Enhancing Long-term Forecasting: Learning from COVID-19 Models

Hazhir Rahmandad
Associate Professor of System Dynamics, Sloan School of Management, MIT
Contact: hazhir@mit.edu, (617) 258-8912
Address: 100 Main Street, Cambridge, MA 02142

Ran Xu
Assistant Professor, Department of Allied Health Sciences, University of Connecticut
Contact: ran.2.xu@uconn.edu, (860) 486-2945
Address: Koons Hall, 358 Mansfield Rd, Unit 1101, Storrs, CT 06269

Navid Ghaffarzadegan
Corresponding author
Associate Professor, Industrial and Systems Engineering, Virginia Tech
Contact: navidg@vt.edu, (703) 538-8434
Address: 7054 Haycock Road, Falls Church, VA 22043
Enhancing Long-term Forecasting: Learning from COVID-19 Models

Abstract

Long-term forecasts are hard, but also indispensable in personal and policy planning. How could long-term predictions of complex phenomena, such as COVID-19 contagion, be enhanced? While much effort has gone into building predictive models of the pandemic, some have argued that early exponential growth combined with the stochastic nature of epidemics make the long-term prediction of contagion trajectories impossible. We leverage the diverse models contributing to CDC repository of COVID-19 death trajectories to identify factors associated with prediction accuracy across different projection horizons. We find that better long-term predictions correlate with (1) capturing the physics of transmission (instead of using black-box models); (2) projecting human behavioral reactions to an evolving pandemic; and (3) resetting state variables to account for randomness not captured in the model before starting projection. A very simple model, SEIRb, that incorporates these features and includes few other nuances offers predictions comparable with the most accurate models in the CDC set. Key to the long-term predictive power of multi-wave COVID-19 trajectories is endogenously capturing behavioral responses: balancing feedbacks where the perceived risk of death continuously changes transmission rate through the adoption, and relaxation, of various Non-Pharmaceutical Interventions (NPIs).

Keywords: epidemics, forecast, SEIR, behavior change, system dynamics

Introduction

From public policy to strategic management, forecasting long-term trajectories is inevitable: planning for future rests on forecasts [1, 2] and models, mental or formal, have to be used for producing those forecasts. Yet long-term forecasts are hard especially when systems’ dynamics are regulated by human action [3, 4]. What features of predictive models for such systems enhance their long-term accuracy? The COVID-19 pandemic offers a natural testbed to assess this broad question in the context of a high-stake problem. Long-term projections of pandemic trajectory have been used to inform planning intensive care capacity, selecting clinical trial locations, deciding on economic policy packages, and many other consequential decisions, as well as individual choices on careers, vacations, home purchases, investments, and others. Prediction is not the primary purpose for many models [5], but a large number of simulation and machine learning models have been developed to predict the trajectory of the disease at different geographic resolutions in order to inform individual and policy responses [6-11]. These models have offered significantly different predictions over time: some early models predicted millions of deaths during the first few months in the United States [7] while others expected an end to the pandemic by May 2020 [8, 12]. Predictions started to become more comparable with the increased knowledge about the disease, nevertheless, policy makers seeking guidance on tough, urgent, questions would get divergent answers from the existing models, leading to criticism of the established modeling approaches [13-15]. In fact, some studies suggested that attempts at predicting epidemic trajectories over longer term will inevitably fail due to the combination of exponential growth and stochastic shocks inherent to dynamics of contagion [16, 17]. Moreover, the complexity of predicting what policies will be adopted in future by various communities have remained as a major challenge to longer term epidemic projections. In response, and with a few exception [e.g., 8, 9, 18], most later models focused on predictions spanning only a few weeks, leaving out the long-term multi-wave dynamics observed across nations and central to many policy and personal questions.

Existing predictive models of COVID-19 vary significantly in their methodological approach, level of complexity, the mechanisms explicitly modeled, how human behavior is modeled, use of data, and their
estimation and projection frameworks. This diversity signifies differing working theories across the research community on what contributes to a good predictive model. Thus, building or even selecting an appropriate model is confounded by the lack of agreement about the features of models critical for useful predictions. Understanding the relative value of different modeling assumptions and constructs would also inform the more promising areas of research that would enhance future models of epidemics and beyond.

In this paper we take a step towards systematically examining what modeling choices regulate forecasting accuracy. First, we use data from the Center for Disease Control (CDC) repository of COVID-19 projections to understand the drivers of forecast accuracy. We focus on 490,210 point forecasts for weekly death incidents (across 57 locations, forecast dates over the span of a year (4/13/2020 to 3/29/2021), 20 forecast horizons (1-week-ahead predictions to 20-week-ahead predictions), and 61 models; not every model offers a forecast for every location, forecast date, or horizon). Prior research has used these data to assess the quality of probabilistic forecasts from various models [19, 20]. We build on this work by examining the association between the modeling approach and prediction quality using hand-coded architectural features of forecasting models.

Those associations inform a set of hypotheses on the features that enhance predictive accuracy of models. We then test those hypotheses by designing a simple model that focuses only on capturing the identified features. We assess the predictive performance of the resulting model and quantify the marginal value of each proposed feature by comparing projections from an alternative model that excludes the feature. This simple model, referred to as SEIR$b$, enhances the conventional models by explicitly capturing how behavioral responses of individuals and societies condition the evolution of a pandemic, and it can be used as a base for developing more accurate models for epidemic forecasting.

Data and Methods

Our research includes two distinct parts: A) Comparing the predictive accuracy of various model architectures; and B) Designing a new predictive model that builds on, and tests, the insights from the comparison. The data for the first part comes from the CDC collection of COVID-19 forecasts. In April 2020, and with the growing need to forecast COVID-19 trajectories in the US, CDC partnered with a research laboratory at the University of Massachusetts Amherst to create a forecast hub that collects and synthesizes COVID-19 trajectory predictions [19]. As of May 2021, more than 70 modeling groups from academic institutions, research laboratories, and the private sector have contributed by submitting their simulation-based trajectory predictions for various locations in the USA (typically state and national level, but also some county level). Every week, each individual model’s predictions (and confidence intervals) and an ensemble of the collective results were reported on CDC’s website. Some of these teams only provided short-term predictions in the range of 1-4 weeks, while others provided longer projections, most up to 8 weeks and a few up to 20 weeks ahead. 61 models provide death projections. These models and their weekly death predictions are the primary data we use for the first part of this study.

Forecast Data- We collected the forecast data from the inception of the project until 3/29/2021 from the GitHub site for the project. Given the better accuracy of death data and the fact that more models predict deaths, we focus our analysis on the accuracy of death projections. We also focus on accuracy of point estimates rather than confidence intervals because interpreting confidence interval accuracy is not straightforward: a model may have poor predictions but accurate confidence intervals (e.g. the 95% confidence interval may be exceedingly wide but still incorporate the true values 95% of the time). We collected all state and national level forecasts we could find from the project website. Therefore, our data structure includes one prediction (in expected number of deaths) for each model (a total of 61 models), location (a total of 57 locations spanning USA states, territories, and national), forecast date (4/13/2020 to 3/29/2021), and
With these guidelines in mind, we start with the basic SEIR model and add to it a seasonality factor (w) and the behavioral response feedback. This feedback reflects how transmission intensity (β) endogenously

Coding Model and Estimation Architectures- We need to identify relevant features of studied models to associate with their predictive accuracy. However, the models included on the CDC hub span different methodological approaches, use different data sources, have different levels of complexity (in number of variables, computational costs, and mechanisms they explicitly represent), and vary in their estimation and projection approaches. This heterogeneity complicates the task of specifying standard assessment dimensions applicable to all models: features relevant for some model types, e.g. whether hospital capacity allocation is explicitly modeled, may be relevant for a mechanistic model but arguably irrelevant for a prediction model using neural networks. Moreover, as one zooms into modeling details the number of features that could potentially associate with predictive performance proliferate. Any statistical analysis associating performance of some sixty models with dozens of features is likely to generate spurious correlations rather than generalizable insights. A more pragmatic challenge in identifying relevant features of the models is the large variation in quality of documentation for the existing models: while some include extensive documentation and opensource codebases, others may have only a few lines of explanation on the underlying model and estimation architecture. With these challenges in mind we focused on identifying high level architectural features that could be identified for most models on the hub. Specifically, we coded each model into one of the four categories of Mechanistic compartmental (with and without state resetting), Non-mechanistic, Ensemble (of multiple models), and other models (the later included two individual-level models, only one providing death projection). Other coded features were data inputs, output variables, approach to project future trajectory of the reproduction number or policy interventions, and capturing mobility, as well as general information such as modelers’ affiliation (academic or non-academic) technical background, and the availability of technical documentation. For mechanistic models with adequate documentation, we looked into the details of the compartmental structure in terms of details and whether they include coupled compartments with commuting across regions or age-related structures. Furthermore, we documented their parameter estimation approaches. For non-mechanistic models with adequate documentation, we separated simple regression models from more sophisticated curve-fitting approaches and machine learning techniques. For ensemble models we coded the type of models used in the ensemble. Appendix 1 provides additional details.

A New Model for Predicting COVID-19 Deaths- In designing a model to test the proposed features for better predictions we only incorporate those features and exclude any others common to predictive models of COVID-19 which may enhance predictions from other channels. Appendix 4 provides a more complete listing of simplifications of SEIRb; a few include: 1) We keep the model very simple (only the four states of Susceptible, Exposed, Infectious, and Removed are modeled for each state/territory in the USA) ignoring differences in acuity, symptoms, hospitalization, age distribution, travel networks, or disaggregation to county level. 2) While more sophisticated methods (such as Kalman and Particle filtering) exist, we heuristically reset state variables once just before the start of projection. 3) We only use data on deaths and cases, ignoring data sources on hospitalization, age distribution of cases, and mobility patterns among others. 4) Calibration is used to find a handful of model parameters for each state and estimation/projection date independent of findings from other locations or projection dates. 5) We do not fine-tune the uncalibrated parameters of the model (e.g. disease duration), or the model structure, to enhance predictive power nor do we change model structure across states or over time. Noting the role of seasonality among better predictive models in the CDC set we do include input for the impact of weather on transmission rates from research that was publicly available since May 2020 [22] and thus usable in fair comparisons across models.

With these guidelines in mind, we start with the basic SEIR model and add to it a seasonality factor (w) and

Electronic copy available at: https://ssrn.com/abstract=3906690
responds to perceived risk \((\beta': a \text{ first order lag of per capita death rate})\) and includes two free parameters \(\alpha\) and \(\gamma\):

\[
\beta = \beta_0 w \frac{1}{(1 + \alpha f')^\gamma}
\]

We include two lag times for upward and downward adjustment of perceived risk reflecting the possibility that individuals and governments respond to increasing risk levels faster than they abandon precautions when deaths go down. These lag times along with \(\beta_0, \alpha, \gamma\), and infection fatality rate are the primary parameters we estimate for each location and calibration date. Figure 1 provides a graphical overview of the model and full equation listing is in the online Appendix 4.

![Figure 1 - A conceptual representation of SEIRb.](https://ssrn.com/abstract=3906690)

Estimation is pursued through maximum likelihood, using a Negative Binomial likelihood function to match data and simulations for daily deaths and cases. The model does not project cases in future, only using case data to find more accurate estimates of historical transmission intensity, which then informs the estimation of the behavioral response function. State-resetting happens once at the end of the estimation period, when the state variables for exposed and infected are reset to values consistent with the recent death data and death and case trends (see Appendix 4 for details).

**Results**

**Models Architectures and Prediction Accuracy**- Models on the CDC Hub (CDC Model Set) varied on several architectural features (Figure 2-A, see Appendix 1 for more details). Of more than seventy modeling groups listed on the CDC’s website, about half used conventional SEIR-based (mechanistic) architectures. These models are distinguished by their use of variables and mechanisms that mimic the corresponding real-world processes. They varied in their aggregation assumptions, spanning simple SIR (three compartment models) to those incorporating documented vs. undocumented cases, details of hospitalization and intensive care, disaggregation to county level, and commuting patterns across different locations. Only two models used individual-level disaggregation. For calibration most models used death data (for death projections), many also used case data, and some utilized testing, mobility, and hospitalization data as well. Mechanistic
models endogenously simulate estimates for the measured variables which could inform both historical comparisons and future projections. They could also reset the state variables in the model at some regular interval to better correspond to measured data. Such “state-resetting” is especially important at the start of projection, aligning the state variables in the simulation model with observed history to date. State-resetting was pursued by 9% of models, either explicitly (e.g. using heuristics or more formal methods such as Kalman filtering) or implicitly when the mechanistic model informed regressions that utilized historical data to estimate the parameters of a dynamic model. 15% of models included coupled mechanistic structures for different age-groups to capture variation in transmission intensity and fatality, or for different regions to represent spatial dynamics of the spread of the disease. Calibration of mechanistic models to data spanned Bayesian methods, heuristic error terms, and machine learning approaches.

About 38% of modeling groups relied on models that lacked a mechanistic representation of the spread of an epidemic; we label these as non-mechanistic models that incorporated different curve-fitting approaches such as time-series models to fit a function of time to the daily cases or deaths, linear or non-linear regression models which could include regional variables, temperature, and COVID-related information on a daily basis, and non-parametric machine learning models. About 12% of the modeling groups employed ensemble models combining results of several non-mechanistic and/or mechanistic models. Some of the teams changed their modeling approach overtime, in which case we are considering their most recent approach in our analysis.

Next, we focused on the 61 models that offered death projections, and systematically investigated the associations between forecast accuracy and methodological approaches using regressions. We used the absolute prediction error normalized by a location’s (state or national) population as the primary measure of prediction quality, and compared each type of the model in Figure 2-A with a constant model using linear regressions with location-forecast date-forecast horizon fixed effects (see details in Appendix 2). Here a constant model is a naïve predictor which projects all future deaths to be the same as the true deaths last observed. On average forecast accuracy deteriorated by 8% for each week extension of the forecast horizon but much heterogeneity in deterioration rates existed across models. Figure 2-B summarizes the results of regressions. In the short-term non-mechanistic and ensemble models perform better than compartmental models that do not benefit from state-resetting; but the orders reversed beyond 4-5 weeks of projection horizon. Many models outperformed the constant model in the mid to long terms, and on average compartmental models with state-resetting outperformed all others in both the short- and the long-term. The differences were statistically significant (see Appendix 2 for full regression reports).

These results may be intuitive: mechanistic compartmental models use structural features of the phenomenon to zero in on the more promising regions of model specification space; absent those constraints the space of possible models is extremely large. While the flexibility afforded by non-mechanistic models may enable good short-term predictions, predicting the longer-term dynamics becomes increasingly hard and prone to the risk of overfitting. State-resetting enhances predictions in compartmental models by reducing the gap between model and data due to accumulation of errors (between model and reality) prior to projection start. We therefore hypothesize that the use of compartmental models and state resetting enhance the quality of longer-term predictions.
We developed a third hypothesis based on the qualitative review of the more accurate mechanistic models (see Appendix 1 for details). Specifically, a key issue in projection is how transmission rates change in the coming weeks in light of changing policies and individual behaviors. Several models with good prediction performance leveraged data on government policies and mobility patterns to fit historical transmission rates.
We design a simple predictive model that incorporates the three promising features: it is mechanistic (following the SEIR framework); it uses state-resetting before projection; and it incorporates the behavioral feedback loops where NPIs and thus transmission rates change endogenously in response to the perceived risk of death. We include in this model a seasonality factor, another feature common to the top forecasting models in the CDC set. The model excludes many other promising features and is not fine-tuned for this forecasting task; as such it is indicative of what can be achieved with only the hypothesized features but leaves much room for improvement. We call this model SEIRb to highlight the importance of endogenous behavioral responses to risk (see Methods section, and Appendix 4). We then compare the predictive power of this model, and its variants that omit each feature (state resetting (-NoRst), behavioral response (-NoB), and weather effect on transmission (-NoW)) to quantify the importance for prediction accuracy of each feature. We also compare the baseline model with existing models on CDC forecast hub to inform the overall value of such a simple model in comparison with a host of alternatives.

Figure 3 shows the results. Panel A shows a sample of national predictions of SEIRb (obtained by summing up across all state projections) starting from a few different projection times. In panel B we compare the predictive power of different variants of SEIRb model. Every Sunday from 5/3/2020 until 3/13/2021 (total of 46 projection dates) each model, calibrated to data until the previous day, provides 1 to 20-week-ahead weekly death projections for all USA states and territories with more than 200,000 population (53 locations). We compare these projections with actual data for deaths for each location until 5/17/2021, a total of 46,116 weekly death projections for all USA states and territories with more than 200,000 population (53 locations). Every Sunday from 5/3/2020 until 3/13/2021 (total of 46 projection dates) each model, calibrated to data until the previous day, provides 1 to 20-week-ahead weekly death projections for all USA states and territories with more than 200,000 population (53 locations). We compare these projections with actual data for deaths for each location until 5/17/2021, a total of 46,116 weekly death projections for all USA states and territories with more than 200,000 population (53 locations).
Figure 3
Comparison of forecasting quality among different models. A) SEIRb national forecasts at a few points in time. B) Forecasting error per capita for SEIRb and its variants (without seasonality: -NoW; without behavioral feedback: -NoB; without resetting: -NoRst) compared with the median ensemble forecast from CDC model set and SEIRb group. C) Forecast quality ranks for CDC model set and SEIRb based on regressing Ln(Per capita projection error) against models, controlling for location-horizon-week combinations. Color codes: compartmental models without state-resetting (blue); with state-resetting (black); non-mechanistic (red); agent-based (green); and ensemble (yellow).
In panel C we compare SEIRb against all models in the CDC group that include at least 50 death predictions for a given projection horizon. We rank models based on each model's fixed effect in regressing the (log of) per capita absolute error (of death predictions). Not every model provides projection for every location, week, or horizon. Thus, a fair comparison across models requires controlling for idiosyncratic challenges of forecasting any specific location, on a specific date, and for a given horizon. We include a control for each location-horizon-projection date combination ensuring fair comparisons across models. That said, results are qualitatively robust to alternative performance measures such as head-to-head win fraction and errors normalized against constant model (those are reported in online Appendix 3 along with raw population-normalized errors). About half of the models stop projections at 4-week horizon, making it impossible to include their their longer-term performance in comparisons. Using this measure, in the short-run the models “UMass-MechBayes”, “YYG-ParamSearch”, “USC-SI-kjalpha”, and “Karlen-pypm” are among top performers, while SEIRb remains in the top ten. Top performing longer-term projections are offered by “IHME-CurveFit” and “YYG-ParamSearch”, both mechanistic models with state-resetting, with SEIRb offering the second-best performance for horizons over 5 weeks. Note that only a handful of models have consistently submitted very long-term (>12 weeks) projections so those rankings are not particularly informative yet comparisons against the constant model (see Appendix 3) show that SEIRb predictions remain informative all the way through 20-week ahead. Note that our focus in this paper is not on comparing individual models in the CDC set so we only report further performance measures on individual models’ in the online appendix.

**Discussion**

Forecasting over long horizons is hard; and yet extended forecasts, whether from implicit mental or explicit formal models, are indispensable for many individual and policy choices. Epidemics are complex forecasting problems [15, 16]; they include reinforcing contagion mechanisms that create exponential growth, are moderated by randomness in environmental and social determinants of transmission, and are subject to endogenous human responses to evolving risk perceptions. The high stakes of forecasting during the COVID-19 pandemic has created a unique opportunity to better understand how we can enhance forecasts of contagion dynamics. Correlating the features of the CDC hub’s predictive models for COVID-19 deaths in the USA against their forecast accuracy, we propose that mechanistic models that reset state variables at projection time are promising for epidemic forecasting. Critical in the performance of models is how they predict changes in behaviors and NPIs that condition transmission rates. Furthermore, including an impact of weather factors on virus transmission can notably improve COVID-19 projections. We provide evidence in support of these hypotheses by building a simple model that incorporates (only) these features and offers predictions on par with the best models in the CDC set.

Among these features, capturing the endogenous changes in transmission rates offers the largest benefit (see Figure 3-B), and is arguably the least appreciated among the models we reviewed. While several approaches have been used to exogenously predict future changes in transmission rates, except for one model in the CDC set [23], we could not establish that any other models had captured this feedback process endogenously. This observation is somewhat surprising, as the importance of behavioral feedbacks in dynamics of contagion has been known both from historical analysis [29] and in the modeling literature [18, 26, 30, 31] and is at the heart of models that integrate economic agent into contagion dynamics [24, 25, 32]. In fact, that feedback provides the primary path to replicating the observed waves of the epidemic (without resorting to exogenous drivers such as weather or holidays): a rise in deaths elevates risk perception, strengthens NPIs, and thus brings down cases, until the now reduced deaths bring down perceived risk, relaxing NPIs and starting the next wave. Overall, endogenizing behavioral responses to perceived risk points to a great opportunity for enhancing predictive models in the context of epidemics.
The behavioral response feedback also informs the surprisingly good performance of a naïve, constant model which outperforms many models in the CDC set over the longer forecast horizons (See Figure 2-B and Appendix 3). Specifically, cases will grow in a community until this feedback brings down transmission rates and cuts reproduction rate below one; on the other hand, a reproduction rate well below one will not be sustained for long: dwindling cases will alleviate risk perception and increase community interactions, raising the transmission rate. So, this feedback process creates a natural attractor for the system at reproduction rate of one. An expected reproduction rate of one entails limited change in the number of deaths in the coming weeks, i.e. the prediction we used for the naïve, constant model.

Are long-term forecasts of epidemics uninformative as some have suggested [16, 17]? Answering this question requires settling on a benchmark for what constitutes an informative forecast. For example, we find that the gap in predictive power across models expands as we look at longer horizons (See Appendix 3). Thus, assuming the benchmark of a typical model, one could argue that predictions from more accurate models become more informative over longer horizons. On the other hand, while the best performing models beat the constant model across all horizons, the gap between the two does not increase beyond 12-week horizons, suggesting an upper bound on the value of long-term predictions beyond a cognitively simple forecast heuristic. Overall, we find no evidence to suggest a decline in the informativeness of (good) forecasts within a 20-week horizon.

Results should be interpreted in light of a few caveats. First, epidemics across US states have been interdependent and therefore prediction performances are correlated: a model that gets the Fall 2020 peak right may do well in our comparisons, but that may be due to luck rather than an intrinsically valuable feature of that model. Indeed, we noted that comparisons of predictive power over dozens of data points, or data from single locations or limited time windows, could be misleading. Second, comparisons using other measures could offer different insights; for example, prediction confidence intervals are informative but we have not used them. Third, we used projections from before April 2020, when vaccination and new variants had not yet changed death trajectories qualitatively; SEIRb would need additional details to account for those factors to offer continued quality forecasts. Fourth, we have not assessed projections for cases, a task that may require more complex models due to variable ascertainment rates across time and states. Fifth, SEIRb was made after the fact, benefiting from various hindsight biases the authors may not be aware of; on the other hand it excludes many features that could be valuable in forecasting performance (See Appendix 4), as such, its primary role is in offering a baseline for building better predictive models.

These results are most informative for epidemic forecasting. Extrapolating, one may speculate that bringing together the physics of a problem with behaviorally realistic representations of human decision making can provide useful insights for predictive models beyond epidemics. Specifically, from climate change and demographic projections to economic planning, feedback processes crossing physical and biological systems and human choice are regulating long-term dynamics. Predictive models that explicitly capture these feedbacks (rather than using exogenous predictions for one to drive the other) may offer promising avenues to enhance longer-term predictions.
Acknowledgements

The authors received no specific funding for this work. We are thankful to Simon Levin (Princeton University), Mohammad Jalali (Harvard University), Paulo Goncalves (Universita della Svizzera italiana), Jeroen Struben (emlyon business school), Bob Pagano, Ajitesh Srivastava (USC), Richard Larson (MIT), Jack Homer (Homer Consulting), Youyang Gu (MIT), Tse Yang Lim (MIT), Dean Karlen (University of Victoria), and John Sterman (MIT) for their thoughtful feedback on earlier drafts of this paper.

Contributions

Study design: HR, NG; Model conceptualization: HR, NG; Coding: HR; Simulation analysis: HR; Data collection and curation: RX, NG; Statistical analysis: RX; Writing: HR, RX, NG.

Code availability statement

The code to replicate the analyses and figures in the manuscript and Supplementary Information is provided.

Competing interests

The authors declare no competing interests.

Supplementary information

The online version contains supplementary material available.
References

1. Armstrong, J.S., *Principles of forecasting: a handbook for researchers and practitioners*. 2001, Boston, MA: Springer Science & Business Media.
2. Hofman, J.M., A. Sharma, and D.J. Watts, *Prediction and explanation in social systems*. Science, 2017. 355(6324): p. 486-488.
3. Forrester, J.W., *System dynamics—the next fifty years*. System Dynamics Review, 2007. 23(2-3): p. 359-370.
4. Salganik, M.J., et al., *Measuring the predictability of life outcomes with a scientific mass collaboration*. Proc Natl Acad Sci U S A, 2020. 117(15): p. 8398-8403.
5. Holmdahl, I. and C. Buckee, *Wrong but useful—what covid-19 epidemiologic models can and cannot tell us*. New England Journal of Medicine, 2020. 383(4): p. 303-305.
6. Dehning, J., et al., *Inferring change points in the spread of COVID-19 reveals the effectiveness of interventions*. Science, 2020. 369(6500): p. eabb9789.
7. Ferguson, N., et al., *Report 9: Impact of non-pharmaceutical interventions (NPIs) to reduce COVID19 mortality and healthcare demand*. 2020.
8. IHME and C.J. Murray, *Forecasting COVID-19 impact on hospital bed-days, ICU-days, ventilator-days and deaths by US state in the next 4 months*. medRxiv, 2020.
9. Kissler, S.M., et al., *Projecting the transmission dynamics of SARS-CoV-2 through the postpandemic period*. Science, 2020. 368(6493): p. 860-868.
10. Wu, J.T., K. Leung, and G.M. Leung, *Nowcasting and forecasting the potential domestic and international spread of the 2019-nCoV outbreak originating in Wuhan, China: a modelling study*. The Lancet, 2020. 395(10225): p. 689-697.
11. Lalmanawma, S., J. Hussain, and L. Chhakchhuak, *Applications of machine learning and artificial intelligence for Covid-19 (SARS-CoV-2) pandemic: A review*. Chaos, Solitons & Fractals, 2020. 139: p. 110059.
12. Chimmula, V.K.R. and L. Zhang, *Time series forecasting of COVID-19 transmission in Canada using LSTM networks*. Chaos, Solitons & Fractals, 2020. 135: p. 109864.
13. Ioannidis, J.P., S. Cripps, and M.A. Tanner, *Forecasting for COVID-19 has failed*. International journal of forecasting, 2020.
14. Chin, V., et al., *A case study in model failure? COVID-19 daily deaths and ICU bed utilisation predictions in New York State*. European Journal of Epidemiology, 2020. 35(8): p. 733-742.
15. Cirillo, P. and N.N. Taleb, *Tail risk of contagious diseases*. Nature Physics, 2020. 16(6): p. 606-613.
16. Castro, M., et al., *The turning point and end of an expanding epidemic cannot be precisely forecast*. Proceedings of the National Academy of Sciences, 2020. 117(42): p. 26190-26196.
17. Wilke, C.O. and C.T. Bergstrom, *Predicting an epidemic trajectory is difficult*. Proceedings of the National Academy of Sciences, 2020. 117(46): p. 28549-28551.
18. Rahmandad, H., T. Lim, and J. Sterman, *Behavioral dynamics of COVID-19: estimating underreporting, multiple waves, and adherence fatigue across 92 nations*. System Dynamics Review, Forthcoming.
19. Cramer, E.Y., et al., *Evaluation of individual and ensemble probabilistic forecasts of COVID-19 mortality in the US*. medRxiv, 2021.
20. Ray, E.L., et al., *Ensemble Forecasts of Coronavirus Disease 2019 (COVID-19) in the us*. MedRXiv, 2020.
21. Dong, E., H. Du, and L. Gardner, *An interactive web-based dashboard to track COVID-19 in real time*. Lancet Infect Dis, 2020. 20(5): p. 533-534.

Electronic copy available at: https://ssrn.com/abstract=3906690
22. Xu, R., et al., *Weather Conditions and COVID-19 Transmission: Estimates and Projections*. medRxiv, 2020.
23. Reiner, R.C., et al., *Modeling COVID-19 scenarios for the United States*. Nature Medicine, 2021. **27**(1): p. 94-105.
24. Morin, B.R., et al., *Economic Incentives in the Socially Optimal Management of Infectious Disease: When $R_0$ Is Not Enough*. EcoHealth, 2018. **15**(2): p. 274-289.
25. Perrings, C., et al., *Merging Economics and Epidemiology to Improve the Prediction and Management of Infectious Disease*. EcoHealth, 2014. **11**(4): p. 464-475.
26. Struben, J., *The coronavirus disease (COVID-19) pandemic: simulation-based assessment of outbreak responses and postpeak strategies*. System Dynamics Review, 2020. **36**(3): p. 247-293.
27. Reich, N.G., et al., *A collaborative multiyear, multimodel assessment of seasonal influenza forecasting in the United States*. Proceedings of the National Academy of Sciences, 2019. **116**(8): p. 3146-3154.
28. Reich, N.G., et al., *Accuracy of real-time multi-model ensemble forecasts for seasonal influenza in the U.S*. PLOS Computational Biology, 2019. **15**(11): p. e1007486.
29. Bauch, C., A. d’Onofrio, and P. Manfredi, *Behavioral epidemiology of infectious diseases: an overview*. Modeling the interplay between human behavior and the spread of infectious diseases, 2013: p. 1-19.
30. Rahmandad, H. and J. Sterman, *Heterogeneity and network structure in the dynamics of diffusion: Comparing agent-based and differential equation models*. Management Science, 2008. **54**(5): p. 998-1014.
31. Ghaffarzadegan, N. and H. Rahmandad, *Simulation-based estimation of the early spread of COVID-19 in Iran: actual versus confirmed cases*. System Dynamics Review, 2020. **36**(1): p. 101-129.
32. Farboodi, M., G. Jarosch, and R. Shimer, *Internal and external effects of social distancing in a pandemic*. 2020, National Bureau of Economic Research.
Online supplementary

Online supplementary material to accompany:

Enhancing Epidemic Forecasting: Learning from COVID-19 Models

By Hazhir Rahmandad (hazhir@mit.edu), Ran Xu (ran.2.xu@uconn.edu), Navid Ghaffarzadegan (navidg@vt.edu)

Contents

Appendix 1: Coding the CDC hub models ................................................................. 1
I - Coding details ........................................................................................................ 1
II - Common features of the models ......................................................................... 4
III - Features of top performing models .................................................................... 5

Appendix 2 - Details of Statistical Analyses ............................................................. 7

Appendix 3: Comparison of models based on other measures ................................. 10

Appendix 4: Model Documentation ......................................................................... 15
I - Model formulation ............................................................................................... 15
II - Model Calibration .............................................................................................. 16
III - State resetting .................................................................................................. 17
IV - Important simplifications and improvement opportunities ............................. 18

Appendix 5- Data and Replication Instructions ......................................................... 20

Appendix 1: Coding the CDC hub models

I - Coding details

A primary list of models that contributed to the CDC Covid forecast hub was obtained from CDC’s website (https://www.cdc.gov/coronavirus/2019-ncov/cases-updates/forecasts-cases.html). The list was verified by comparing with two pre-print manuscripts on medRxiv co-authored by the contributors to the hub ([1, 2]). The list was then compared with available forecast data on the hub, and as a result five missing models were added to the list, yielding a total of 74 models. We then narrowed down the list to the models that provided COVID-19 death forecast, which included a large majority of the models (n=61). The final list of models included:

Alpert-pwllnod, BPagano-RtDriven, Caltech-CS156, CEID-Walk, CMU-TimeSeries, 
Columbia_UNC-SurvCon, Covid19Sim-Simulator, CovidActNow-SEIR_CAN, CovidAnalytics-
DELPHI, COVIDhub-ensemble, CU-select, DDS-NBDS, epiforecasts-ensemble1, Geneva-
DetGrowth, Google_Harvard-CPF, GT_CHHS-COVID19, GT-DeepCOVID, IEM_MED-
CovidProject, IHME-CurveFit, IowaStateLW-STEM, IUPUI-HkPrMobiDyR, JCB-PRM,
JHU_CSSE-DECOM, JHU_IDD-CovidSP, JHUAPL-Bucky, Karlen-pypm, LANL-GrowthRate,
LNQ-ens1, Microsoft-DeepSTIA, MIT_CritData-GBCF, MIT_ISOLOT-Mixtures,
MITCovAlliance-SIR, MOBS-GLEAM_COVID, MSRA-DeepST, NotreDame-FRED, NotreDame-

Electronic copy available at: https://ssrn.com/abstract=3906690
Two researchers (NG and RX) analyzed the models based on any available information, and coded their methodological approaches. Specifically detailed notes were taken about modeling approaches based on documentations on websites, related journal publications, and in a few cases upon contacting modelers with clarifying questions. Other important sources of information included a webpage (https://zoltardata.com/project/44/forecasts) which includes self-reported brief information (about 1-2 paragraphs) on methodological approaches of each model. This website particularly helped with several models that lacked any further technical documentation. We also consulted the information on a related GitHub repository (https://github.com/cdcepi/COVID-19-Forecasts/blob/master/COVID-19_Forecast_Model_Descriptions.md). We further checked the modelers’ websites, blogs, or twitter links, for updates, possible changes in methods, or more methodological details. A few groups changed their models throughout the forecast for which we considered their most recent approach in our coding.

The primary coding question was related to the modeling approach. Initially five mutually exclusive and exhaustive groups of modeling approaches were identified and used to categorize the models:

1) Mechanistic compartmental models: this is the conventional approach in epidemiology to model the spread of an infectious disease in which the population is represented by different compartments. A common example is the Susceptible-Exposed-Infected-Removed (SEIR) model.

2) Non-mechanistic models: these models do not capture the physics of the spread of the disease, and instead, by using different data-analytic approaches, try to uncover association between death incidents and other variables. These models include different forms of parametric and non-parametric regression models and machine learning techniques.

3) Ensemble models: these models provide estimates based on combining two or more distinct models’ (which could be each mechanistic and non-mechanistic) forecasts.

4) Others: Only two models that used agent-based modeling approaches belonged to the “others” group. We thus name the group agent-based models. These models are mechanistic and representing individuals explicitly rather than lumping them together in mixed compartments.

Among the first group of mechanistic compartmental models, we further categorized them into mechanistic models with state-resetting vs mechanistic models without state-resetting based on how simulation outcomes were combined with the data. A sub-group of mechanistic models used state-resetting procedures to improve their forecast accuracy. Simply put state-resetting is a procedure to combine simulation outcomes from the model with observed data to come up with more accurate values for the state variables in the model, and then to reset the state variables to those more likely values which would potentially enhance the quality of both parameter estimation and predictions. An example of state-resetting is an SIR-based model that periodically estimates the number of active cases from the case data, feeds it to the “I” variable, and then simulates the model for the purpose of projection. In this example one could also use a backward estimation of active cases based on reported death. In fact a more sophisticated method could combine both estimates into a better estimate for “I.” While explicit and structured methods for state-resetting, such as Kalman and Particle filtering, are well known, for simpler models one can use simple heuristic resetting with much lower computational costs. Moreover, a few models did state-resetting implicitly. For example, they estimated a regression model that was based on an SEIR-type mechanistic formulation. Such regressions would calculate the state variables based on observed data every period and as such are doing state resetting implicitly.
We further examined models’ structures looking into methodological details. A major challenge however was the large variation of the quality of documentation of the models. While some of the models had reported sufficient details for replication of their models and findings, others may only had short documents, or a few lines of explanations about the underlying models and estimation techniques. Nevertheless we coded the models based on:

- data inputs:
  - variable type (e.g., death data, case data, hospitalized data);
  - approach to use data: data as an exogenous input vs. data for model calibration;
- output variable:
  - the main predictions of the models (e.g., death, case);
  - the time horizon of the predictions;
- approach to estimate transmission intensity:
  - is transmission intensity constant or changing;
  - are they modeling social distancing explicitly or implicitly, and if so how;
- approach to project future trajectory of transmission intensity:
  - are they assuming transmission intensity (and the reproduction number) is going to stay constant, or change;
  - if changes, do they model change in transmission intensity (and the reproduction number), or only do scenario analysis (constant varying transmission intensity);
  - if they model change in transmission intensity (and the reproduction number), does it include an endogenous mechanism or it is an exogenous time series based on expected time to reopen;
- modeling mobility:
  - are they modeling change in mobility;
  - are they using mobility data;
- General information such as:
  - modelers’ affiliation (academic or non-academic);
  - technical background; and
  - the availability of technical documentation.

Furthermore, for different methodological approaches we specifically looked for the following criteria:

- For mechanistic models with adequate documentation:
  - details of the compartmental structure: compartments (Simple SIR vs. SEIR vs. more compartments for capturing different stages of illness and symptoms);
  - do they include coupled age-structure;
  - do they have a coupled compartments with commuting across regions;
  - parameter estimation (model calibration):
    - sources of parameter values;
    - do they calibrate their model with the data, and if so what is the payoff function and methods to find optimal parameters;
  - weather impact:
    - do they include any estimate of weather impact on transmission intensity or the reproduction number;
- For non-mechanistic models with adequate documentation:
  - Specifics of the method:
    - From simple regression models to more sophisticated curve-fitting approaches and machine learning techniques;
Online supplementary

- weather impact:
  - do they include any estimate of weather impact in their model;
- For ensemble models:
  - the type of models used in the ensemble.

Moreover, we made note of any interesting observation such as change in method of forecast and models or change in parameter values or attempts for fine-tunings.

After NG and RX independently coded the models, they shared and discussed their results. The initial inter-rater reliability (percent agreement between the two raters) was 90%, high enough that did not require any changes in the coding process. The coders converged on the final results after a discussion and those results inform the relevant regressions. All three authors discussed major lessons learned through reading the documents.

II- Common features of the models

A few initial observations were noteworthy for the research team:

1) **Only two (<4%) models used agent-based architectures.** In contrast to our initial expectation, only 2 models used agent-based individual-level approaches, and they seemed to have stopped projecting after a few rounds. Only one of them provided death projection. On the other hand, the majority of the models preferred to model at US state- or county-levels, using compartmental or non-mechanistic approaches. Lack of ABM approaches may partially be explained by the computational costs of these methods in light of the calibration requirements and large parameters spaces they typically include.

2) **About 38% of the models used non-mechanistic approaches.** With the growing attention to data-driven methods across various fields we observed a considerable number of non-mechanistic models. Particularly about 16% of the models used machine-learning techniques for projection confirming a growing trend in the application of AI. Many of these models were developed by computer science and engineering researchers.

3) **About half of the groups used conventional SIR-like models with modest modifications.** Given the growing alternatives for modeling the dynamics of transmission it was interesting that still many modelers start with the classical architectures. The prevalence of S(E)IR models, some including more details about asymptomatic cases or hospitalized cases, and a few using detailed coupled compartmental structures where people travel between different regions puts these methods at the heart of the existing approaches.

4) **Among mechanistic models, the majority used simple techniques for parameter estimation.** Most mechanistic models tried to utilize recent documented measures about COVID-19 (such as infection fatality rate or the disease duration) from other research publications. They then estimated a few unknown parameters such as the basic reproductive number (or transmission intensity), often by fitting the simulation with data in a nonlinear optimization. The process of parameter estimation was often simple, minimizing the mean square error between simulation and data on daily or weekly deaths/cases. The search strategy for optimal parameters ranged from simple algorithms to more advanced machine-learning techniques. Only a handful of groups used more sophisticated estimation approaches with explicit likelihood functions and state-resetting (e.g. Markov Chain Monte Carlo simulations and Kalman Filtering).
For fitting simulation with the historical trends, mechanistic models commonly considered non-constant reproductive numbers. A large number of models tried to incorporate change in the reproductive number (or transmission intensity of $\beta$). Some of them used different mobility data, and estimated change in transmission intensity as a function of change in mobility. Others used estimation of the reproductive number from daily cases. A few groups used data on when each US state started their social distancing policies. Such data were fed into the model to better estimate change in the reproductive number. For example, a few models assumed a specific percentage decline in the reproductive number after implementation of lockdown policies.

For the purpose of projection, mechanistic models commonly assumed constant reproductive numbers. Most models lacked techniques of projecting the reproductive number (or transmission intensity of $\beta$). A large majority used their latest estimate of the reproductive number from the past data for projecting the future cases.

Modelers updated their models through the course of the pandemic. Like any other social setting, modelers learned from the past projections and tried to incorporate new ideas to improve their future projections. Several of them updated their parameter values as more data became available about the nature of the disease. A few groups dropped out after a few projections, and a few other joined the hub several months after the starting date. We noted that a few groups changed their modeling approaches too. The common direction of changing modeling methods was from curve-fitting to mechanistic compartmental models.

III- Features of top performing models

Our primary analysis uses consistent coding applicable across all models. Given the significant heterogeneity in the documentation of CDC model set this analysis does not inform more detailed features of models beyond a few aggregate categories. We therefore studied the top 3 models in the long-term prediction performance in more depth (short-hands: IHME, YYG, BPagano) to learn about more specific features that might have improved performance beyond those measurable across all models.

Importantly, we noted that the assumption of constant vs. changing reproductive number is essential in long-term projections. Among mechanistic models the challenge of modeling a pandemic primarily boils down to the prediction of societal reactions and policy decisions. Two particular models of YYG and IHME are good examples: the former used available reports on states’ plans for opening and the modelers’ best judgment for extrapolating those in future. Specifically, YYG estimated the reproductive number ($R(t)$) by four main values of ($R_0$, $R$(post-mitigation), $R$(post-opening), $R$(equilibrium [sometime after post-opening])=$\approx 1$), and used a sigmoid function for the transition between $R_0$, and $R$(post-mitigation), and possibly other $R$-values, where the slope of the function was also estimated through model calibration. The mitigation and opening were based on a New York Times dataset. IHME used a more detailed approach: data on state policies (severe travel restrictions, closing of public educational facilities, closure of nonessential businesses, stay-at-home orders, and restrictions on gathering size) were gathered from press release or state government official orders. The model then estimated the policy effects on mobility and their effect on transmission intensity. In addition, IHME modeled policy change endogenously: when a threshold of 8 death per million was passed social-distancing measures were reactivated [3].

Another observation was about state-resetting techniques. For example, in the BPagano the number of daily infections was estimated by shifting daily death backward, and dividing it by the most recent estimate of the infection fatality rate. Then the current active cases $(I)$ was estimated as the sum of daily infections for the duration of the infectious period. The IHME model took a similar approach by using death-based estimation of daily infection as data inputs (rather than simulated outcomes) in the SEIR model.
Some other factors were also noted in the models. High-performing models incorporated the weather effect. IHME for example used flu season as factor in modeling transmission intensity. Moreover, YYG modeled lockdown fatigue which considers that \( R(\text{post-mitigation}) \) may increase before opening. This model also considered change in infected fatality rate which might be due to healthcare systems’ learning over time or changes in composition of infected towards younger cohorts. Such mechanisms are potentially helpful in better projections.
Appendix 2- Details of Statistical Analyses

Analysis of CDC models

We used data from the Center for Disease Control (CDC) repository of COVID-19 projections, which included 490,210 point forecasts for weekly death incidents across 57 locations, forecast dates over the span of a year (4/13/2020 to 3/29/2021), 20 forecast horizons (1-week-ahead predictions to 20-week-ahead predictions), and 61 models. We chose the normalized/per-capita absolute prediction error as the basis of comparison for model performance, i.e. the absolute difference between a model’s prediction of weekly incident death and the true weekly incident death, divided by location’s population. As this measure is highly skewed we log-transformed the measure and included 1-99 percentile of the data for further analysis. We further excluded two agent-based models and COVID-hub ensemble (the ensemble of all other eligible CDC models) from the analysis, which resulted in a final sample with 463,305 predictions made by 58 models for each state of the United States, with target end dates ranging from 4/18/2020 to 4/3/2021. We compared how each type of the model performed in each forecast horizon with a constant model – a model that predicts future weekly incident death to be the same as the weekly incident death last observed. Specifically, we included the weekly incident death predictions from the constant model for each unique combination of location, forecast date and forecast horizon in our sample frame as the baseline, and we included each model type (ensemble, non-mechanistic, mechanistic without state-resetting, and mechanistic with state-resetting) as a key independent variable in a linear regression with location-time fixed effects, and we ran separate regression analyses for each prediction horizon. As depicted in Table A1, the coefficient for each model-type represents the average differences in log-transformed normalized error between that model-type and the baseline (constant) model. Results showed that (1) in one-week-ahead predictions the constant model performs better than all other model types on average (but several individual models outperform the constant one); many models and model types outperform the constant model in mid to long term predictions, with mechanistic model with state-resetting performing the best starting from 2-week-ahead predictions; (2) In short-term non-mechanistic and ensemble models perform better than mechanistic models without state-resetting, but that ordering reversed beyond 4-5 weeks of projection horizon. However, on average mechanistic models with state-resetting outperformed all others in both the short- and the long-term. As a robustness check we also included more model characteristics, i.e. whether the modelers are affiliated with academia, whether the model has detailed documentation, and the interaction of the two, as covariates and reran the aforementioned analysis (we excluded the constant model from the analysis and used the mechanistic (excluding state-resetting) model as the baseline model). Results are reported in Table A2 and are consistent with our main results. Interestingly models with documentation performed worse than those without within 1-8 weeks-ahead predictions, but that effect was attenuated (and reversed in 5-8 weeks-ahead predictions) if the modelers were from academia.

Table A1

| Outcome: Log(normalized absolute error) | (1) | (2) | (3) | (4) | (5) | (6) | (7) |
|----------------------------------------|-----|-----|-----|-----|-----|-----|-----|
|                                        | Week 1 | Week 2 | Week 3 | Week 4 | Week 5 | Week 6 | Week 7 |
| Ensemble                               | 0.125*** | -0.00640 | -0.0581** | -0.0449* | 0.0457 | 0.0808** |
|                                        | (0.0170) | (0.0176) | (0.0182) | (0.0186) | (0.0275) | (0.0270) |
| Non-mechanistic                        | 0.141*** | 0.0350** | 0.000288 | -0.00241 | -0.00914 | -0.0298 | 0.0170 |
|                                        | (0.0113) | (0.0118) | (0.0122) | (0.0125) | (0.0179) | (0.0176) | (0.0181) |
| Mechanistic with state-resetting       | 0.0544*** | -0.0751*** | -0.136*** | -0.174*** | -0.171*** | -0.184*** | -0.105*** |
|                                        | (0.0126) | (0.0130) | (0.0134) | (0.0136) | (0.0147) | (0.0143) | (0.0155) |
| Mechanistic (excluding state-resetting)| 0.227*** | 0.0988*** | 0.0214 | -0.00429 | -0.0355** | -0.0580*** | -0.0415** |
|                                        | (0.0110) | (0.0113) | (0.0116) | (0.0119) | (0.0134) | (0.0135) | (0.0153) |
| Constant                               | -12.32*** | -12.02*** | -11.81*** | -11.63*** | -11.51*** | -11.37*** | -11.36*** |
|                                        | (0.00985) | (0.0101) | (0.0104) | (0.0106) | (0.0108) | (0.0105) | (0.0110) |
| Observations                           | 90,453 | 86,459 | 83,279 | 80,676 | 38,180 | 34,576 | 25,398 |
Online supplementary

Table A2

| Outcome: | (1) | (2) | (3) | (4) | (5) | (6) | (7) |
|----------|-----|-----|-----|-----|-----|-----|-----|
| Log(normalized absolute error) | Week 8 | Week 9 | Week 10 | Week 11 | Week 12 | Week 13 | Week 14 |
| Non-mechanistic | 0.0292*** | 0.0629*** | 0.107*** | 0.0823*** | 0.0097*** | 0.122*** | 0.152*** |
| | (0.0193) | (0.0251) | (0.0255) | (0.0255) | (0.0262) | (0.0273) | (0.0288) |
| Mechanic with state-resetting | -0.244*** | -0.231*** | -0.177*** | -0.169*** | -0.128*** | -0.120*** | -0.151*** |
| | (0.0179) | (0.0177) | (0.0179) | (0.0185) | (0.0191) | (0.0205) | (0.0220) |
| Mechanic (excluding state-resetting) | -0.0158 | -0.125*** | -0.0588** | -0.0352 | -0.0071 | 0.0661** | 0.0737*** |
| | (0.0165) | (0.0202) | (0.0210) | (0.0211) | (0.0231) | (0.0248) | (0.0271) |
| Constant | -11.27*** | -11.23*** | -11.22*** | -11.18*** | -11.12*** | -11.12*** | -11.06*** |
| | (0.0120) | (0.0123) | (0.0127) | (0.0129) | (0.0134) | (0.0142) | (0.0153) |
| Observations | 21,339 | 13,545 | 12,534 | 11,743 | 10,791 | 9,709 | 8,931 |
| Within R-squared | 0.014 | 0.020 | 0.015 | 0.013 | 0.010 | 0.012 | 0.018 |
| Number of location-time combination | 2,273 | 2,195 | 2,129 | 1,951 | 1,898 | 1,830 | 1,724 |

Standard errors in parentheses
*** p<0.001, ** p<0.01, * p<0.05
Within R-squared

| Number of location-time combination | 0.018 | 0.016 | 0.018 | 0.016 | 0.012 | 0.021 | 0.027 |
|-------------------------------------|-------|-------|-------|-------|-------|-------|-------|
| 2,271                               | 2,192 | 2,128 | 1,948 | 1,894 | 1,825 | 1,721 |

Standard errors in parentheses
*** p<0.001, ** p<0.01, * p<0.05
Appendix 3: Comparison of models based on other measures

In the paper we reported the model rankings based on regressions conducted for predictions at every projection horizon between 1 and 20 weeks. Those regressions include fixed effects for every combination of location and projection date, ensuring that idiosyncratic challenges in projecting specific locations and weeks is not driving the differences in prediction errors across different models. After controlling for those fixed-effects the coefficient for each model represents the distinct contribution of that model to prediction errors. In fact, most models do not offer projections for every location, prediction date, or horizon, making such controls important for fair comparisons across models. Nevertheless, more direct comparisons of measures of prediction accuracy could inform more familiar ways to read the prediction data, and therefore present three of those comparisons below, followed by a replication of the ranking graph including the model names which were not part of the graph in the main paper. Whereas the ranking graph includes only models with 50 predictions for a given horizon, the three graphs below include all models regardless of number of predictions. This may lead to some outliers, e.g. QJHong-Encounter has submitted fewer than 50 in any horizon, and thus does not show up in the rankings graph, but performs very well where it has submitted a prediction as can be seen in the following graphs.

Head-to-Head Win Fraction

For each location-week-horizon combination a few models may offer predictions, offering opportunities to see how models compare against each other in head-to-head battles. For example, if 5 different models are predicting deaths for the week of March 14, 2021 as part of a 10-week ahead horizon, those comparisons offer 4 win/lose options for each model in that set. A model that wins 3 of those 4 head-to-head comparisons gets a score of 0.75 from this location-week. For each model such win fractions, when averaged across all such comparisons for the 10-week horizon, would inform the quality of the model’s predictions at 10-week ahead horizon. The following graph reports those average fractions across CDC model set and SEIRb family. Using this measure SEIRb outperforms other models in several longer time horizons while QJHong-Encounter (when it submitted a prediction), YYG-ParamSearch and IHME also perform very well (Figure A1).
The next performance measure compares various models against the naïve (constant) benchmark. As discussed in the paper, the constant benchmark is not that naïve after all: it is the straightforward prediction that accounts for endogenous behavioral feedback keeping the reproduction number around 1; it also beats many models both in the short and long-term horizons. Specifically, for each model we go through the following calculations: for each prediction (for a given location, week, and horizon) the per capita error for the constant model is deducted from the model’s per capita error to offer a comparative normalized error;

Figure A1: Comparison of performance of models based on head-to-head win fraction

Normalized Error Relative to Constant Model

The next performance measure compares various models against the naïve (constant) benchmark. As discussed in the paper, the constant benchmark is not that naïve after all: it is the straightforward prediction that accounts for endogenous behavioral feedback keeping the reproduction number around 1; it also beats many models both in the short and long-term horizons. Specifically, for each model we go through the following calculations: for each prediction (for a given location, week, and horizon) the per capita error for the constant model is deducted from the model’s per capita error to offer a comparative normalized error;
then the median across all those comparative errors for each projection horizon is mapped (median is used
given the fat-tailed distribution of these errors). Using this measure IHME and SEIRb are the top performers
in longer time horizons (Figure A2) while Caltech-CS156, MSRA-DeepST, and QJHong-Encounter (when
submitting a prediction), offer the best short-term predictions.

![Normalized Error Relative to Constant Model](image)

Figure A2: Comparison of performance of models based on normalized error relative to constant model

**Absolute Prediction Error (Population Normalized)**
The next detailed graph reports the absolute prediction error (normalized by location populations, in Death/Million/Week). For each model and each horizon, we report the median error across all locations and projection dates for which that model has submitted a prediction. It is noteworthy that this metric leads to somewhat different rankings compared to other measures: because each model has submitted predictions for only a subset of locations, projection dates, and horizons some may be competing on harder forecast tasks than others. This problem was partially addressed in the first two graphs by comparing models against each other (in Win fraction measure) or comparing against a constant model (in the second graph above). It was also more explicitly addressed by including fixed effects (comparing against mean) for each location-projection date-horizon in the primary regressions (used in the main ranking graph).

Figure A3: Comparison of performance of models based on population-normalized absolute error
Online supplementary

**Model Rankings with Model Names**

Here we replicate the primary rankings graph in the body of the paper (Figure 2-C) and include the individual model names which were not included in the primary analysis to avoid focusing on individual model comparisons (which is not the primary purpose of this analysis).

**Figure A4:** Ranking of performance of models based on model fixed effects in regressing LN(Error) against models, controlling for location-horizon-projection date combinations.

Electronic copy available at: https://ssrn.com/abstract=3906690
Appendix 4: Model Documentation

I- Model formulation

For the purpose of parsimony, we develop a very simple model. Consistent with conventional SEIR models, the population \((N)\) is represented in four stocks of Susceptible \((S)\), Exposed \((E)\), Infectious \((I)\), and Removed \((R)\) (eq. 1-4).

\[
\begin{align*}
\frac{dS}{dt} &= -\frac{\beta SI}{N} \\
\frac{dE}{dt} &= \frac{\beta SI}{N} - \frac{E}{\tau_1} \\
\frac{dI}{dt} &= \frac{E}{\tau_1} - \frac{I}{\tau_2} \\
\frac{dR}{dt} &= \frac{I}{\tau_2}
\end{align*}
\]

(1-4)

where \(\beta\) is transmission intensity, \(\tau_1\) is exposure period, and \(\tau_2\) is infection period from symptom onset to recovery or death. In this simple representation, daily death of \(f\) can be represented as a fraction of removal rate, where the fraction, \(i\), is referred as infection fatality rate (eq. 5)

\[
f = i \frac{l}{\tau_2}
\]

(5)

The transmission intensity of \(\beta\) which determines the speed of the spread of the disease, and the reproductive number, should change over time, and in fact that is the main difference between our SEIRb model and others. Specifically, we expect \(\beta\) to decline and people practice more NPIs as perceived risk of death \((f')\) increases, i.e., \(\frac{d\beta}{df'} < 0\). Consistent with the literature that finds change in weather influences transmission [4], we include weather impact of \(w\) in the formulation of transmission intensity. Equation 6 shows how we represented this relation in one of its simplest formats:

\[
\beta = \beta_0 w \frac{1}{(1+af')^\gamma}
\]

(6)

For the weather impact of \(w\), projections from a previous study [4] is used in our comparisons. Specifically, we start with the “Covid-19 Risk factor due to Weather” (CRW) that was publicly released in May 2020, and use a transformation of that factor \((w = CRW^{2.64})\) based on other modeling work [5] that had found the CRW factor to be conservative in reflecting the impact of weather on transmission.

Equation 6 closes a balancing feedback loop from daily death rate to future transmission intensity and consequently future exposure, onset, and death. In this relation, \(f'\) is simply modeled as a lagged variable of \(f\), daily deaths, assuming public risk perception is a lagged function of confirmed death cases. Perception adjustment for increasing and decreasing death may be different, thus we include two lag times for upward \((\tau_U)\) and downward \((\tau_D)\) adjustment of \(f'\), both estimated from model calibration:

\[
\frac{df'}{dt} = \frac{f-f'}{\tau_R}
\]
Online supplementary

\[ \tau_R = \begin{cases} \tau_U & \text{if } f > f' \\ \tau_D & \text{if } f \leq f' \end{cases} \]  

(7)

II- Model Calibration

Model calibration is done separately and independently for each location (53 states and territories of USA with population over 200,000) and estimation/projection date \( T_f \): the end of each Saturday starting on May 2, 2020 and ending on March 13, 2021. Estimation is pursued maximizing the likelihood of observed Cases and Deaths for each location starting from an initial time for data inclusion \( T_0 \): when the official number of cases/deaths respectively exceeds \( 1e^{-6}/1e^{-8} \) per day as a fraction of location’s population, or the beginning of May 2020, whichever comes first) until the estimation/projection date \( T_f \). We use a Negative Binomial likelihood function for both cases and deaths \( (x_{vt}, y_{vt}(\theta)) \): where \( x \) is data, \( y(\theta) \) is model predictions for data given the unknown parameter vector \( \theta \); \( t \) is the day and \( v \in [i, d] \) denotes the cases/deaths; we smooth out weekly cycles by using 7-day moving averages for death data):

\[ L_L(\theta, \lambda_v) = \sum_v \sum_{t=T_0}^{T_f} \ln \left( \frac{1}{1 + \lambda_v x_{tv}} + \Gamma \left( x_{tv} + \frac{1}{\lambda_v} \right) - \Gamma \left( \frac{1}{\lambda_v} \right) - \left( x_{tv} + \frac{1}{\lambda_v} \right) \ln(1 + \lambda_v y_{tv}(\theta)) + x_{tv}(\ln(y_{tv}(\theta)) + \ln(\lambda_v)) \right) \]  

(8)

In this function \( \Gamma(z) \) represents the natural logarithm of the generalized factorial function for \( z-1 \) \( (\Gamma(z + 1) = z! \) for integer \( z \)). Predicted deaths \( (y_{td}(\theta)) \) come directly from the SEIR model described above.

Two additional features inform the estimation process (but not projections, which come purely from the SEIR model described above). First, before the projection date \( T_f \) the perceived risk, \( f' \), uses the actual data \( (x_{td}) \) rather than simulated values for deaths. That is we use the following equation instead of equation 7:

\[ \frac{df'}{dt} = \frac{x_{td} - f'}{\tau_R} \]  

(9)

Second, we use the following equations to predict cases \( (y_{ti}(\theta)) \):

\[ y_{ti}(\theta) = \frac{\beta S_{it} I_{ti}}{N} \]  

(10)

\[ \frac{dI_D}{dt} = x_{ti} - \frac{l_D}{\tau_2} \]  

(11)

Essentially, the data for measured infections flows into a stock \( (I_D) \) that parallels the model-simulated infection rate (that is, it flows out with the same time constant of \( \tau_2 \)), and this stock of “measured” infectious population is used to predict expected “measured” infections based on model generated transmission intensity and susceptible fraction. This approach enables using measured case data to inform the parameters going into transmission intensity (most notably the response function parameters) without worrying about ascertainment rates which likely are far below 100% and vary across locations.

The vector of estimated model parameters \( (\theta) \) and the ranges we use for each in the calibration are listed in Table A3.

| Parameter | Range | Units | Explanation |
|-----------|-------|-------|-------------|
| \( \beta_0 \) | [0.1-4] | 1/Day | Basic Transmission Intensity |

Table A3

Electronic copy available at: https://ssrn.com/abstract=3906690
We also estimate two parameters regulating the shape of the negative binomial distribution \( \lambda \), leading to a total of 9 estimated parameters for each location and estimation/projection date for the SEIRb model. Other variants include the same or fewer parameters (SEIRb-NoB: 5; SEIRb-NoW: 9; SEIRb-NoRst: 9) but otherwise follow the same exact calibration process.

Maximization of the likelihood function in equation 8 is pursued using Powell Direction Set method built into Vensim™ simulation software. For each location we conduct one initial calibration for the last estimation date (March 13, 2021) with 15 different random start points for unknown parameters. For all other estimation dates we use 5 different start points but also include the parameter setting found in the next estimation date. This process enhances our confidence in finding good optimization solutions while keeping the computational costs to a minimum. Overall all the 2436 (=53*46) estimations for SEIRb model could be completed in about 4 hours on a regular desktop when compiled and parallelized over 10 cores.

### III- State resetting

The basic idea of state resetting is to ensure projections start from the right level given the most recent data on cases and deaths. Various data fusion, smoothing, and filtering methods exist to leverage current data to offer good, even optimal, estimates for state variables in a model. Those methods can enhance both model estimation and projections, however, they are computationally expensive and their elaborate setup may mask the basic benefits achievable from more simple state resetting schemes. We therefore opt for using a simpler approach in which only once, at \( T_0 \), we reset the two relevant state variables of \( E \) and \( I \) to their likely values, \( E^* \) and \( I^* \), given recent deaths and cases. Specifically, we use the following equations to calculate \( E^* \) and \( I^* \):

\[
E^* = \frac{x_d'(1+s_E)}{i}\tau_1
\]

\[
I^* = \frac{x_d'(1+s_I)}{i}\tau_2
\]

\[
\frac{dx_E'}{dt} = \frac{x_{tw}-x_E'}{\tau_a}; \tau_a = 7 \text{ days}
\]

\[
S_E = w_{dE}\sigma_d\left(\frac{\tau_1}{2} + \tau_2\right) + (1-w_{dE})\sigma_i\left(\frac{\tau_1}{2} + \tau_a\right)
\]

\[
S_I = w_{di}\sigma_d\left(\frac{\tau_2}{2} + \tau_a\right) + (1-w_{di})\sigma_i\left(\tau_1 + \frac{\tau_2}{2} + \tau_a\right)
\]

\[
w_{di} = \frac{\frac{1}{\tau_1} + \frac{\tau_2}{2} + \tau_a}{\tau_2 + \frac{1}{\tau_1} + \frac{\tau_2}{2}}
\]

\[
w_{dE} = \frac{\frac{1}{\tau_1} + \frac{\tau_2}{2} + \tau_a}{\tau_1 + \frac{1}{\tau_2} + \tau_a}
\]
Online supplementary

\( \sigma_v = \frac{x_{tv} - x_{tv}^*}{r_{al}[x_{tv}]^*} \) 

The basic idea behind these equations is to calculate expected E and I state variables based on (recent) death rate \( (x_d^*) \) and adjust that approximation based on the expected slope of E and I \( (s_E \text{ and } s_I) \) calculated using the observed slopes of cases \( (\sigma_t) \) and deaths \( (\sigma_d) \).

**IV- Important simplifications and improvement opportunities**

The SEIRb model is very simple. It is built only to test the usefulness of three features we find correlate with the predictive quality of various models, and by design, to exclude various other features which could further enhance a predictive model. Here we provide a partial list of those missing features, focusing on mechanistic models (elaborating on alternative curve-fitting models goes beyond the scope of this paper). Since we have not tested the features below we cannot comment on their relative value in terms of enhancing predictive power, but we suspect several from this list could improve upon SEIRb’s performance. Indeed, the model “IHME-CurveFit” outperform SEIRb over longer time horizons, and benefit from incorporating a few of these features. However, several other models do benefit from a subset of these features and yet do not show notable improvements over SEIRb, thus we do not imply that incorporating all these features would tend to enhance a model’s predictive power.

**Model Structure**

- Capturing operational mechanisms of relevance
  - Hospitalization, treatment, and critical care capacity
  - Testing, changes in testing capacity, and its impact on ascertainment and risk response
  - Changes in demand for testing based on recent cases and deaths
  - Prioritization of testing and treatment capacity based on symptoms and other factors
  - Incorporating travel networks between different locations and importation of cases from abroad
- Modeling at more granular levels
  - Modeling at county (vs. state) level
  - Disaggregating based on age groups and high vs. low risk groups
  - Disaggregating based on severity of disease, including asymptomatic transmission
- Capturing additional feedback mechanisms
  - Changes (reductions) in Infection Fatality Rate with accumulation of deaths due to changes in behavior among higher risk groups (e.g. elderly), improved treatment, and depletion of most at-risk populations (e.g. nursing homes).
  - Changes in behavioral response due to adherence fatigue
  - Emergence of new variants and endogenous changes in transmissibility of the SARS-CoV-2 virus

**Data Sources**

- Including data for testing
- Including data for hospitalization and ICU visit
- Including data for mobility changes in each location
- Incorporating data for mobility across locations
- Incorporating data for government policies and mandates, and their removal over time

Electronic copy available at: https://ssrn.com/abstract=3906690
Online supplementary

Model Estimation

- Representing delays in reporting of cases and deaths
- Estimating various assumed model parameters (e.g. $\tau_1, \tau_2, \tau_3$)
- Estimating the impact of weather factors on transmission directly
- Jointly estimating model parameters across states, using hierarchical Bayesian methods
- Using more sophisticated likelihood functions to account for interdependency over time and across locations in the observed data
- Using more sophisticated optimization algorithms and more computational power to decrease the chances of converging to local peaks in the parameter space

Fine tuning for prediction

- Testing alternative model structures to pick the one that offers better predictions
- Testing alternative assumed parameters to pick the set offering better predictions
- Adopting different model structures for different locations to enhance prediction
- Adopting ensembles of models to increase predictive robustness

State resetting

- Using particle filters, extended or unscented Kalman filters, or other filtering methods for state resetting
- Also resetting other state variables (e.g. perceived risk) based on recent cases and deaths
Appendix 5 - Data and Replication Instructions

A zip file provides both the models and the data used in this paper, offering opportunity to replicate and extend our results.
Reference

1. Cramer, E.Y., et al., *Evaluation of individual and ensemble probabilistic forecasts of COVID-19 mortality in the US*. medRxiv, 2021: p. 2021.02.03.21250974.
2. Ray, E.L., et al., *Ensemble Forecasts of Coronavirus Disease 2019 (COVID-19) in the U.S*. medRxiv, 2020: p. 2020.08.19.20177493.
3. Reiner, R.C., et al., *Modeling COVID-19 scenarios for the United States*. Nature Medicine, 2021. 27(1): p. 94-105.
4. Xu, R., et al., *Weather Conditions and COVID-19 Transmission: Estimates and Projections*. medRxiv, 2020.
5. Rahmandad, H., T. Lim, and J. Sterman, *Behavioral dynamics of COVID-19: estimating underreporting, multiple waves, and adherence fatigue across 92 nations*. System Dynamics Review, Forthcoming.
6. Chin, V., et al., *A case study in model failure? COVID-19 daily deaths and ICU bed utilisation predictions in New York State*. European Journal of Epidemiology, 2020. 35(8): p. 733-742.