An Assessment of the IHME Covid-19 Model: US Fatalities in 2020

Thomas R. Robbins
East Carolina University

The Covid-19 forecasting model published by the Institute for Health Metrics and Evaluation (IHME) is one of the most influential, consequential, and controversial forecasting models ever published. The model has been widely cited by policy makers, but it has also been widely criticized. In this paper we make an objective, external review of the model. We evaluate the outputs of the model, how they changed, and how they compared to actual results. The model was subject to frequent adjustment that resulted in wide swings back and forth. The errors associated with the model were high and actual results fell inside the model's reported 95% confidence interval far less than 95% of the time. Overall, we find the accuracy of the model was poor, and the model's predictions were unrealistically precise.

Keywords: COVID-19, forecasting, forecast accuracy, health metrics

INTRODUCTION

The Covid-19 forecasting model published by the Institute for Health Metrics and Evaluation (IHME) is one of the most influential, consequential, and controversial forecasting models ever published. The IHME model is the model most widely cited by US government policy makers. It has greatly influenced policy makers at both the federal and state levels. The projections coming out of the model have been used to help justify drastic interventions that have shut down vast sections of the US economy and led to an unprecedented economic contraction that led to a record rise in unemployment. While the IHME model is widely cited by policy makers, it has also been widely criticized from all sides.

In this paper we conduct an external review of the model; external in the sense that we do not concern ourselves directly with the methodology or algorithm used be the model, only with the output of the model. We effectively take the point of view of a policy maker and treat the model as a black box. Our focus is on evaluating the outputs of the model, how they varied over time, and how they compared to actual reported data. While the model forecasts multiple quantities, including hospitalizations, ICU usage, and ventilator usage, we focus solely on forecasted fatalities. We also limit the analysis to the United States, examining both national and state level forecasts, focusing on the 10 states with the highest death totals.

In the remainder of this paper we review the IHME and the model they have put forward including the key criticisms of the model, we then briefly review related literature, and summarize the sources for the data used in this analysis. We then focus on our evaluation of the model, including:

- Evolution of the Model: we analyze and critique the series of rapid and often significant changes in the model's forecasts.

- Daily Deaths: in this section we analyze day by day forecasts of fatalities. We compare the forecasted fatality count with the actual results recorded. We focus on the predictive accuracy of each model iteration over a 14-day forecasting window.
- Cumulative Deaths: we evaluate the overall fatality forecast on specific target dates. We compare how accurate each iteration of the model was in predicting cumulative fatalities at the national and state level.

We analyze the model in three separate time buckets; roughly corresponding to the three waves of the pandemic.

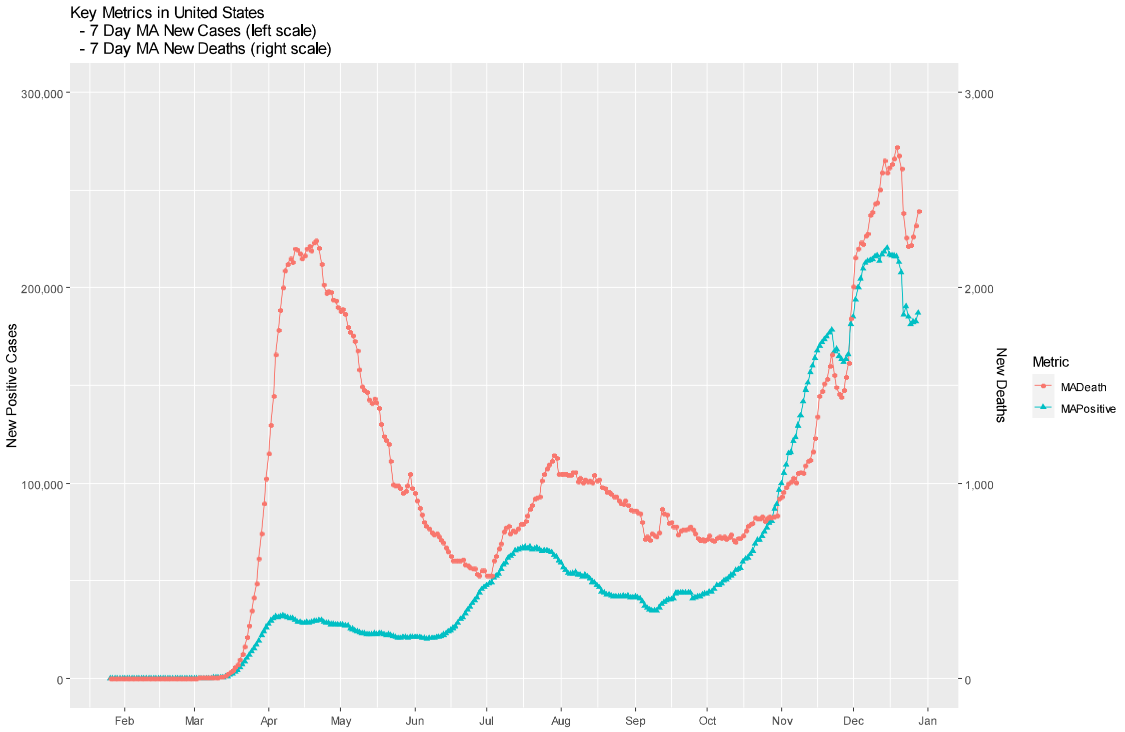
- Wave 1: the initial outbreak of the pandemic in the US. In this section we analyze models released between 3-25-2020 and 6-10-20, and actual fatalities through the mid-June. During this period the virus spread rapidly, fatalities rose to above 2,000 a day then dropped steadily to about 500 per day.
- Wave 2: the second wave of the pandemic that occurred over the summer. In this section we analyze models released between 6-10-2020 and 9-24-2020, and actual fatalities through the end of September. During this period cases rose dramatically, and death rose to a peak somewhat lower than the first peak. Death rose above 1,000 per day before dropping slowly to the 750 range in the early fall.
- Wave 3: the third wave of the pandemic that began in the Fall and was on-going at the end of the year. In this section we analyze models released between 9-24-2020 and 12-23-20, and actual fatalities through the end of 2020. During this period fatalities rose sharply reaching a level of over 2,500 per day. At the end of 2020 it is unclear if the daily death count has peaked yet.

THE COVID-19 PANDEMIC IN THE US

Covid-19 is a highly contagious respiratory disease caused by the SARS-CoV-2 virus. SARS-CoV-2 is thought to spread from person to person through droplets released when an infected person coughs, sneezes, or talks (CDC, 2020). Most people with COVID-19 recover without needing special treatment, but other people are at higher risk of serious illness or death. Covid-19 is believed to have originated in Wuhan, China and has subsequentially spread worldwide. The first confirmed case in the United States was reported on January 20th, 2020 (Holshue et al., 2020). The first reported death attributed to Covid-19 occurred in Seattle, Washington on Feb 28, 2020, although subsequent investigation indicates Covid fatalities likely occurred in early February (CIDRAP, 2020).

Throughout 2020 Covid-19 has spread throughout the world and throughout the United States. By the end of the year there have been more than 19 million confirmed cases and 346,421 deaths. Figure 1 shows the 7-day moving average for new cases and deaths throughout 2020 based on data reported by Johns Hopkins.

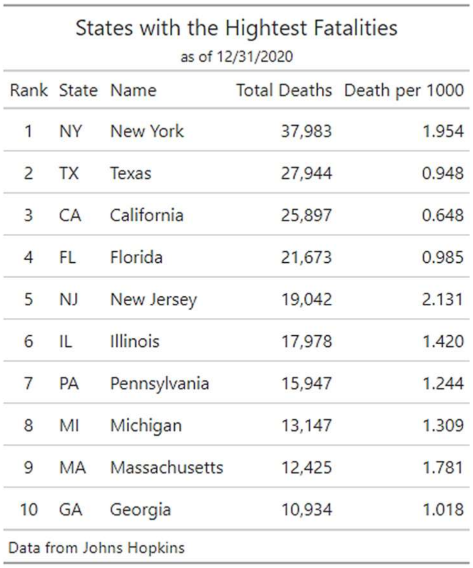
FIGURE 1
KEY METRICS IN THE UNITED STATES



Cases numbers are likely significantly under-reported, especially in the early stages of the pandemic, due to limited testing capabilities. Fatalities statistics are likely more accurate, though they are also subject to uncertainty (Robbins, 2020b). The pandemic appears to have spread through 3 successive waves. The first wave of the pandemic began with the rapid spread of the pandemic during March. Fatalities peaked in April at more than 2,000 per day before declining throughout May and June. A second, smaller, wave occurred over the summer. Confirmed cases began to grow rapidly in mid-June and deaths began to increase in early July, growing from about 500 a day to more than 1,000 per day in late July and early August, before a slow decline. The death rate was relatively flat during most of September and October. A third wave appears to have started in the fall with confirmed cases increasing sharply starting in October. Deaths increased sharply during November and by early December the death rate was higher than the peak of wave 1. The daily death rate was in excess of 2,500 per day through much of December. Wave 3 deaths may have peaked in late December, though as of the end of December it is unclear if this is a true peak, a temporary decline, or a reporting anomaly related to the Christmas holiday.

While confirmed cases and deaths have been reported in all 50 states, the impact has been uneven. Table 1 lists the states with the highest fatality counts. Together these states account for 202,971 deaths, 58.6% of total US deaths. Throughout the remainder of this paper we will focus on fatalities in the US as a whole, and in these high fatality states.

TABLE 1
STATE FATALITIES



States with the Hightest Fatalities
as of 12/31/2020

Rank	State	Name	Total Deaths	Death per 1000
1	NY	New York	37,983	1.954
2	TX	Texas	27,944	0.948
3	CA	California	25,897	0.648
4	FL	Florida	21,673	0.985
5	NJ	New Jersey	19,042	2.131
6	IL	Illinois	17,978	1.420
7	PA	Pennsylvania	15,947	1.244
8	MI	Michigan	13,147	1.309
9	MA	Massachusetts	12,425	1.781
10	GA	Georgia	10,934	1.018

Data from Johns Hopkins

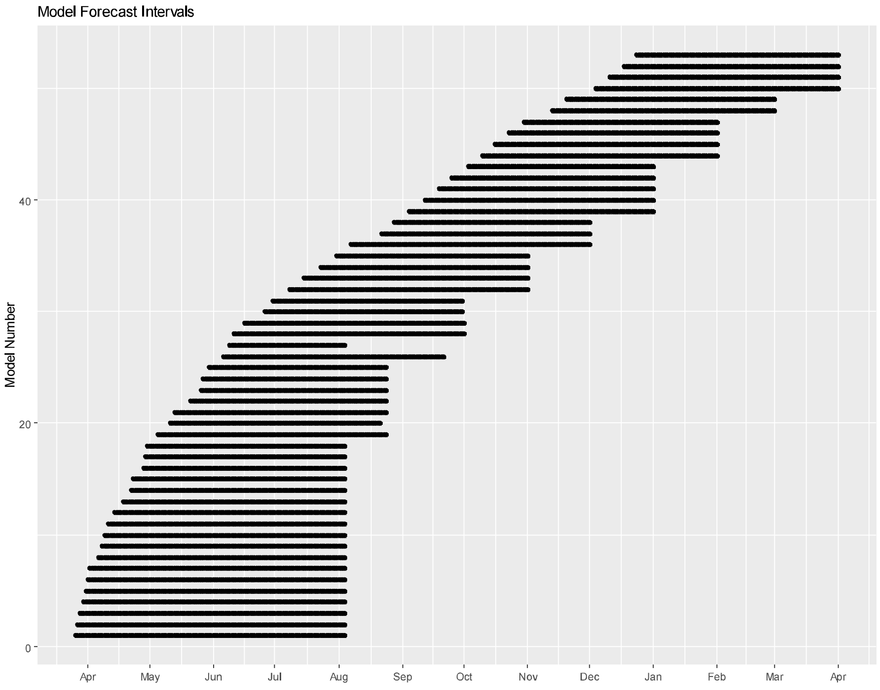
THE IHME AND THE IHME MODEL

The Institution of Health Metrics and Evaluation (IHME) is an independent population health research center associated with the University of Washington (IHME, 2020). The organization gathers information on the world's largest health issues and evaluates the strategies used to address them. IHME’s research stems from three different questions.

1. What are the world's major health problems?
2. How well is society addressing these problems?
3. How do we best dedicate resources to maximize health improvements?

The IHME developed COVID-19 models and projections in response to requests from the University of Washington and other hospitals and government agencies concerned that Covid-19 would overwhelm hospital capacity. Their model forecasts daily and cumulative requirements for key hospital resources, as well as Covid-19 related fatalities. The IHME’s stated goal is to release updates as frequently as possible, updating their model with the latest available information. The IHME provides access to their forecasts via an interactive visualization tool on their website. They also allow the model’s forecasts to be downloaded from their website in a .csv file format. The first set of projections posted on the website is dated March 25, 2020 and by the end of the year 52 different version of the model have been posted. Figure 2 shows the forecast period for each of the 52 models released in 2020.

FIGURE 2
FORECAST INTERVALS



The first iteration of the IHME model provided a four month forecast on the impact of Covid-19 pandemic (Murray, 2020). That first model included forecasts for total hospital beds, ICU beds, ventilators, and deaths. Multiple models were released throughout March and April, with the first 18 versions of the model released by the end of April. Each of these models had a forecast that ended on August 4th. Over time, the release schedule for the model became less frequent and more regular, with the end period of the forecast being extended by a month approximately every 4 weeks.

The original forecast predicted a total of 81,114 deaths with a 95% confidence interval of [38,242 , 162,106]. In that first iteration of the model the pandemic was expected to have been largely over by the end of May and the mean projected death count for that time was 78,320, 96.6% of the total.

The IHME model has been widely cited by US public health officials, “Dr. Deborah Birx, the White House coronavirus response coordinator, has repeatedly cited the IHME model at press conferences, and journalists have often asked officials about it, too.” (Azad, 2020). Dr. Birx again specifically mentioned the IHME model in early April when the model’s forecasted fatalities dropped significantly, from 81,766 to 60,415 a change she justified based on increased social distancing policies (Carvajal, 2020).

Criticism of the IHME Model

The IHME model has been widely criticized for multiple reasons in the mainstream media, on social media, and in the academic literature. Critiques of the model include:

- Overly optimistic projections
- Overly pessimistic projections
- No epidemiologically basis
- Too unstable – large changes in the forecast over short time periods.
- Too uncertain – confidence levels that are too wide for practical use.
- Too short a time horizon – only a four-month time horizon with no allowance for a second wave

- Forecasting zero deaths at the end of the forecast period
- Based on bad data and assumptions

The model has been criticized for being too optimistic, especially after it lowered estimated fatalities in early April and generated forecasts inconsistent with and lower than other models (Wan & Johnson, 2020). The same downward revision caused others to criticize the original model as too pessimistic. Critics pointed out the lack of a national standard model but noted the IHME model was receiving significant support from the White House. Other critics have also identified the disparity across models and the wide range of uncertainty (Boice, 2020).

The model has been criticized because it is not based on a traditional Susceptible-Infectious-Recovered (SIR) model, but rather it is a data fitting model that attempts to extrapolate, at least initially, from Chinese data. Critics charged that the comparison to China was "inherently optimistic because it assumes that all states respond as swiftly as China" (Wan, Dawsey, Parker, & Achenbach, 2020). Dr. Anthony Fauci, the longtime director for the National Institute of Allergy and Infectious Diseases, discounted the use of models in general when he said models are only as good as the assumptions put into them (Hatmaker, 2020). The models are also dependent on the quality of the data they use, especially a curve fitting model like the IHME model. The issues associated with Covid-19 data are significant and are addressed in (Ioannidis, 2020) and reviewed in detail in Robbins (2020b). Perhaps the most significant issue is the under-reporting of actual deaths in China, since the early versions of the model relied so heavily on the Chinese data. A non-academic review in early May posted on Vox.com focused on the optimistic nature of the model relative to actuals at that point as well as its divergence from other models and questioned the rationale for continued reliance on the model by policy makers (Piper, 2020). A different critique, posted a few weeks earlier at issueinsights.com, criticized the model for being too pessimistic, noting the significant downgrades in projections in early April (Fumento, 2020). That author questioned all the models and suggested "It's Time to Permanently Dump Epidemic Models."

Harsh criticism of the model for rapidly changing forecasts is documented in Begley (2020). A quote from that article summarizes the view of some in the traditional epidemiological modeling community, "That the IHME model keeps changing is evidence of its lack of reliability as a predictive tool," said epidemiologist Ruth Etzioni of the Fred Hutchinson Cancer Center, "….That it is being used for policy decisions and its results interpreted wrongly is a travesty unfolding before our eyes."

A paper posted on the non-peer reviewed site Medium.com, but co-authored by Wesley Pedgen of Carnegie Mellon, was highly critical of the Covid-19 models in general, and the IHME model in particular, for the short term forecasting horizon, in particular the assumption of a zero death-period and no second wave (Chikina & Pegden, 2020). This paper argues that the strict social distancing imposed by lockdowns across the United States may flatten the curve but are likely pushing deaths into a second wave outside the forecast window of the models. Another report, written by academics but posted without peer review, criticized the IHME model for the high number of forecasts that fall outside the model's 95% confidence intervals (Marchant, Rosen, Tanner, & Cripps, 2020). This review was dated April 8th, still early in the outbreak. It examined forward predictions of one to four days made at the end of March and found that 49%-73% of the forecasts are outside the confidence level and they conclude the model dramatically underestimates uncertainty. An assessment of the early stages of the IHME model found similar results (Robbins, 2020a).

RELATED LITERATURE

This paper is not an epidemiological paper or even a modelling paper. We do not consider the internal working of the IHME model. Background information on epidemic models can be found in a review format (Chowell, Sattenspiel, Bansal, & Viboud, 2016) and a tutorial format (Dimitrov & Meyers, 2010). Our focus is more on the assessment of forecasting predictions.

Multiple streams of research address issues of forecasting and forecasting accuracy. In the statistical community forecasting competitions are common. These competitions generally involve the evaluation of specific methodologies, rather than individual models. A somewhat dated review of empirical assessment of forecasting models is provided in Fildes and Makridakis (1995). A more recent evaluation of forecasting methods is the M-3 competition, featured in a special issue of *The International Journal of Forecasting* (Ord, Hibon, & Makridakis, 2000). The results and findings of the competition are summarized and reviewed in Makridakis and Hibon (2000). This was the third in a series of forecasting competitions that compare the performance of a large number of time series forecasting methods. Another forecasting competition, this one related specifically to the tourism industry, is chronicled in Athanasopoulos, Hyndman, Song, and Wu (2011).

A different type of forecasting competition is discussed in Tetlock (2017). Tetlock organized multiple forecasting tournaments over an extended period of time that focused on individual forecasts of major world events. He generally found forecasters did little better than randomized predictions and typically did worse than algorithms. The characteristics of good forecasters are further detailed in (Tetlock & Gardner, 2015). However, Tetlock's work is primarily oriented toward individual human forecasters rather than model-based forecasters.

An alternative stream of research examines the predictive ability of group-based prediction methods. In the book *Infotopia*, Cass Sunstein compares the predictive ability of group based methods such as polls and prediction markets (Sunstein, 2006). Sunstein makes the case that predictive markets can, under the right circumstances, outperform other predictive methods. Other research examines market-based forecasts, such as the prediction of commodity prices by futures markets (Alquist & Kilian, 2010; Chernenko, Schwarz, & Wright, 2004; Chinn & Coibion, 2009). This research tends to find that futures markets are subject to significant errors but are more accurate than other model-based approaches.

We were unable to find any peer reviewed detailed evaluations of specific models in general, or of the Covid-19 models in particular. Only the general press and non-peer reviewed papers identified previously. But a peer-reviewed paper in JAMA does outline the general principle and value of projections for Covid-19 models (Jewell, Lewnard, & Jewell, 2020). The paper argues that epidemiological models are best suited to evaluate "the relative effect of various interventions in reducing disease burden rather than to produce precise quantitative predictions about extent or duration of disease burdens." The paper does identify, and criticize, specific aspects of the IHME model. The paper states "[IHME model] has received considerable attention and has been widely quoted by government officials. On the surface, the model yields specific predictions of the day on which COVID-19 deaths will peak in each state and the cumulative number of deaths expected over the next 4 months (with substantial uncertainty intervals). However, caveats in these projections may not be widely appreciated by the public or policy makers because the model has some important but opaque limitations." It goes on to criticize the IHME model for the volatility of fatality forecasts, citing the large change in NY projections at the beginning of April.

This paper provides an important contribution to the literature by providing a detailed, and fact-based assessment of the IHME model. Rather than pointing out individual flaws, out of context, we provide an objective assessment of the model, looking at all the different iterations provided, and comparing those projections to actual reported results. Any model as impactful as the IHME Covid-19 deserves careful scrutiny and evaluation so as to provide both modelers and policy makers with insights that can be applied in future forecasting situations.

DATA AND DATA SOURCES

Our analysis in this paper relies on two primary data sources. We utilize data on the IHME model projections along with actual data reported concerning the impacts of the pandemic. IHME is highly transparent and provides data on their model on their website (IHME, 2020). Each iteration of the model is provided as a separate downloadable .csv file, and release notes are posted for each version. Our data set includes every model posted by the IHME from March 25th to December 23rd. Each model is identified with the date identified on the Covid-19 estimate download page. The data provided by the model has evolved along with the model, and while initial versions of the model focused only on the United States, later versions added international data. For our purposes we use only data for the US. The IHME projections include data for each state as well as the US as a whole. Later iterations of the model added data for some

US territories such as Puerto Rico, but we restrict our analysis to the 50 states and the District of Columbia. Early versions of the model provided projections through August 4th. Beginning with the May 4th iteration of the model the projections for some metrics extend to August 24th, though the forecasts for US deaths only extended through August 4th. While additional data has been added over time, for this analysis we focus on the following data: deaths: the average number of deaths projected by day, by location, and totaldea: the cumulative number of deaths. Each forecast parameter has a mean, and 95% confidence intervals for each location for each day. Additional measures we do not analyze in this paper include projected overages for beds and ICUs. We compare these models to actual data, and we use data sources for actuals consistent with the data used by the IHME modelers.

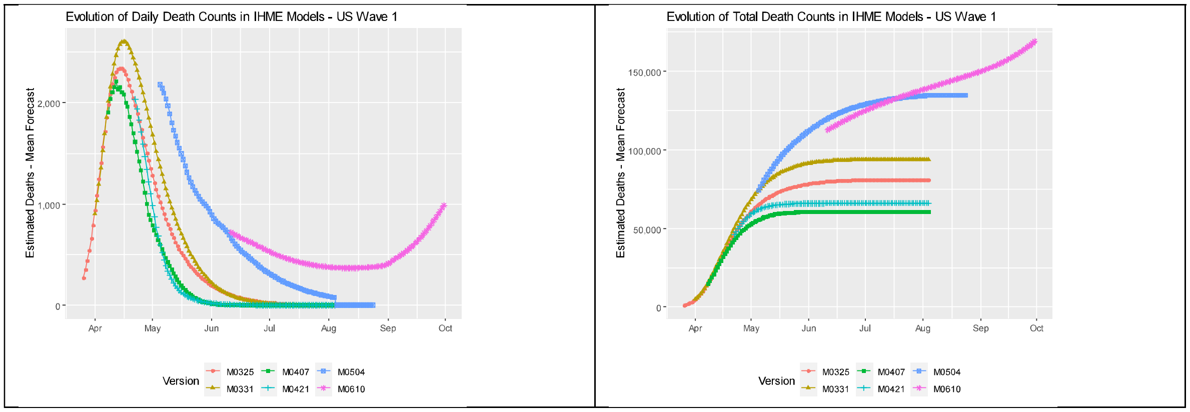
Our second major source of data is the Johns Hopkins Coronavirus Resource Center (Johns_Hopkins, 2020). The resource center provides a number of different analysis and visualization tools. For our purposes we focus on the time series data they publish on Github (GitHub, 2020). The downloadable data sets include time series data for US counties. For this analysis we use only the deaths file, which documents total deaths. We use this data and calculate new deaths by a simple difference operation. We also aggregate this data to the state and national level. The site provides additional data on recoveries, as well as international cases that we do not include in this analysis.

WAVE 1 MODELS

In this section we will consider the models released through early June, specifically June 10th, as models corresponding to what we can now consider to be the first wave of the pandemic. During this 77-day period a total of 28 models were released, roughly one release every 2.75 days. As can be seen in Figure 2, the releases were frequent and irregular. In fact, between 3-25 and 4-1, a period of 8 days, 7 different versions of the model were released.

Figure 3 shows the daily and total death projections for a subset of the models released between late May and early June. These early versions of the model all made projections through the beginning of August 2020. The first 26 versions of the model issued between 3-25 and 6-5, all had the death rate dropping to 0 by the end of the forecast period, implicitly predicting a single wave that would end by mid-summer. It was not until the 27th iteration of the model, released on 6-8, that deaths were forecasted to continue beyond August 4th. So, it was only after roughly 2 ½ months and 27 version of the model, that the predictions showed the pandemic not ending in the US by early summer.

FIGURE 3
WAVE 1 FORECASTS



The total forecasted fatalities varied significantly in these wave-1 models, and not in a single direction. The first iteration of the model called for 81.1K deaths. That rose as high as 93.8k in the 3-31 version, before dropping significantly to 60.4k in the 4-7 version. By the 5-4 version the estimate had risen to 134.5K. By the 6-10 version the total death estimate had risen to 169.9K and was projected to continue. Estimates at the state level also varied considerably as we will investigate later.

Model Accuracy – Daily Deaths

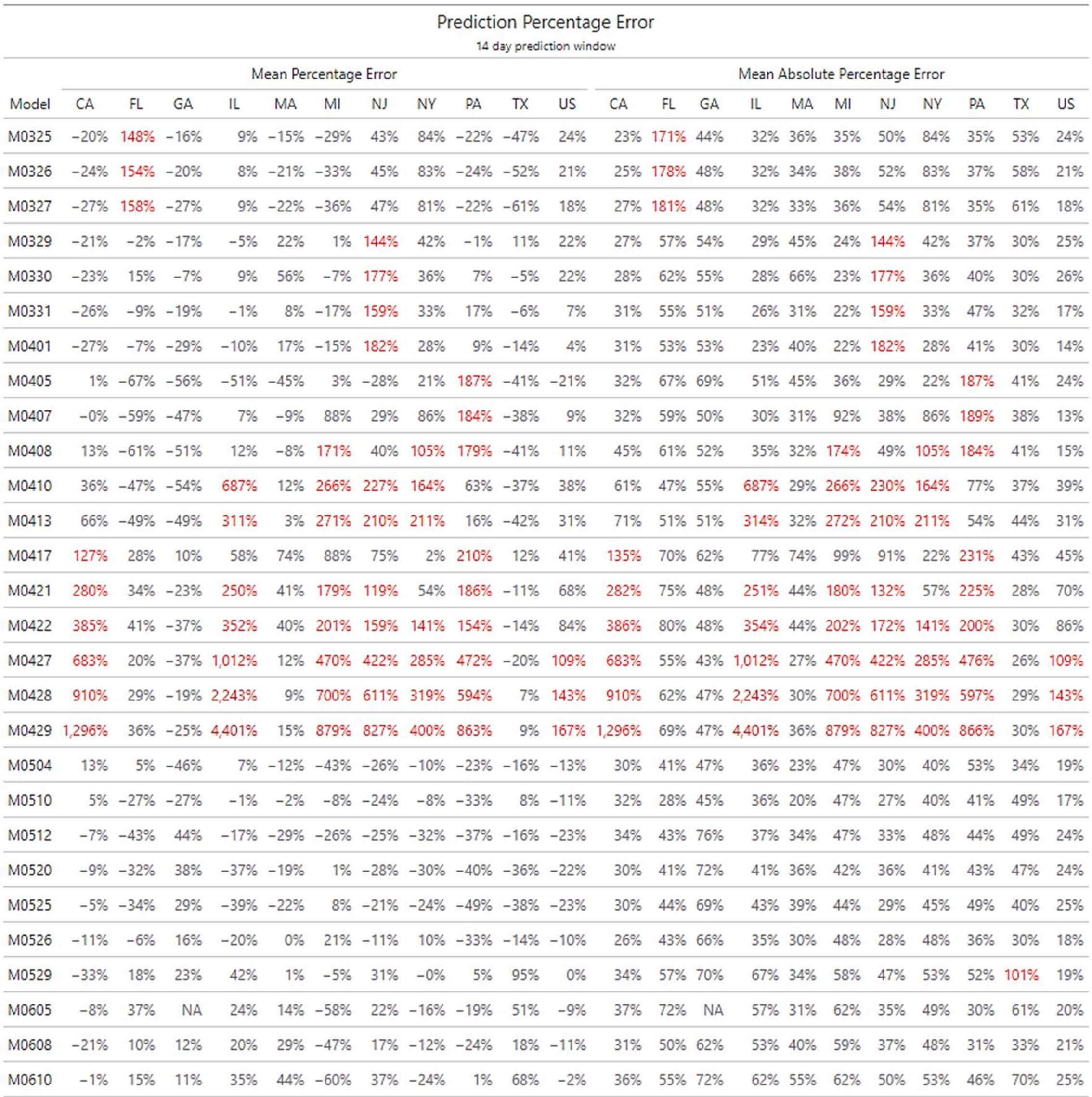
We now consider the accuracy of the model forecasts for daily fatalities. We will evaluate the forecasts for each model version of a 14-day forecast window and evaluate the accuracy of the model for each of the top-10 fatality states listed in Table 1. We will examine the forecasts along several dimensions. First, we consider the Mean Percentage Error (MPE), and the Mean Absolute Percentage Error (MAPE), defined as

$$MPE = \frac{1}{n}\sum_{i=1}^{n}\frac{A_i - F}{A_i} \tag{1}$$

$$MAPE = \frac{1}{n}\sum_{i=1}^{n}\frac{|A_i - F_i|}{A_i} \tag{2}$$

The MAPE is a widely used measure of predictive accuracy that gives an overall assessment of the magnitude of errors. We use a percentage error to account for the wide range of values that occur between states and over the course of the pandemic. We also evaluate the MPE as it gives us an indication of a bias in the forecasts. A positive MPE indicates forecasts that are biased low (more deaths than forecasted), while a negative MPE indicates forecasts that are biased high (less deaths than forecasted). A disadvantage of the percentage metrics is they are only defined for observations where the actual metric is non-zero which is an issue for some of our measures that we will address on a case-by-case basis. Table 2 shows the MPE and MAPE for the wave 1 models for the US as a whole and the 10 highest fatality states. When calculating these metrics, we excluded any data points were the actual number of deaths was zero. In wave 1 this only effects reports from MA from 4-16 to 4-20.

TABLE 2
WAVE 1 – PREDICTION ERRORS



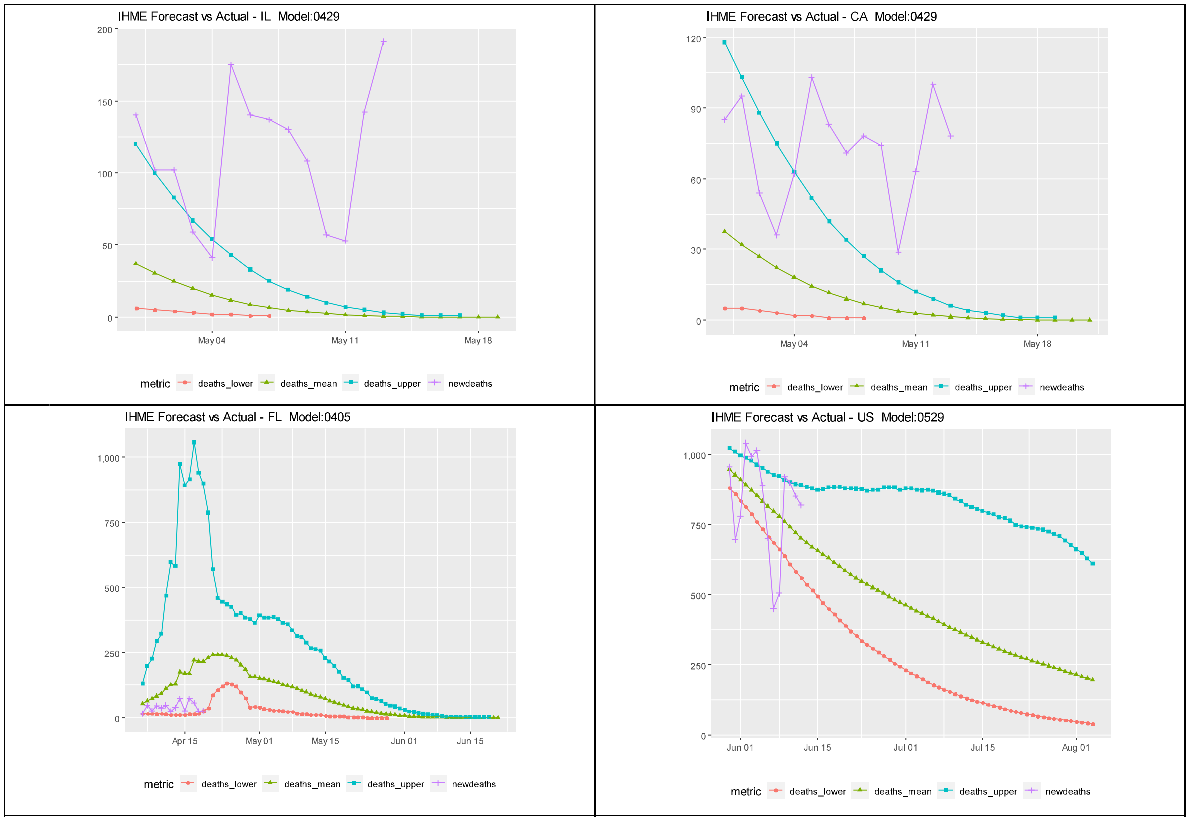
Prediction Percentage Error

14 day prediction window

	Mean Percentage Error											Mean Absolute Percentage Error										
Model	CA	FL	GA	IL	MA	MI	NJ	NY	PA	TX	US	CA	FL	GA	IL	MA	MI	NJ	NY	PA	TX	US
M0325	−20%	148%	−16%	9%	−15%	−29%	43%	84%	−22%	−47%	24%	23%	171%	44%	32%	36%	35%	50%	84%	35%	53%	24%
M0326	−24%	154%	−20%	8%	−21%	−33%	45%	83%	−24%	−52%	21%	25%	178%	48%	32%	34%	38%	52%	83%	37%	58%	21%
M0327	−27%	158%	−27%	9%	−22%	−36%	47%	81%	−22%	−61%	18%	27%	181%	48%	32%	33%	36%	54%	81%	35%	61%	18%
M0329	−21%	−2%	−17%	−5%	22%	1%	144%	42%	−1%	11%	22%	27%	57%	54%	29%	45%	24%	144%	42%	37%	30%	25%
M0330	−23%	15%	−7%	9%	56%	−7%	177%	36%	7%	−5%	22%	28%	62%	55%	28%	66%	23%	177%	36%	40%	30%	26%
M0331	−26%	−9%	−19%	−1%	8%	−17%	159%	33%	17%	−6%	7%	31%	55%	51%	26%	31%	22%	159%	33%	47%	32%	17%
M0401	−27%	−7%	−29%	−10%	17%	−15%	182%	28%	9%	−14%	4%	31%	53%	53%	23%	40%	22%	182%	28%	41%	30%	14%
M0405	1%	−67%	−56%	−51%	−45%	3%	−28%	21%	187%	−41%	−21%	32%	67%	69%	51%	45%	36%	29%	22%	187%	41%	24%
M0407	−0%	−59%	−47%	7%	−9%	88%	29%	86%	184%	−38%	9%	32%	59%	50%	30%	31%	92%	38%	86%	189%	38%	13%
M0408	13%	−61%	−51%	12%	−8%	171%	40%	105%	179%	−41%	11%	45%	61%	52%	35%	32%	174%	49%	105%	184%	41%	15%
M0410	36%	−47%	−54%	687%	12%	266%	227%	164%	63%	−37%	38%	61%	47%	55%	687%	29%	266%	230%	164%	77%	37%	39%
M0413	66%	−49%	−49%	311%	3%	271%	210%	211%	16%	−42%	31%	71%	51%	51%	314%	32%	272%	210%	211%	54%	44%	31%
M0417	127%	28%	10%	58%	74%	88%	75%	2%	210%	12%	41%	135%	70%	62%	77%	74%	99%	91%	22%	231%	43%	45%
M0421	280%	34%	−23%	250%	41%	179%	119%	54%	186%	−11%	68%	282%	75%	48%	251%	44%	180%	132%	57%	225%	28%	70%
M0422	385%	41%	−37%	352%	40%	201%	159%	141%	154%	−14%	84%	386%	80%	48%	354%	44%	202%	172%	141%	200%	30%	86%
M0427	683%	20%	−37%	1,012%	12%	470%	422%	285%	472%	−20%	109%	683%	55%	43%	1,012%	27%	470%	422%	285%	476%	26%	109%
M0428	910%	29%	−19%	2,243%	9%	700%	611%	319%	594%	7%	143%	910%	62%	47%	2,243%	30%	700%	611%	319%	597%	29%	143%
M0429	1,296%	36%	−25%	4,401%	15%	879%	827%	400%	863%	9%	167%	1,296%	69%	47%	4,401%	36%	879%	827%	400%	866%	30%	167%
M0504	13%	5%	−46%	7%	−12%	−43%	−26%	−10%	−23%	−16%	−13%	30%	41%	47%	36%	23%	47%	30%	40%	53%	34%	19%
M0510	5%	−27%	−27%	−1%	−2%	−8%	−24%	−8%	−33%	8%	−11%	32%	28%	45%	36%	20%	47%	27%	40%	41%	49%	17%
M0512	−7%	−43%	44%	−17%	−29%	−26%	−25%	−32%	−37%	−16%	−23%	34%	43%	76%	37%	34%	47%	33%	48%	44%	49%	24%
M0520	−9%	−32%	38%	−37%	−19%	1%	−28%	−30%	−40%	−36%	−22%	30%	41%	72%	41%	36%	42%	36%	41%	43%	47%	24%
M0525	−5%	−34%	29%	−39%	−22%	8%	−21%	−24%	−49%	−38%	−23%	30%	44%	69%	43%	39%	44%	29%	45%	49%	40%	25%
M0526	−11%	−6%	16%	−20%	0%	21%	−11%	10%	−33%	−14%	−10%	26%	43%	66%	35%	30%	48%	28%	48%	36%	30%	18%
M0529	−33%	18%	23%	42%	1%	−5%	31%	−0%	5%	95%	0%	34%	57%	70%	67%	34%	58%	47%	53%	52%	101%	19%
M0605	−8%	37%	NA	24%	14%	−58%	22%	−16%	−19%	51%	−9%	37%	72%	NA	57%	31%	62%	35%	49%	30%	61%	20%
M0608	−21%	10%	12%	20%	29%	−47%	17%	−12%	−24%	18%	−11%	31%	50%	62%	53%	40%	59%	37%	48%	31%	33%	21%
M0610	−1%	15%	11%	35%	44%	−60%	37%	−24%	1%	68%	−2%	36%	55%	72%	62%	55%	62%	50%	53%	46%	70%	25%

There is a lot of information in this table but one of the key observations is that many of the error rates are quite high, and most of the large errors are positive, indicating an actual fatality level higher than the forecasted level. For most of these models the MAPE is equal to the MPE indicating the model was biased so that every forecast was below the actual. The models released in mid to late April are the most error prone. Recall from Figure 3 that this the time when death rates were forecasted to be in sharp decline. These under forecasts led to the sharp upward revision in early May. Errors as high as 1,296% are reported. Of the 308 instances, 55 (17.9%) have a MAPE greater than 100%. To further explore this phenomenon, we look at several individual instances in Figure 4.

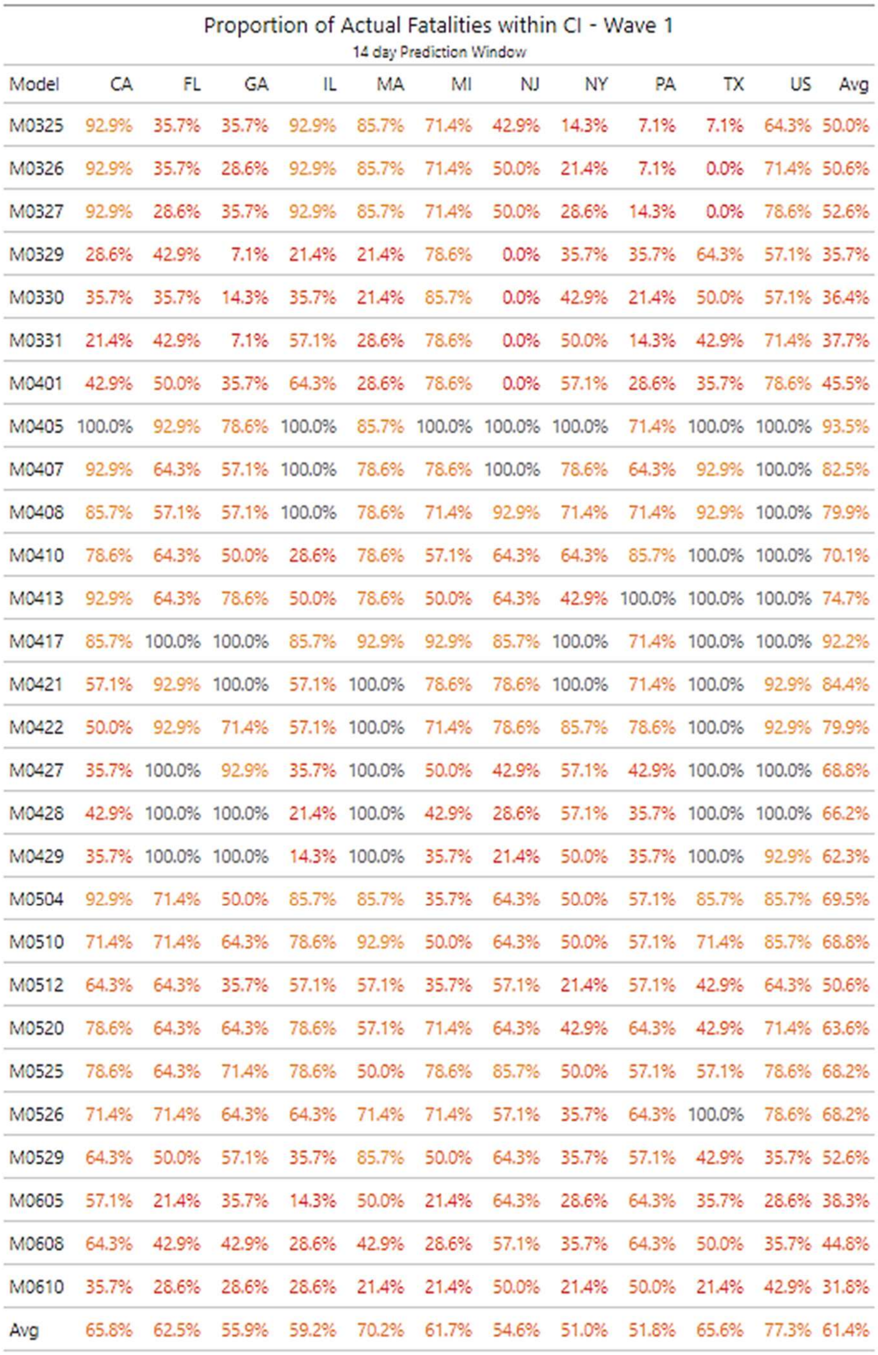
FIGURE 4
WAVE 1 – FORECAST VS ACTUAL



The two upper panes show IL and CA for the 4-29 model. In these two states the forecast called for deaths to drop off toward zero, while in actuality the death rates remained high; well above the mean and mostly outside the 95% confidence interval. The lower left panel presents a different condition altogether, the state of FL in early April. Here a significant spike in deaths was forecasted to occur as it had in other states, while in actuality deaths remained very low. The lower right panel shows the aggregate results for the US as whole from a model in late May. Interestingly this model has an MPE of 0%, but a MAPE of 19%. The actual results varied widely from day to day, sometimes over and other times under the forecast. While it could be argued that this fluctuation represents noise in the data, an important observation is that of the 14 daily forecasts 8 of them, or 57% are outside the 95% confidence level associated with the forecast. Stated differently, only 43% of the forecasts were inside the region that we would expect on average to contain 95% of the outcomes. We examine this issue in more detail in Table 3.

The data in Table 3 summarizes the proportion of the 14 days of forecasts were inside the reported 95% confidence interval for the forecast. Forecasts were the count is below 95% are colored. Of the 308 cells in Table 3, only 42 (13.6%) have all 14 days within the 95% interval. This clearly implies that the model's confidence intervals are too narrow. Or, stated differently, the model was overly precise with its forecasts.

TABLE 3
WAVE 1 – 14 DAY CONFIDENCE INTERVALS



Proportion of Actual Fatalities within CI - Wave 1

14 day Prediction Window

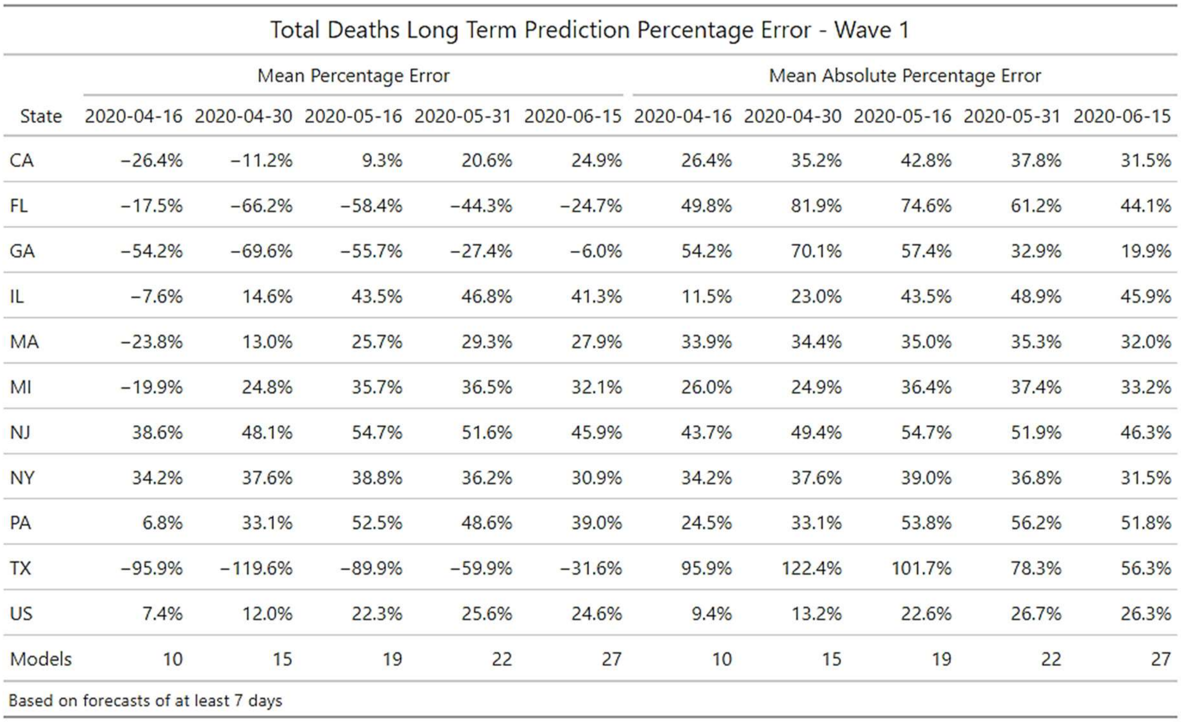
Model	CA	FL	GA	IL	MA	MI	NJ	NY	PA	TX	US	Avg
M0325	92.9%	35.7%	35.7%	92.9%	85.7%	71.4%	42.9%	14.3%	7.1%	7.1%	64.3%	50.0%
M0326	92.9%	35.7%	28.6%	92.9%	85.7%	71.4%	50.0%	21.4%	7.1%	0.0%	71.4%	50.6%
M0327	92.9%	28.6%	35.7%	92.9%	85.7%	71.4%	50.0%	28.6%	14.3%	0.0%	78.6%	52.6%
M0329	28.6%	42.9%	7.1%	21.4%	21.4%	78.6%	0.0%	35.7%	35.7%	64.3%	57.1%	35.7%
M0330	35.7%	35.7%	14.3%	35.7%	21.4%	85.7%	0.0%	42.9%	21.4%	50.0%	57.1%	36.4%
M0331	21.4%	42.9%	7.1%	57.1%	28.6%	78.6%	0.0%	50.0%	14.3%	42.9%	71.4%	37.7%
M0401	42.9%	50.0%	35.7%	64.3%	28.6%	78.6%	0.0%	57.1%	28.6%	35.7%	78.6%	45.5%
M0405	100.0%	92.9%	78.6%	100.0%	85.7%	100.0%	100.0%	100.0%	71.4%	100.0%	100.0%	93.5%
M0407	92.9%	64.3%	57.1%	100.0%	78.6%	78.6%	100.0%	78.6%	64.3%	92.9%	100.0%	82.5%
M0408	85.7%	57.1%	57.1%	100.0%	78.6%	71.4%	92.9%	71.4%	71.4%	92.9%	100.0%	79.9%
M0410	78.6%	64.3%	50.0%	28.6%	78.6%	57.1%	64.3%	64.3%	85.7%	100.0%	100.0%	70.1%
M0413	92.9%	64.3%	78.6%	50.0%	78.6%	50.0%	64.3%	42.9%	100.0%	100.0%	100.0%	74.7%
M0417	85.7%	100.0%	100.0%	85.7%	92.9%	92.9%	85.7%	100.0%	71.4%	100.0%	100.0%	92.2%
M0421	57.1%	92.9%	100.0%	57.1%	100.0%	78.6%	78.6%	100.0%	71.4%	100.0%	92.9%	84.4%
M0422	50.0%	92.9%	71.4%	57.1%	100.0%	71.4%	78.6%	85.7%	78.6%	100.0%	92.9%	79.9%
M0427	35.7%	100.0%	92.9%	35.7%	100.0%	50.0%	42.9%	57.1%	42.9%	100.0%	100.0%	68.8%
M0428	42.9%	100.0%	100.0%	21.4%	100.0%	42.9%	28.6%	57.1%	35.7%	100.0%	100.0%	66.2%
M0429	35.7%	100.0%	100.0%	14.3%	100.0%	35.7%	21.4%	50.0%	35.7%	100.0%	92.9%	62.3%
M0504	92.9%	71.4%	50.0%	85.7%	85.7%	35.7%	64.3%	50.0%	57.1%	85.7%	85.7%	69.5%
M0510	71.4%	71.4%	64.3%	78.6%	92.9%	50.0%	64.3%	50.0%	57.1%	71.4%	85.7%	68.8%
M0512	64.3%	64.3%	35.7%	57.1%	57.1%	35.7%	57.1%	21.4%	57.1%	42.9%	64.3%	50.6%
M0520	78.6%	64.3%	64.3%	78.6%	57.1%	71.4%	64.3%	42.9%	64.3%	42.9%	71.4%	63.6%
M0525	78.6%	64.3%	71.4%	78.6%	50.0%	78.6%	85.7%	50.0%	57.1%	57.1%	78.6%	68.2%
M0526	71.4%	71.4%	64.3%	64.3%	71.4%	71.4%	57.1%	35.7%	64.3%	100.0%	78.6%	68.2%
M0529	64.3%	50.0%	57.1%	35.7%	85.7%	50.0%	64.3%	35.7%	57.1%	42.9%	35.7%	52.6%
M0605	57.1%	21.4%	35.7%	14.3%	50.0%	21.4%	64.3%	28.6%	64.3%	35.7%	28.6%	38.3%
M0608	64.3%	42.9%	42.9%	28.6%	42.9%	28.6%	57.1%	35.7%	64.3%	50.0%	35.7%	44.8%
M0610	35.7%	28.6%	28.6%	28.6%	21.4%	21.4%	50.0%	21.4%	50.0%	21.4%	42.9%	31.8%
Avg	65.8%	62.5%	55.9%	59.2%	70.2%	61.7%	54.6%	51.0%	51.8%	65.6%	77.3%	61.4%

Model Accuracy – Cumulative Deaths

It is possible that some of the problems with day to day results is a result of noise and reporting seasonality in the data. While day to day forecasts are important, we can argue that the aggregate level of deaths is a more important metric and the basis on which we should evaluate the model. So, we now turn to the total death forecast for particular days.

The IHME model was, at least initially, updated very frequently and all models had an end date of 8-4. We can think of each update to the model as an update forecast for the total fatalities that would occur by 8-4, or any other date for that matter. So, to examine the model's accuracy in predicting total fatalities at fixed points in time we construct the data in Table 4 and Table 5. These tables evaluate the models at a series of dates, generally the mid-point and end-point of each month. The number of models that had a prediction for that date is shown at the bottom of the table. The MPE and MAPE represent the mean error of that series of predictions. So, for example, a total of 10 models had predictions for the total deaths that would occur in CA by 4-16 that were made at least 7 days prior to that date. The average error for those predictions was 26.4% and all the errors were negative, indicating that in every case the model over-forecast deaths. In PA the average error was 52.5% for 5-16, while the average absolute error was 53.68%. This indicates most, but not all, of the predictions were under-forecasts.

TABLE 4
WAVE 1 – LONG TERM PREDICTION ERROR



Total Deaths Long Term Prediction Percentage Error - Wave 1

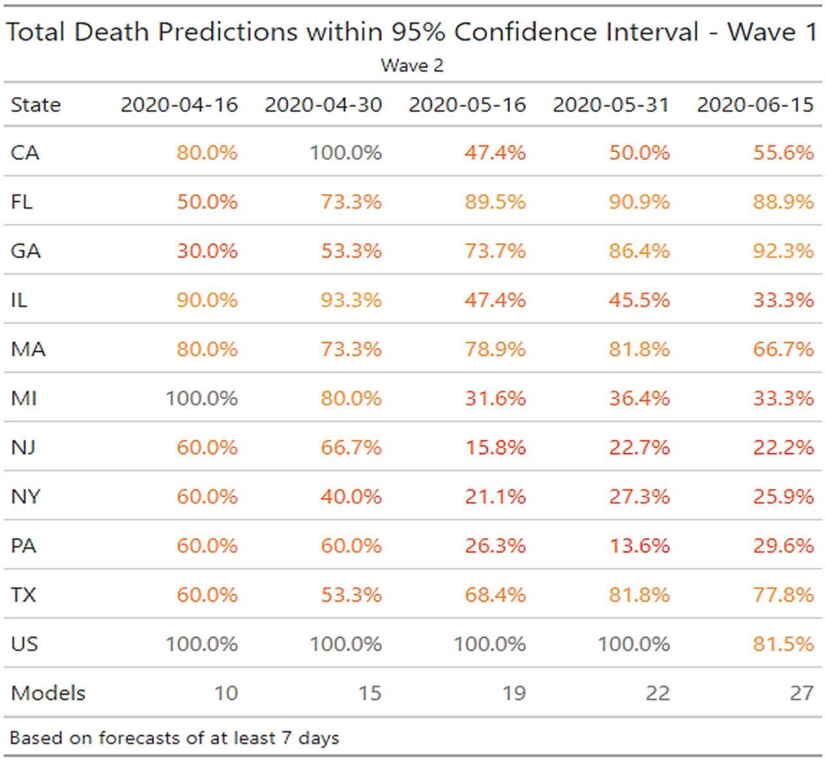
	Mean Percentage Error					Mean Absolute Percentage Error				
State	2020-04-16	2020-04-30	2020-05-16	2020-05-31	2020-06-15	2020-04-16	2020-04-30	2020-05-16	2020-05-31	2020-06-15
CA	−26.4%	−11.2%	9.3%	20.6%	24.9%	26.4%	35.2%	42.8%	37.8%	31.5%
FL	−17.5%	−66.2%	−58.4%	−44.3%	−24.7%	49.8%	81.9%	74.6%	61.2%	44.1%
GA	−54.2%	−69.6%	−55.7%	−27.4%	−6.0%	54.2%	70.1%	57.4%	32.9%	19.9%
IL	−7.6%	14.6%	43.5%	46.8%	41.3%	11.5%	23.0%	43.5%	48.9%	45.9%
MA	−23.8%	13.0%	25.7%	29.3%	27.9%	33.9%	34.4%	35.0%	35.3%	32.0%
MI	−19.9%	24.8%	35.7%	36.5%	32.1%	26.0%	24.9%	36.4%	37.4%	33.2%
NJ	38.6%	48.1%	54.7%	51.6%	45.9%	43.7%	49.4%	54.7%	51.9%	46.3%
NY	34.2%	37.6%	38.8%	36.2%	30.9%	34.2%	37.6%	39.0%	36.8%	31.5%
PA	6.8%	33.1%	52.5%	48.6%	39.0%	24.5%	33.1%	53.8%	56.2%	51.8%
TX	−95.9%	−119.6%	−89.9%	−59.9%	−31.6%	95.9%	122.4%	101.7%	78.3%	56.3%
US	7.4%	12.0%	22.3%	25.6%	24.6%	9.4%	13.2%	22.6%	26.7%	26.3%
Models	10	15	19	22	27	10	15	19	22	27

Based on forecasts of at least 7 days

Most of the state forecasts have sizeable error rates, and the forecasts are often biased in that the MAPE and MPE are nearly equal in magnitude. In NY and NJ, states where the pandemic hit early, deaths exceeding the forecast by more than 30% for each period. Conversely, in states such as TX and GA, where the pandemic hit later, the models significantly over-forecasted the death toll. Over and under forecasts at the state level cancelled out to some degree and the aggregate forecast for the US had lower error rates than individual states. That being said, the error rates for the US as a whole are still quite high, with percentage errors in the 20%-25% range throughout May and June. The majority of the aggregate forecasts under-predicted the total death rate. Given that the Covid-19 pandemic is an unprecedented event it is not unexpected that forecasts would be difficult, especially in the early stage of the pandemic, and the forecasts would be subject to significant uncertainty. The IHME was explicit about that uncertainty and included confidence intervals for all forecasts. Table 5 evaluates how frequently the actual cumulative death count was within the interval at the same dates.

At the aggregate level the forecasts intervals were reasonably accurate with 100% of the predictions for the evaluation dates in April and May including the final total. At the state level the forecasts are much less accurate.

TABLE 5
WAVE 1 – LONG TERM CONFIDENCE INTERVALS



Total Death Predictions within 95% Confidence Interval - Wave 1

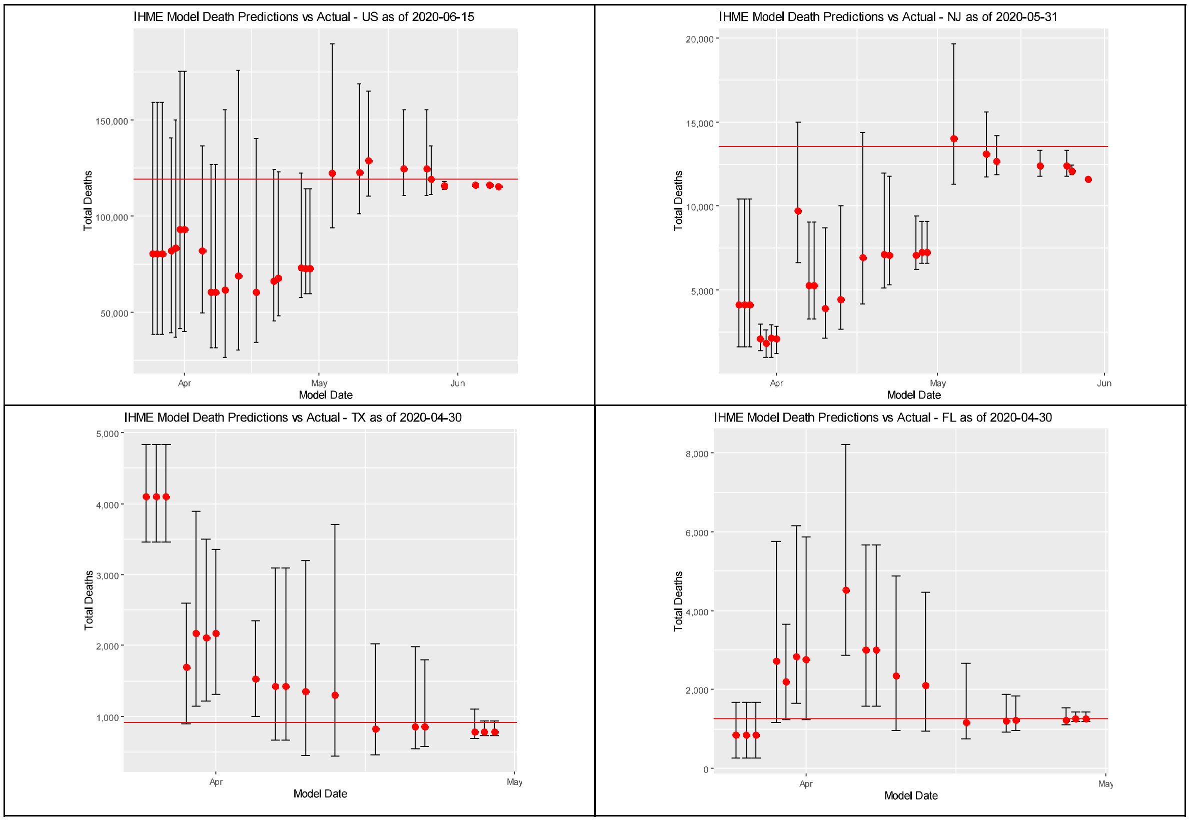
Wave 2

State	2020-04-16	2020-04-30	2020-05-16	2020-05-31	2020-06-15
CA	80.0%	100.0%	47.4%	50.0%	55.6%
FL	50.0%	73.3%	89.5%	90.9%	88.9%
GA	30.0%	53.3%	73.7%	86.4%	92.3%
IL	90.0%	93.3%	47.4%	45.5%	33.3%
MA	80.0%	73.3%	78.9%	81.8%	66.7%
MI	100.0%	80.0%	31.6%	36.4%	33.3%
NJ	60.0%	66.7%	15.8%	22.7%	22.2%
NY	60.0%	40.0%	21.1%	27.3%	25.9%
PA	60.0%	60.0%	26.3%	13.6%	29.6%
TX	60.0%	53.3%	68.4%	81.8%	77.8%
US	100.0%	100.0%	100.0%	100.0%	81.5%
Models	10	15	19	22	27

Based on forecasts of at least 7 days

We illustrate some examples in Figure 5. The upper left panel shows the aggregate results for the US on 6-15, near the end of the first wave. The graph illustrates how the models released through March and April significantly under forecast fatalities, but with wide intervals most of these models contain the final count. Beginning in May the model forecast have shifted up dramatically. The upper right panel is for same time frame but in the state of NJ. The March and April models significantly under-forecasted the death count and most of the CIs do not contain the final level. Even the models released in late May, with relatively short forecast timelines did not have CIs that contain the final level. Perhaps the most striking issue with these forecasts is the low level of uncertainty expressed in many of the forecasts. The CI in 3-27 model the forecast was 4,108 with an interval was [1607, 10,409] but dropped to 2,096 with an interval of [1408 , 2,979] on 3-39 and stayed relatively unchanged for 4 versions of the model. The actual total deaths on 5-31 was 13,518, 6.5 times the 3-29 forecast and 4.5 times the upper limit of the interval. Clearly the model was putting forth forecasts that were not only inaccurate, but unrealistically precise. The lower left panel shows the case of TX for the end of April. Here forecasts start dramatically high, more than 4 times the actuals, then drops over time. Finally, the lower right panel shows that the FL forecasts start out quite accurate before becoming overly pessimistic and then returning to the correct level. Each of these panels illustrate how variable both the point forecasts and intervals are over time.

FIGURE 5
WAVE 1 – SELECT PREDICTION VS ACTUAL



Summary – Wave 1

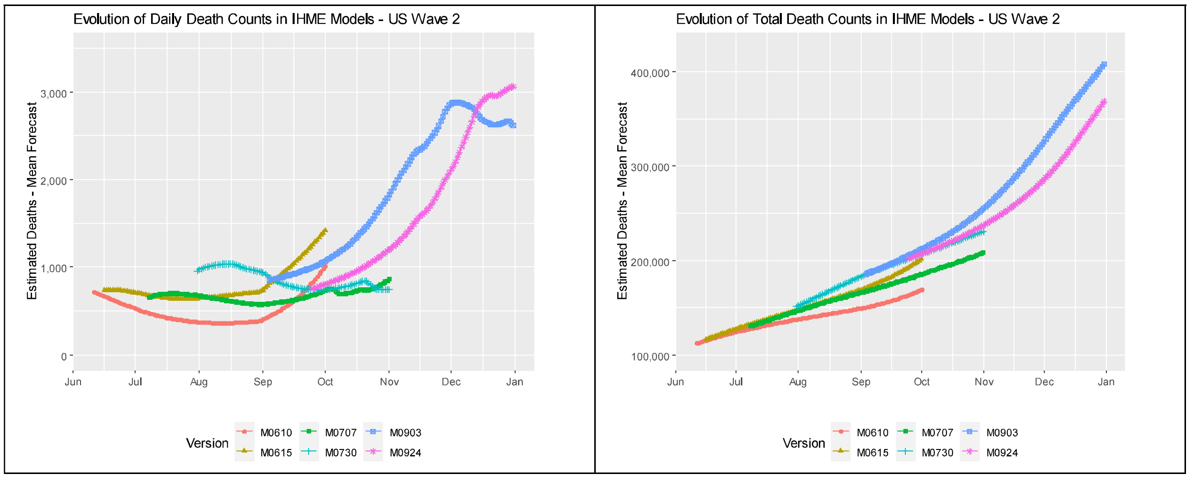
The initial versions of the IHME model were arguably the most consequential. The Covid pandemic was new and the potential impact was uncertain. The model had significant impact on policy makers at the highest levels of the government in the United States. By objective measures the model's forecasts were inaccurate. The model implicitly forecast a pandemic that would end by mid-summer with fatalities below 100 thousand. By mid-June, roughly the end of the first wave, deaths were over 199 thousand and daily deaths stood not at 0, but over 500 per day. It was only then that the model began to incorporate a second wave and an unknown end to the pandemic. The accuracy of the models in these first few months was poor with high error rates and results consistently outside of the stated 95% confidence intervals. The confidence intervals for daily and aggregate deaths were in fact far too narrow, and the model's predictions implied a level of precision that was unwarranted.

WAVE 2 MODELS

In this section we will consider wave 2 of the pandemic. For this analysis we will look at models beginning with the 6-10 model, the last model we examined for wave 1, and ending with the model released on 9-24. We will consider actual deaths that occur up to 9-30. During this period the IHME released 15 versions of the model, one version approximately every 7.2 days. Models were released on a more regular schedule, about once per week, but on some occasions the time between models was as short as 5 days, or as long as 15 days. The end date for these models also extended on a fairly regular basis, extending for a month generally every 4th release. The last model in this set extended until the end of 2020. Figure 6 summarizes a set of models selected from this wave.

The predictions in this wave are significantly different from those in wave 1. The pandemic is no longer shown as ending within the forecast horizon. These models began to recognize a second wave and most saw the daily death rate higher at the end of the forecast than at the beginning. The most pessimistic of these models (9-11) projected 415k fatalities by the end of the year, though by the last model of this set (9-24) that number had dropped to 372K. All of these models recognized the reality of a second wave and were significantly more pessimistic than early models which saw the pandemic ending in the late spring. Most of these models implicitly predicted the wave 2 peak was beyond the forecast horizon.

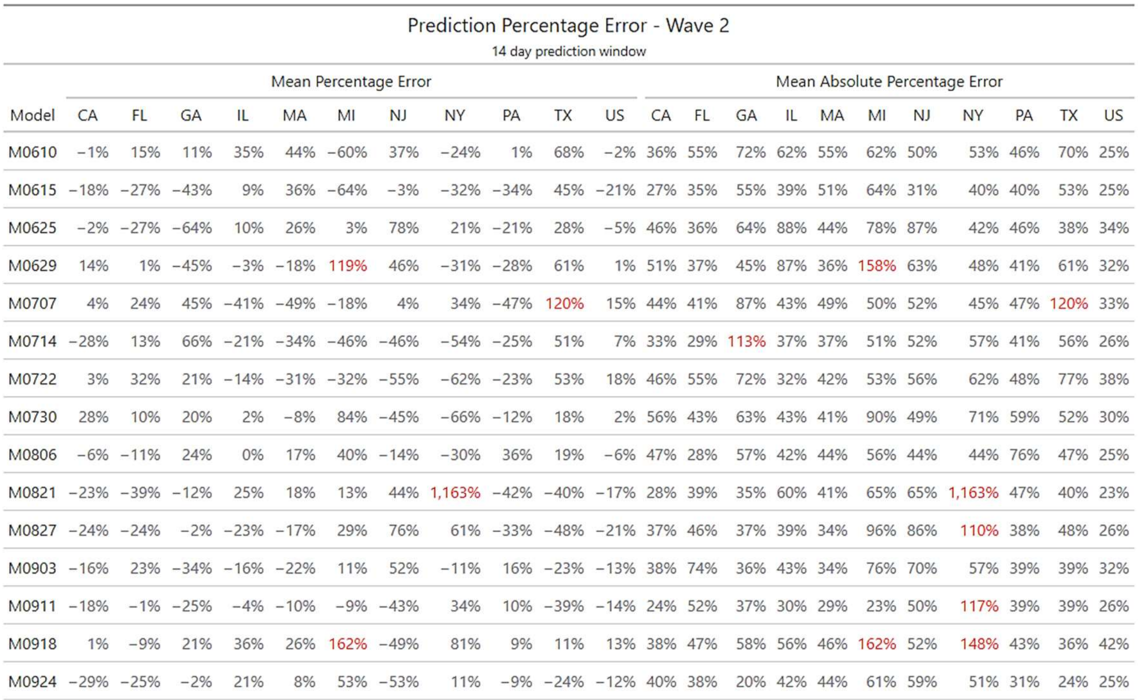
FIGURE 5
WAVE 2 FORECASTS



Model Accuracy – Daily Deaths

As we did for the wave 1 models, we now consider the accuracy for each model over a 14-day forecast horizon for the US as a whole, and for the top 10 fatality states. Table 6 presents the MAP and MAPE scores.

TABLE 6
WAVE 2 – PREDICTION ERRORS



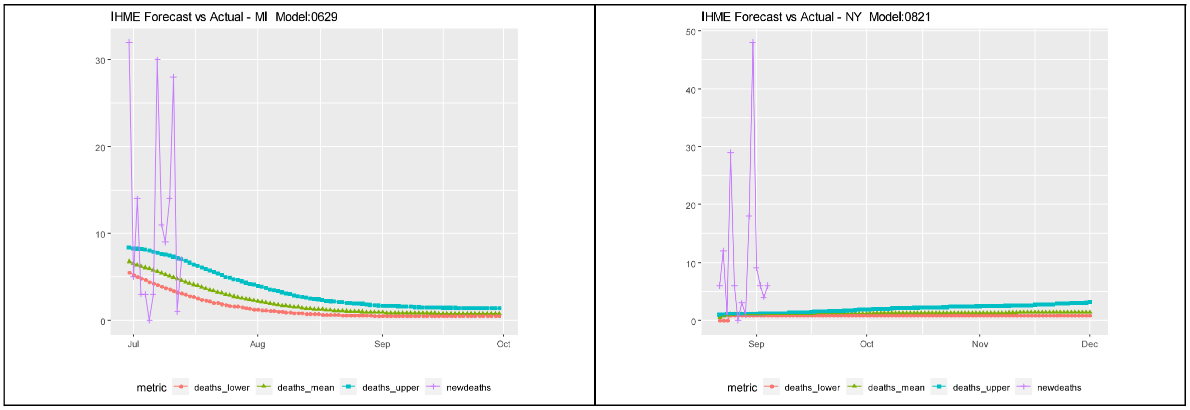
Prediction Percentage Error - Wave 2

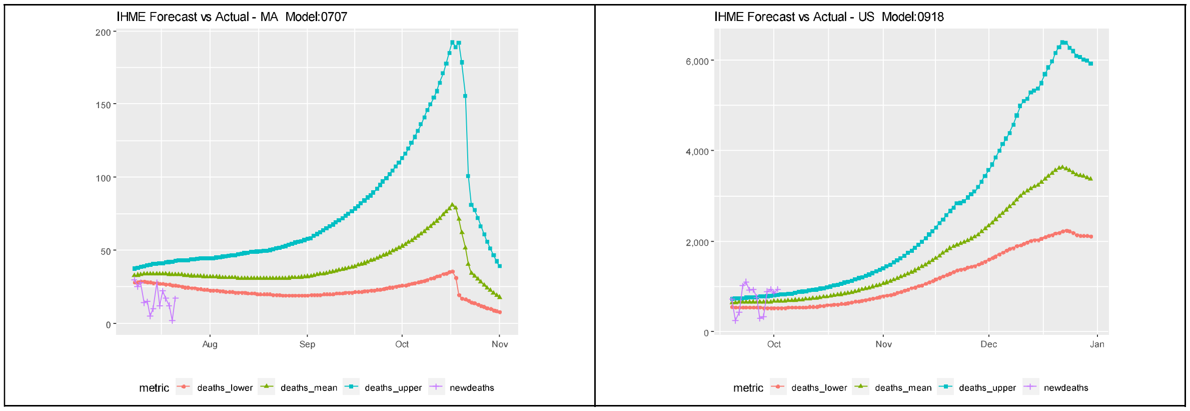
14 day prediction window

	Mean Percentage Error											Mean Absolute Percentage Error										
Model	CA	FL	GA	IL	MA	MI	NJ	NY	PA	TX	US	CA	FL	GA	IL	MA	MI	NJ	NY	PA	TX	US
M0610	−1%	15%	11%	35%	44%	−60%	37%	−24%	1%	68%	−2%	36%	55%	72%	62%	55%	62%	50%	53%	46%	70%	25%
M0615	−18%	−27%	−43%	9%	36%	−64%	−3%	−32%	−34%	45%	−21%	27%	35%	55%	39%	51%	64%	31%	40%	40%	53%	25%
M0625	−2%	−27%	−64%	10%	26%	3%	78%	21%	−21%	28%	−5%	46%	36%	64%	88%	44%	78%	87%	42%	46%	38%	34%
M0629	14%	1%	−45%	−3%	−18%	119%	46%	−31%	−28%	61%	1%	51%	37%	45%	87%	36%	158%	63%	48%	41%	61%	32%
M0707	4%	24%	45%	−41%	−49%	−18%	4%	34%	−47%	120%	15%	44%	41%	87%	43%	49%	50%	52%	45%	47%	120%	33%
M0714	−28%	13%	66%	−21%	−34%	−46%	−46%	−54%	−25%	51%	7%	33%	29%	113%	37%	37%	51%	52%	57%	41%	56%	26%
M0722	3%	32%	21%	−14%	−31%	−32%	−55%	−62%	−23%	53%	18%	46%	55%	72%	32%	42%	53%	56%	62%	48%	77%	38%
M0730	28%	10%	20%	2%	−8%	84%	−45%	−66%	−12%	18%	2%	56%	43%	63%	43%	41%	90%	49%	71%	59%	52%	30%
M0806	−6%	−11%	24%	0%	17%	40%	−14%	−30%	36%	19%	−6%	47%	28%	57%	42%	44%	56%	44%	44%	76%	47%	25%
M0821	−23%	−39%	−12%	25%	18%	13%	44%	1,163%	−42%	−40%	−17%	28%	39%	35%	60%	41%	65%	65%	1,163%	47%	40%	23%
M0827	−24%	−24%	−2%	−23%	−17%	29%	76%	61%	−33%	−48%	−21%	37%	46%	37%	39%	34%	96%	86%	110%	38%	48%	26%
M0903	−16%	23%	−34%	−16%	−22%	11%	52%	−11%	16%	−23%	−13%	38%	74%	36%	43%	34%	76%	70%	57%	39%	39%	32%
M0911	−18%	−1%	−25%	−4%	−10%	−9%	−43%	34%	10%	−39%	−14%	24%	52%	37%	30%	29%	23%	50%	117%	39%	39%	26%
M0918	1%	−9%	21%	36%	26%	162%	−49%	81%	9%	11%	13%	38%	47%	58%	56%	46%	162%	52%	148%	43%	36%	42%
M0924	−29%	−25%	−2%	21%	8%	53%	−53%	11%	−9%	−24%	−12%	40%	38%	20%	42%	44%	61%	59%	51%	31%	24%	25%

The daily predictions in wave 2 are more accurate than those in wave 1, although error rates above 20% are common. A smaller number of forecasts have errors of more than 100%. Unlike wave 1 there is no time period with very large error rates. The errors in wave 2 are less biased than in wave 1, with more of a mix of over and under forecasts. Again, the error rates tend to be higher for individual states than for the US as a whole. Figure 7 illustrates some of the individual forecast and Table 7 shows how often the forecasts were in the 95% intervals.

FIGURE 7
WAVE 2 – FORECAST VS ACTUAL

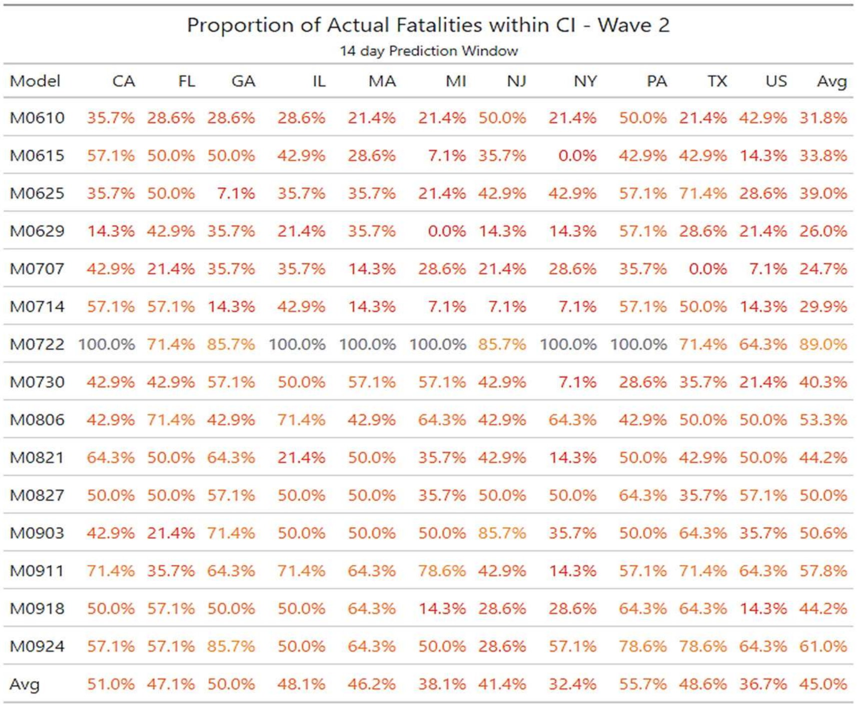




The upper left panel shows MI in late June – early July forecast from the 6-29 model. Here the death count was once again projected to drop toward zero. While the actual data appears to be quite noisy, the actual deaths are well above forecast on average. A more extreme example is shown in the upper right. Here in NY the model forecast a very low death count near zero, while the actual death was much higher. A different scenario is shown in the lower left, where the early July death counts are all well below forecasts and below the lower limit of the forecast. The last panel shows the US as a whole in mid-September. The MAPE here is 42%, while the MPE is only 13%, and only 14% of the predictions are within the 95% interval, this would again seem to indicate that the model's confidence intervals are unrealistically narrow.

The unrealistic precision implied in the model's confidence intervals is further illustrated in Table 7. With the exception of the 7-22 model, no model has 95% of the results inside the 95% interval. There are several instances where none of the outcomes are within the interval, including the MI example above. The best model has 61% of the outcomes in the interval, and overall, only 45% of the forecasts, for a two-week period, are in the 95% interval.

TABLE 7
WAVE 2 – 14 DAY CONFIDENCE INTERVALS



Proportion of Actual Fatalities within CI - Wave 2

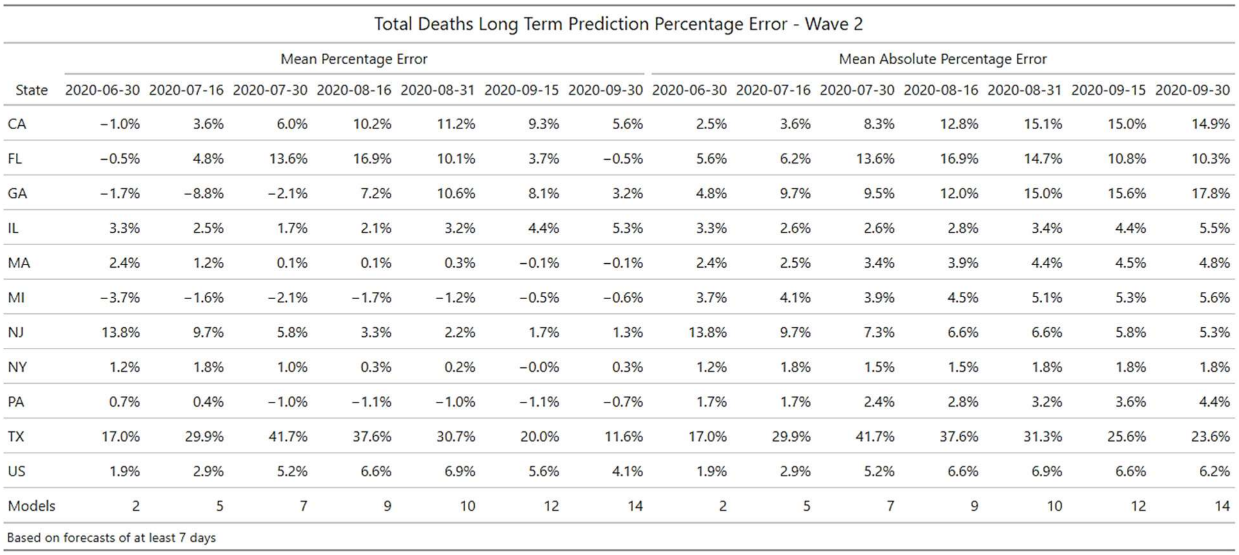
14 day Prediction Window

Model	CA	FL	GA	IL	MA	MI	NJ	NY	PA	TX	US	Avg
M0610	35.7%	28.6%	28.6%	28.6%	21.4%	21.4%	50.0%	21.4%	50.0%	21.4%	42.9%	31.8%
M0615	57.1%	50.0%	50.0%	42.9%	28.6%	7.1%	35.7%	0.0%	42.9%	42.9%	14.3%	33.8%
M0625	35.7%	50.0%	7.1%	35.7%	35.7%	21.4%	42.9%	42.9%	57.1%	71.4%	28.6%	39.0%
M0629	14.3%	42.9%	35.7%	21.4%	35.7%	0.0%	14.3%	14.3%	57.1%	28.6%	21.4%	26.0%
M0707	42.9%	21.4%	35.7%	35.7%	14.3%	28.6%	21.4%	28.6%	35.7%	0.0%	7.1%	24.7%
M0714	57.1%	57.1%	14.3%	42.9%	14.3%	7.1%	7.1%	7.1%	57.1%	50.0%	14.3%	29.9%
M0722	100.0%	71.4%	85.7%	100.0%	100.0%	100.0%	85.7%	100.0%	100.0%	71.4%	64.3%	89.0%
M0730	42.9%	42.9%	57.1%	50.0%	57.1%	57.1%	42.9%	7.1%	28.6%	35.7%	21.4%	40.3%
M0806	42.9%	71.4%	42.9%	71.4%	42.9%	64.3%	42.9%	64.3%	42.9%	50.0%	50.0%	53.3%
M0821	64.3%	50.0%	64.3%	21.4%	50.0%	35.7%	42.9%	14.3%	50.0%	42.9%	50.0%	44.2%
M0827	50.0%	50.0%	57.1%	50.0%	50.0%	35.7%	50.0%	50.0%	64.3%	35.7%	57.1%	50.0%
M0903	42.9%	21.4%	71.4%	50.0%	50.0%	50.0%	85.7%	35.7%	50.0%	64.3%	35.7%	50.6%
M0911	71.4%	35.7%	64.3%	71.4%	64.3%	78.6%	42.9%	14.3%	57.1%	71.4%	64.3%	57.8%
M0918	50.0%	57.1%	50.0%	50.0%	64.3%	14.3%	28.6%	28.6%	64.3%	64.3%	14.3%	44.2%
M0924	57.1%	57.1%	85.7%	50.0%	64.3%	50.0%	28.6%	57.1%	78.6%	78.6%	64.3%	61.0%
Avg	51.0%	47.1%	50.0%	48.1%	46.2%	38.1%	41.4%	32.4%	55.7%	48.6%	36.7%	45.0%

Model Accuracy – Cumulative Deaths

We now consider forecasts for cumulative deaths made by the wave 2 models. Table 8 lists the MPE and MAPE for forecasted deaths in half month intervals from the end of June to the end of September. Table 9 lists the proportion of the models were the 95% interval contained the final results.

TABLE 8
WAVE 2 – LONG TERM PREDICTION ERROR



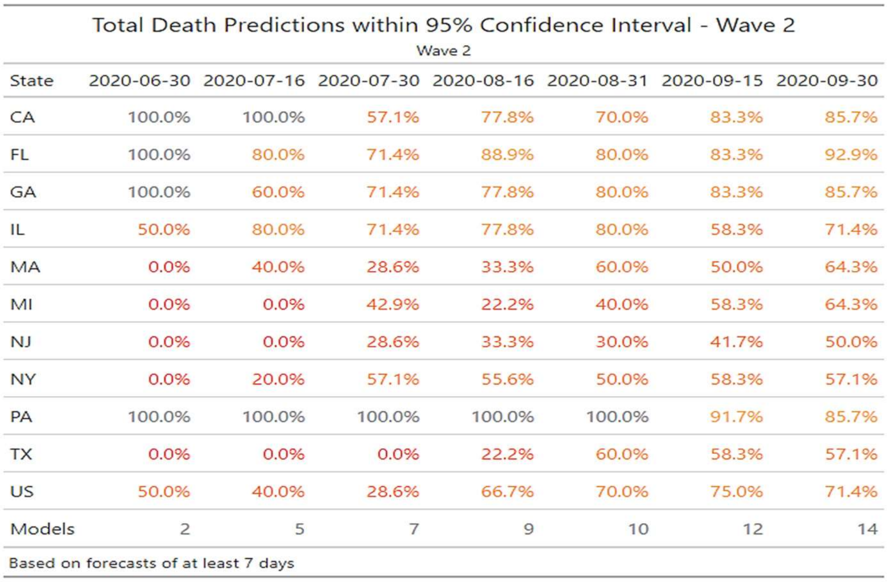
Total Deaths Long Term Prediction Percentage Error - Wave 2

	Mean Percentage Error							Mean Absolute Percentage Error						
State	2020-06-30	2020-07-16	2020-07-30	2020-08-16	2020-08-31	2020-09-15	2020-09-30	2020-06-30	2020-07-16	2020-07-30	2020-08-16	2020-08-31	2020-09-15	2020-09-30
CA	−1.0%	3.6%	6.0%	10.2%	11.2%	9.3%	5.6%	2.5%	3.6%	8.3%	12.8%	15.1%	15.0%	14.9%
FL	−0.5%	4.8%	13.6%	16.9%	10.1%	3.7%	−0.5%	5.6%	6.2%	13.6%	16.9%	14.7%	10.8%	10.3%
GA	−1.7%	−8.8%	−2.1%	7.2%	10.6%	8.1%	3.2%	4.8%	9.7%	9.5%	12.0%	15.0%	15.6%	17.8%
IL	3.3%	2.5%	1.7%	2.1%	3.2%	4.4%	5.3%	3.3%	2.6%	2.6%	2.8%	3.4%	4.4%	5.5%
MA	2.4%	1.2%	0.1%	0.1%	0.3%	−0.1%	−0.1%	2.4%	2.5%	3.4%	3.9%	4.4%	4.5%	4.8%
MI	−3.7%	−1.6%	−2.1%	−1.7%	−1.2%	−0.5%	−0.6%	3.7%	4.1%	3.9%	4.5%	5.1%	5.3%	5.6%
NJ	13.8%	9.7%	5.8%	3.3%	2.2%	1.7%	1.3%	13.8%	9.7%	7.3%	6.6%	6.6%	5.8%	5.3%
NY	1.2%	1.8%	1.0%	0.3%	0.2%	−0.0%	0.3%	1.2%	1.8%	1.5%	1.5%	1.8%	1.8%	1.8%
PA	0.7%	0.4%	−1.0%	−1.1%	−1.0%	−1.1%	−0.7%	1.7%	1.7%	2.4%	2.8%	3.2%	3.6%	4.4%
TX	17.0%	29.9%	41.7%	37.6%	30.7%	20.0%	11.6%	17.0%	29.9%	41.7%	37.6%	31.3%	25.6%	23.6%
US	1.9%	2.9%	5.2%	6.6%	6.9%	5.6%	4.1%	1.9%	2.9%	5.2%	6.6%	6.9%	6.6%	6.2%
Models	2	5	7	9	10	12	14	2	5	7	9	10	12	14

Based on forecasts of at least 7 days

These forecasts tend to be a bit more accurate than the wave 1 forecasts, though large errors are still common. Most of the MPE statistics, and virtually all the large ones, are positive, indicating an under forecasting of the final death count. Of the 77 cases shown in Table 9, only nine include the actual result in the confidence interval at least 95% of the time.

TABLE 9
WAVE 2 – LONG TERM CONFIDENCE INTERVALS



Total Death Predictions within 95% Confidence Interval - Wave 2

Wave 2

State	2020-06-30	2020-07-16	2020-07-30	2020-08-16	2020-08-31	2020-09-15	2020-09-30
CA	100.0%	100.0%	57.1%	77.8%	70.0%	83.3%	85.7%
FL	100.0%	80.0%	71.4%	88.9%	80.0%	83.3%	92.9%
GA	100.0%	60.0%	71.4%	77.8%	80.0%	83.3%	85.7%
IL	50.0%	80.0%	71.4%	77.8%	80.0%	58.3%	71.4%
MA	0.0%	40.0%	28.6%	33.3%	60.0%	50.0%	64.3%
MI	0.0%	0.0%	42.9%	22.2%	40.0%	58.3%	64.3%
NJ	0.0%	0.0%	28.6%	33.3%	30.0%	41.7%	50.0%
NY	0.0%	20.0%	57.1%	55.6%	50.0%	58.3%	57.1%
PA	100.0%	100.0%	100.0%	100.0%	100.0%	91.7%	85.7%
TX	0.0%	0.0%	0.0%	22.2%	60.0%	58.3%	57.1%
US	50.0%	40.0%	28.6%	66.7%	70.0%	75.0%	71.4%
Models	2	5	7	9	10	12	14

Based on forecasts of at least 7 days

FIGURE 8
WAVE 2 – SELECT PREDICTION VS ACTUAL

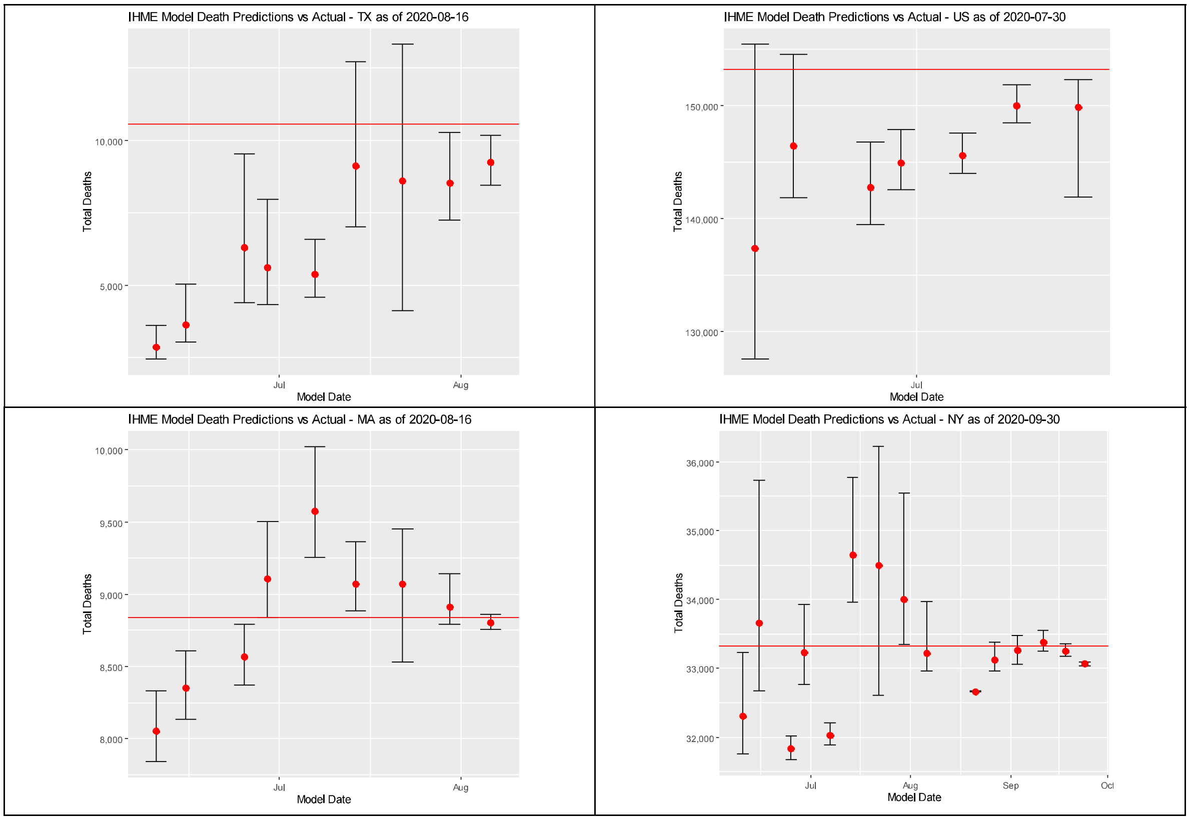


Figure 8 shows several examples of the evolution of the model in this time period. The upper left panel shows the forecast for total deaths in TX on 6-16. A total of 9 models made predictions for this date, and only two contained the final outcome. The first two models were highly inaccurate, with forecasts less than half the ultimate total, and highly precise. The interval increased significantly with the third model in this series, but narrowed again in the fifth, only to increase sharply in the next two versions. From a policy maker's perspective, the large fluctuations in the forecasts and forecast precision over such short time frames dramatically limit the utility of the forecast. The upper right panel shows similar gyrations for the US as a whole. The lower left panel shows large swings up, and then down for MA. The lower right panel illustrates NY at the end September and illustrates major changes in the width of the interval, from very large to very small and back several times. While it is impossible to go through every example, these graphs demonstrate the volatility and inaccuracy of some of the models in this time period.

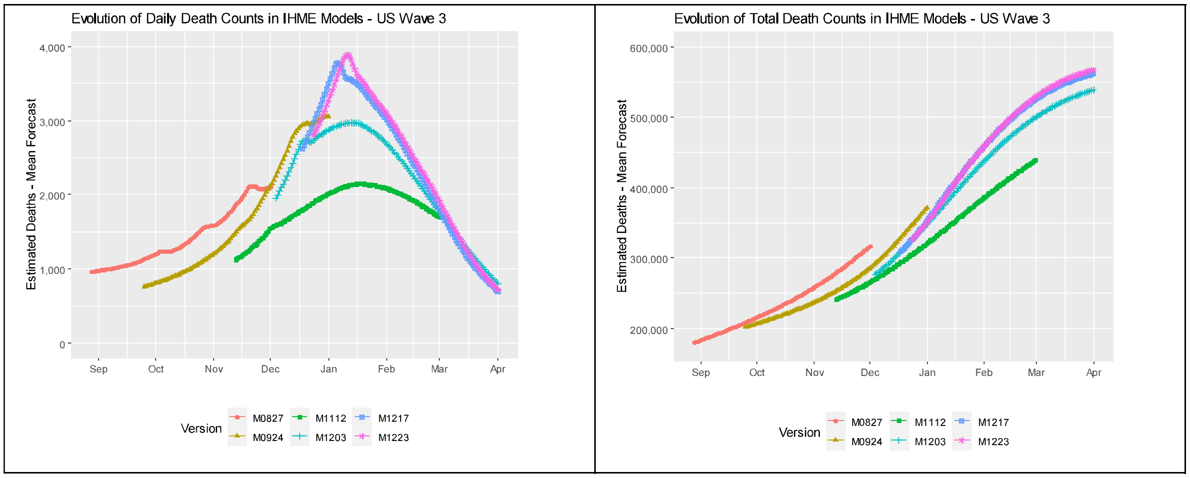
Summary – Wave 2

The wave 2 models recognized that the pandemic was going to extend beyond the forecast horizon and most of the models did not predict a peak. The models were released in a more regular fashion tended to be somewhat less prone to rapid changes, though there were multiple instances where predictions would jump dramatically from release to release. Error rates for these models tended to be somewhat lower than wave 1, but still fairly high. Error rates of more than 30% for daily deaths were common and in multiple cases errors were more than 10%. Errors for cumulative deaths naturally tended to be lower and were often in the single digits, but in some cases were much higher. The cumulative error for TX was as high as 41.7% in late July. The accuracy of the model's confidence intervals continued to be questionable both for daily and cumulative deaths. Far less than 95% of actual results fell within the intervals. The aggregate death level for the US as a whole at the end of July was only captured in half the 95% intervals.

WAVE 3 MODELS

In this section we will consider wave 3 of the pandemic. For this analysis we will look at models beginning with the 9-24 model, the last model we examined for wave 2, and ending with the model released on 12-23, the last model released in 2020. Wave 3 of the pandemic is on-going, but we are tracking actual deaths through 12-31. During this period the IHME released 12 versions of the model, one version approximately every 9 days. During this period models were released on a more consistent basis of one per week, with two weeks skipped due to holidays. The forecasts in this period are very different from the previous forecast set. In this time period the models again predict a peak and rapid decline. Most of the models predict the peak will occur sometime in January thought the size of the peak varies considerably. The 11-12 model, for example, forecasts a peak of just over 2,000 deaths per day. The model released a little more than a month later shows peak death nearing 4,000 per day. All these models show a rapid decline through February and March, but deaths remain above 500 at the end of March so the models are implicitly forecasting the pandemic to continue at least into the 2nd half of 2021, and therefore no final death toll is forecast. The latest models have a total death estimate over 567K while the pandemic is still on-going. This is 7 times the total forecast from the first version of the model in March. Figure 9 shows daily and cumulative deaths for select models in this group.

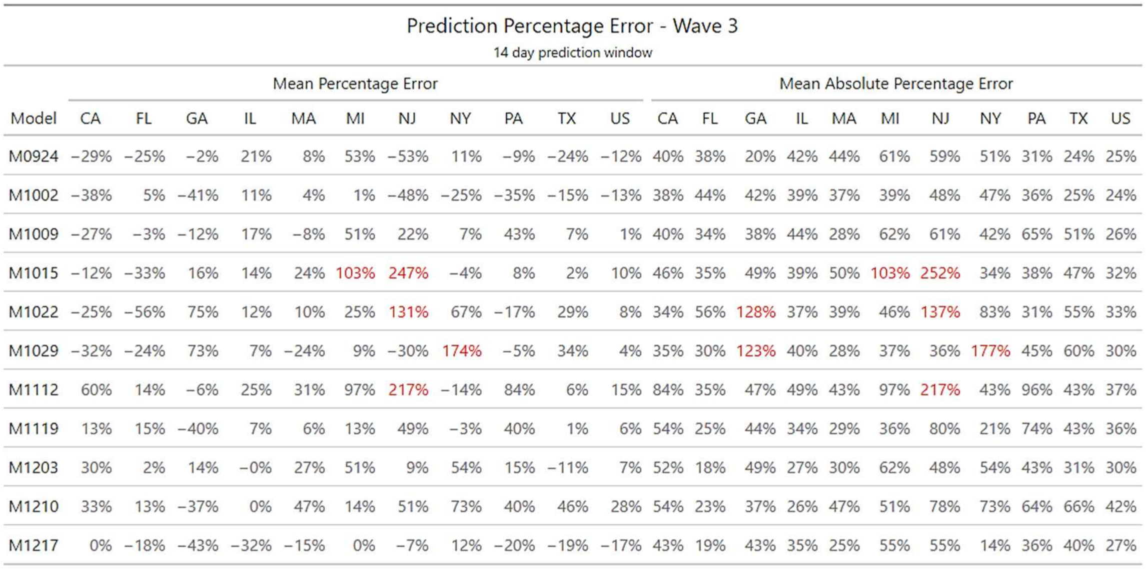
FIGURE 9
WAVE 3 FORECASTS



Model Accuracy – Daily Deaths

We again consider the accuracy of each of the models over a 14-day horizon. Table 10 summarizes the MPE and MAPE for these predictions. Note: this table does not include the 12-23 model since a 14-day window of actuals is not available.

TABLE 10
WAVE 3 – PREDICTION ERRORS



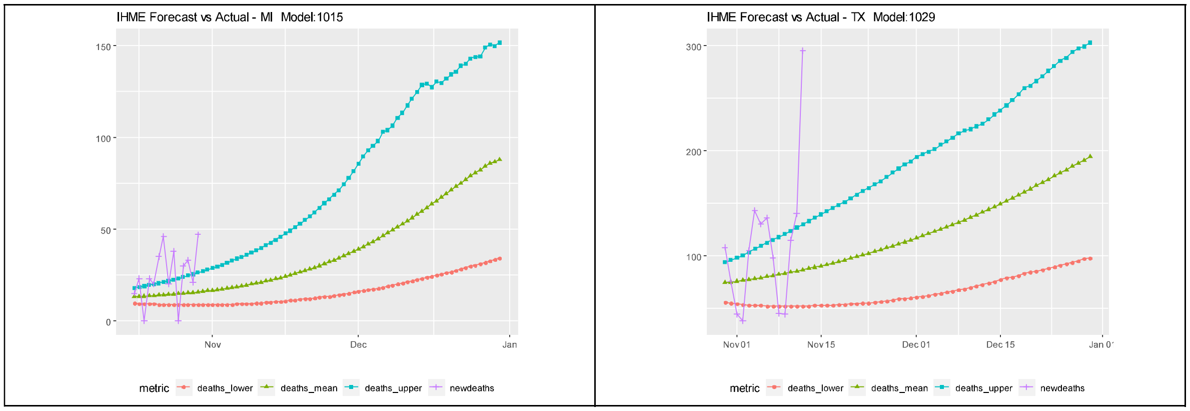
Prediction Percentage Error - Wave 3

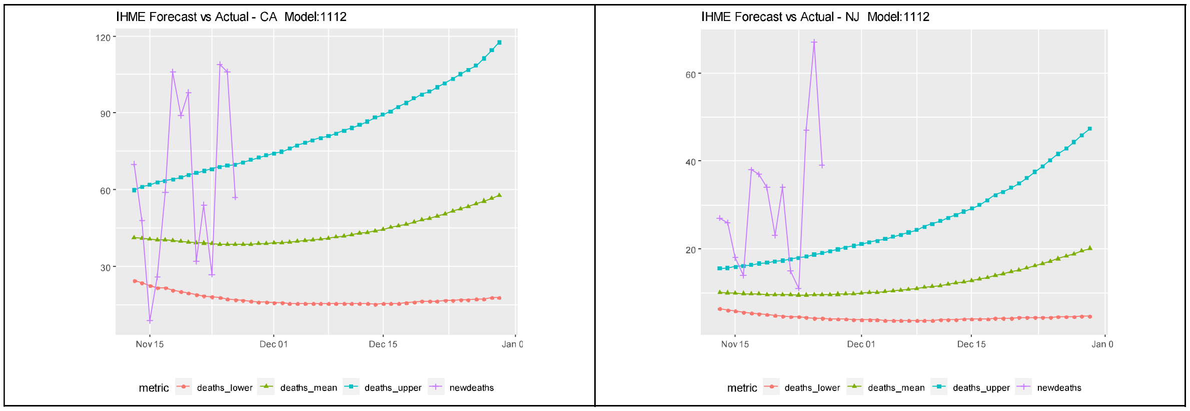
14 day prediction window

	Mean Percentage Error											Mean Absolute Percentage Error										
Model	CA	FL	GA	IL	MA	MI	NJ	NY	PA	TX	US	CA	FL	GA	IL	MA	MI	NJ	NY	PA	TX	US
M0924	−29%	−25%	−2%	21%	8%	53%	−53%	11%	−9%	−24%	−12%	40%	38%	20%	42%	44%	61%	59%	51%	31%	24%	25%
M1002	−38%	5%	−41%	11%	4%	1%	−48%	−25%	−35%	−15%	−13%	38%	44%	42%	39%	37%	39%	48%	47%	36%	25%	24%
M1009	−27%	−3%	−12%	17%	−8%	51%	22%	7%	43%	7%	1%	40%	34%	38%	44%	28%	62%	61%	42%	65%	51%	26%
M1015	−12%	−33%	16%	14%	24%	103%	247%	−4%	8%	2%	10%	46%	35%	49%	39%	50%	103%	252%	34%	38%	47%	32%
M1022	−25%	−56%	75%	12%	10%	25%	131%	67%	−17%	29%	8%	34%	56%	128%	37%	39%	46%	137%	83%	31%	55%	33%
M1029	−32%	−24%	73%	7%	−24%	9%	−30%	174%	−5%	34%	4%	35%	30%	123%	40%	28%	37%	36%	177%	45%	60%	30%
M1112	60%	14%	−6%	25%	31%	97%	217%	−14%	84%	6%	15%	84%	35%	47%	49%	43%	97%	217%	43%	96%	43%	37%
M1119	13%	15%	−40%	7%	6%	13%	49%	−3%	40%	1%	6%	54%	25%	44%	34%	29%	36%	80%	21%	74%	43%	36%
M1203	30%	2%	14%	−0%	27%	51%	9%	54%	15%	−11%	7%	52%	18%	49%	27%	30%	62%	48%	54%	43%	31%	30%
M1210	33%	13%	−37%	0%	47%	14%	51%	73%	40%	46%	28%	54%	23%	37%	26%	47%	51%	78%	73%	64%	66%	42%
M1217	0%	−18%	−43%	−32%	−15%	0%	−7%	12%	−20%	−19%	−17%	43%	19%	43%	35%	25%	55%	55%	14%	36%	40%	27%

Relative to wave 1, the 14-day accuracy of the predictions is better, thought the errors are still large. There are less of the massive error then we saw in wave 1, though there are 7 instances, where MAPE is greater than 100%. The largest errors occurred between mid-October and mid-November. As can be seen in Figure 1, this is when the daily death rate began a sharp increase. Even though the positive rate had been climbing since mid-September, the models did not accurately forecast the rising death rates. This is further illustrated in Figure 10, which shows several examples, and Table 11 which shows the in-interval levels for these models.

FIGURE 10
WAVE 3 – FORECAST VS ACTUAL

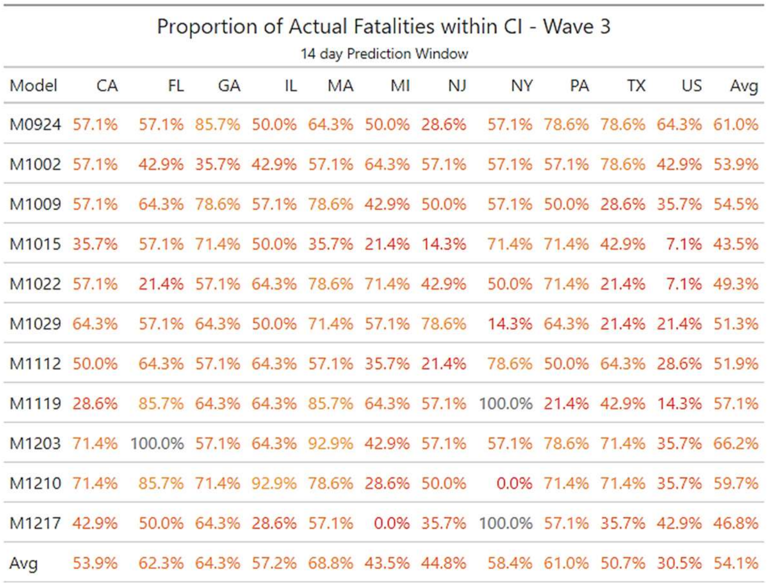




These graphs show how the death rate started to spike ahead of the gradual forecast predicted by the models. All these graphs illustrate the significant noise in the data, with large day to day swings. The upper panels show the gradual upswing expected in mid-late October in MI and TX. In both cases the actual data is a bit noisy, with some observations below the lower limit, but many observations above the upper limit. Each of these forecasts had 21.4% of the observations inside the 95% interval. The lower two panels show mid-November forecasts for CA and NJ. The CA forecast has a 60% MPE and 84% MAPE with half the actuals inside the interval. The numbers for NJ are worse, with an MPE and MAPE of 217%. Forecasts in NJ were problematic throughout this wave with MPE levels of 131%, 217%, and 247% reported.

The in-interval data shows that only 3 of the 121 scenarios had at least 95% of the outcomes in the interval. For the 11 forecasts evaluated (top 10 states plus the US) only two models had more than 60% of the outcomes inside the 95% interval. Similarly, only 3 states had more than 60% of forecasts in the interval. For the US as a whole only 30.5% of the forecasts were inside the 95% interval.

TABLE 11
WAVE 3 – 14 DAY CONFIDENCE INTERVALS



Proportion of Actual Fatalities within CI - Wave 3

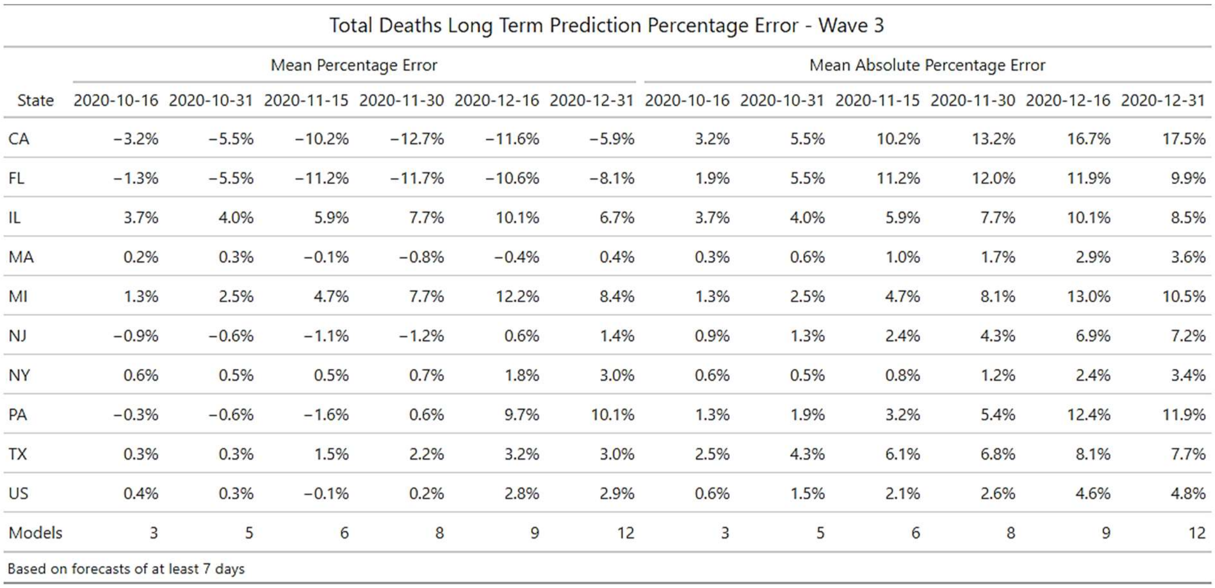
14 day Prediction Window

Model	CA	FL	GA	IL	MA	MI	NJ	NY	PA	TX	US	Avg
M0924	57.1%	57.1%	85.7%	50.0%	64.3%	50.0%	28.6%	57.1%	78.6%	78.6%	64.3%	61.0%
M1002	57.1%	42.9%	35.7%	42.9%	57.1%	64.3%	57.1%	57.1%	57.1%	78.6%	42.9%	53.9%
M1009	57.1%	64.3%	78.6%	57.1%	78.6%	42.9%	50.0%	57.1%	50.0%	28.6%	35.7%	54.5%
M1015	35.7%	57.1%	71.4%	50.0%	35.7%	21.4%	14.3%	71.4%	71.4%	42.9%	7.1%	43.5%
M1022	57.1%	21.4%	57.1%	64.3%	78.6%	71.4%	42.9%	50.0%	71.4%	21.4%	7.1%	49.3%
M1029	64.3%	57.1%	64.3%	50.0%	71.4%	57.1%	78.6%	14.3%	64.3%	21.4%	21.4%	51.3%
M1112	50.0%	64.3%	57.1%	64.3%	57.1%	35.7%	21.4%	78.6%	50.0%	64.3%	28.6%	51.9%
M1119	28.6%	85.7%	64.3%	64.3%	85.7%	64.3%	57.1%	100.0%	21.4%	42.9%	14.3%	57.1%
M1203	71.4%	100.0%	57.1%	64.3%	92.9%	42.9%	57.1%	57.1%	78.6%	71.4%	35.7%	66.2%
M1210	71.4%	85.7%	71.4%	92.9%	78.6%	28.6%	50.0%	0.0%	71.4%	71.4%	35.7%	59.7%
M1217	42.9%	50.0%	64.3%	28.6%	57.1%	0.0%	35.7%	100.0%	57.1%	35.7%	42.9%	46.8%
Avg	53.9%	62.3%	64.3%	57.2%	68.8%	43.5%	44.8%	58.4%	61.0%	50.7%	30.5%	54.1%

Model Accuracy – Cumulative Deaths

While the daily predictions of new deaths are subject to significant noise effects, that is less of an issue with cumulative death forecasts. Table 12 summarizes errors on total deaths for 6 dates from mid-October until the end of the year.

TABLE 12
WAVE 3 – LONG TERM PREDICTION ERROR



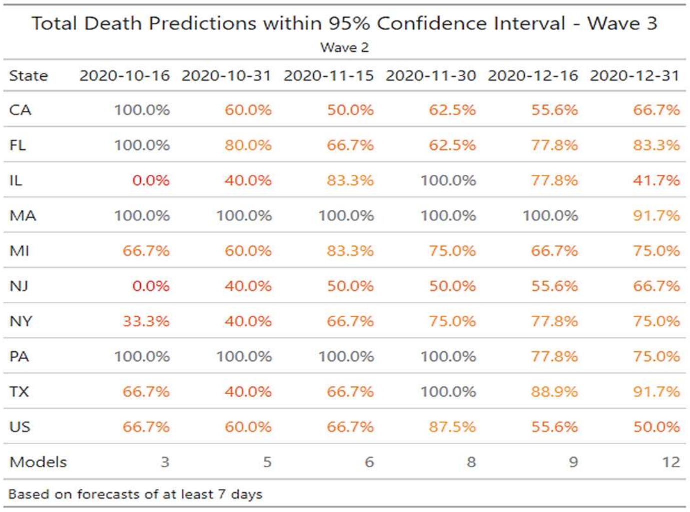
Total Deaths Long Term Prediction Percentage Error - Wave 3

	Mean Percentage Error						Mean Absolute Percentage Error					
State	2020-10-16	2020-10-31	2020-11-15	2020-11-30	2020-12-16	2020-12-31	2020-10-16	2020-10-31	2020-11-15	2020-11-30	2020-12-16	2020-12-31
CA	−3.2%	−5.5%	−10.2%	−12.7%	−11.6%	−5.9%	3.2%	5.5%	10.2%	13.2%	16.7%	17.5%
FL	−1.3%	−5.5%	−11.2%	−11.7%	−10.6%	−8.1%	1.9%	5.5%	11.2%	12.0%	11.9%	9.9%
IL	3.7%	4.0%	5.9%	7.7%	10.1%	6.7%	3.7%	4.0%	5.9%	7.7%	10.1%	8.5%
MA	0.2%	0.3%	−0.1%	−0.8%	−0.4%	0.4%	0.3%	0.6%	1.0%	1.7%	2.9%	3.6%
MI	1.3%	2.5%	4.7%	7.7%	12.2%	8.4%	1.3%	2.5%	4.7%	8.1%	13.0%	10.5%
NJ	−0.9%	−0.6%	−1.1%	−1.2%	0.6%	1.4%	0.9%	1.3%	2.4%	4.3%	6.9%	7.2%
NY	0.6%	0.5%	0.5%	0.7%	1.8%	3.0%	0.6%	0.5%	0.8%	1.2%	2.4%	3.4%
PA	−0.3%	−0.6%	−1.6%	0.6%	9.7%	10.1%	1.3%	1.9%	3.2%	5.4%	12.4%	11.9%
TX	0.3%	0.3%	1.5%	2.2%	3.2%	3.0%	2.5%	4.3%	6.1%	6.8%	8.1%	7.7%
US	0.4%	0.3%	−0.1%	0.2%	2.8%	2.9%	0.6%	1.5%	2.1%	2.6%	4.6%	4.8%
Models	3	5	6	8	9	12	3	5	6	8	9	12

Based on forecasts of at least 7 days

Since these models are forecasting the total cumulative deaths after the pandemic has raged for several months, and deaths have accumulated, we should expect these forecasts to be more accurate than those early in the pandemic, and that is the case. The cumulative error rates are much lower than wave 1, and somewhat lower than wave 2. US errors are all lower than 5%, while in wave 2 they went as high as 6.9% and in wave 1 they were often above 20% and as high as 26.7%. Error rates for individual states are mostly in the single digits, with the exception of CA and FL, two states where the models tended to forecast too high. Errors in wave 1 were often above 30% and as high as 119.6%. Errors for wave 2 were better than wave 1, but often in double digits and as high 41.7%. Table 13 shows how often the end date forecasts were in the 95% interval.

TABLE 13
WAVE 3 – LONG TERM CONFIDENCE INTERVALS



Total Death Predictions within 95% Confidence Interval - Wave 3

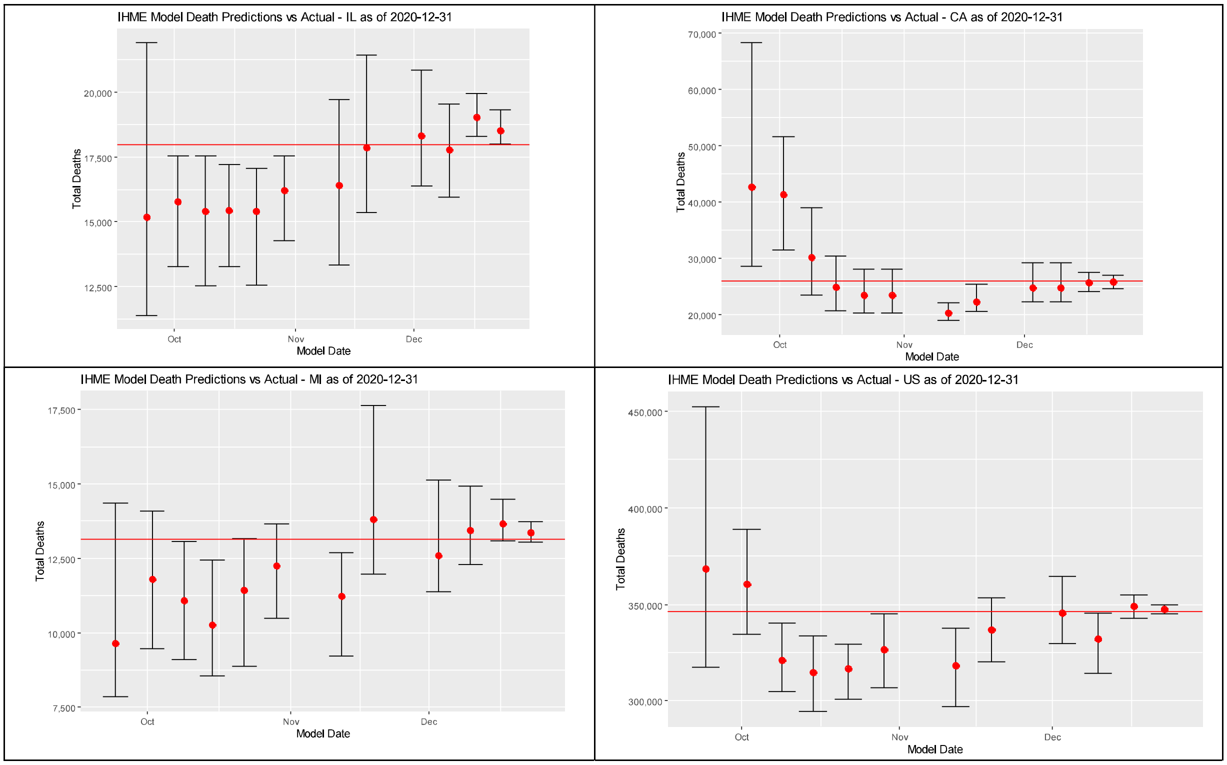
Wave 2

State	2020-10-16	2020-10-31	2020-11-15	2020-11-30	2020-12-16	2020-12-31
CA	100.0%	60.0%	50.0%	62.5%	55.6%	66.7%
FL	100.0%	80.0%	66.7%	62.5%	77.8%	83.3%
IL	0.0%	40.0%	83.3%	100.0%	77.8%	41.7%
MA	100.0%	100.0%	100.0%	100.0%	100.0%	91.7%
MI	66.7%	60.0%	83.3%	75.0%	66.7%	75.0%
NJ	0.0%	40.0%	50.0%	50.0%	55.6%	66.7%
NY	33.3%	40.0%	66.7%	75.0%	77.8%	75.0%
PA	100.0%	100.0%	100.0%	100.0%	77.8%	75.0%
TX	66.7%	40.0%	66.7%	100.0%	88.9%	91.7%
US	66.7%	60.0%	66.7%	87.5%	55.6%	50.0%
Models	3	5	6	8	9	12

Based on forecasts of at least 7 days

Again, these numbers are better, but even this far into the pandemic, only 13 out of the 66 scenarios met the stated CI goal. Of the 12 different models predicting the end of year death count for the US as a whole, only 6 has the final death count in the 95% interval. Several examples are shown in Figure 11.

FIGURE 11
WAVE 3 – SELECT PREDICTION VS ACTUAL



The case of the US is shown in the lower right panel. The earliest model in this group, issued more than 3 months out, was reasonably accurate and the interval contained the final count. That model predicted a final death count for the year of 368,457 against an actual final death count of 346,421, and error of only 6.3%. Over time the model adjusted the total forecast downward, while significantly lowering the interval width. By 10-9 the forecast had been lowered to 321,140 for a slightly higher error of 7.3%. But the interval had been lowered from ± 50,999 to ±16,122. That model, and the next 4, failed to contain the actual outcome within the interval. A similar issue is shown in TX, where the 10-02 model has a significantly narrower interval that no longer contained the actual. In CA, the early model intervals did not contain the actual, but after subsequent adjustment they did. But then the 11-12 model created a very small, and inaccurate interval.

Summary – Wave 3

In the wave 3 models, the pandemic was once again predicted to reach a peak and then decline. These models tend to show peaks in the January time frame, beyond the point this analysis concludes. Model accuracy on a daily basis is still problematic, especially in the mid-October to mid-November time frame where there are multiple instances of errors in excess of 100% in MI, NJ, and NY. The error rate on cumulative deaths was lower, at least in part to the accumulated deaths lowering overall percentage errors. That being said, the forecasts for total deaths by mid-December was off on average by more than 10% in CA, FL, Il and MI, and was off by 9.7% on average for PA. The model's confidence intervals continued to be overly precise. Of the 12 models that forecast a final death count for the US, only 6 have a 95% interval that covers the actual count.

CONCLUSIONS AND LIMITATIONS

The analysis presented in this paper is purposely limited. While the IHME model forecasts multiple parameters globally, we focused only on fatalities in the United States. Additional analysis is possible on the other metrics, or on other regions of the world. An analysis could be performed on the relative accuracy of different metrics, or on the relative accuracy of forecasts in different countries or regions. We chose to focus on deaths as they are arguably the most important overall metric where reasonably complete and accurate information is available. It could be argued that hospitalization metrics are of high importance, but the data on hospitalizations is not available in every state, not is it reported consistently from state to state. We chose the end of the calendar year as a reasonable time to take a check on the model's results to date. Since the pandemic is on-going, additional analysis may be performed later when the pandemic is complete.

The developers of the IHME model faced an extraordinarily difficult challenge. Forecasting a new pandemic, early in the outbreak, based at least initially on limited and suspect data from China. While the forecasting challenge was difficult, the impact and reach of the model demands a critical analysis of the results.

As we have shown in this analysis, the early results of the model were significantly flawed. Early critics of the model claimed it was too optimistic, and this is objectively true. The original model forecasts a pandemic that would die out after causing 81 thousand deaths, with a worst case of 162 thousand deaths. At the end of the year the death count stands in excess of 346 thousand with unfortunately many more to come. It took several months for the model to recognize what many critics predicted; deaths being pushed into a second wave beyond the forecast horizon of the initial model. During these early days the model was updated frequently and in a manner that caused wide swings in both the point forecast and the confidence intervals. While the model become more accurate over time, many of these problems continued. Accuracy on a relatively short planning horizon of two weeks was poor for new deaths. The predictions for total fatalities at fixed points in time was also poor, though it naturally improved over time as new deaths represented a smaller fraction of the total. The model was more accurate at the aggregate level than it was at the detailed state level, which is a common feature of forecasts. Forecasts at lower levels of detail are generally more difficult and subject to higher levels of error.

Arguably the most significant problem with the model's predictions has to do with the integrity of the predicted confidence intervals. The output of the model may accurately represent a 95% confidence interval internally; that is, the range spanned by 95% of the model's sample paths. But the output is not consistent with an externally valid confidence interval, a forecast in which the final output lies within the interval 95% of the time. We have demonstrated that throughout the pandemic the daily and cumulative forecast intervals covered the actual results far less than 95% of the time. While the model was criticized for being too uncertain – confidence levels too wide for practical use, they were in fact clearly not wide enough. It seems reasonable to conclude that the model has been overly confident with its forecasts throughout.

An additional criticism of the model was its instability. The model was updated frequently. As we have demonstrated point forecasts often shifted radically up and down in short time periods, while intervals expanded and contracted. These changes were often made, and reversed, too quickly to reflect any real changes in the evolution of the pandemic or in the health system's response to the pandemic.

While the model was often cited by public policy makers, at least early in the pandemic, it is impossible to know how much the output of the model directly influenced decision makers. If policy decisions were made based on an unwarranted level of confidence in the model's predictions, the consequences may have been severe. We hope that providing a fair and impartial assessment of this model, and its results, will help future modelers and policy makers better understand the difficulty and uncertainty inherent in forecasting such an unpredictable event.

REFERENCES

Alquist, R., & Kilian, L. (2010). What do we learn from the price of crude oil futures? *Journal of Applied Econometrics*, *25*(4), 539-573. doi:10.1002/jae.1159

Athanasopoulos, G., Hyndman, R.J., Song, H., & Wu, D.C. (2011). The tourism forecasting competition. *International Journal of Forecasting*, *27*(3), 822-844. doi:https://doi.org/10.1016/j.ijforecast.2010.04.009

Azad, A. (2020). *An influential model projects coronavirus deaths will stop this summer, but experts are skeptical*. Retrieved from https://www.fox10tv.com/news/coronavirus/an-influential-model-projects-coronavirus-deaths-will-stop-this-summer-but-experts-are-skeptical/article_f308b3c2-9752-5367-9d06-fd078e309ee0.html

Begley, S. (2020). *Influential Covid-19 model uses flawed methods and shouldn't guide U.S. policies, critics say*. Retrieved from https://www.statnews.com/2020/04/17/influential-covid-19-model-uses-flawed-methods-shouldnt-guide-policies-critics-say/

Boice, J. (2020). Best-Case and Worst-Case Coronavirus Forecasts Are Very Far Apart. *FiveThirtyEight*. Retrieved from https://fivethirtyeight.com/features/best-case-and-worst-case-coronavirus-forecasts-are-very-far-apart/

Carvajal, N. (2020). *Birx says drop in US death projection is due to Americans changing their behavior through social distancing*. Retrieved from https://www.cnn.com/2020/04/08/politics/deborah-birx-social-distancing-models/index.html

CDC. (2020). *About COVID-19*. Retrieved from https://www.cdc.gov/coronavirus/2019-ncov/cdcresponse/about-COVID-19.html

Chernenko, S., Schwarz, K., & Wright, J. (2004). The Information Content of Forward and Futures Prices: Market Expectations and the Price of Risk. *SSRN Electronic Journal*. doi:10.2139/ssrn.560386

Chikina, M., & Pegden, W. (2020). A call to honesty in pandemic modeling. *Medium*. Retrieved from https://medium.com/@wpegden/a-call-to-honesty-in-pandemic-modeling-5c156686a64b

Chinn, M., & Coibion, O. (2009). The Predictive Content of Commodity Futures. *Journal of Futures Markets, 34*. doi:10.2139/ssrn.1490043

Chowell, G., Sattenspiel, L., Bansal, S., & Viboud, C. (2016). Mathematical models to characterize early epidemic growth: A review. *Physics of Life Reviews*, *18*, 66-97. doi:10.1016/j.plrev.2016.07.005

CIDRAP. (2020). *Coroner: First US COVID-19 death occurred in early February*. Retrieved from https://www.cidrap.umn.edu/news-perspective/2020/04/coroner-first-us-covid-19-death-occurred-early-february

Dimitrov, N.B., & Meyers, L.A. (2010). Mathematical Approaches to Infectious Disease Prediction and Control. *Risk and Optimization in an Uncertain World*, pp. 1-25.

Fildes, R., & Makridakis, S. (1995). The Impact of Empirical Accuracy Studies on Time Series Analysis and Forecasting. *International Statistical Review / Revue Internationale de Statistique, 63*(3), 289-308. doi:10.2307/1403481

Fumento, M. (2020). *After Repeated Failures, It's Time to Permanently Dump Epidemic Models.* Retrieved from https://issuesinsights.com/2020/04/18/after-repeated-failures-its-time-to-permanently-dump-epidemic-models/

GitHub. (2020). *Johns Hopkins Coronoavirus Time Series data*. Retrieved from https://github.com/CSSEGISandData/COVID-19/tree/master/csse_covid_19_data/csse_covid_19_time_series

Hatmaker, T. (2020). *Fauci: US can expect more than 100,000 COVID-19 deaths, millions of cases.* Retrieved from https://techcrunch.com/2020/03/29/fauci-how-many-coronavirus-deaths-in-us-estimate/

Holshue, M.L., DeBolt, C., Lindquist, S., Lofy, K.H., Wiesman, J., Bruce, H., & Pillai, S.K. (2020). First Case of 2019 Novel Coronavirus in the United States. *New England Journal of Medicine*, *382*(10), 929-936. doi:10.1056/NEJMoa2001191

IHME. (2020). Health Data.org. Retrieved from https://covid19.healthdata.org/projections

Ioannidis, J.P. (2020). A fiasco in the making? As the coronavirus pandemic takes hold, we are making decisions without reliable data. *STAT*. Retrieved from https://www.statnews.com/2020/03/17/a-fiasco-in-the-making-as-the-coronavirus-pandemic-takes-hold-we-are-making-decisions-without-reliable-data/

Jewell, N.P., Lewnard, J.A., & Jewell, B.L. (2020). Predictive Mathematical Models of the COVID-19 Pandemic: Underlying Principles and Value of Projections. *JAMA*, *323*(19), 1893-1894. doi:10.1001/jama.2020.6585

Johns Hopkins. (2020). *Coronavirus Resource Center*. Retrieved from https://coronavirus.jhu.edu/

Makridakis, S., & Hibon, M. (2000). The M3-Competition: Results, conclusions and implications. *International Journal of Forecasting*, *16*(4), 451-476. doi:https://doi.org/10.1016/S0169-2070(00)00057-1

Marchant, R., Samia, N.I., Rosen, O., Tanner, M.A., & Cripps, S. (2020). *Learning as We Go: An Examination of the Statistical Accuracy of COVID19 Daily Death Count Predictions.* Retrieved from https://arxiv.org/abs/2004.04734

Murray, C.J. (2020). Forecasting COVID-19 impact on hospital bed-days, ICU-days, ventilator-days and deaths by US state in the next 4 months. *medRxiv*, 2020.2003.2027.20043752. doi:10.1101/2020.03.27.20043752

Ord, K., Hibon, M., & Makridakis, S. (2000). The M3-Competition. *International Journal of Forecasting*, *16*(4), 433-436. doi:https://doi.org/10.1016/S0169-2070(00)00078-9

Piper, K. (2020). *This coronavirus model keeps being wrong. Why are we still listening to it?* Retrieved from https://www.vox.com/future-perfect/2020/5/2/21241261/coronavirus-modeling-us-deaths-ihme-pandemic

Robbins, T.R. (2020a). *A Preliminary Assessment of the IHME Covid-19 Model*. Paper presented at the 51st Annual Conference of The Decision Sciences Institute. Retrieved from https://www.researchgate.net/publication/341913503_A_Preliminary_Assessment_of_the_IHME_Covid-19_Model

Robbins, T.R. (2020b). *Real Time Tracking of a Global Pandemic: Data Issues and Implications.* Retrieved from https://www.researchgate.net/publication/340772468_Real_Time_Tracking_of_a_Global_Pandemic_Data_Issues_and_Implications

Sunstein, C.R. (2006). *Infotopia: How Many Minds Produce Knowledge.*

Tetlock, P.E. (2017). *Expert Political Judgment, How Good is it? How can we tell?*

Tetlock, P.E., & Gardner, D. (2015). *Superforecasting: The Art and Science of Prediction.*

Wan, W., Dawsey, J., Parker, A., & Achenbach, J. (2020). Experts and Trump's advisers doubt White House's 240,000 coronavirus deaths estimate. *The Washington Post*. Retrieved from https://www.washingtonpost.com/health/2020/04/02/experts-trumps-advisers-doubt-white-houses-240000-coronavirus-deaths-estimate/

Wan, W., & Johnson, C.Y. (2020, April 7). America's most influential coronavirus model just revised its estimates downward. But not every model agrees. *The Washington Post*. Retrieved from https://www.washingtonpost.com/health/2020/04/06/americas-most-influential-coronavirus-model-just-revised-its-estimates-downward-not-every-model-agrees/