/COVIDAnalytics/DELPHI	SEIR	158 Countries*	November 15th
Imperial-LMIC	https://github.com/mrc-ide/global-lmic-reports	SEIR	160 Countries	November 14th
LANL-Growthrate	https://covid-19.bsvgateway.org/	Dynamic Growth	142 Countries*	October 6th
SIKJalpha	https://github.com/scc-usc/ReCOVER-COVID-19	SEIR	177 Countries*	November 29th

Table 1. Models Included in the Study

All eight 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

medRxiv preprint doi: https://doi.org/10.1101/2020.07.13.20151233; this version posted August 26, 2020. 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 4.0 International license .

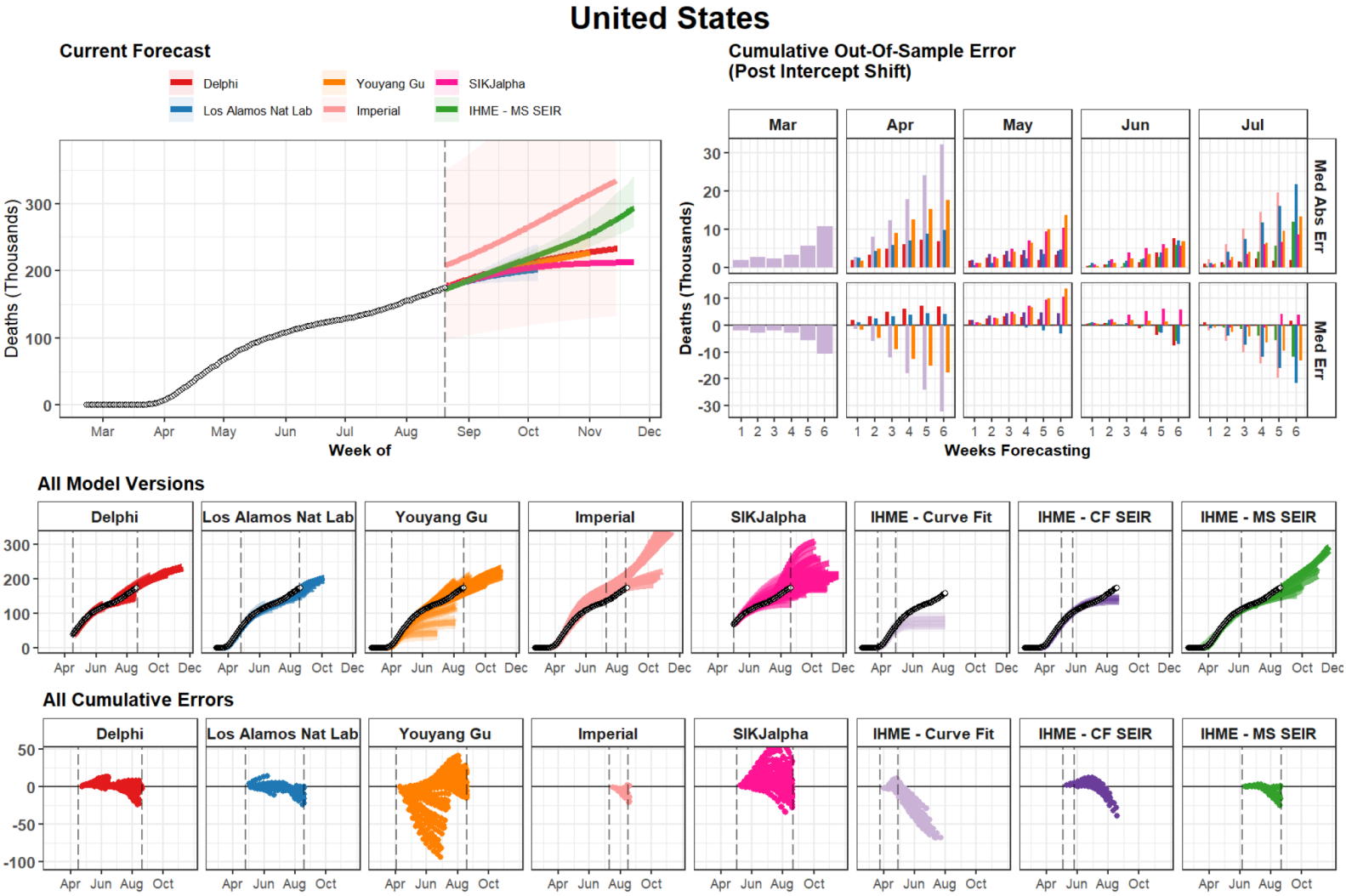

Figure 1. Cumulative Mortality Forecasts and Prediction Errors by Model – Example for United States

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. 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 appendix.

medRxiv preprint doi: https://doi.org/10.1101/2020.07.13.20151233; this version posted August 26, 2020. 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 4.0 International license .

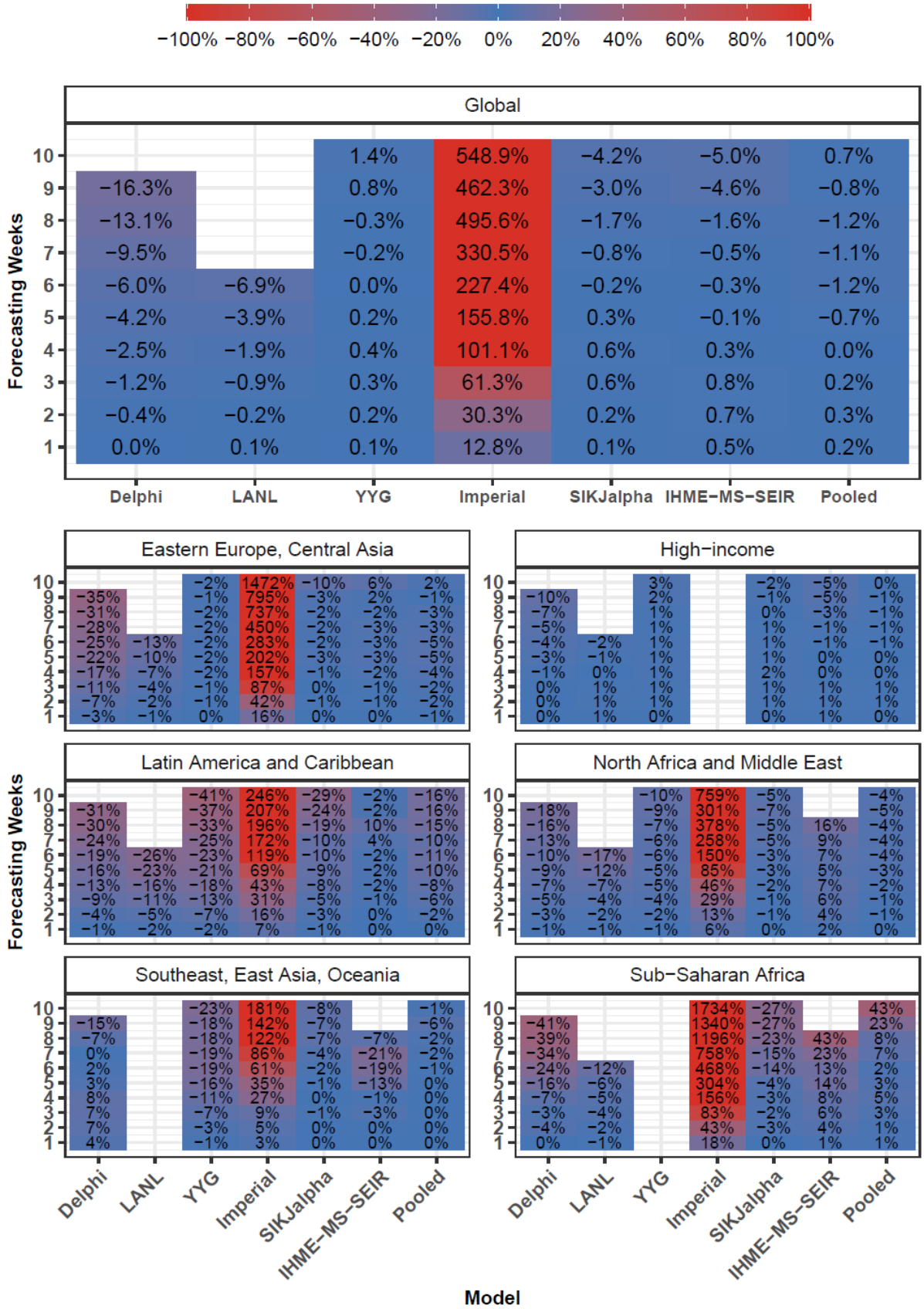

Figure 2. Cumulative Mortality Bias - Median Percent Error

Median percent error values, a measure of bias, were calculated across all observed errors at weekly intervals, for each model, by weeks of forecasting and geographic region. 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. Results are shown here for models released in June, and results from other months are shown in the appendix.

medRxiv preprint doi: https://doi.org/10.1101/2020.07.13.20151233; this version posted August 26, 2020. 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 4.0 International license .

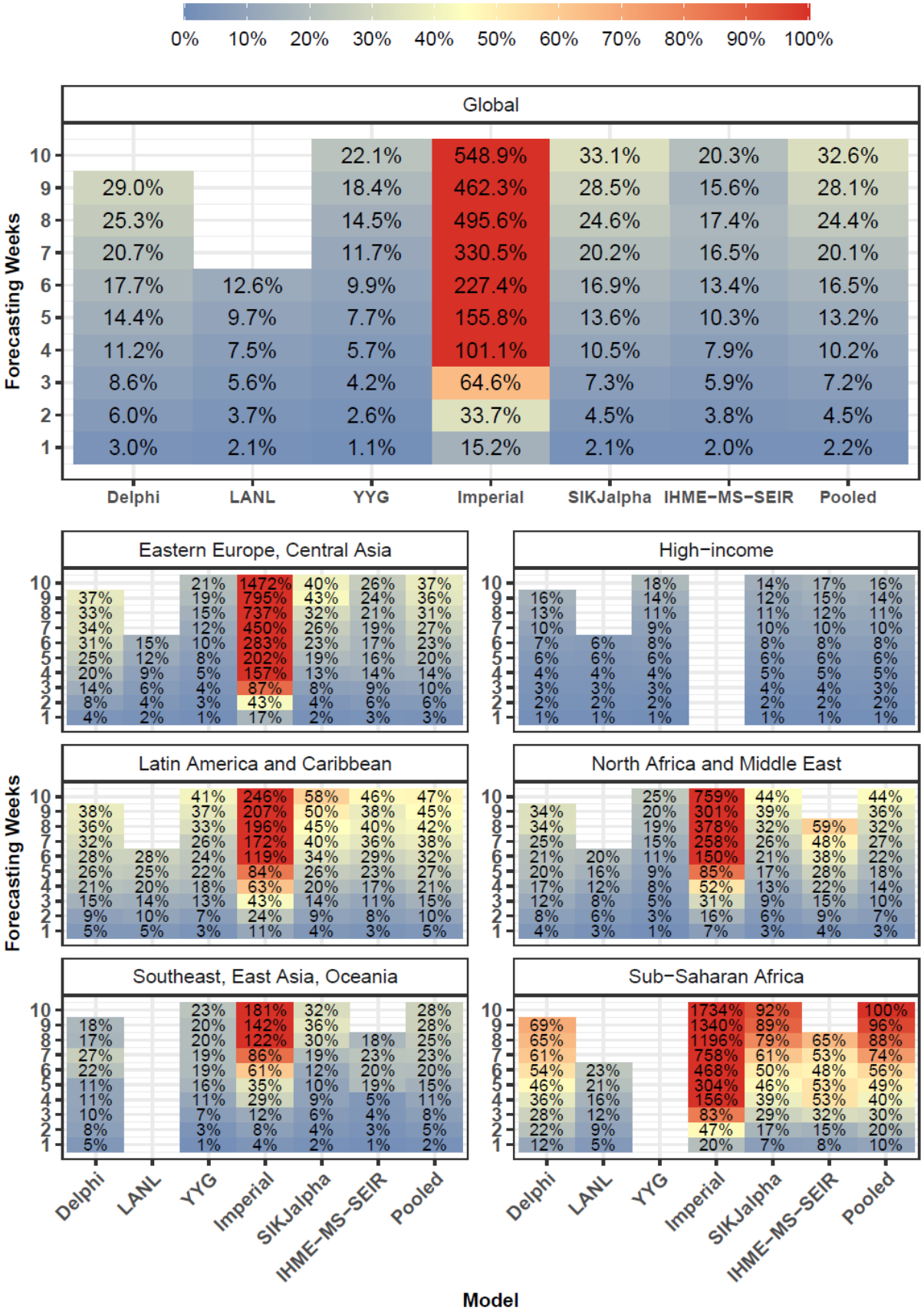

Figure 3. Cumulative Mortality Accuracy – Median Absolute Percent Error

Median absolute percent error values, a measure of accuracy. were calculated across all observed errors at weekly intervals, for each model by weeks of forecasting and geographic region. 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. Results are shown here for models released in June, and results from other months are shown in the supplement.

medRxiv preprint doi: https://doi.org/10.1101/2020.07.13.20151233; this version posted August 26, 2020. 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 4.0 International license .

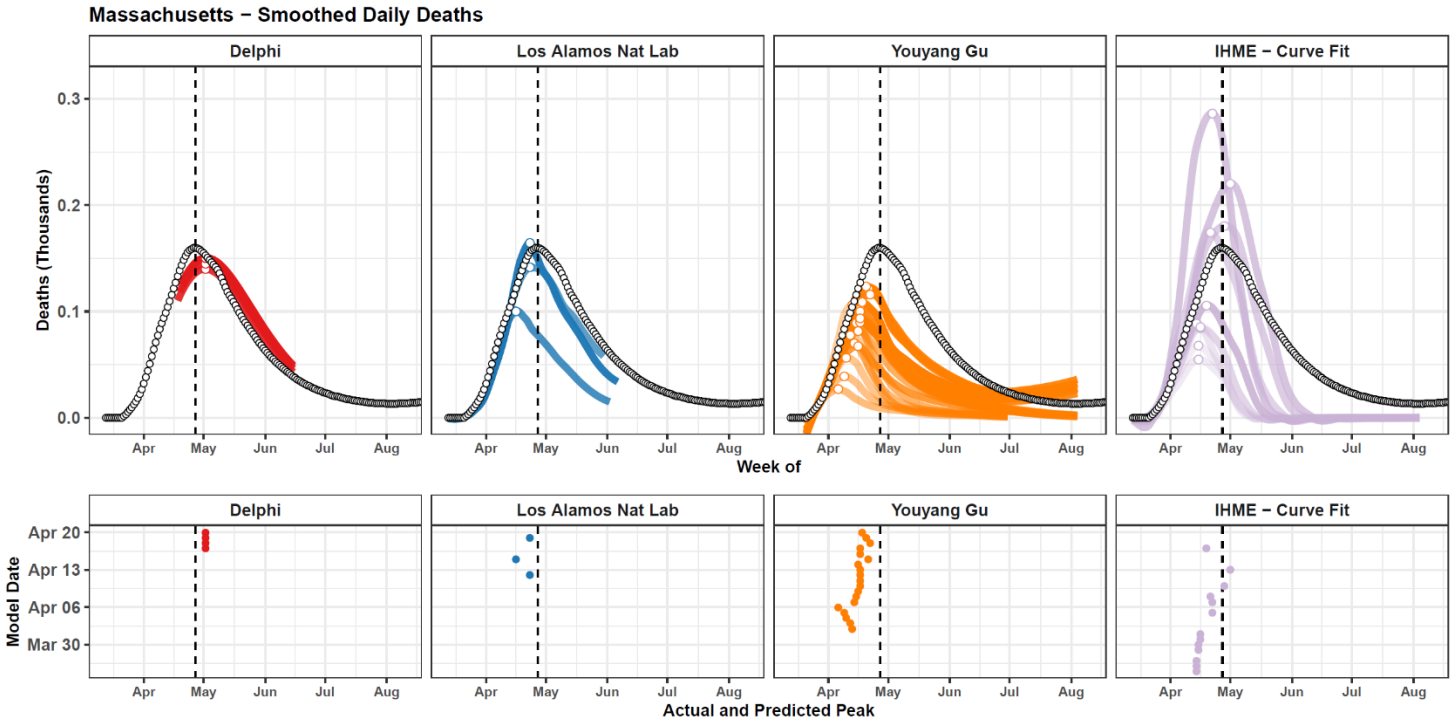

Figure 4. Observed vs Predicted Peak in Daily Deaths– Example for Massachusetts

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 appendix.

medRxiv preprint doi: https://doi.org/10.1101/2020.07.13.20151233; this version posted August 26, 2020. 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 4.0 International license .

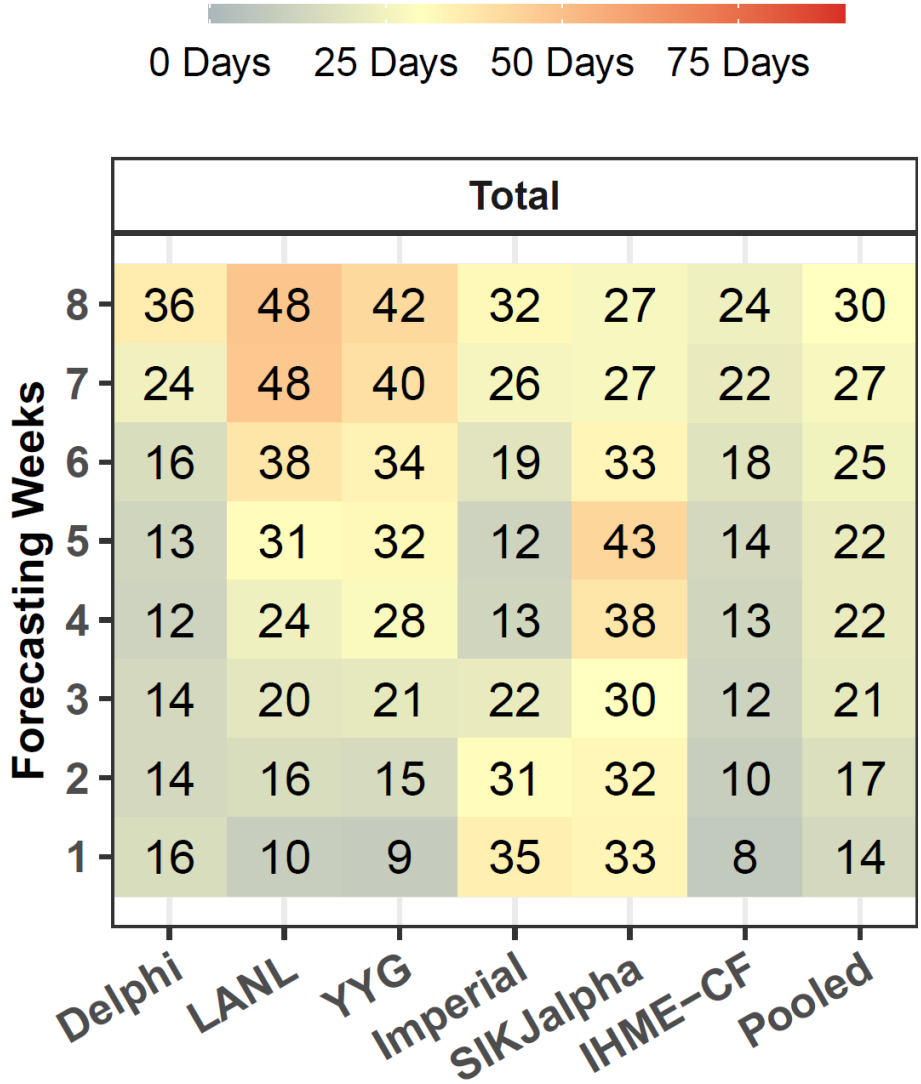

Figure 5. Peak Timing Accuracy – Median Absolute Error in Days

Median absolute error in days is shown by model and number of weeks of forecasting. Models that are not available for at least 25 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. Errors are pooled across month of estimation, as we found little evidence of change in peak timing performance by month (see appendix).