Online supplementary material to accompany:

Enhancing Epidemic Forecasting: Learning from COVID-19 Models

Contents

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 ([38, 39]). 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:

> AIpert-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_ISOLAT-Mixtures, MITCovAlliance-SIR, MOBS-GLEAM_COVID, MSRA-DeepST, NotreDame-FRED, NotreDame-mobility, OliverWyman-Navigator, PSI-DRAFT, QJHong-Encounter, RobertWalraven-ESG, RPI_UW-Mob_Collision, SigSci-TS, SteveMcConnell-CovidComplete, STH-3PU, SWC-TerminusCM, TTU-squider, UA-EpiCovDA, UChicagoCHATTOPADHYAY-UnIT, UChicago-CovidIL, UCLA-SuEIR, UCM_MESALab-FoGSEIR, UCSB-ACTS, UCSD_NEU-DeepGLEAM, UMass-MechBayes, UMich-RidgeTfReg, UpstateSU-GRU, USACE-ERDC_SEIR, USC-SI_kJalpha, UT-Mobility, Wadhwani_AI-BayesOpt, and YYG-ParamSearch.

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);
    - disciplinary 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;
    - 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).

5) **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 ß). 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.
6) **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 ß). A large majority used their latest estimate of the reproductive number from the past data for projecting the future cases.
7) **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[1], 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])=~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

[1] It is important to note that the IHME model of COVID-19 pandemic started as a non-mechanistic model but moved away from their initial curve-fitting approach towards a detailed, mechanistic model, which offered substantially better predictions than their curve-fitting approach (see reference 24. Reiner, R.C., et al., *Modeling COVID-19 scenarios for the United States.* Nature Medicine, 2021. **27**(1): p. 94-105.). Since the first incarnation of the model was non-mechanistic, many may not have realized this important change.

orders. The model then estimated the policy effects on mobility and their effect on transmission intensity. In addition, IHME modeled future policy changes endogenously, with a binary feedback mechanism: they assumed that there is a threshold for daily death at 8 per million population, and if simulation forecasts for death pass the threshold, infectivity will decline due to possible implementation of social distancing measures. [24].

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(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. Note that the within R-squared (excluding the variation accounted for by the fixed effects) were small for all of the models. This is somewhat expected as there was much heterogeneity across model performance [19], and the purpose of this analysis was not to make causal inference about factors influencing model performance but to descriptively compare the model performance based on the broad categories that their methodological approaches fall into. These analyses generate model building hypotheses which we can further test in the simulation. To further investigate model heterogeneity we also conducted more detailed qualitative review for each model in Appendix 1.

Table A1

Outcome: Log(normalized absolute error)	(1) Week 1	(2) Week 2	(3) Week 3	(4) Week 4	(5) Week 5	(6) Week 6	(7) 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
Within R-squared	0.007	0.004	0.003	0.004	0.005	0.007	0.003
Number of location-time combination	2,847	2,790	2,738	2,681	2,623	2,566	2,331

Standard errors in parentheses
*** p<0.001, ** p<0.01, * p<0.05

Outcome: Log(normalized absolute error)	(1) Week 8	(2) Week 9	(3) Week 10	(4) Week 11	(5) Week 12	(6) Week 13	(7) Week 14
Non-mechanistic	0.0292	0.0629*	0.107***	0.0823**	0.0907***	0.122***	0.152***
	(0.0193)	(0.0251)	(0.0255)	(0.0255)	(0.0262)	(0.0273)	(0.0288)
Mechanistic 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)	(0.0190)	(0.0205)	(0.0220)
Mechanistic (excluding state-resetting)	-0.0158	-0.125***	-0.0588**	-0.0352	-0.00771	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

Table A2

Outcome: Log(normalized absolute error)	(1) Week 1	(2) Week 2	(3) Week 3	(4) Week 4	(5) Week 5	(6) Week 6	(7) Week 7
Ensemble	-0.0571***	-0.0667***	-0.0464**	-0.0151	-0.237***	0.0167	
	(0.0156)	(0.0162)	(0.0168)	(0.0172)	(0.0413)	(0.0456)	
Non-mechanistic	-0.0564***	-0.0356***	0.00495	0.0254**	0.0356*	0.0342*	0.107***
	(0.00766)	(0.00823)	(0.00859)	(0.00879)	(0.0170)	(0.0172)	(0.0200)
Mechanistic with state-resetting	-0.217***	-0.225***	-0.203***	-0.212***	-0.176***	-0.177***	-0.0473**
	(0.0103)	(0.0107)	(0.0111)	(0.0113)	(0.0140)	(0.0141)	(0.0161)
Academia	0.00394	0.0131	0.0204	0.0343*	0.285***	0.0728	0.125**
	(0.0121)	(0.0126)	(0.0134)	(0.0138)	(0.0372)	(0.0409)	(0.0440)
Model documentation	0.183***	0.192***	0.175***	0.172***	0.159***	0.195***	0.348***
	(0.0128)	(0.0134)	(0.0138)	(0.0141)	(0.0196)	(0.0194)	(0.0250)
Academia*Model documentation	-0.0864***	-0.121***	-0.117***	-0.127***	-0.553***	-0.376***	-0.524***
	(0.0156)	(0.0162)	(0.0170)	(0.0174)	(0.0400)	(0.0438)	(0.0482)
Constant	-12.18***	-12.00***	-11.85***	-11.71***	-11.51***	-11.39***	-11.47***
	(0.00976)	(0.0102)	(0.0107)	(0.0109)	(0.0156)	(0.0156)	(0.0235)
Observations	81,754	77,803	74,725	72,336	30,273	26,858	18,678
Within R-squared	0.008	0.008	0.006	0.007	0.019	0.021	0.033
Number of location-time combination	2,843	2,787	2,736	2,679	2,621	2,565	2,328

Standard errors in parentheses
*** p<0.001, ** p<0.01, * p<0.05

Outcome:	(1) Week 8	(2) Week 9	(3) Week 10	(4) Week 11	(5) Week 12	(6) Week 13	(7) Week 14

Log(normalized absolute error)							
Non-mechanistic	0.108***	0.157***	0.145***	0.103***	0.0897**	0.0625	0.0771*
	(0.0221)	(0.0300)	(0.0305)	(0.0304)	(0.0313)	(0.0329)	(0.0354)
Mechanistic with state-resetting	-0.141***	-0.127***	-0.139***	-0.127***	-0.105***	-0.169***	-0.218***
	(0.0208)	(0.0249)	(0.0254)	(0.0254)	(0.0266)	(0.0283)	(0.0305)
Academia	0.157***	0.0568	0.0862	-0.0232	-0.0589	-0.289	0.177**
	(0.0474)	(0.0501)	(0.0493)	(0.0510)	(0.0944)	(0.172)	(0.0576)
Model documentation	0.174***	-0.0686	-0.103	0.112	0.154	0.445*	
	(0.0319)	(0.0575)	(0.0570)	(0.0631)	(0.104)	(0.180)	
Academia*Model documentation	-0.321***						
	(0.0556)						
Constant	-11.36***	-11.31***	-11.22***	-11.27***	-11.19***	-11.18***	-11.12***
	(0.0273)	(0.0389)	(0.0396)	(0.0429)	(0.0448)	(0.0515)	(0.0607)
Observations	15,527	8,883	8,303	7,769	7,097	6,297	5,822
Within R-squared	0.018	0.016	0.018	0.016	0.012	0.021	0.027
Number of location-time combination	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, for completeness 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).

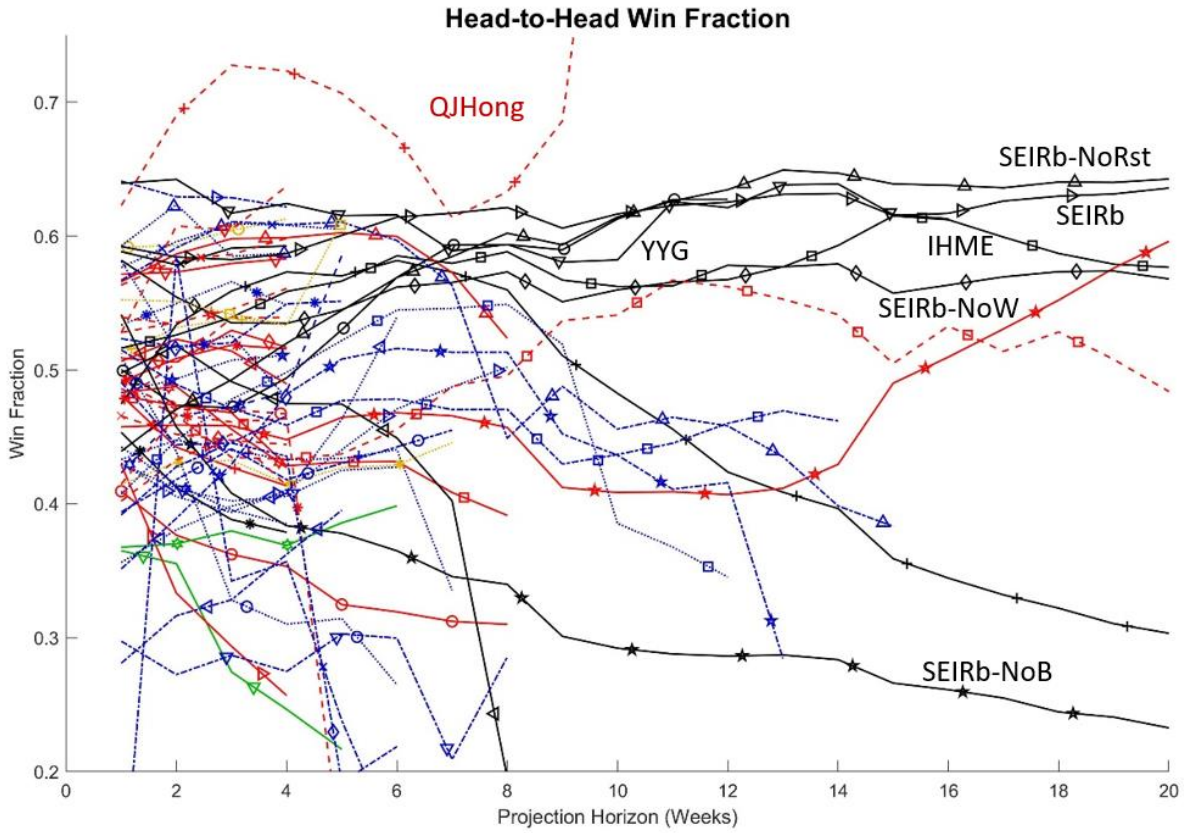

Figure A1: Comparison of performance of models based on head-to-head win fraction

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.

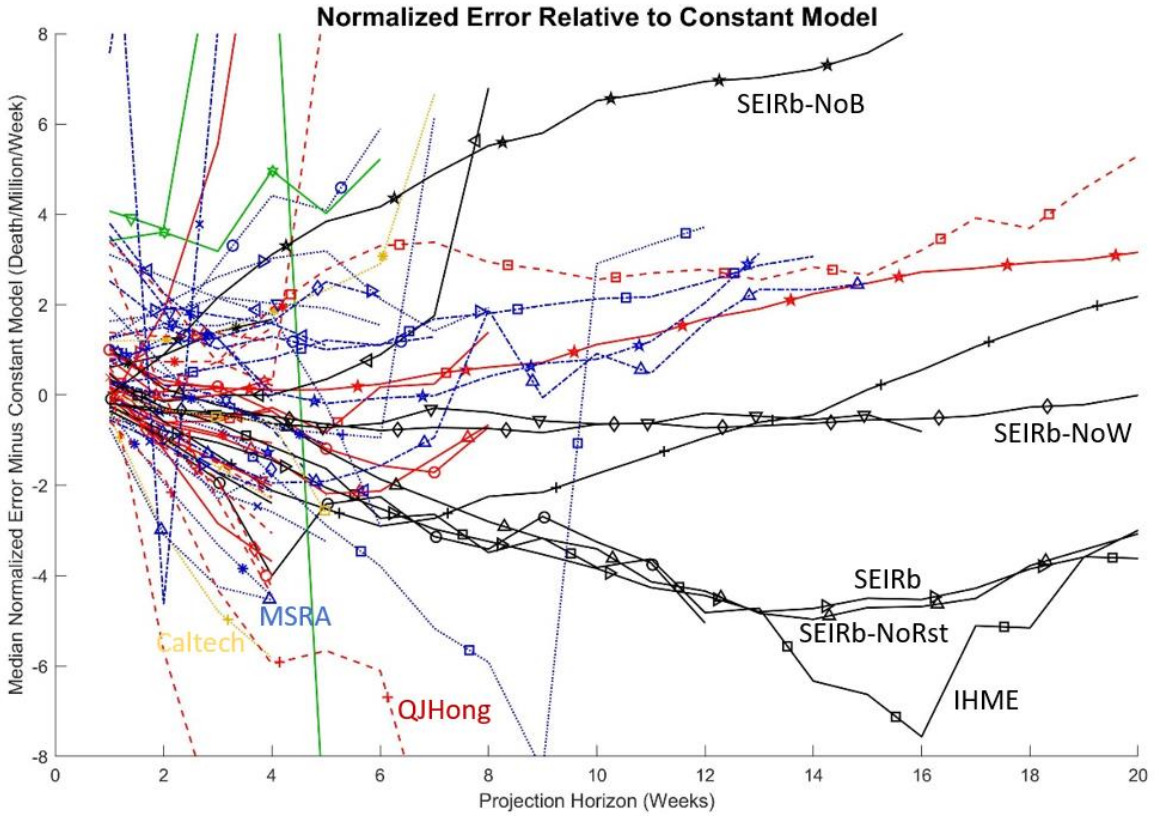

Figure A2: Comparison of performance of models based on normalized error relative to constant model

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).

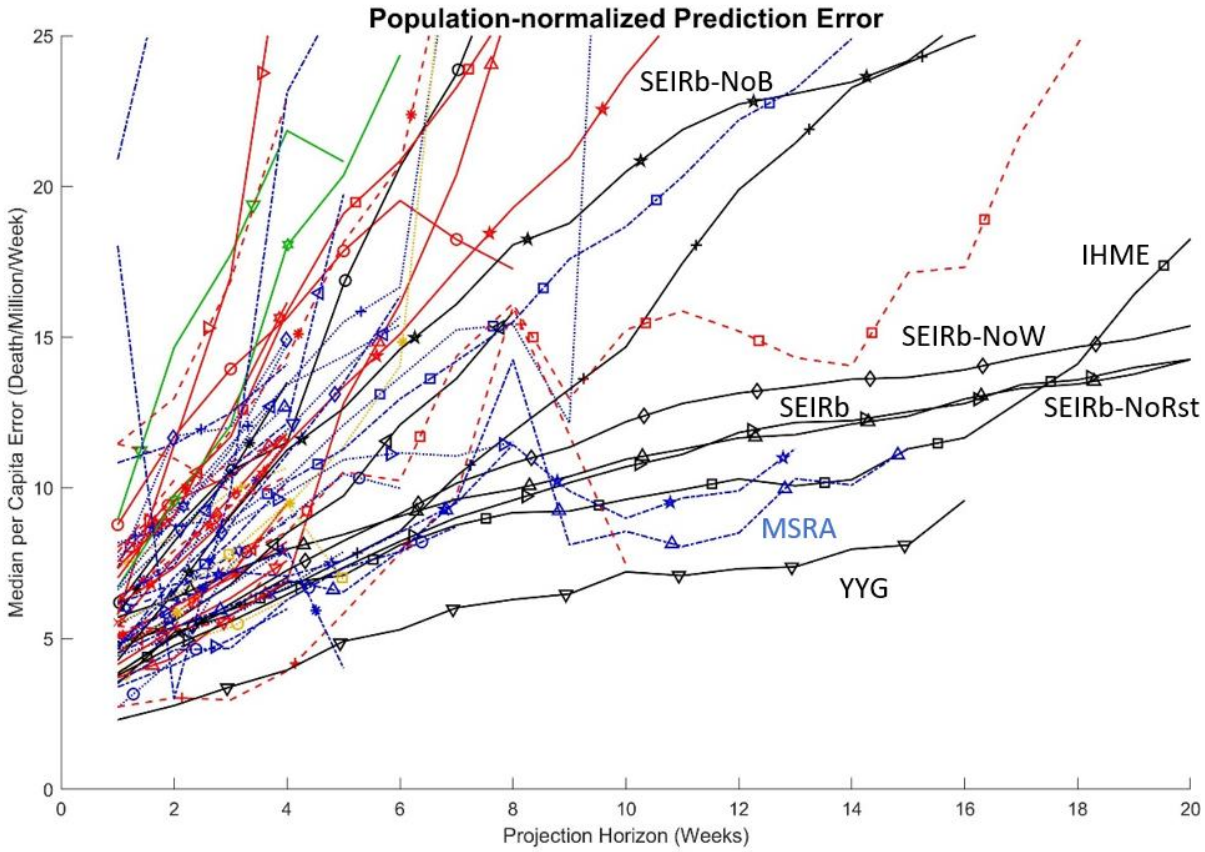

Figure A3: Comparison of performance of models based on population-normalized absolute error

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).

$$\frac{dS}{dt} = -\frac{\beta SI}{N}$$

$$\frac{dE}{dt} = \frac{\beta SI}{N} - \frac{E}{\tau_1}$$

$$\frac{dI}{dt} = \frac{E}{\tau_1} - \frac{I}{\tau_2}$$

$$\frac{dR}{dt} = \frac{I}{\tau_2} \qquad (1\text{-}4)$$

where β is transmission intensity, τ_1 is exposure period, and τ_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)

$$f = i\frac{I}{\tau_2} \qquad (5)$$

The transmission intensity of β which determines the speed of the spread of the disease, and the reproductive number, should change overtime, and in fact that is the main difference between our SEIRb model and others. Specifically, we expect β 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 [30], 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:

$$\beta = \beta_0 w \frac{1}{(1+\alpha f')^{\gamma}} \qquad (6)$$

For the weather impact of w, projections from a previous study [30] 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 [18] 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 (τ_U) and downward (τ_D) adjustment of f', both estimated from model calibration:

$$\frac{df'}{dt} = \frac{f - f'}{\tau_R}$$

$$\tau_R = \begin{cases} \tau_U \ \ if \ \ f > f' \\ \tau_D \ \ if \ f \leq f' \end{cases} \quad (7)$$

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} and $y_{vt}(\theta)$: where x is data, $y(\theta)$ is model predictions for data given the unknown parameter vector θ; 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):

$$LL(\theta, \lambda_v) = \sum_v \sum_{t=T_0}^{t=T_f} -\frac{\ln(1+\lambda_v x_{tv})}{\lambda_v} + ln\Gamma\left(x_{tv} + \frac{1}{\lambda_v}\right) - ln\Gamma\left(\frac{1}{\lambda_v}\right) - \left(x_{tv} + \frac{1}{\lambda_v}\right)\ln\left(1 + \lambda_v y_{tv}(\theta)\right) + x_{tv}(\ln\left(y_{tv}(\theta)\right) + \ln(\lambda_v)) \quad (8)$$

In this function $\Gamma(z)$ represents the natural logarithm of the generalized factorial function for z-1 ($ln\Gamma(z+1) = \ln(z!)$ 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 SEIRb 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:

$$\frac{df'}{dt} = \frac{x_{td} - f'}{\tau_R} \quad (9)$$

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

$$y_{ti}(\theta) = \frac{\beta S I_D}{N} \quad (10)$$

$$\frac{dI_D}{dt} = x_{ti} - \frac{I_D}{\tau_2} \quad (11)$$

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 τ_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 (θ) and the ranges we use for each in the calibration are listed in Table A3.

Table A3

Parameter	Range	Units	Explanation
β_0	[0.1-4]	1/Day	Basic Transmission Intensity
α	[0.01,100]	Day/Person	Sensitivity to death
γ	[0,5]		Death risk diminishing impact
τ_U	[1,100]	Day	Time to adjust risk perception upwards
τ_D	[10,400]	Day	Time to adjust risk perception downwards
T_0	Oct 15, 2019-Mar 3, 2020	Day	Patient zero arrival time
i	[0.003-0.01]		Infection fatality rate

We also estimate two parameters regulating the shape of the negative binomial distribution (λ_v), 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_f, 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*:

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

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

$$\frac{dx_v^*}{dt} = \frac{x_{tv}-x_v^*}{\tau_a};\ \tau_a = 7 \text{ days}$$

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

$$s_I = w_{dI}\sigma_d\left(\frac{\tau_2}{2}+\tau_a\right) + (1-w_{dI})\sigma_i\left(\tau_1+\frac{\tau_2}{2}+\tau_a\right)$$

$$w_{dI} = \frac{\frac{2}{\tau_2}}{\frac{1}{\tau_1+\frac{\tau_2}{2}}+\frac{2}{\tau_2}}$$

$$w_{dE} = \frac{\frac{1}{\tau_1+\frac{\tau_2}{2}}}{\frac{1}{\tau_1+\frac{\tau_2}{2}}+\frac{2}{\tau_2}}$$

$$\sigma_v = \frac{x_{tv}-x_v^*}{\tau_a|x_v^*|}$$

(12-19)

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 and s_I) calculated using the observed slopes of cases (σ_i) and deaths (σ_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
    - Loss of immunity among those recovered, reinfections, and potential reduction in severity of disease in future infections
    - 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

Supplementary materials

- 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

Model Estimation

- Representing delays in reporting of cases and deaths
- Estimating various assumed model parameters (e.g. τ_1, τ_2, τ_a)
- 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- Additional Analyses on Model Performance

In this section we further assessed the models' performance using alternative metrics as well as by different periods. First we adopted Interval Score (IS) based on the 95% prediction interval as the outcome [22], which summarizes each 95% prediction interval as a single number and is penalized by not containing true death as well as wide intervals. We do not use this metric in the main analysis because determinants of confidence interval accuracy are not the topic of our study. We replicated our main analysis for study 1 (see details in Appendix 2) comparing IS (normalized by state population) for each model in the CDC repository.[2] Results were summarized in Figure A4, which were largely consistent with the one in the main paper, but showed a larger advantage in model performance for compartmental models with state-resetting (compared to other model categories).

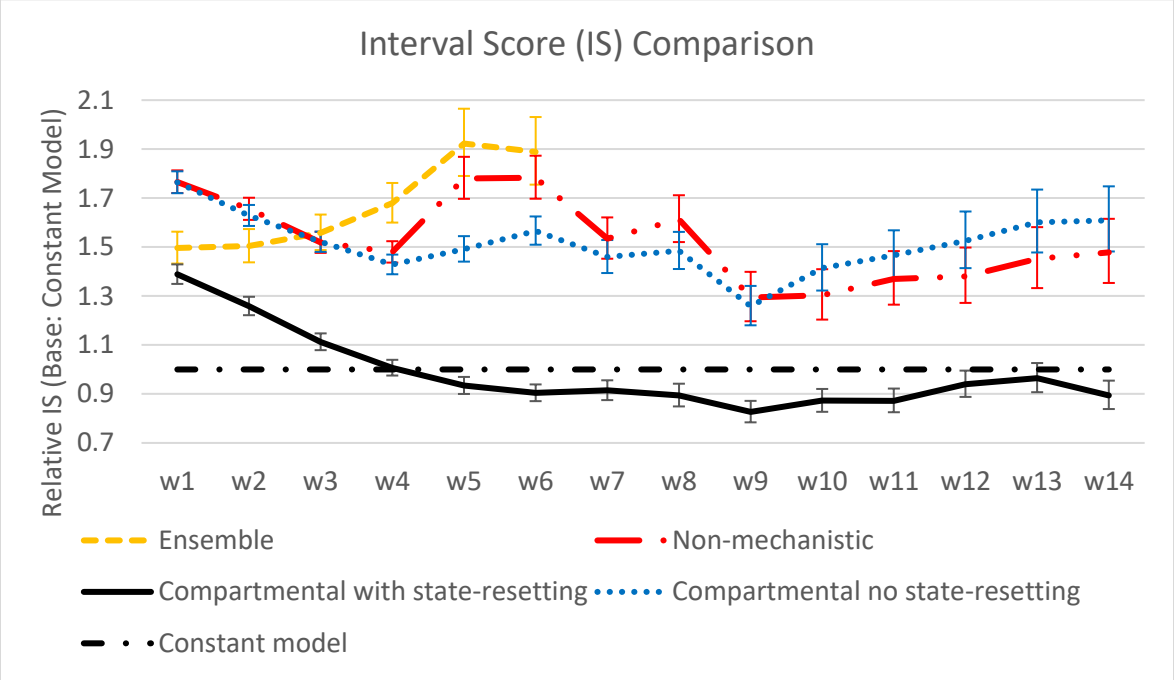

Figure A4: Death projection performance of the CDC model set over different time horizons compared to a constant model based on Interval Score.

To provide insights on whether timing of the prediction impact each model's performance and if certain models are better at predicting turning points (e.g., emergence of a new peak), we explored each model's performance for two subsets of forecasts, based on whether they were made pre- or post- turning points. Specifically, we divided the death time series for each location into different segments marked by turning-points in the (smoothed) death rates: each segment starts from one turning-point (maximum/minimum) and ends at the next. We then divided all predictions into two groups: those where the prediction date (the date at which prediction is done) and target date (the date for which a model is predicting the number of deaths) are within the same segment (pre-turning) or in different segments (post-turning). We replicated our analyses in Fig. 1B and Fig. 3C for each group. This approach allows us to quantitatively assess the types of models that are better in predicting death before/after reaching the turning points (extremums).

[2] As constant model only contains point estimates, we constructed the width of its prediction interval in each location-horizon-projection date combination to be the median of the prediction interval width for all CDC models in the same location-horizon-projection date combination. This choice does not impact how the performance of CDC models in different categories compare to each other.

Fig. A5a and A5b compared the performance of all models in CDC repository and the results showed that: (1) as expected, all models performed worse when there were turning points in upcoming death trends, but compartmental models with state-resetting still performed best overall; (2) as before compartmental models without state-resetting performed worse than non-mechanistic models in the short term, and their advantage grew with forecast horizon; however, this trend was more salient for predictions post-turning points. We concluded that there might be some evidence about additional benefits of compartmental models in forecasts that should foresee upcoming turning points but this evidence is at best indicative.

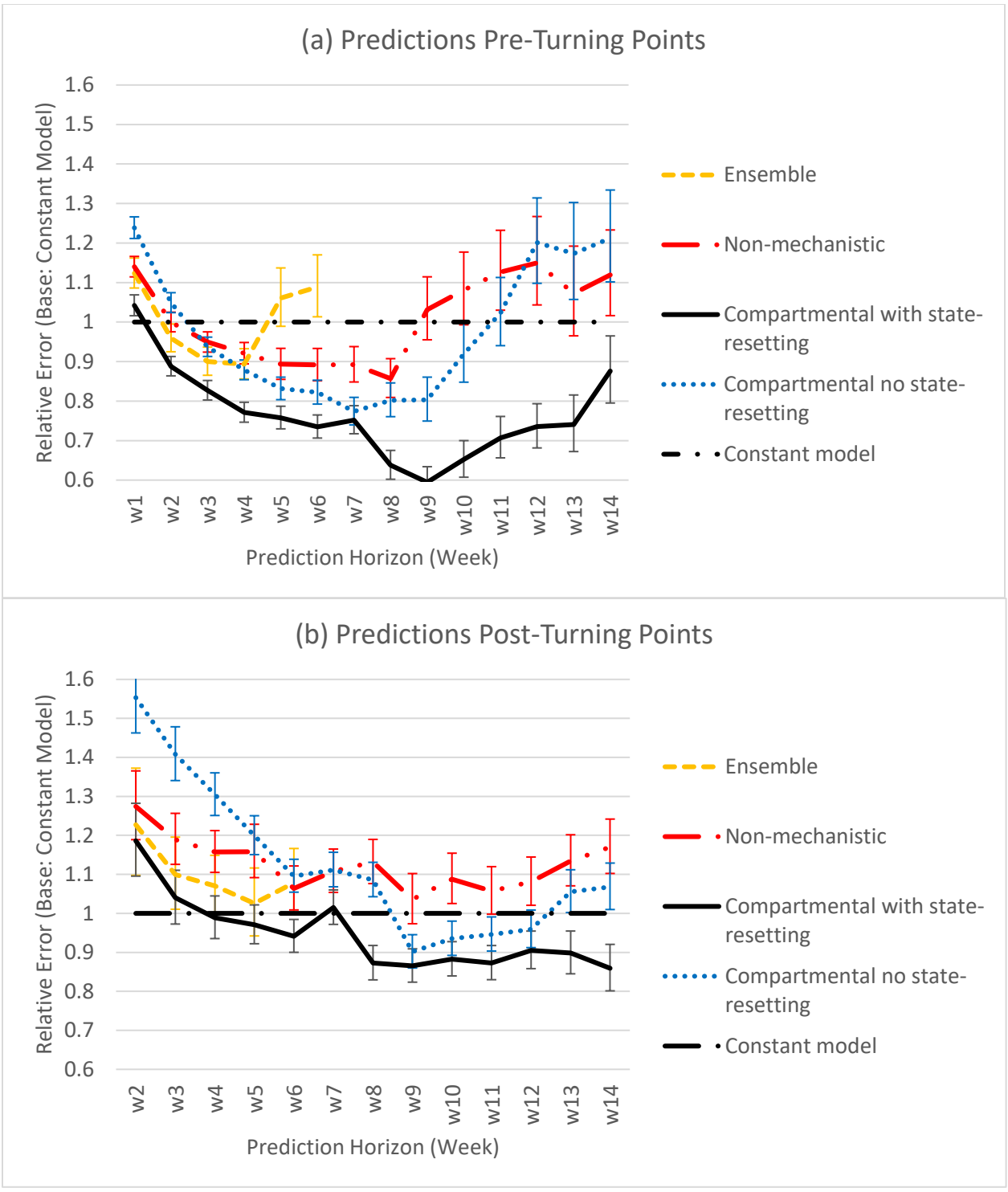

Figure A5. Death projection performance of the CDC model set over different time horizons compared to a constant model before turning points (a) and after turning points (b) (week 1 was excluded from Fig. A5b for consistent scales on Y-axis across figures).

We also replicated the analysis in Figure 3C to evaluate how performance of SEIR-b and related models compared with other models in CDC repository based on the same definitions of segments and pre- vs. post-turning point forecasts. The results in Fig. A6a and A6b showed that the top performing models (IHME and SEIRb) remain at the top for the longer time horizons in both pre- and post-turning points. Thus we found no strong support for the hypothesis that behavioral models are distinctly better post-turning points.

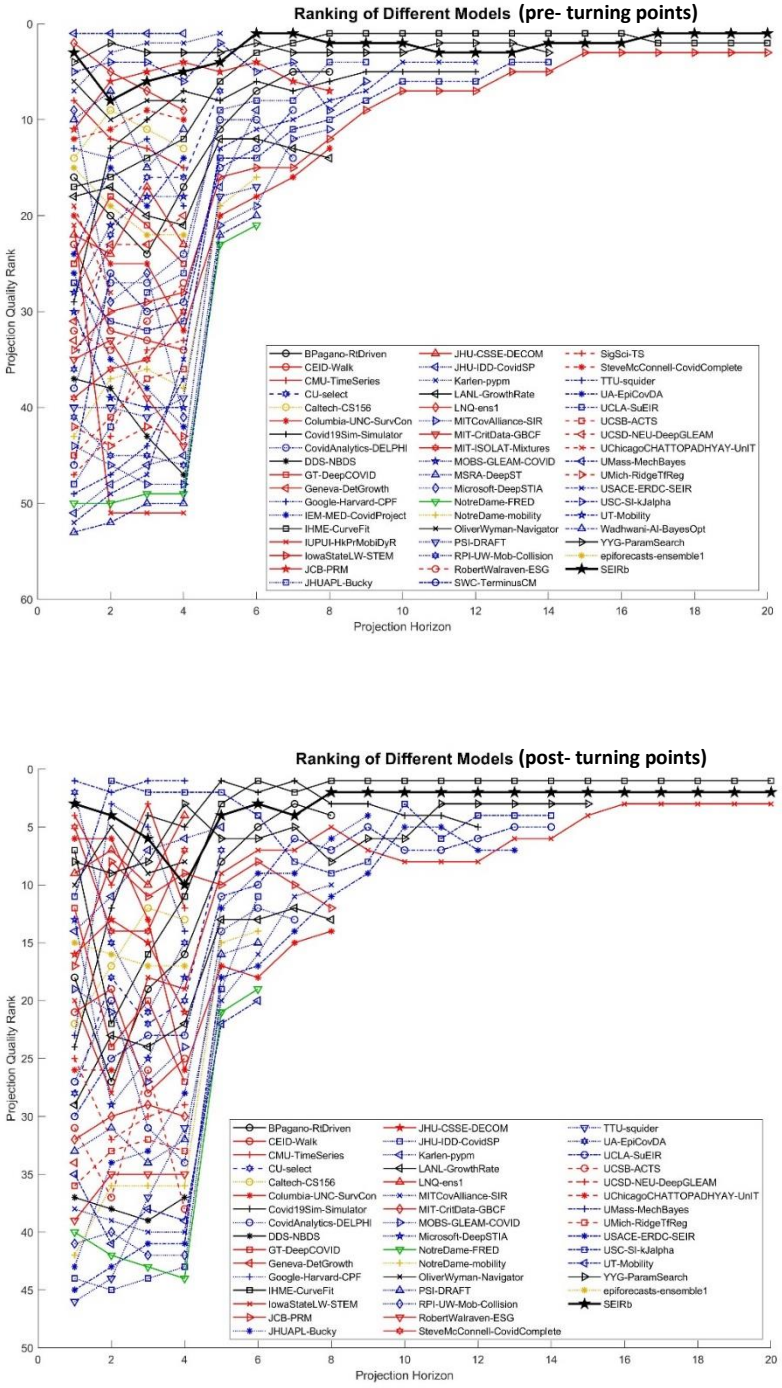

Figure A6. Forecast quality ranks for CDC model set and SEIRb based on regressing Ln(Per capita projection error) against models, controlling for location-horizon-week combinations (a) pre- and (b) post- turning points. Color codes: compartmental models without state-resetting (blue); with state-resetting (black); non-mechanistic (red); agent-based (green); and ensemble (yellow).

Appendix 6- 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.