Interpretable Sequence Learning for COVID-19 Forecasting

Sercan Ö. Arık, Chun-Liang Li, Jinsung Yoon, Rajarishi Sinha, Arkady Epshteyn, Long T. Le, Vikas Menon, Shashank Singh, Leyou Zhang, Nate Yoder, Martin Nikoltchev, Yash Sonthalia, Hootan Nakhost, Elli Kanal, and Tomas Pfister

Google Cloud AI

ABSTRACT

We propose a novel approach that integrates machine learning into compartmental disease modeling to predict the progression of COVID-19. Our model is explainable by design as it explicitly shows how different compartments evolve and it uses interpretable encoders to incorporate covariates and improve performance. Explainability is valuable to ensure that the model's forecasts are credible to epidemiologists and to instill confidence in end-users such as policy makers and healthcare institutions. Our model can be applied at different geographic resolutions, and here we demonstrate it for states and counties in the United States. We show that our model provides more accurate forecasts, in metrics averaged across the entire US, than state-of-the-art alternatives, and that it provides qualitatively meaningful explanatory insights. Lastly, we analyze the performance of our model for different subgroups based on the subgroup distributions within the counties.

1 Introduction

The rapid spread of COVID-19, the disease caused by the SARS-CoV-2 virus, has had and continues to have a significant impact on humanity. Accurately forecasting the progression of COVID-19 can help (i) healthcare institutions to ensure sufficient supply of equipment and personnel to minimize fatalities, (ii) policy makers to consider potential outcomes of their policy decisions, (iii) manufacturers and retailers to plan their business decisions based on predicted attenuation or recurrence of the pandemic, and (iv) the general population to have confidence in the choices made by the above actors.

Data is one of the greatest assets of the modern era, including for healthcare[1]. We aim to exploit the abundance of available data online to generate more accurate COVID-19 forecasts. From available healthcare supply to mobility indices, many information sources are expected to have predictive value for forecasting the spread of COVID-19. Data-driven time-series forecasting has enjoyed great success, particularly with advances in deep learning[2–4]. However, several features of the current pandemic limit the success of standard time-series forecasting methods:

- Because there is no close precedent for the COVID-19 pandemic, it is necessary to integrate existing data with priors based on epidemiological knowledge of disease dynamics.
- The data-generating processes are non-stationary because progression of the disease influences public policy and individuals' public behaviors and vice versa.
- There are many available data sources, but their causal impact on the progression of the disease is unknown.
- The problem is non-identifiable as most infections can be undocumented.
- Data sources are noisy due to reporting issues and data collection problems.
- In addition to accuracy, explainability is important – users from healthcare, policy and businesses need to be able to interpret the results in a meaningful way to help them with strategic planning.

Compartmental models, such as the SIR and SEIR[5] models, are widely used for disease modeling by healthcare and public authorities. Such models represent the number of people in each of the compartments (see Fig. 1) and model the transitions between them with differential equations. Compartmental models often have several shortcomings: (i) they have few parameters and hence low representational capacity, (ii) the modeled dynamics are stationary due to static rates in the differential equations; (iii) they do not use covariates to extract information; (iv) they assume well-mixed compartments, i.e. each individual is statistically identical to others in the same compartment[6]; (v) there is no efficient mechanism for sharing information across time or geography, and (vi) they suffer from non-identifiability – identical results may arise from different parametrizations[7].

In this work we develop a model that provides highly accurate forecasts that preserve interpretability that go beyond the capabilities of standard compartmental models by utilizing rich datasets with high temporal and spatial granularity. Our approach is based on integrating covariate encoding into compartment transitions to extract relevant information via end-to-end learning (Fig. 1). In this way, we provide an inherently interpretable model that reflects the inductive biases of epidemiology.

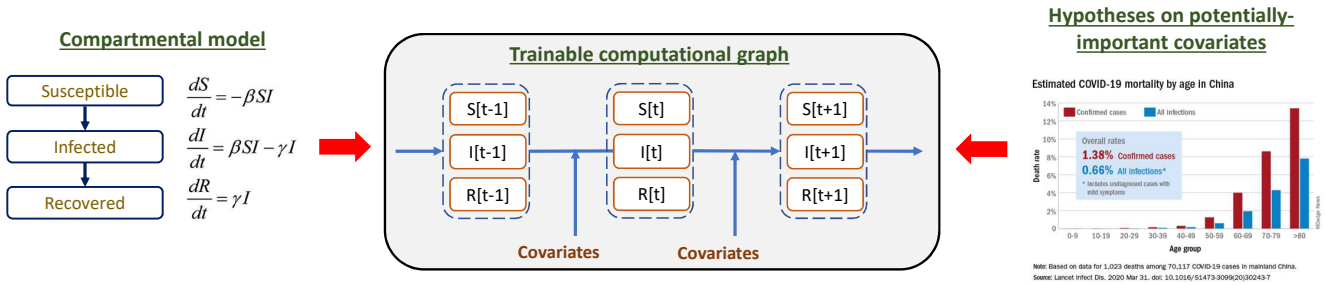


Figure 1. Our approach is based on distilling the inductive bias from compartmental models (as exemplified here for the popular SIR, Susceptible-Infected-Recovered, model) into a computational graph, where the transitions depend on the related covariates.

To get high accuracy, we introduce several innovative contributions:

1. We propose an extension to the standard SEIR model that includes additional compartments for undocumented cases and hospital resource usage. Our end-to-end modeling framework can infer meaningful estimates for undocumented cases even if there is no direct supervision for them.
2. The disease dynamics vary over time – e.g. as mobility reduces, the spreading decays. To accurately reflect such dynamics, we propose time-varying encoding of the covariates.
3. We propose learning mechanisms to improve generalization while learning from limited training data, using (i) masked supervision from partial observations, (ii) partial teacher-forcing to minimize error propagation, (iii) regularization, and (iv) cross-location information-sharing.

COVID-19 has affected some United States (US) counties in a more severe way than others[8–12]: in particular, counties with proportionally large populations of racial and ethnic minorities have suffered disproportionately more from COVID-19. In the US, a COVID-19 forecast that demonstrates consistent accuracy rates across groups will still be off by a higher absolute amount in its prediction for African American and Hispanic communities because of the underlying fact that they experience greater overall impacts from the disease. Consequently, when developing the model we paid close attention to how this disproportionate impact would be reflected in the forecasts, and in the analysis we show that our model does not exacerbate the inherit disparities in the ground truth data further.

We demonstrate our approach for COVID-19 forecasting for the US, the country that has suffered from the highest number of confirmed cases and deaths as of August 2020. For state-level and county-level forecasting, we outperform the alternatives by a large margin. In addition to achieving accurate performance, we show that our model can be used to obtain insights that can help better understand the COVID-19 pandemic. The explainability of our model is two-fold: (1) Our model shows how different compartments evolve over time and provide insights on disease dynamics (e.g. when the number of cases for the undocumented infect peaks), and (2) It shows how different covariates are effecting transitions between compartments (e.g. how mobility or interventions affect the transition from the susceptible to exposed compartment).

2 Related Work

Compartmental models: Compartmental models are commonly used for infectious diseases[13], starting with the notable work by Kermack and McKendrick[5]. The fundamentals of compartmental modeling of infectious diseases are nearly a hundred years old – Kermack and McKendrick[5] proposes a simple compartmental model with differential equations that assign individuals to be susceptible, infected or recovered compartments. Several infectious diseases, including COVID-19, manifest an incubation period during which an individual is infected but is not yet a spreader. To this end, an Exposed (E) compartment can be added, which results in an SEIR model[14]. In addition to these basic types, several other extensions to compartment models have been proposed, such as granular compartments for infections[15] and undocumented compartments[16]. Compartment models are also related to state-space models[17] from Bayesian machine learning. Fitting probabilistic state-space models to compartmental equations are also considered [18]. Overall, the significant advance we bring over standard compartmental modeling is a systematic framework to integrate information-bearing covariates using learnable encoders.

Integrating covariates into compartmental models: Policy changes such as travel bans or public restrictions have a marked, if local, effect on the disease progression. For example, the effect of travel restrictions on the disease spread in China is studied[19] . SEIR model can be modified with mobility covariates[20] and the impact of interventions in the US are shown. Flaxman et al.[21] presents a Bayesian hierarchical model for the effect of non-pharmaceutical interventions on COVID-19 in Europe. Such studies have typically been limited to the impact of one or two covariates, unlike our method which models

numerous static and time-varying covariates simultaneously.

Disease modeling using machine learning: Apart from compartmental models, a wide variety of methods exist for modeling infectious disease. These include sparse identification of non-linear dynamics[22], diffusion models[23], agent-based models[24], and cellular automata[25]. With the motivation of data-driven learning, some also integrate covariates into disease modeling, e.g. using LSTM-based models[26–28].

Learning from data and equations: Strong inductive biases can improve machine learning. One type of such bias is the set of equations between input and output, particularly common in physics or chemistry. To incorporate the inductive bias of equations, parametric approaches have been studied[29–32], as in our paper, where trainable models are incorporated to model only certain terms in the equations, while the equations still govern the end-to-end relationships.

3 Proposed Compartmental Model for COVID-19

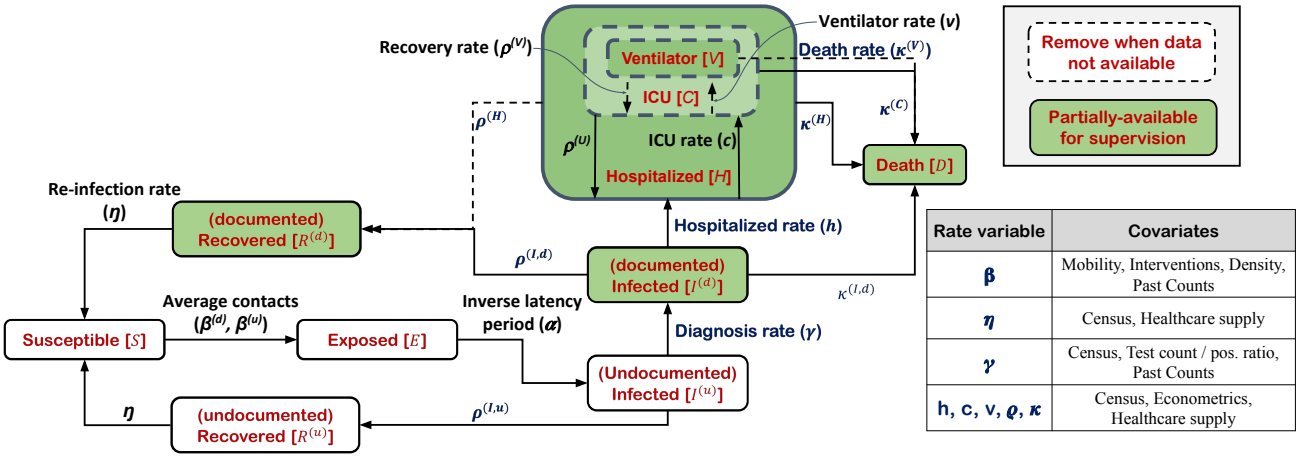


Figure 2. The modeled compartments and the corresponding covariates, with the legend on the right.

We adapt the standard SEIR model with some major changes, as shown in Fig. 2:

1. **Introduction of compartments for undocumented infected and recovered cases:** Integration of undocumented compartments is motivated by the recent studies[33–35] that suggest that majority of the infected people are not detected[1] and they dominate disease spreading[36,37] [2]. An undocumented infected individual is able to spread the disease, until they are diagnosed or they recover while remaining undocumented, while documented cases may be either self-isolated or hospitalized.
2. **Introduction of hospitalized, ICU and ventilator compartments:** We introduce compartments for the people who are hospitalized, in the ICU, or on a ventilator, as there is a demand to model these[38] and there is partially-available observed data to be used for supervision. We assume that the patients who are in the ICU are still hospitalized, and the patients who are on ventilators are still in the ICU.
3. **Partial immunity:** To date, there is no scientific consensus on what fraction of recovered cases demonstrate immunity to future infection. Due to reports of reinfection[39] we model the rate of reinfection from recovered compartments (though our model infers low reinfection rates).
4. **No death from undocumented infected:** We assume the published COVID-19 death counts are coming from documented cases, not undocumented.
5. **Invariant population:** We assume that the entire population is invariant, i.e. births and non-Covid deaths are negligible in comparison to the entire population.

We assume a fixed sampling interval of 1 day (motivated by data publishing frequency), and model each compartment in Table 1. For a compartment X, $X_i[t]$ denotes the number of individuals in that compartment at location i and time t. $N[t]$ denotes the total population. Table 2 describes transition rate variables used to relate the compartments via the equations (we omit the

[1] This might be especially evident for locations that lack sufficient healthcare testing and reporting resources.
[2] Previous work[16] estimates that $> 80\%$ of cases in China were undocumented during the early phase of the pandemic.

Table 1. Modeled compartments. We obtain the total number of confirmed cases, Q, from $I^{(d)}+R^{(d)}+H+D$.

Compartment	Description	Compartment	Description
S	Susceptible	$R^{(u)}$	Recovered undocumented
E	Exposed	H	Hospitalized
$I^{(d)}$	Infected documented	C	In intensive care unit (ICU)
$I^{(u)}$	Infected undocumented	V	On ventilator
$R^{(d)}$	Recovered documented	D	Death

Table 2. Variables and the covariates that affect them. (doc.: documented, undoc.: undocumented)

Variable	Description	Covariates
β	**Average contacts** of doc. infected ($\beta^{(d)}$) / undoc. infected ($\beta^{(u)}$)	Mobility, Interventions, Density
η	Re-infected rate	Census, Healthcare
α	Inverse latency period	-
γ	Diagnosis rate	Census, Test info
h	Hospitalization rate for infected	Census, Income, Healthcare
c	ICU rate for hospitalized	
v	Ventilator rate from ICU	
ρ	**Recovery rate** for doc. infected ($\rho^{(I,d)}$), undoc. infected ($\rho^{(I,u)}$), hospitalized ($\rho^{(H)}$), ICU ($\rho^{(U)}$), ventilator ($\rho^{(V)}$)	
κ	**Death rate** for doc. infected ($\kappa^{(I,d)}$), hospitalized ($\kappa^{(H)}$), ICU ($\kappa^{(C)}$), ventilator ($\kappa^{(V)}$)	

index i for conciseness):

$$
\begin{aligned}
S[t] &= S[t-1]-(\beta^{(d)}\cdot I^{(d)}[t-1]+\beta^{(u)}\cdot I^{(u)}[t-1])\cdot S[t-1]/N[t-1]+\eta\cdot(R^{(d)}[t-1]+R^{(u)}[t-1]),\\
E[t] &= E[t-1]+(\beta^{(d)}\cdot I^{(d)}[t-1]+\beta^{(u)}\cdot I^{(u)}[t-1])\cdot S[t-1]/N[t-1]-\alpha\cdot E[t-1],\\
I^{(u)}[t] &= I^{(u)}[t-1]+\alpha\cdot E[t-1]-(\rho^{(I,u)}+\gamma)\cdot I^{(u)}[t-1],\\
I^{(d)}[t] &= I^{(d)}[t-1]+\gamma\cdot I^{(u)}[t-1]-(\rho^{(I,d)}+\kappa^{(I,d)}+h)\cdot I^{(d)}[t-1],\\
R^{(u)}[t] &= R^{(u)}[t-1]+\rho^{(I,u)}\cdot I^{(u)}[t-1]-\eta\cdot R^{(u)}[t-1],\\
R^{(d)}[t] &= R^{(d)}[t-1]+\rho^{(I,d)}\cdot I^{(d)}[t-1]+\rho^{(H)}\cdot(H[t-1]-C[t-1])-\eta\cdot R^{(d)}[t-1],\\
H[t] &= H[t-1]+h\cdot I^{(d)}[t-1]-(\kappa^{(H)}+\rho^{(H)})\cdot(H[t-1]-C[t-1])-\kappa^{(C)}\cdot(C[t-1]-V[t-1])-\kappa^{(V)}\cdot V[t-1],\\
C[t] &= C[t-1]+c\cdot(H[t-1]-C[t-1])-(\kappa^{(C)}+\rho^{(C)}+v)\cdot(C[t-1]-V[t-1])-\kappa^{(V)}\cdot V[t-1],\\
V[t] &= V[t-1]+v\cdot(C[t-1]-V[t-1])-(\kappa^{(V)}+\rho^{(V)})\cdot V[t-1],\\
D[t] &= D[t-1]+\kappa^{(V)}\cdot V[t-1]+\kappa^{(C)}\cdot(C[t-1]-V[t-1])+\kappa^{(H)}\cdot(H[t-1]-C[t-1])+\kappa^{(I,d)}\cdot I^{(d)}[t-1],
\end{aligned}
$$

Corollary: Effective Reproduction Number. An analysis of our compartmental model using the Next-Generation Matrix method[40] yields the effective reproductive number (spectral radius) as:

$$R_e = \frac{\beta^{(d)}\gamma+\beta^{(u)}(\rho^{(I,d)}+\kappa^{(I,d)}+h)}{(\gamma+\rho^{(I,u)})\cdot(\rho^{(I,d)}+\kappa^{(I,d)}+h)}. \tag{1}$$

Please see Appendix for derivations. Note that when $\gamma=0$, our compartmental model reduces to the standard SEIR model with the undocumented infected and recovered. In this case, $R_e=\beta^{(u)}/\rho^{(I,u)}$. Later on, the derived analytical R_e not only allows us to gain insights of the trained model, we will show it can also be used as an effective regularization to improve the future forecasting (generalization) in Sec. 5.

4 Encoding Covariates

We next discuss the model design choices made in order to successfully be able to utilize the wide range of covariates needed to accurately predict COVID-19.

Time-varying modeling of variables: Ideally, instead of using static rate variables across time to model compartment transitions as in standard compartmental models, there should be time-varying functions that map them from known observations. For example, if human mobility decreases over time, the $S \to E$ transition should reflect that. Consequently, we propose replacing all static rate variables with learnable functions that output their value from the related static and time-varying covariates at each location and timestep. We list all the covariates used for each rate variable in the Appendix. We note that learnable encoding of variables can still preserve the inductive bias of the compartmental modeling framework while increasing the model capacity via learnable encoders.

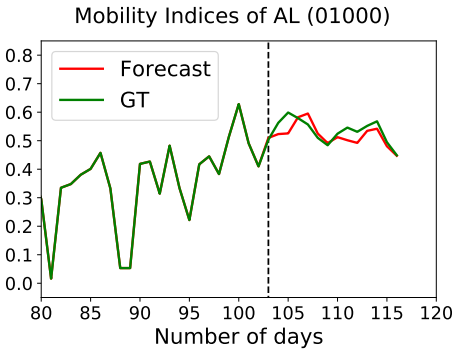


Figure 3. The forecasted normalized mobility covariate of AL (01000), where we train until day 103 (dashed vertical line) and forecast from day 104 to day 117.

Interpretable encoder architecture: In addition to making accurate forecasts, it is valuable to understand how each covariate affects the model. Such explanations can greatly help users from healthcare and public sector to understand the disease dynamics better, and also can help model developers to ensure the model is learning appropriate dynamics via sanity checks with known scientific studies or common knowledge. To this end we adopt a generalized additive model[41] for each variable v_i from Table 2 based on additional *covariates* $\text{cov}(v_i, t)$ at different time t. The covariates we consider include (1) the set of static covariates $\mathscr{S}$, such as population density, and (2) $\{f[t-j]\}_{f \in \mathscr{F}_i, j=1,\dots,k}$ the set of time-varying covariates (features) $\mathscr{F}_i$ with the observation from $t-1$ to $t-k$, such as mobility. Omitting individual feature interactions and applying additive aggregation, we obtain

$$v_i[t] = v_{i,L} + (v_{i,U} - v_{i,L}) \cdot \sigma\left(c + b_i + \mathbf{w}^\top \text{cov}(v_i, t)\right), \tag{2}$$

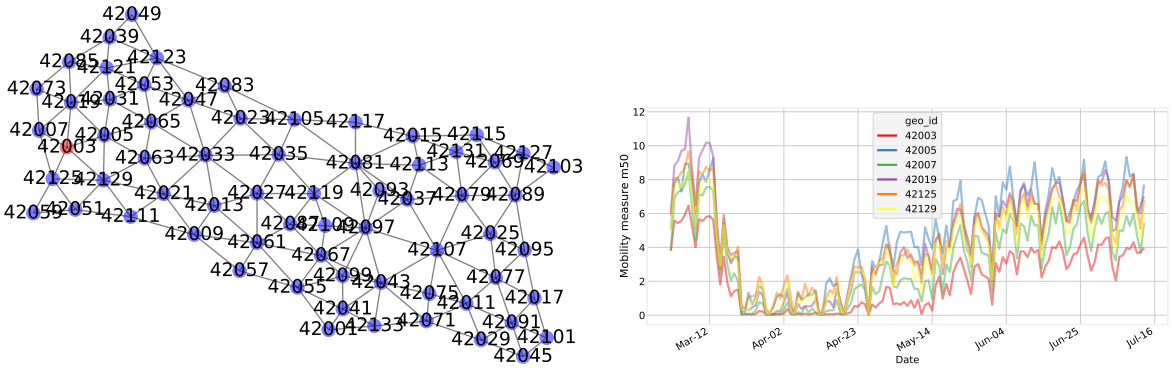
where $v_{i,L}$ and $v_{i,U}$ are the lower and upper bounds of v_i for all t, c is the global bias, b_i is the location-dependent bias. $\sigma()$ is the sigmoid function to limit the range to $[v_{i,L}, v_{i,U}]$[3] and $\mathbf{w}$ is the set of trainable parameters. We observe limiting the ranges to be important to stabilize training and avoid overfitting. We note that although Eq. (2) denotes a linear decomposition for $v_i[t]$ at each timestep, the overall behavior can still be highly non-linear due to the relationships between compartments.

Covariate forecasting: The challenge of using (2) for future forecasting is that some time-varying covariates are not available for the entire forecasting horizon. Assume we have the observations of covariates and compartments until T, and we want to forecast from $T+1$ to $T+\tau$. To forecast $v_i[T+\tau]$, we need the time varying covariates $f[T+\tau-k : T+\tau-1]$ for $f \in \mathscr{F}_i$, but some of them are not observed when $\tau > k$. To solve this issue, we propose to forecast $f[T+\tau-k : T+\tau-1]$ based on their own past observations until T, which is a standard one dimensional time series forecasting for a given covariate f at a given location. In this paper, we adopt a simple linear autoregressive model based on the past ζ (forecasted) observations

$$f[t] = \mathbf{w}_f^\top \left[f[t-\zeta : t-1], \frac{\max(f[t-\zeta : t-1])}{\text{mean}(f[t-\zeta : t-1])} \right]. \tag{3}$$

In our experiments, we keep $\zeta = 14$ to allow the model learn the weekly periodic patterns. The last crafted feature is called *peak-to-mean* ratio, which is commonly used in practice to encode increasing/decreasing trends. An example is shown in Fig. 3. Although we use a simple linear autoregressive model, it can capture the trends and periodic patterns. We note that there are other advanced models[42] can be adopted, which could potentially boost the accuracy of forecasting of all compartments.

[3] We use $v_{i,L}$=0 for all variables, $v_{i,U} = 10$ for β, 0.2 for α and 0.1 for others.



(a) Network graph

(b) Mobility covariate

Figure 4. (a) Network graph of Pennsylvania with Allegheny County, PA (42003) highlighted. (b) Mobility covariate for Allegheny County and its neighboring counties.

Features from the location network graph: Movement across location borders between proximate geographic regions can influence the spread of the disease[43]. For example, people who work in one county and live in a neighboring county could inadvertently communicate the disease between the two counties. However, modeling this effect would require capturing travel dynamics across the land border. Most mobility datasets do not offer such information, and are localized to each county (or geographic region) without capturing inter-regional effects (Figure 4).

One way to incorporate inter-county effects is to represent the counties as nodes in a network graph with edges representing land borders. A simple model of the inter-county effects is one or more aggregation functions applied to the immediate neighborhood of each county, and the results incorporated as new features for that county. Specifically, given the network graph of counties $G \equiv (V,E)$ where V is the set of vertices representing the counties, and E is the set of edges representing land borders between neighboring counties, we define a set of aggregation functions $\mathscr{A}$ over the neighborhood $\mathscr{N}(V_i)$ of county i. Then the set of new features $\mathscr{F}$ for county i are:

$$\mathscr{F} = \mathbf{A}(\mathbf{C})\ \forall \mathbf{A} \in \mathscr{A},\ \forall \mathbf{C} \in \mathscr{C} \tag{4}$$

where $\mathscr{C}$ is the set of existing features for vertices V. Aggregated features are used as extra time-varying covariates for the county-level models. We do not use such features for state-level models as mobility between neighboring state seems to be a less predictive signal.

Information-sharing across locations: Some aspects of the disease dynamics are location-dependent while others are not. In addition, data availability varies across locations – there may be limited observations to learn the impact of a covariate. A model able to learn both location dependent and independent dynamics is desirable. Our encoders in Eq. (2) partially capture location-shared dynamics via shared $\mathbf{w}$ and the global bias c. To allow the model capture remaining location-dependent dynamics, we introduce the local bias b_i. A challenge is that the model could ignore the covariates by encoding all information into b_i during training. This could hurt generalization as there would not be any information-sharing on how static covariates affect the outputs across locations. Thus, we introduce a regularization term $L_{ls} = \lambda_{ls} \sum_i |b_i|^2$ to encourage the model to leverage covariates and c for information-sharing instead of relying on b_i. Without L_{ls}, we observe that the model can try to use the local bias more than the encoded covariates, and can suffer from poorer generalization.

5 End-to-End Training

Finally we discuss the end-to-end learning mechanisms we developed to improve generalization while learning from limited training data.

Learning from partially-available observations: Fitting would have been easy with observations for all compartments, however, we only have access to some. For instance, $I^{(d)}$ is not given in the ground truth of US data but we instead have, Q, the total number of confirmed cases, that we use to supervise $I^{(d)}+R^{(d)}+H+D$. Note that $R^{(ud)}, I^{(ud)}, S, E$ are not given as well.

Algorithm 1 Pseudo-code for training the proposed model

Inputs: Forecasting horizon τ, compartment observations I_i^D, H_i, C_i, V_i, D_i until T, the number of fine tuning iterations F, loss coefficients λ_{comp}, λ_{smooth}, and λ_{ls}.
Initialize trainable parameters θ, including $\mathbf{w_i}$, c, b_i, and initial conditions for the compartments $\hat{E}[0]$, $\hat{I}^{(d)}[0]$, $\hat{I}^{(u)}[0]$, $\hat{R}^{(d)}[0]$, $\hat{R}^{(u)}[0]$, $\hat{H}[0]$, $\hat{C}[0]$, $\hat{V}[0]$, $\hat{D}[0]$
Split τ days for validation $Y_i[T-\tau:T]$ for all locations i, where $Y \in \{Q,H,C,V,D,R^{(d)}\}$
while until convergence **do**
  Sample initial conditions $E_i[0]$, $I_i^{(d)}[0]$, $I_i^{(u)}[0]$, $R_i^{(d)}[0]$, $R_i^{(u)}[0]$, $H_i[0]$, $C_i[0]$, $V_i[0]$, $D_i[0]$
  $\theta \leftarrow \theta - \text{RMSProp}(\nabla_\theta \mathscr{L}(T-M, T-\tau-1))$
  Update the optimal parameters: $\theta_{opt} = \theta$ and $L_{fit}[T-\tau:T]$ is the currently best
Final fine-tuning: fine-tune the selected with joint training and validation data:
$\theta \leftarrow \theta_{opt}$
for F iterations **do**
  $\theta \leftarrow \theta - \text{RMSProp}(\nabla_\theta \mathscr{L}(T-M, T))$
  Update the optimal parameters: $\theta_{opt} = \theta$ if $L_{fit}[T-\tau:T]$ is the currently best
Output: Return θ_{opt}

Formally, we assume availability of the observations $Y[T-M:T]$[4], for $Y \in \{Q,H,C,V,D,R^{(d)}\}$, and consider forecasting the next τ days, $\hat{Y}[T+1:T+\tau]$.

Fitting objective: There is no direct supervision for the training of encoders, and they should be learning in an end-to-end way via the aforementioned partially-available observations. We propose the following objective function for range $[T_s, T_e]$:

$$L_{fit}[T_s:T_e] = \sum_{Y \in \{Q,H,C,V,D,R^{(d)}\}} \lambda_Y \sum_{t=T_s}^{T_e-\tau} \sum_{i=1}^{\tau} \mathbb{I}(Y[t+i]) \cdot q(t+i-T_s;z) \cdot L(Y[t+i], \hat{Y}[t+i]). \tag{5}$$

$\mathbb{I}(\cdot) \in \{0,1\}$ indicates the availability of the Y to allow the training to focus only on available observations. $L(,)$ is the loss between the ground truth and the predicted values (e.g., ℓ_2 or quantile loss), and λ_Y are the importance weights to balance compartments due to scale differences (e.g., Q is much larger than D). Lastly, $q(t;z) = \exp(t \cdot z)$ is a time-weighting function (when $z = 0$, there is no time weighting) to allow the fitting to favor more recent observations and z is a hyperparameter.

Constraints and regularization: Given the limited dataset size, overfitting can be a concern for high-capacity encoders trained on insufficient data. In addition limiting the model capacity with the epidemiological inductive bias, we further apply regularization to improve generalization to unseen future data. First, we penalize infeasible ranges – based on how much the summation of the rates from a particular compartment (e.g. $\kappa^{(V)} + \rho^{(V)}$) exceed 1:

$$\begin{aligned} L_{comp}[T_s:T_e] = \sum_{t=T_s}^{T_e} (\rho^{(I,d)}[t] + \kappa^{(I,d)}[t] + h[t] - 1)_+^2 + (\rho^{(I,u)}[t] + \gamma[t] - 1)_+ + (c[t] + \kappa^{(H)}[t] + \rho^{(H)}[0] - 1)_+^2 \\ + (v[t] + \kappa^{(C)}[t] + \rho^{(C)}[0] - 1)_+^2 + (\kappa^{(V)}[t] + \rho^{(V)}[0] - 1)_+^2, \end{aligned} \tag{6}$$

where $(\cdot)_+ = \max(\cdot, 0)$. Second, we encourage temporal smoothness of variables by penalizing first-order time-derivative inconsistency of the predictions: $L_{smooth}[T_s:T_e] = \sum_{t=T_s+1}^{T_e-1} \sum_{\hat{Y}} (\hat{Y}[t-1] + \hat{Y}[t+1] - 2 \cdot \hat{Y}[t])$.

Lastly, an effective regularization is constraining effective reproduction number R_e as derived in Eq. (1). There are rich literature in epidemiology on R_e to give us good prior knowledge on the range of the number should be. For a reproduction number $R_e[t]$ at time t, instead of enforcing a hard upper bound, we consider the regularization

$$L_{R_e}[T_s:T_e] = \sum_{t=T_s}^{T_e} \exp\left((R_e[t] - R)_+\right),$$

where R is a prespecified *soft* upper bound. The regularization favors the model with R_e in a reasonable range in addition to good absolute forecasting numbers. In the experiment, we set $R = 5$ without further tuning. The final objective function is

$$\mathscr{L}(T_s, T_e) = L_{fit}[T_s:T_e] + \lambda_{comp} \cdot L_{comp}[T_s:T_e] + \lambda_{smooth} \cdot L_{smooth}[T_s:T_e] + \lambda_{ls} \cdot L_{ls} + \lambda_{R_e} \cdot L_{R_e}[T_s:T_e], \tag{7}$$

[4]We use the notation $S_i[T_s:T_e]$ to denote all timesteps between T_s (inclusive) and T_e (inclusive).

where $L_{ls} = \lambda_{ls} \sum_i |b_i|^2$ as discussed in Sec. 4.

Partial teacher forcing: The compartmental model presented in Sec. 3 produces the future propagated values from the current timestep. During training, we have access to the observed values for $Y \in \{Q, H, C, V, D, R^{(d)}\}$ at every timestep, which we could condition the propagated values on, commonly-known as teacher forcing[44] to mitigate error propagation. At inference time, however, ground truth beyond the current timestep t is unavailable, hence the predictions should be conditioned on the future estimates. Using solely ground-truth to condition propagation would create a train-test mismatch. In the same vein of past research to mix the ground truth and predicted data to condition the projections on[45], we propose partial teacher forcing, simply conditioning $(1 - \lambda \mathbb{I}\{Y[t]\})Y[t] + \lambda \mathbb{I}\{Y[t]\})\hat{Y}[t]$, where $\mathbb{I}\{Y[t]\} \in \{0, 1\}$ indicates whether the ground truth $Y[t]$ exists and $\lambda \in [0, 1]$. In the first stage of training, we use teacher forcing with $\lambda = 0.5$. In the fine-tuning (please see below), we use $\lambda = 1$ to unroll the last τ steps to mimic the real forecasting scenario.

Model fitting and selection: The pseudo-code for training is presented in Algorithm 1. We split the observed data into training and validation, setting the last τ timesteps for validation to mimic a testing scenario. We use the training data for optimization of the trainable degrees of freedom, collectively represented as θ, while the validation data is used for early stopping and model selection. To determine the loss coefficients and initial conditions, we run hyperparameter tuning and pick the best model according to the validation loss.

Once the model is selected, we fix the hyperparameters and run fine-tuning on joint training and validation data, to not waste valuable recent information by using it only for model selection. For optimization we use RMSProp as it is empirically observed to yield lower losses compared to other optimization algorithms and providing the best generalization performance. Compartmental models can be sensitive to initial values; hence we treat them as hyperparameters and select them via aforementioned hyperparameter tuning mechanism.

6 Experiments

We perform several experiments to compare our model to other publicly available COVID-19 models, demonstrate that our model enables explainable insights into the functioning of the model, and perform analysis of our model's performance among various population subgroups including racial and ethnic minorities.

Ground truth data: We conduct all experiments on US COVID-19 data. The primary ground truth data for the progression of the disease, for Q and D, are from[46] as used by several others, e.g.[47]. They obtain the raw data from the state and county health departments. Because of the rapid progression of the pandemic, past data has often been restated, or the data collection protocols have been changed. Ground truth data for the H, C and V (see Fig. 1) are obtained from[48].

Covariates: The progression of COVID-19 is influenced by a multitude of factors, including relevant properties of the population, health, environmental, hospital resources, demographics and econometrics indicators. Time-varying factors such as population mobility, hospital resource usage and public policy decisions can also be important. However, indiscriminately incorporating a data source may have deleterious effects. Thus, we curate our data sources to limit them to one source in each category of factors that may have predictive power at the corresponding transition. We use datasets from public sources (see Appendix). A summary of the selected covariates is presented in the Appendix. We apply forward- and backward-filling imputation (respectively) for time-varying covariates, and median imputation for static covariates. Then, all covariates are normalized to be in [0, 1], considering statistics across all locations and time-steps.

Training: We implement Algorithm 1 in TensorFlow at state- and county-levels, using ℓ_2 loss for point forecasts. We employ Bayesian optimization based hyperparameter tuning (including all the loss coefficients, learning rate, and initial conditions) with the objective of optimizing for the best validation loss, with 400 trials and we use $F = 300$ fine-tuning iterations. We choose the compartment weights $\lambda^D = \lambda^Q = 0.1$, $\lambda^H = 0.01$ and $\lambda^{R^{(d)}} = \lambda^C = \lambda^V = 0.001$.[5] At county granularity, we do not have published data for C and V, so, we remove them along with their connected variables.

Evaluation: Our model is capable of forecasting all the modeled compartments, but we focus on the number of deaths for benchmarking as it is known to be the most reliable ground truth data to assess accuracy[476]. In Appendix, we also present results for the forecasting of the number of hospitalized. Note that for each experiment, we reserve the last τ for testing, and we do not use any of the testing data for model development – our model selection is automated and is entirely based on the validation performance. We present our results in Root Mean Square Error (RMSE) and Mean Absolute Error (MAE) metrics,

[5] These make the loss terms roughly within the same range, our results are not highly sensitive to these.

[6] One attributed reason is the limited testing capacity as tests are usually prioritized for the more severely-ill.

to reflect the aggregated performance across the whole nation and along the entire trajectory. In Sec. 15, we also show results in other metrics. We quantify the results average over τ days (e.g. Tables 3 and 4), whenever the benchmark models publish predictions over the entire forecasting horizon. Otherwise we quantify the results for τ day ahead prediction performance(e.g. Tables 5 and 6), as the intermediate prediction values are missing for the other benchmarks.

6.1 Results

Table 3. τ-day average RMSE for forecasting the number of deaths at state level granularity. Since benchmark models from covid19-forecast-hub repository release forecasts at different dates and horizons, not all models have predictions for all prediction dates/horizons (indicated by "—"). **Bold** indicates the best.

Pred. horizon τ (days)	Pred. date	Ours	IHME	LANL	UT	MIT	YYG	UCLA
5	04/20/2020	**146.5**	220.5	155.8	832.5	—	192.1	—
	04/27/2020	**79.3**	259.2	125.2	908.2	505.5	103.2	—
	05/04/2020	**78.6**	133.2	172.0	985.7	603.1	153.7	—
	05/11/2020	**58.7**	—	166.4	186.7	—	137.7	—
	05/18/2020	**43.4**	—	115.6	135.2	156.7	118.1	—
	05/24/2020	**51.4**	—	—	—	—	142.4	132.9
	05/25/2020	**36.6**	—	131.3	139.5	164.3	138.2	—
	05/31/2020	**37.7**	—	—	—	—	152.9	154.5
	06/01/2020	**28.5**	—	148.2	162.0	176.1	153.4	—
	06/07/2020	**33.8**	—	—	—	—	157.2	155.2
	06/08/2020	**27.5**	—	154.7	169.5	168.9	150.8	—
	06/15/2020	**39.6**	—	—	178.8	167.7	151.5	—
	06/20/2020	**106.2**	—	—	—	—	197.1	—
	06/22/2020	**184.3**	—	—	244.6	265.9	255.3	—
	06/27/2020	**75.0**	—	—	—	—	158.3	—
	06/29/2020	**90.8**	—	—	151.6	174.8	158.4	—
7	04/20/2020	**142.9**	225.7	217.8	864.6	—	196.9	—
	04/27/2020	**88.6**	292.9	146.5	931.6	472.6	102.5	—
	05/04/2020	**79.5**	158.9	171.6	1004.5	587.6	149.1	—
	05/11/2020	**52.5**	—	167.4	198.2	—	135.4	—
	05/18/2020	**36.3**	—	114.3	141.0	170.7	119.8	—
	05/24/2020	**41.5**	—	136.9	—	—	143.9	132.1
	05/25/2020	**31.3**	—	131.8	145.5	171.0	140.6	—
	05/31/2020	**37.6**	—	—	—	—	154.4	156.6
	06/01/2020	**34.7**	—	148.9	166.8	182.3	155.2	–
	06/07/2020	**32.8**	—	—	—	—	159.1	158.8
	06/08/2020	**30.3**	—	157.0	173.2	176.2	152.1	—
	06/15/2020	**31.4**	—	—	182.1	175.3	153.3	—
	06/20/2020	**154.4**	—	—	—	—	228.2	—
	06/22/2020	**200.8**	—	—	258.3	284.5	270.1	—
	06/27/2020	**79.2**	—	—	—	—	160.2	—
	06/29/2020	**110.1**	—	—	154.0	185.8	160.1	—
14	04/20/2020	**203.7**	306.5	—	—	—	—	—
	04/27/2020	**254.9**	482.4	—	—	394.4	—	—
	05/04/2020	**125.2**	226.7	—	1079.2	551.5	146.5	—
	05/11/2020	**109.3**	—	178.5	244.7	—	136.9	—
	05/18/2020	**104.1**	—	—	175.0	221.2	138.4	—
	05/24/2020	**76.7**	—	144.4	—	—	158.3	135.9
	05/25/2020	**61.5**	—	133.9	165.3	195.9	150.8	—
	05/31/2020	**53.2**	—	—	—	—	163.1	168.1
	06/01/2020	**44.9**	—	152.9	185.1	204.1	161.7	—
	06/07/2020	**50.0**	—	—	—	—	165.3	171.6
	06/08/2020	**42.5**	—	166.1	187.4	203.5	157.6	—
	06/15/2020	**139.6**	—	—	241.9	263.5	211.0	—
	06/22/2020	**224.9**	—	—	279.3	323.1	296.3	—
	06/29/2020	**96.6**	—	—	161.0	217.5	168.3	—

Table 4. τ-day average MAE for forecasting the number of deaths at state level granularity. Since benchmark models from covid19-forecast-hub repository release forecasts at different dates and horizons, not all models have predictions for all prediction dates/horizons (indicated by "—"). **Bold** indicates the best.

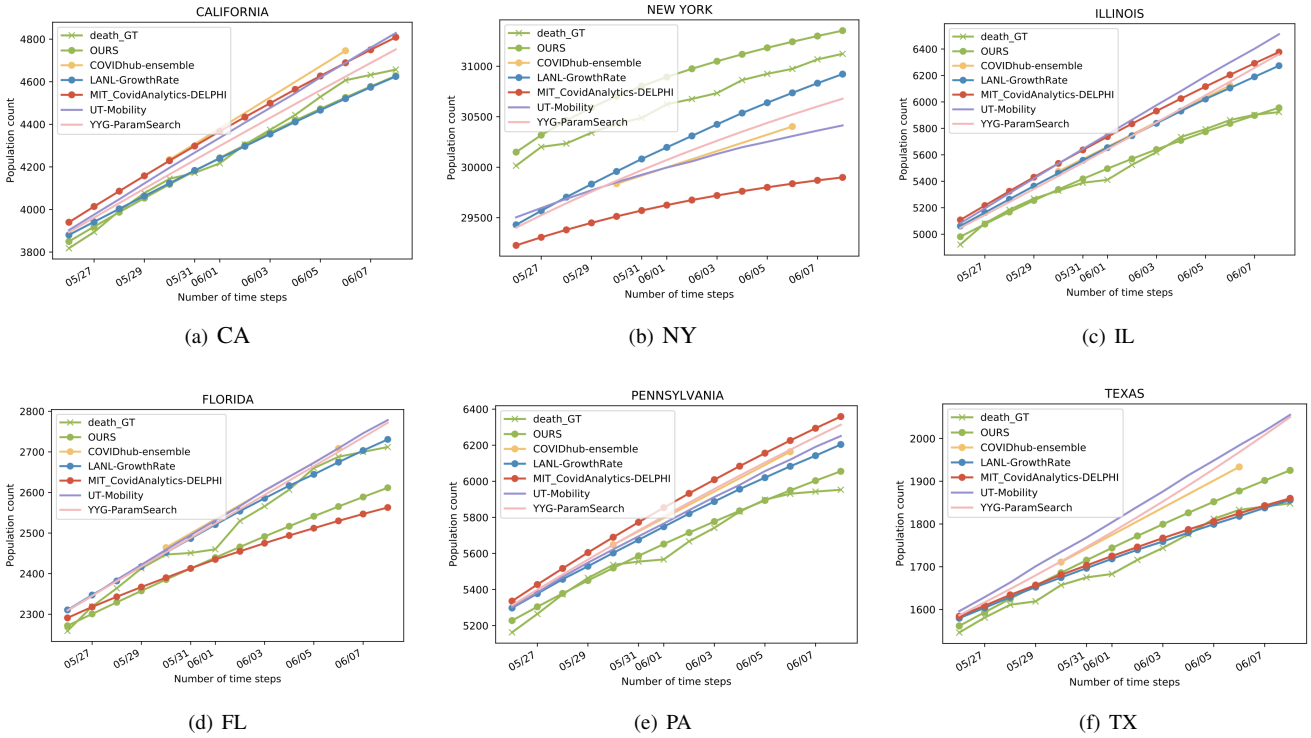
Pred. horizon τ (days)	Pred. date	Ours	IHME	LANL	UT	MIT	YYG	UCLA
5	04/20/2020	**42.8**	68.6	60.0	162.9	—	62.6	—
	04/27/2020	**39.2**	89.0	60.0	169.9	111.4	48.6	—
	05/04/2020	**36.6**	70.1	63.9	175.9	154.8	56.6	—
	05/11/2020	**27.9**	—	60.6	65.4	—	50.8	—
	05/18/2020	**25.3**	—	49.9	54.9	81.8	49.5	—
	05/24/2020	**31.2**	—	—	—	—	63.4	54.2
	05/25/2020	**20.9**	—	52.1	60.8	65.1	55.7	—
	05/31/2020	**23.6**	—	—	—	—	55.3	54.6
	06/01/2020	**17.7**	—	53.9	55.6	61.1	52.8	—
	06/07/2020	**21.8**	—	—	—	—	54.9	53.3
	06/08/2020	**13.8**	—	52.9	54.7	61.1	46.6	—
	06/15/2020	**19.6**	—	—	60.1	64.3	46.7	—
	06/20/2020	**23.3**	—	—	—	—	62.6	—
	06/22/2020	**37.1**	—	—	72.5	80.9	73.2	—
	06/27/2020	**41.4**	—	—	—	—	52.6	—
	06/29/2020	**37.0**	—	—	48.7	65.8	49.5	—
7	04/20/2020	**58.5**	72.9	74.2	172.5	—	66.6	—
	04/27/2020	**43.7**	103.3	70.0	177.5	111.8	50.3	—
	05/04/2020	**39.3**	81.3	66.7	181.1	161.8	56.8	—
	05/11/2020	**25.7**	—	64.3	70.5	—	52.4	—
	05/18/2020	**22.0**	—	—	57.3	90.3	52.7	—
	05/24/2020	**24.4**	—	63.2	—	—	65.5	55.5
	05/25/2020	**19.4**	—	53.9	65.4	69.7	58.7	—
	05/31/2020	**21.8**	—	—	—	—	57.1	57.0
	06/01/2020	**21.3**	—	56.1	57.4	65.5	54.3	–
	06/07/2020	**19.9**	—	—	—	—	57.0	57.4
	06/08/2020	**16.9**	—	56.3	57.9	68.3	49.2	—
	06/15/2020	**18.0**	—	—	63.4	70.8	49.9	—
	06/20/2020	**34.7**	—	—	—	—	70.9	—
	06/22/2020	**44.8**	—	—	79.1	90.5	78.5	—
	06/27/2020	**39.7**	—	—	—	—	54.5	—
	06/29/2020	**42.6**	—	—	52.4	74.4	51.7	—
14	04/20/2020	**112.7**	116.7	—	—	—	—	—
	04/27/2020	126.3	**121.1**	—	—	175.6	—	—
	05/04/2020	70.3	108.8	—	213.8	184.9	**65.3**	—
	05/11/2020	**55.9**	—	77.5	90.8	—	59.4	—
	05/18/2020	**52.8**	—	56.7	74.0	116.8	66.4	—
	05/24/2020	**46.7**	—	69.9	—	—	76.1	61.9
	05/25/2020	**36.4**	—	57.9	78.0	83.2	67.9	—
	05/31/2020	**32.7**	—	—	—	—	66.4	68.5
	06/01/2020	**26.7**	—	65.3	66.8	80.4	60.9	—
	06/07/2020	**29.4**	—	—	—	—	67.0	71.9
	06/08/2020	**24.5**	—	68.2	69.4	90.9	57.8	—
	06/15/2020	**38.2**	—	—	85.0	104.2	70.4	—
	06/22/2020	**51.6**	—	—	94.1	112.3	91.8	—
	06/29/2020	**40.7**	—	—	61.4	94.5	63.3	—

Table 5. τ-day ahead RMSE for forecasting the number of deaths at state level granularity. Since benchmark models from covid19-forecast-hub repository release forecasts at different dates and horizons, not all models have predictions for all prediction dates/horizons (indicated by "—"). The result is summarized at a single horizon prediction τ. **Bold** indicates the best.

Pred. horizon τ (days)	Pred. date	Ours	IHME	LANL	UT	MIT	YYG	UCLA
5	06/15/2020	**28.8**	153.5	—	183.6	178.9	153.8	—
	06/22/2020	**244.3**	326.2	—	290.5	319.9	303.7	—
	06/29/2020	**64.8**	162.6	163.0	154.7	191.3	160.7	—
	07/06/2020	**92.1**	194.6	158.2	156.1	—	158.5	—
	07/13/2020	**52.9**	—	165.0	153.6	—	154.2	—
12	06/15/2020	**229.7**	321.7	—	319.2	364.9	285.4	—
	06/22/2020	**243.3**	342.8	—	301.5	365.2	324.6	—
	06/29/2020	**125.0**	172.3	193.2	170.1	250.1	180.7	—
	07/06/2020	225.8	226.0	237.2	205.5	—	220.9	—
	07/13/2020	**106.1**	—	235.8	179.3	—	187.9	—

Table 6. τ-day ahead MAE for forecasting the number of deaths at state level granularity. Since benchmark models from covid19-forecast-hub repository release forecasts at different dates and horizons, not all models have predictions for all prediction dates/horizons (indicated by "—"). The result is summarized at a single horizon prediction τ. **Bold** indicates the best.

Pred. horizon τ (days)	Pred. date	Ours	IHME	LANL	UT	MIT	YYG	UCLA
5	06/15/2020	**19.2**	63.0	—	63.4	74.1	49.7	—
	06/22/2020	**48.5**	100.5	—	89.7	104.0	89.8	—
	06/29/2020	**33.7**	57.8	65.1	52.2	81.2	52.2	—
	07/06/2020	**34.1**	71.0	56.0	58.0	—	59.5	—
	07/06/2020	**34.3**	—	65.8	55.4	—	56.3	—
12	06/15/2020	**69.1**	120.9	—	118.5	152.8	101.4	—
	06/22/2020	**66.7**	120.7	—	112.6	120.7	107.7	—
	06/29/2020	**64.9**	73.0	99.8	73.4	117.9	80.8	—
	07/06/2020	**88.2**	98.5	101.0	92.9	—	95.1	—
	07/13/2020	**71.1**	—	110.4	87.6	—	78.5	—



(a) CA (b) NY (c) IL

(d) FL (e) PA (f) TX

Figure 5. Ground-truth vs. predicted number death forecasts for 6 states: CA, NY, IL, FL, PA, and TX on prediction date of 05/25/2020. Our model outperforms other alternatives across most of the states.

State-level forecasts:

We compare our method to various widely-used benchmarks for state-level predictions of the number of deaths in each US state. Forecasts from the top-performing and/or most widely-cited models (according to https://covid19-projections.com/about/#historical-performance) are retrieved from the `covid19-forecast-hub` repository[7]. Specifically, we report comparisons with the CurveFit model from the Institute for Health Metrics and Evaluation (IHME)[47], the GrowthRate model[49] from Los Alamos National Laboratory (LANL), the DELPHI model[50] from the Massachusetts Institute of Technology (MIT), UCLA SuEIR (UCLA), UT Mobility (UT) and the YYG model[51]. We ensure a fair comparison by making sure that all models use the same amount of data for model development (including training and validation), and forecast the same τ days. We focus on the prediction dates on which most models publish their forecasts for the τ day horizon. Note that, in contrast to usual machine learning benchmarks, these models may change significantly between forecast dates. E.g., on May 4, IHME switched from a primarily simulation-based model to a multi-stage hybrid model incorporating simulation and statistical modeling components. Errors (RMSE or MAE, over all 50 US states and Washington, DC) are computed relative to the Johns Hopkins[46] ground truth data.

Fig. 5 exemplifies our forecasting on different states for reported deaths. We observe that the comparison models, especially those rely on simple SEIR dynamics such as the MIT one, suffer from having a model bias while trying to fit the training data with low model capacity, and end up yielding big inconsistencies at the beginning of the prediction date. We observe that our model fits the ground truth trajectory more accurately compared to the alternatives, consistently across different states.

Table 3 has comparisons of our models RMSE for different prediction dates and forecasting horizons τ and Table 4 compares MAE. Overall, our model outperforms the other methods by a large margin. Accuracy of our model is consistent across different phases of the pandemic and across the time horizons considered. Notably, we observe that the performance of our model improves over time without any modifications, as it learns from more training data. With its larger learning capacity and appropriate inductive bias, model takes advantage of more training data in ways other models cannot as the learnable encoders can generalize better to unseen data and they contribute to accurate forecasts.

County-level forecasts:

County-level forecasting has been studied much less than state-level forecasting. We compare our method to Berkeley Yu

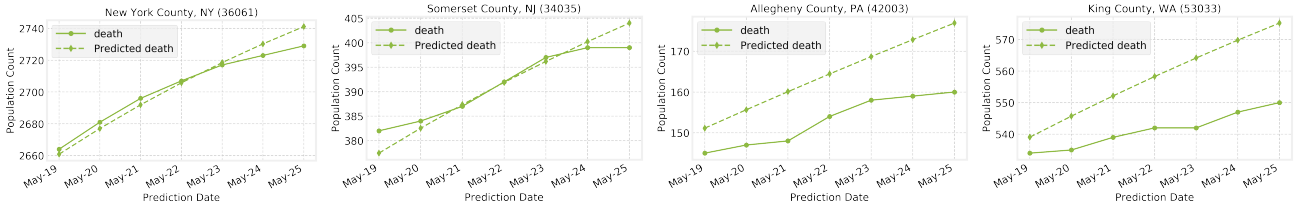
[7] https://github.com/reichlab/covid19-forecast-hub

Table 7. τ-day average MAE and RMSE for county-level forecasts.

Pred. horizon τ (days)	Pred. date	Death MAE	Death RMSE	Hospitalized MAE	Hospitalized RMSE	Confirmed MAE	Confirmed RMSE
7	05/11/2020	1.45	5.33	4.99	20.48	26.01	78.41
	05/18/2020	1.13	3.68	3.61	12.79	18.82	46.82
	05/25/2020	1.16	3.66	4.77	16.38	20.83	85.96
	05/30/2020	0.90	3.01	11.50	32.44	21.19	79.05
	06/08/2020	1.21	3.26	5.34	14.17	20.77	79.46
	06/20/2020	1.16	4.68	2.05	5.21	23.72	114.29
	06/27/2020	2.06	11.94	5.66	12.03	34.46	163.31
14	05/11/2020	2.41	8.44	15.10	65.53	66.75	272.87
	05/18/2020	2.30	7.26	9.23	31.08	46.48	152.81
	05/25/2020	1.96	8.58	8.75	29.41	40.66	203.96
	05/30/2020	1.74	4.63	20.34	63.67	41.49	174.16
	06/08/2020	1.50	5.10	4.62	14.14	39.10	186.96
	06/20/2020	1.97	10.68	6.09	17.07	60.08	375.98
	06/27/2020	2.30	13.26	6.49	13.62	73.25	464.14

Group's predictions for the number of deaths in each US county[52]. The group produces several different predictors (including county-specific exponential and linear predictors) and combine their forecasts using ensembling techniques, resulting in an ensemble called Combined Linear and Exponential Predictors (CLEP). The setting to compare the county is same as state-level comparisons. Table 8 demonstrates that our model yields much lower error compared to the Berkeley CLEP model.

Fig. 6 exemplifies the prediction for a few counties. Table 7 shows the performance of our model on all (more than 3000) US counties. Compared to state-level forecasting, county-level forecasting is more challenging due to sparse observations and less consistent ground truth data. We observe consistently accurate predictions across all dates, and indeed the forecasts get more accurate over time with more training data.



(a) New York County, NY (b) Somerset County, NJ (c) Allegheny County, PA (d) King County, WA

Figure 6. Ground-truth vs. predicted number of death forecasts for 4 counties: New York, Somerset, Allegheny and King. We observe highly-accurate fitting for New York as a high case count county, as well as Somerset and Allegheny as medium case count counties. Somerset suffers from mispredictions towards the end of the horizon, caused by flattening of the trend in the middle.

6.2 Extracting explainable insights

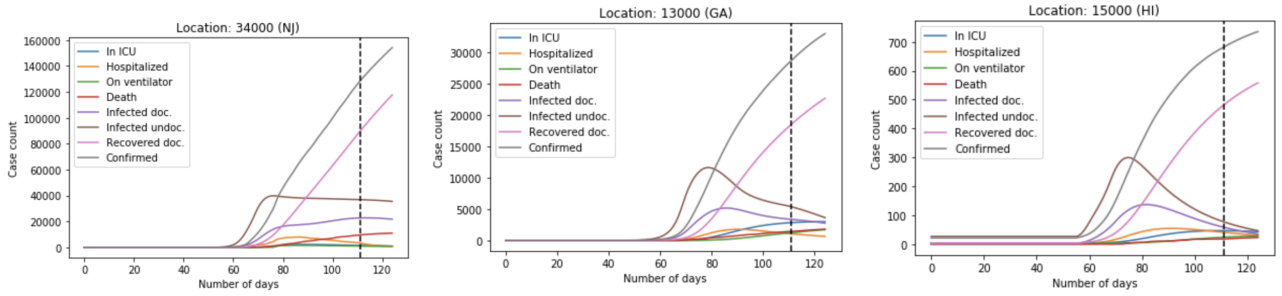
The interpretability of our model is two fold.

First, our approach is based on modeling the compartments explicitly, thus it provides insights into how the disease evolves. Fig 7 shows the fitted curves that can be used to infer important insights on where the peaking occurs, or the current decay trends. We observe the ratio of undocumented to documented infected at different phases, as well as the amount of increase/decrease for each compartment.

Second, our model uses interpretable encoders, as discussed in Sec. 4. Ignoring co-linearity (e.g. interdependence of covariates like the median income and the number of people on food-stamps), rough insights can be inferred. Fig. 8 shows the learned weights of the time-varying covariates for $\beta^{(u)}$. The weights of the past days seem similar – the model averages them with slight decay in trends. For intervention covariates, the largest weights seem to occur after a lag of a few days, suggesting their effectiveness after some lag. The positive weights of the mobility index, and negative weights of public interventions are clearly observed. Similar analysis can be performed on other variables as well. E.g. for γ, we observe the positive correlation of the positive ratio of tests. For static covariates, the insights are less apparent, but we observe meaningful learned patterns like

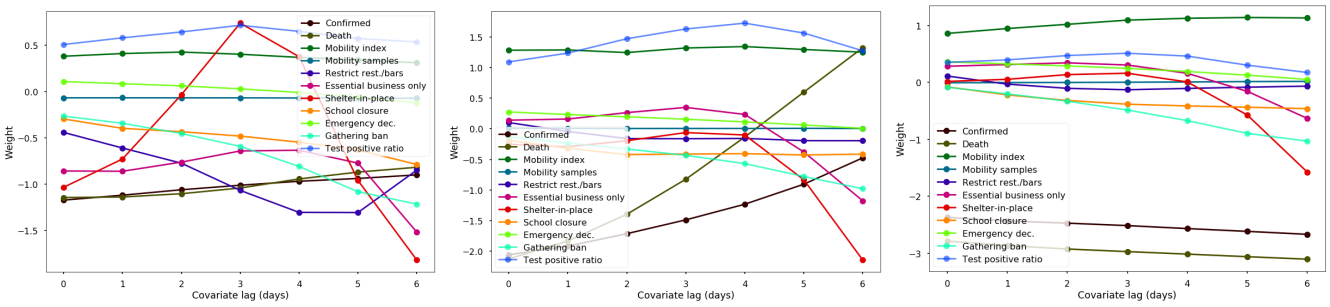
Table 8. Mean Absolute Error (MAE) for predicting cumulative deaths at daily intervals for county-level. Evaluation is "multi-horizon" (i.e., aggregated over time horizons $\leq \tau$)

Pred. horizon τ (days)	Pred. date	Ours	Berkeley CLEP
7	05/11/2020	**1.42**	1.65
	05/18/2020	**1.37**	1.67
	05/25/2020	**1.26**	1.51
	06/08/2020	**1.13**	1.54
	06/20/2020	**1.13**	1.54
14	05/11/2020	**2.41**	3.13
	05/18/2020	**2.30**	3.08
	05/25/2020	**1.96**	2.92
	06/08/2020	**1.50**	2.97
	06/20/2020	**1.97**	3.26



(a) NJ (b) GA (c) HI

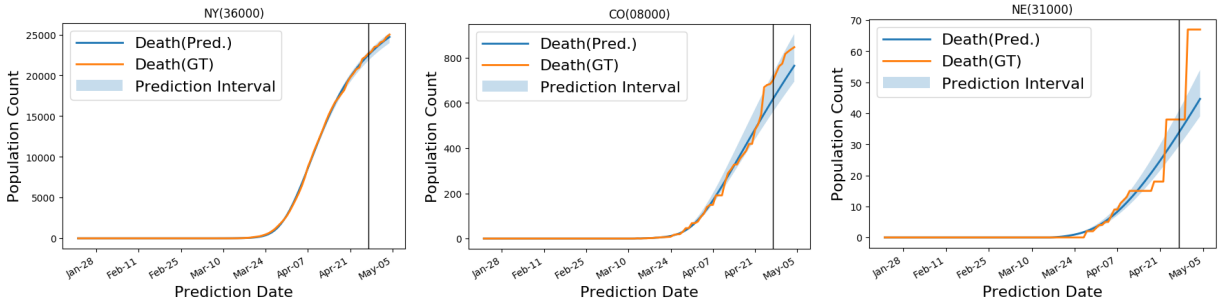
Figure 7. Fitted compartments for (a) NJ, (b) GA and (c) HI, where vertical lines show the forecasting starting timestep. Note that infected values are not cumulative, thus decay over time while the confirmed keeps increasing. These can be used to gain insights in disease evolution, e.g. we observe the increasing trend of the number of confirmed cases more sharply in NJ, whereas it is saturating in HI, due to the sharp decrease in the number of infected people after the peak.



(a) 05/24 (b) 05/31 (c) 06/07

Figure 8. Learned weights of the time-varying covariates for $\beta^{(u)}$, for 7-day state-level forecasting models on 05/24/2020, 05/31/2020 and 06/07/2020. Mobility index consistently has highly-positive impact on $\beta^{(u)}$, while gathering bans, school closures and shelter-in-place interventions have highly-negative effects. We also observe that the weight magnitude of the interventions get larger after a lag of few days.

the positive correlation of the number of households on public assistance or food stamps, population density and 60+ year old population ratio, on death rates.



(a) NY (b) CO (c) NE

Figure 9. Prediction intervals for 7-day forecasting. We use the 10-th and the 90-th quantile prediction as the the lower and the upper bound of prediction intervals, respectively. The solid vertical lines indicate the forecast horizon date. Please see Appendix for details.

6.3 Obtaining prediction intervals

In addition to point forecasts, prediction intervals could be helpful for healthcare and public policy planners to consider a range of possible scenarios. Our framework allows the capability of modeling prediction interval forecasts. To do that, we replace the ℓ_2 loss with quantile loss[53] in Eq. (5) and map the scalar propagated values to the vector of quantile estimates. Fig. 9 exemplifies well-calibrated prediction interval forecasts with slight degradation in overall RMSE (when considered for the median). The ranges tend to be wider when there are non-smooth behaviors in data, suggesting the efficacy of the fitted quantiles.

6.4 Error analysis among population subgroups

COVID-19 has affected some US counties and demographic groups more severely than others[8–12]. In this section, we analyze our model's error on US county level forecasts to ensure that it accurately reflects any trends in the ground truth data and does not distort them. Our methodology is based on binning the counties based on the percentage of each subgroups within that county and analyzing the distribution of the error within those groups using boxplots. The demographic data for this investigation was taken from 2018 US census data and our model's results are compared with the Yu Group model's results for three different forecast dates[52].

For fairness analysis, one important aspect is ensuring the distribution of the errors of a prediction method are aligned with the ground truth data. Consequently, we introduce the Normalized Mean Absolute Error (NMAE), which normalizes sum of

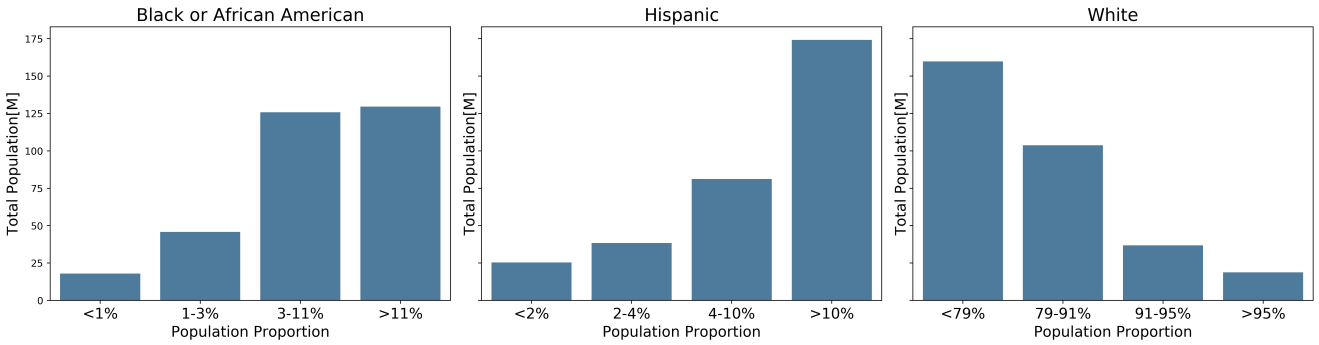


Figure 10. Total number of people in each subgroup for groups with the same number of counties.

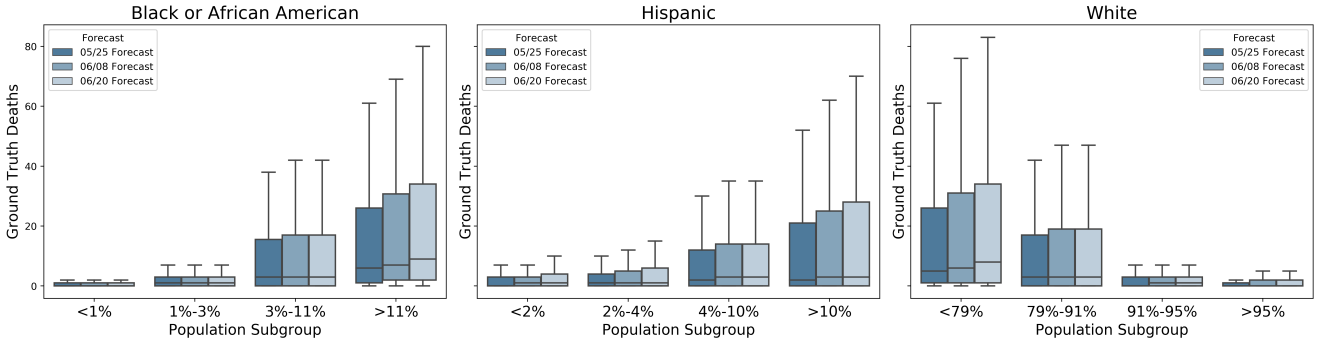


Figure 11. Boxplot of ground truth deaths per county across racial and ethnic subgroups at the end of the 14-day forecast interval for three different forecast dates.

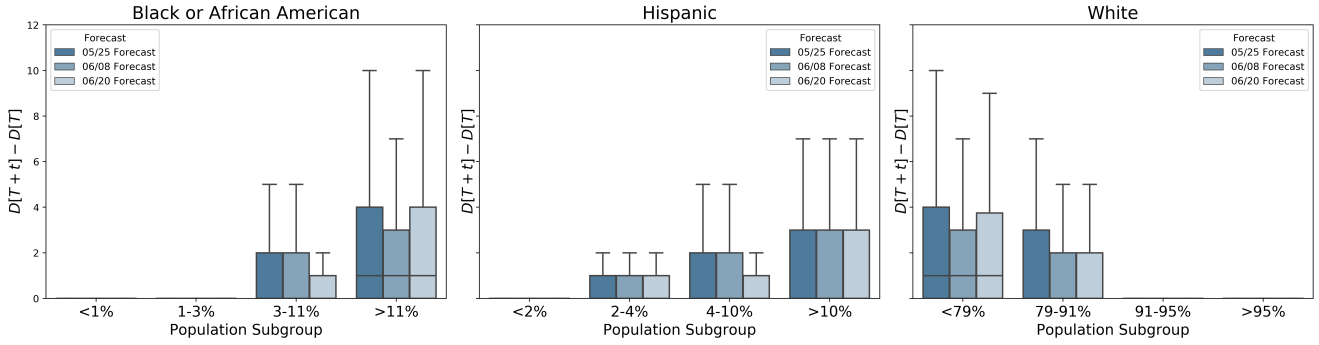


Figure 12. Boxplot of the number of incremental deaths per county for racial and ethnic subgroups over the 14-day forecast horizon for three different forecast dates.

the absolute differences by the cumulative number of deaths in a county over the forecasting horizon:

$$NMAE(T,\tau) = \frac{1}{\tau}\sum_{t=T+1}^{T+\tau}\frac{|\hat{D}[t]-D[t]|}{(D[T+\tau]-D[T+1])_1}, \tag{8}$$

where $(\cdot)_1 = \max(\cdot, 1)$. We note that there are drawbacks to this metric and encourage further exploration of better ways to normalize error metrics for time series forecasting fairness analyses in the presence of highly non-normal ground truth distributions.

6.4.1 Grouping of counties

In addition to the difficulty of defining a satisfactory metric, there are many ways in which the counties can be grouped for the comparison. The counties may be sorted by the proportion of the demographic variable of interest and then grouped into bins that have approximately equal numbers of counties, bins that have approximately equal numbers of people, or bins that have approximately equal numbers of ground truth deaths, just to name a few. For this analysis we have decided to separate the counties into groups that have equal numbers of counties (i.e. the first bin has the quarter of the counties with the lowest proportion of the population, the second bin contains the quarter of the counties whose proportions were between the 25th and 50th quantiles, etc.).

However, because we are creating groups using equal numbers of counties, this partitioning can lead to large disparities in the number of people in each bin which, in turn, can lead to skewed distributions in ground truth data. As an example of this disparity, the number of people who identified in the census as Black or African American Alone, Hispanic, or White Alone are shown in Fig. 10 and boxplot of the ground truth number of deaths for each county are shown in Fig. 11. There is a strong correlation between the number of people within the racial and ethnic subgroups and the total number of ground truth deaths and this correlation also applies to the incremental number of deaths during the forecast horizon (Fig. 12) that is used as our divisor for NMAE.

The wide variation across the subgroups is important because, as is evident from Fig. 13, our model's MAE tends to increase as the number of ground truths and the number incremental deaths increase. This matches intuition because, in general, an estimate of a large number will have a larger amount of absolute error than a small one assuming the model has relatively consistent percentage errors (e.g. a 10% error in 100 deaths is larger than a 150% error on 5 deaths). Our model has a lower MAE than the Yu Group model for all of the groups other than for the 8th decile's 5/25 forecast of ground truth deaths and the 7th decile's 6/20 forecast of the increase in ground truth deaths over the forecast interval. However, the confidence intervals overlap in both cases. In general our model tends to be significantly better for the counties with lower numbers of deaths.

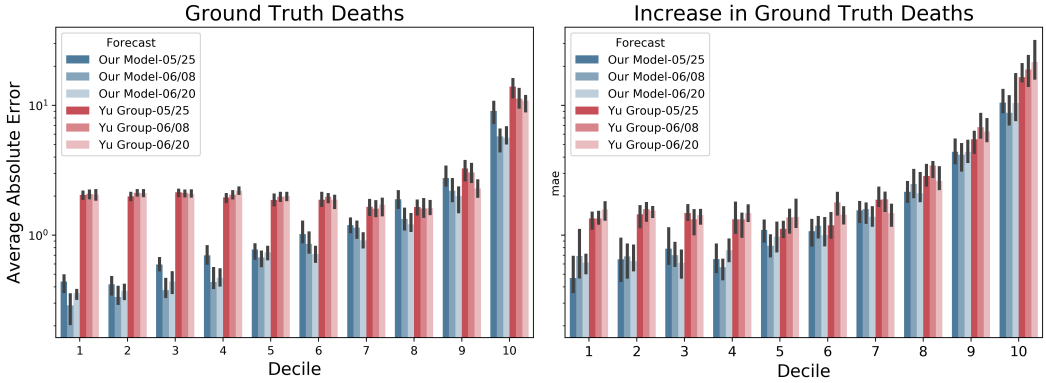


Figure 13. Median model MAE vs total cumulative ground truth deaths (left) and incremental deaths during the forecast interval (right) plotted on a log-scale for three forecast dates for counties with at least one death. The black lines show the 95% confidence interval calculated using bootstrap sampling.

6.4.2 Correlations with race and ethnicity

Our model's MAE plotted against the quartiles of the racial and minority percentiles of the counties, which is shown in Fig. 14, shows increasing error the proportion of racial and ethnic minorities in the county increase and decreasing error as the proportion of the county that is white increases. One contributing factor to this trend is that, as seen in Fig. 13, the absolute value of deaths increases as the proportion of the county that is a racial or ethnic minority increases. Therefore, if our model has a relatively consistent error rate across all groups it will have a larger absolute error for counties with larger percentages of racial and ethnic minorities. This trend is mitigated, but not all together removed, after normalize those errors by computing NMAE, Fig. 15.

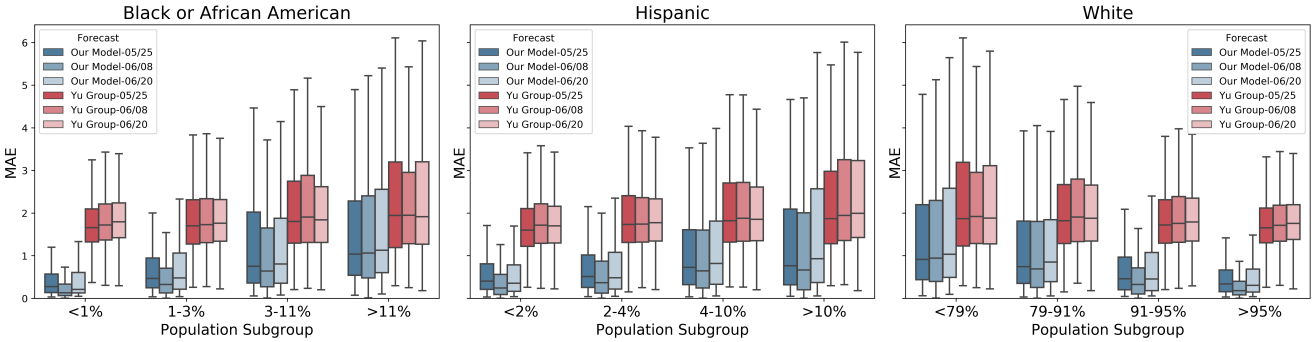


Figure 14. Boxplot of MAE across racial and ethnic minority groups for 14-day county level forecasts.

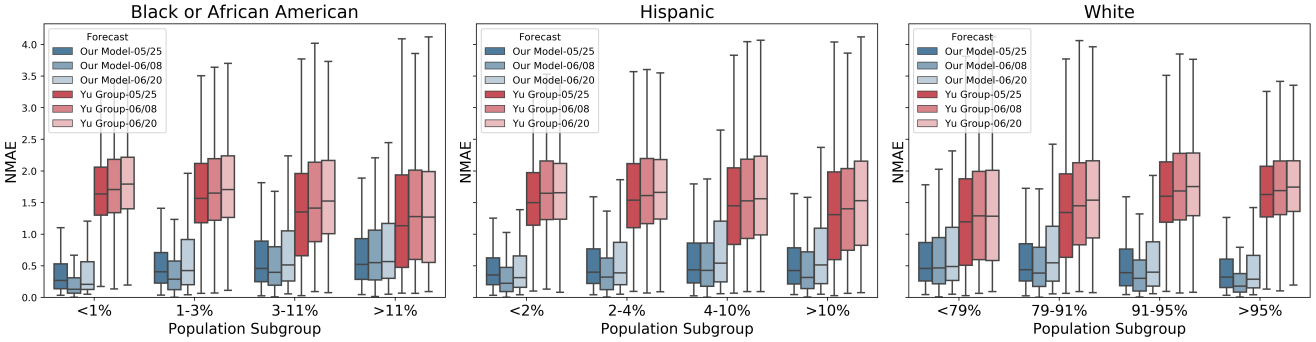


Figure 15. Boxplot of NMAE across racial and ethnic subgroups for 14-day county level forecasts.

The Yu Group model also shows similar trends to our model but to a smaller extent and has larger median error for every forecast date and population subgroup. We believe that this may be due to the fact that the Yu Group model had a lower model capacity than our model and consequently has a larger bias error. This results in larger, but more consistent, errors across the subgroups. Because of our model's higher complexity, and therefore higher variance, we are able to more accurately capture counties across the entire range of death rates but are errors are also more correlated to the magnitude of the value that we are estimating.

We also investigate the ME of the model's forecasts over the forecasting horizon to asses for bias, or over- vs. under-prediction. The positive median ME indicates that our model (Fig. 16) tends to overestimate the average number of cumulative daily deaths. However, the amount that we overestimate the number of COVID-19 deaths was smaller than the Yu Group's model for every population subgroup and forecast date. This difference was the most significant for the counties that had the smallest number of deaths.

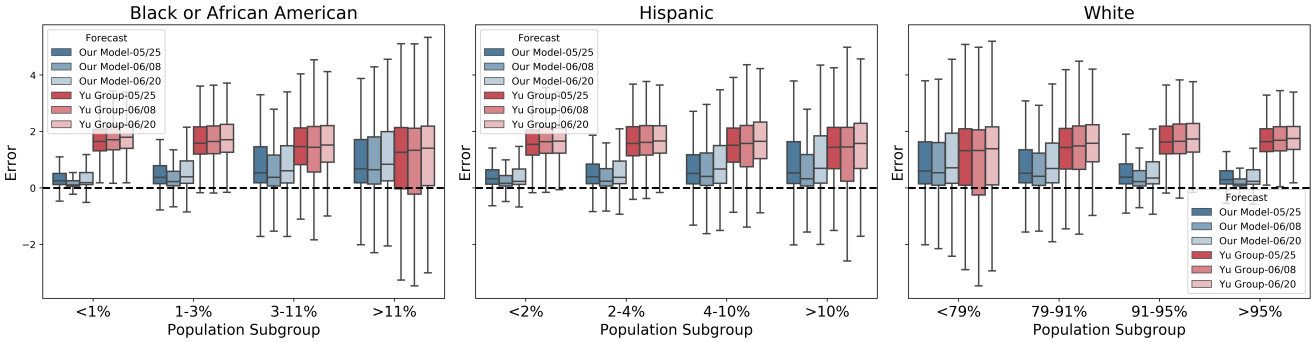


Figure 16. Boxplot of ME across racial and ethnic subgroups for 14-day county level forecasts.

One reason why our model might tend to slightly over-predict the number of cases because of the non-stationarity of the data and the slowdown of the spread during the forecasting horizons after the rapid increase during the beginning of the pandemic. This tendency could be exacerbated by the fact that we use Mean Square Error (MSE) as the loss function to train

our model and the mean of the ground truth data is larger than 86% of the data indicating that the ground truth distribution is positively skewed.

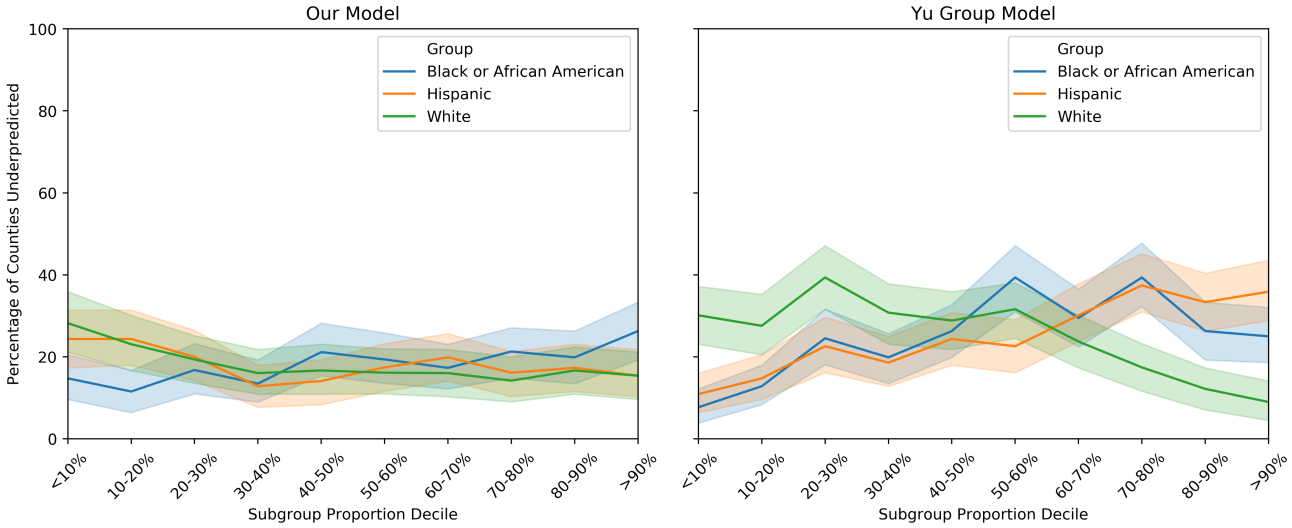


Figure 17. Percentage of counties underpredicted across racial and ethnic subgroups using 6/20/2020 predictions.

To ensure that our model did not underpredict counties in a way that is biased against specific racial or ethnic minorities, the percentage of underpredicted counties is examined for any county with more than one COVID-19 ground truth death at the end of the prediction interval. The horizontal lines in Fig. 17 show the percentage of that decile (10% of the total number of counties) that is under-predicted and the shaded regions around each line show the 95% bootstrapped confidence interval on the estimate of that percentage. The lack of separation and overlapping confidence intervals for our model suggests that our model tends to overestimate the number of deaths independently of the racial and ethnic composition of the county. The Yu Group's model tends to underpredict a larger percentage of counties when there are larger proportions of racial and ethnic minorities in those counties.

6.4.3 Correlations with median age

The relative impact of COVID-19 on counties based on their age distribution was investigated using the median age of the counties. Fig. 18 shows that despite the fact that age has been shown to have a significant impact on the fatality of COVID-19[54,55], the counties with older median ages have tended to have fewer deaths. Because the death rate of COVID-19 gets higher as people get older[56,57] this trend is likely due to other correlated factors that are beyond the scope of this analysis. Just as with in the section, the model's MAE increases with an increase in median age as shown in Fig. 18. The figure also shows that NMAE has a reduced correlation compared to MAE and that there is a slight positive bias is seen for all subgroups.

6.4.4 Correlations with median income

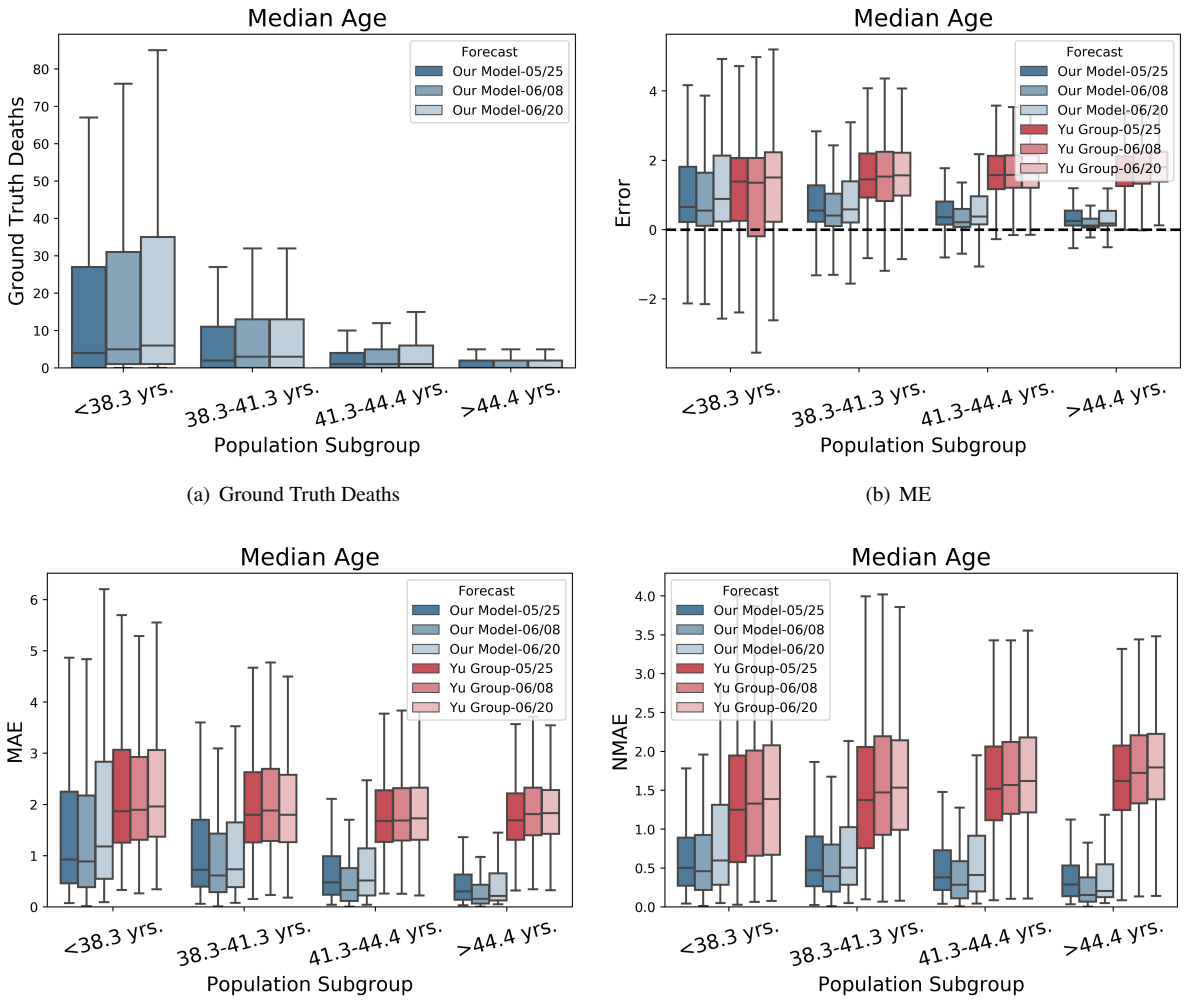
The variation on the impact of COVID-19 based on the counties' median income is shown in Fig. 19. The counties that have had the most deaths are those in the bin with the highest median income levels. The positive correlation of median income level on the prevalence of COVID-19 confirmed cases has been previously reported[58]. Despite this disparity in overall deaths, the normalized model error is more consistent across income groups and trends in the ME are relatively small.

6.4.5 Correlations with gender

The impact of COVID-19 on each county and it's relationship to the model's error is examined with respect to the percentage of the county that reported as female in the census. The death rate of COVID-19 is found to be higher among males compared to females[59], however, we observe higher death counts for counties with higher female ratios, which could be attributed to the fact that most of such counties are in dense metropolitan areas. While the number of deaths increases when the counties' percentage of population that is female increases the model's normalized error is relatively constant (Fig. 20).

6.4.6 Correlations with population density

The large increases in deaths per county for increasing population density in Fig. 21 is larger than those seen for other grouping categories and the quartile with the highest population density has a median number of deaths per county that is over an order of magnitude larger than any of the other quartiles. In this case, both MAE and NMAE show a strong positive correlation with increasing population density. It is interesting to note that in contrast to our model the NMAE for the Yu Group's model shows a decline for the counties with the highest population densities. Despite this decline our model shows lower MAE and NMAE



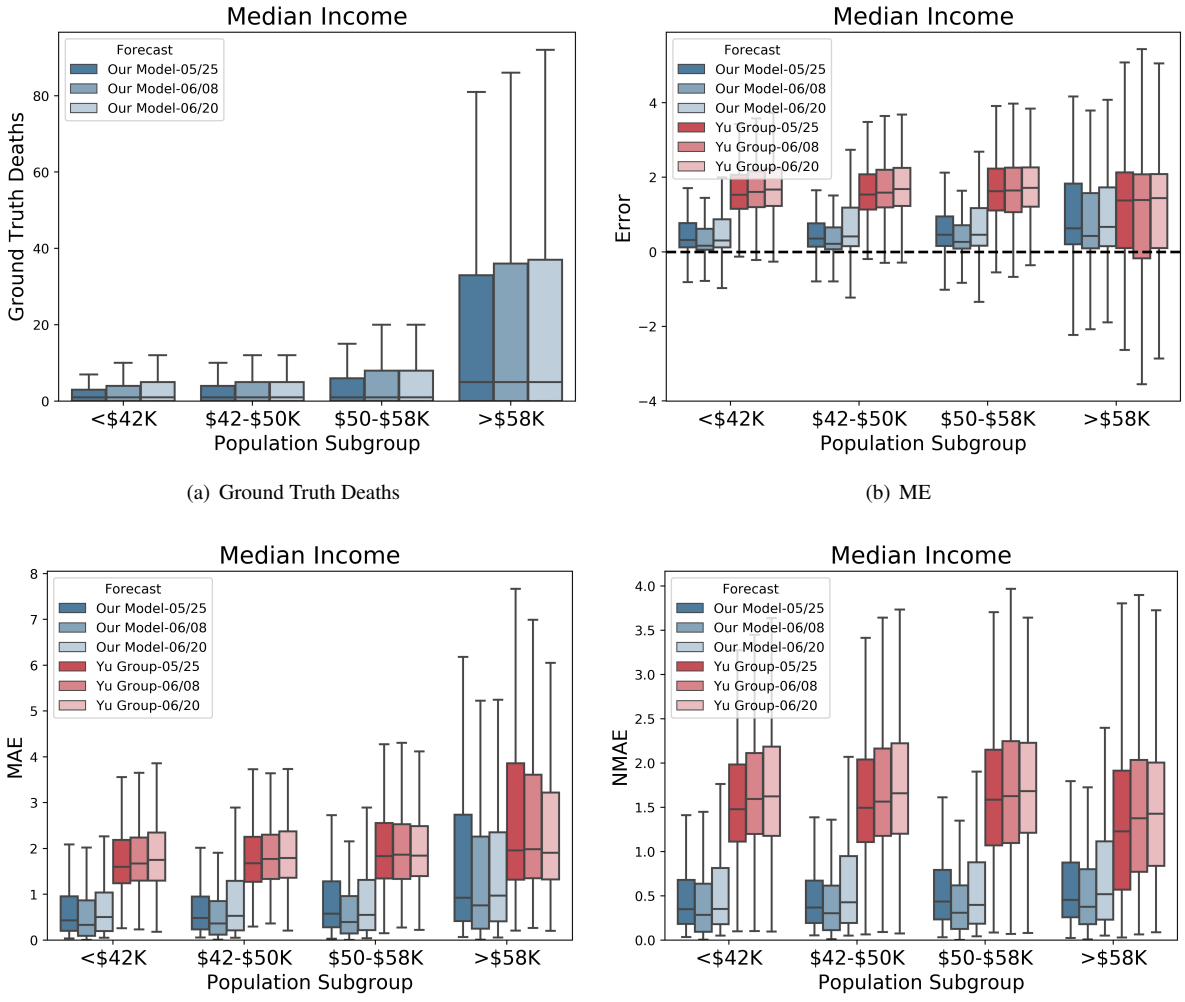
(a) Ground Truth Deaths

(b) ME

(c) MAE

(d) NMAE

Figure 18. Boxplots of COVID-19 impact and model errors across median age subgroups for 14-day predictions.



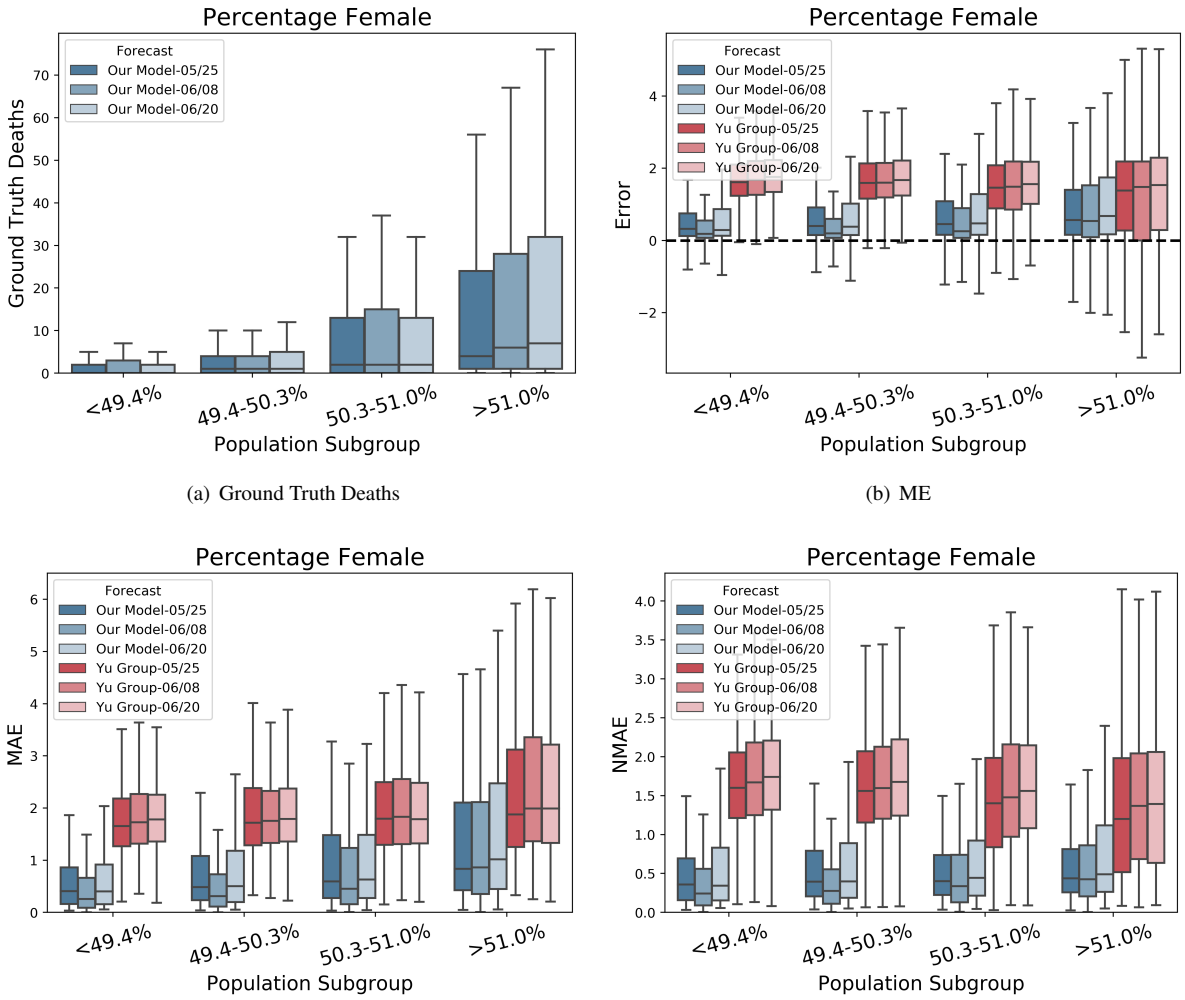
(a) Ground Truth Deaths

(b) ME

(c) MAE

(d) NMAE

Figure 19. Boxplots of COVID-19 impact and model errors across income subgroups for 14-day predictions.



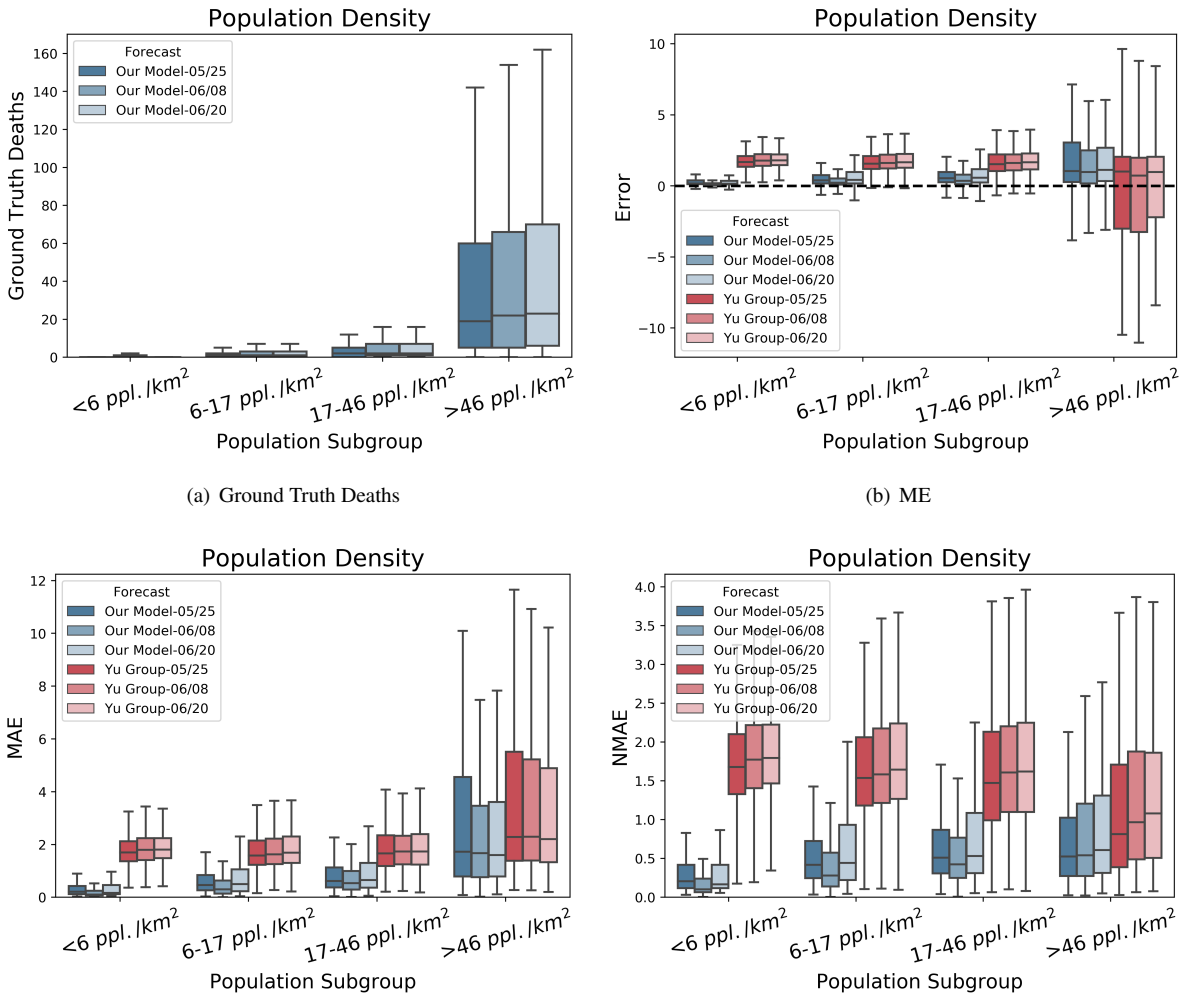
(a) Ground Truth Deaths

(b) ME

(c) MAE

(d) NMAE

Figure 20. Boxplots of COVID-19 impact and model errors across subgroups based on the fraction of the county that was female for 14-day predictions.



(a) Ground Truth Deaths

(b) ME

(c) MAE

(d) NMAE

Figure 21. Boxplots of COVID-19 impact and model errors across subgroups based on the county's population density for the 14-day predictions.

values across all groups. The ME of the our model tends to increase slightly with increasing population density while the Yu Group's model decreases and achieves a lower ME for the counties with the highest population densities.

7 Potential Limitations

In this section, we list the potential limitations and failure cases of our model, to guide those who may use the techniques to build forecasting systems that may effect public health decisions:

- **Ground-truth data issues:** We are using different case counts data to supervise model training. It has been noted that the ground truth case counts might not be completely accurate for various reasons, such as the practices to obtain case counts varying across locations[60,61]. We have weighted optimization to balance supervision from different signals, e.g. we have higher weight on the supervision from the death case count because it is known to be more accurate. Yet, the case counts data quality may also vary across locations and may affect our model's performance.

- **Failure to capture very rapid trend changes:** When the case count curves suddenly becomes very flat or very sharp, our model can fail to capture such dynamics. Some of such trends occur due to modifications in reporting practices, and some due to other factors that are not captured by the covariates we use. We believe more optimal temporal encoding approaches and integration of additional time-varying covariates may further mitigate this.

- **Using equal weight for all locations:** Our goal is to define a nation-wide metric that represents all individuals. We do not apply hand-tuned weighting for different locations, although it would be trivial with our framework. When equal weights are considered, the locations with the high case counts dominate the learning, which is often desired, but if any application requires a different emphasis mechanism, such as more accuracy for locations specifically with higher average age etc., the coefficient of the constituent locations' loss terms can be re-weighted.

- **Having symmetry in loss:** Under- vs. over-prediction have different implications on public policy, socioeconomic dynamics and public health. Our framework allows penalizing them in an asymmetric way, and we have tried different weights but we could not obtain consistent improvements when the overall accuracy is considered. Instead of overall accuracy, if an application needs to focus on under- or over-prediction specifically, the model could be retrained for improved performance.

- **Performance differences among sub-groups:** As COVID-19 is affecting certain subgroups more than others, the case counts are not uniformly distributed among the entire population. As absolute errors tend to be higher for greater case counts, there could be performance differences among different sub-groups. We do not observe our model to exacerbate the inherit case count differences (e.g. for racial and ethnic subgroups, as studied in Sect. 6.4).

- **Overfitting to the past:** Especially in early phases of the disease, our model suffers from overfitting, as the past observations may not have sufficient information content for all the dynamics of the future. We have various mitigation mechanisms to prevent overfitting, but it is impossible to completely get rid of it. We overall observe improved performance relative to the benchmarks with more training data.

- **Prediction intervals and uncertainty:** Our approach is not based on Bayesian approaches[62,63] per se (we do not estimate the posteriors of the parameters). We adapt quantile regression to obtain prediction intervals, but our approach to get prediction intervals cannot decouple the aleatoric (statistical) vs. epistemic (systematic) uncertainty inherent in the data and the model, while training. An ideal forecasting model should be able to decouple those and while providing accurate (point) forecasts, it should be able to tell the range of scenarios accurately (in a well-calibrated way). We leave such Bayesian approaches to improve our base ideas to future work.

8 Conclusions

We propose an approach to modelling infectious disease progression by incorporating covariates into a domain-specific encoding which is understandable by experts. We compare predictions for this novel model with state-of-the-art models and show that disaggregating the infected compartment into sub-compartments relevant to decision-making can make the model more useful to decision-makers.

9 Acknowledgements

Contributions of Dario Sava, Jasmin Repenning, Andrew Moore, Matthew Siegler, Ola Rozenfeld, Isaac Jones, Rand Xie, Brian Kang, Vishal Karande, Shane Glass, Afraz Mohammad, David Parish, Ron Bodkin, Hanchao Liu, Yong Li, Karthik Ramasamy, Priya Rangaswamy, Andrew Max, Tin-yun Ho, Sandhya Patil, Rif A. Saurous, Matt Hoffman, Peter Battaglia, Oriol Vinyals, Jeremy Kubica, Jacqueline Shreibati, Michael Howell, Meg Mitchell, George Teoderci, Kevin Murphy, Helen Wang, Tulsee Doshi, Garth Graham, Karen DeSalvo, David Feinberg, are gratefully acknowledged.

References

1. Health, S. Harnessing the power of data in health (2017).
2. Lim, B., Arik, S. O., Loeff, N. & Pfister, T. Temporal fusion transformers for interpretable multi-horizon time series forecasting. *arXiv:1912.09363* (2019).
3. Salinas, D., Flunkert, V. & Gasthaus, J. Deepar: Probabilistic forecasting with autoregressive recurrent networks. *arXiv:1704.04110* (2017).
4. Oreshkin, B. N., Carpov, D., Chapados, N. & Bengio, Y. N-beats: Neural basis expansion analysis for interpretable time series forecasting. *arXiv:1905.10437* (2019).
5. Kermack, W. O. & McKendrick, A. G. Contributions to the mathematical theory of epidemics—i (1927).
6. Yan, P. & Chowell, G. *Beyond the Initial Phase: Compartment Models for Disease Transmission*, chap. 4, 1–27 (Springer International Publishing, 2019).
7. Roosa, K. & Chowell, G. Assessing parameter identifiability in compartmental dynamic models using a computational approach: application to infectious disease transmission models. *Theor. Biol. Med. Model.* **16** (2019).
8. Yancy, C. W. COVID-19 and African Americans. *JAMA* **323**, 1891–1892 (2020).
9. Webb Hooper, M., Nápoles, A. M. & Pérez-Stable, E. J. COVID-19 and Racial/Ethnic Disparities. *JAMA* **323**, 2466–2467 (2020).
10. Chowkwanyun, M. & Reed, A. L. Racial health disparities and covid-19 — caution and context. *New Engl. J. Medicine* **383**, 201–203 (2020).
11. Bhala, N., Curry, G., Martineau, A. R., Agyemang, C. & Bhopal, R. Sharpening the global focus on ethnicity and race in the time of covid-19. *The Lancet* **395**, 1673–1676 (2020).
12. Henning-Smith, C., Tuttle, M. & Kozhimannil, K. B. Unequal distribution of covid-19 risk among rural residents by race and ethnicity. *The J. rural health* 10.1111/jrh.12463 (2020).
13. Smith, D. & Moore, L. The sir model for spread of disease (2004).
14. Blackwood, J. C. & Childs, L. M. An introduction to compartmental modeling for the budding infectious disease modeler. *Lett. Biomath.* **5**, 195–221 (2018).
15. Grand Rounds. Covid-19 forecasting: Fit to a curve or model the disease in real-time? (2020). https://grandrounds.com/blog/covid-19-forecasting-fit-to-a-curve-or-model-the-disease-in-real-time/, Last accessed on 2020-05-29.
16. Li, R. *et al.* Substantial undocumented infection facilitates the rapid dissemination of novel coronavirus (sars-cov-2). *Science* **368**, 489–493 (2020).
17. Koller, D. & Friedman, N. *Probabilistic graphical models: principles and techniques* (MIT press, 2009).
18. Osthus, D., Hickmann, K. S., Caragea, P. C., Higdon, D. & Del Valle, S. Y. Forecasting seasonal influenza with a state-space SIR model. *The annals applied statistics* **11**, 202 (2017).
19. Chinazzi, M. *et al.* The effect of travel restrictions on the spread of the 2019 novel coronavirus (covid-19) outbreak. *Science* **368**, 395–400, DOI: 10.1126/science.aba9757 (2020).
20. Pei, S., Kandula, S. & Shaman, J. Differential effects of intervention timing on covid-19 spread in the united states. *medRxiv* (2020).
21. Flaxman, S. *et al.* Report 13 - estimating the number of infections and the impact of non-pharmaceutical interventions on covid-19 in 11 european countries (2020).
22. Horrocks, J. & Bauch, C. T. Algorithmic discovery of dynamic models from infectious disease data. *Sci. Reports* **10** (2020).
23. Capasso, V. Reaction-diffusion models for the spread of a class of infectious diseases. In Neunzert, H. (ed.) *Proceedings of the Second European Symposium on Mathematics in Industry*, vol. 3, 181–194 (Springer, Dordrecht, 1988).
24. Hunter, E., Namee, B. M. & Kelleher, J. An open-data-driven agent-based model to simulate infectious disease outbreaks. *PLoS ONE* **13** (2018).
25. White, S. H., del Rey, A. M. & Sánchez, G. R. Modeling epidemics using cellular automata. *Appl. Math. Comput.* **186**, 193–202 (2006).
26. Venna, S. R. *et al.* A novel data-driven model for real-time influenza forecasting. *IEEE Access* **7**, 7691–7701 (2019).

27. Yang, Z. *et al.* Modified seir and ai prediction of the epidemics trend of covid-19 in china under public health interventions. *J. Thorac. Dis.* **12**, 165–174 (2020).

28. Wang, L., Chen, J. & Marathe, M. Tdefsi: Theory guided deep learning based epidemic forecasting with synthetic information (2020). 2002.04663.

29. Greydanus, S., Dzamba, M. & Yosinski, J. Hamiltonian neural networks. *arXiv:1906.01563* (2019).

30. Cranmer, M. *et al.* Lagrangian neural networks (2020). 2003.04630.

31. Lutter, M., Ritter, C. & Peters, J. Deep lagrangian networks: Using physics as model prior for deep learning. *arXiv:1907.04490* (2019).

32. Iclr 2020 workshop on integration of deep neural models and differential equations. http://iclr2020deepdiffeq.rice.edu/. Accessed: 2020-06-04.

33. Chow, C. C., Chang, J. C., Gerkin, R. C. & Vattikuti, S. Global prediction of unreported sars-cov2 infection from observed covid-19 cases. *medRxiv* DOI: 10.1101/2020.04.29.20083485 (2020).

34. Maugeri, A., Barchitta, M., S., B. & A., A. Estimation of unreported novel coronavirus (sars-cov-2) infections from reported deaths: A susceptible-exposed-infectious-recovered-dead model. *J. Clin. Medicine* **9**, 998–1006 (2020).

35. Hortaçsu, A., Liu, J. & Schwieg, T. Estimating the fraction of unreported infections in epidemics with a known epicenter: an application to covid-19. Working Paper 27028, National Bureau of Economic Research (2020). DOI: 10.3386/w27028.

36. Fu, X. Global analysis of daily new covid-19 cases reveals many static-phase countries including us and uk potentially with unstoppable epidemics. *medRxiv* (2020).

37. Long, Y.-S. *et al.* Quantitative assessment of the role of undocumented infection in the 2019 novel coronavirus (covid-19) pandemic. *arXiv:2003.12028* (2020).

38. Biddison, E. *et al.* Scarce resource allocation during disasters: A mixed-method community engagement study. *Chest* **153**, 187–195 (2018).

39. Kirkcaldy, R. D., King, B. A. & Brooks, J. T. Covid-19 and postinfection immunity: Limited evidence, many remaining questions. *JAMA* (2020).

40. van den Driessche, P. & Watmough, J. *Further Notes on the Basic Reproduction Number*, chap. 6, 159–178. Lecture Notes in Mathematics, LNM vol 1945 (Springer, Berlin, Heidelberg, 2008).

41. Hastie, T. & Tibshirani, R. Generalized additive models. *Stat. Sci.* **1**, 297–310, DOI: 10.1214/ss/1177013604 (1986).

42. Hochreiter, S. & Schmidhuber, J. Long short-term memory. *Neural computation* **9**, 1735–1780 (1997).

43. Elliott, P. & Wartenberg, D. Spatial epidemiology: current approaches and future challenges. *Environ. Heal. Perspectives* **112**, 998–1006 (2004).

44. Williams, R. J. & Zipser, D. A learning algorithm for continually running fully recurrent neural networks. *Neural Comput.* **1**, 270–280 (1989).

45. Bengio, S., Vinyals, O., Jaitly, N. & Shazeer, N. Scheduled sampling for sequence prediction with recurrent neural networks. *arXiv:1506.03099* (2015).

46. Ensheng Dong, H. D. & Gardner, L. An interactive web-based dashboard to track covid-19 in real time. *The Lancet Infect. Dis.* **20**, 533–534 (2020).

47. Murray, C. J. Forecasting covid-19 impact on hospital bed-days, icu-days, ventilator-days and deaths by us state in the next 4 months. *medRxiv* (2020).

48. Covid-Tracking. The covid tracking project (2020).

49. GrowthRate model from Los Alamos National Laboratory. https://covid-19.bsvgateway.org/#link%20to%20forecasting%20site. Accessed: 2020-06-04.

50. Bayesian compartmental models for covid-19 from University of Massachussetts, Amherst. https://github.com/dsheldon/covid. Accessed: 2020-06-04.

51. YYG model for COVID-19 forecasting. https://covid19-projections.com. Accessed: 2020-06-04.

52. Altieri, N. *et al.* Curating a covid-19 data repository and forecasting county-level death counts in the united states. *arXiv:2005.07882* (2020).

53. Wen, R., Torkkola, K., Narayanaswamy, B. & Madeka, D. A multi-horizon quantile recurrent forecaster (2017). 1711.11053.

54. Dowd, J. B. *et al.* Demographic science aids in understanding the spread and fatality rates of covid-19. *Proc. Natl. Acad. Sci. U. S. A.* **117**, 9696–9698 (2020).

55. Team, C. C.-. R. Severe outcomes among patients with coronavirus disease 2019 (covid-19)—united states, february 12–march 16, 2020. *MMWR Morb Mortal Wkly Rep. 2020* **69**, 998–1006 (2020).

56. Liu, K., Chen, Y., Lin, R. & Han, K. Clinical features of covid-19 in elderly patients: A comparison with young and middle-aged patients. *J. Infect.* **80** (2020).

57. Onder, G., Rezza, G. & Brusaferro, S. Case-Fatality Rate and Characteristics of Patients Dying in Relation to COVID-19 in Italy. *JAMA* **323**, 1775–1776 (2020).

58. Mollalo, A., Vahedi, B. & Rivera, K. M. Gis-based spatial modeling of covid-19 incidence rate in the continental united states. *Sci. The Total. Environ.* **728** (2020).

59. Jin, J.-M. *et al.* Gender differences in patients with covid-19: Focus on severity and mortality. *Front. Public Heal.* **8**, 152 (2020).

60. Pearce, N., Vandenbroucke, J. P., VanderWeele, T. J. & Greenland, S. Accurate statistics on covid-19 are essential for policy guidance and decisions. *Am. J. Public Heal.* **110**, 949–951 (2020).

61. Fenton, N., Hitman, G. A., Neil, M., Osman, M. & McLachlan, S. Causal explanations, error rates, and human judgment biases missing from the covid-19 narrative and statistics. *PsyArXiv:10.31234/osf.io/p39a4* (2020).

62. Blundell, C., Cornebise, J., Kavukcuoglu, K. & Wierstra, D. Weight uncertainty in neural networks. *arXiv:1505.05424* (2015).

63. Maddox, W., Garipov, T., Izmailov, P., Vetrov, D. P. & Wilson, A. G. A simple baseline for bayesian uncertainty in deep learning. *arXiv:1902.02476* (2019).

64. Bryant, P. & Elofsson, A. Estimating the impact of mobility patterns on covid-19 infection rates in 11 european countries. *medRxiv* DOI: 10.1101/2020.04.13.20063644 (2020).

65. Warren, M. S. & Skillman, S. W. Mobility changes in response to covid-19. *arXiv:2003.14228 [cs.SI]* (2020).

66. Realtime tracking of state-wide npi implementations. https://c19hcc.org/resources/npi-dashboard/. Accessed: 2020-06-04.

67. Wu, X., Nethery, R. C., Sabath, B. M., Braun, D. & Dominici, F. Exposure to air pollution and covid-19 mortality in the united states: A nationwide cross-sectional study. *medRxiv* (2020).

68. Bigquery public datasets. https://cloud.google.com/bigquery/public-data. Accessed: 2020-06-04.

69. Sanche, S. *et al.* High contagiousness and rapid spread of severe acute respiratory syndrome coronavirus 2. *Emerg. Infect. Dis.* **26**, DOI: 10.3201/eid2607.200282 (2020).

10 Datasets

Table 9. Covariates selected for model.

Covariate	Variables that the covariate affect
Per capita income	$\beta^{(d)}$, $\beta^{(u)}$, η, γ, $\rho^{(I,d)}$, $\rho^{(I,u)}$, $\rho^{(H)}$, $\rho^{(C)}$, $\rho^{(V)}$, h, c, v, $\kappa^{(I,d)}$, κ^{H}, κ^{C}, κ^{V}
Population density	$\beta^{(d)}$, $\beta^{(u)}$, η, γ, $\rho^{(I,d)}$
Households on food stamps	η, $\rho^{(I,d)}$, $\rho^{(I,u)}$, $\rho^{(H)}$, $\rho^{(C)}$, $\rho^{(V)}$, h, c, v, $\kappa^{(I,d)}$, κ^{H}, κ^{C}, κ^{V}
Population	All
Number of households	$\beta^{(d)}$, $\beta^{(u)}$, η, γ, $\rho^{(I,d)}$, $\rho^{(I,u)}$, $\rho^{(H)}$, $\rho^{(C)}$, $\rho^{(V)}$, h, c, v, $\kappa^{(I,d)}$, κ^{H}, κ^{C}, κ^{V}
Population ratio above age 60	$\beta^{(d)}$, $\beta^{(u)}$, η, γ, $\rho^{(I,d)}$, $\rho^{(I,u)}$, $\rho^{(H)}$, $\rho^{(C)}$, $\rho^{(V)}$, h, c, v, $\kappa^{(I,d)}$, κ^{H}, κ^{C}, κ^{V}
Hospital rating scale	η, γ, $\rho^{(I,d)}$, $\rho^{(I,u)}$, $\rho^{(H)}$, $\rho^{(C)}$, $\rho^{(V)}$, h, c, v, $\kappa^{(I,d)}$, κ^{H}, κ^{C}, κ^{V}
Available types of hospitals	η, $\rho^{(I,d)}$, $\rho^{(I,u)}$, $\rho^{(H)}$, $\rho^{(C)}$, $\rho^{(V)}$, h, c, v, $\kappa^{(I,d)}$, κ^{H}, κ^{C}, κ^{V}
Hospital patient experience rating	η, $\rho^{(I,d)}$, $\rho^{(I,u)}$, $\rho^{(H)}$, $\rho^{(C)}$, $\rho^{(V)}$, h, c, v, $\kappa^{(I,d)}$, κ^{H}, κ^{C}, κ^{V}
Air quality measures	$\beta^{(d)}$, $\beta^{(u)}$, η, $\kappa^{(I,d)}$; also for state model: h, c, v, κ^{H}, κ^{C}, κ^{V} and for county model: γ, $\rho^{(I,d)}$, $\rho^{(I,u)}$
Mobility indices	$\beta^{(d)}$, $\beta^{(u)}$
Non-pharmaceutical interventions (state model)	$\beta^{(d)}$, $\beta^{(u)}$
Total tests (state model)	γ, h
Confirmed per Total tests	$\beta^{(d)}$, $\beta^{(u)}$, γ, h
Confirmed Cases (lagged)	$\beta^{(d)}$, $\beta^{(u)}$, γ, h
Deaths (lagged)	$\beta^{(d)}$, $\beta^{(u)}$, γ, h

As discussed in the main body of this work, vast numbers of candidate datasets exist that could be related to the problem of COVID-19 forecasting. However, these datasets cannot be used indiscriminately. We select data sources based on whether they could have a predictive signal for the disease outcomes. Selecting multiple datasets from the same class of causes can obfuscate their predictive power. Therefore, we select datasets, one each from the classes of econometrics, demographics, mobility, non-pharmaceutical interventions, hospital resource availability, historical air quality. From each of these datasets, we further select covariates that could have an impact on the model compartments. We allow covariates to influence only those compartments (and hence transition rates) on which we posit that there exists a causal relationship (Table 9).

Ground Truth. We obtain primary ground truth for this work from the Johns Hopkins COVID-19 dataset[46]. Additional ground truth data that is used in the models for US states are obtained from the Covid Tracking Project[48].

Mobility. We posit that human mobility with a region, for work and personal reasons, has an effect on the average contact rates[64]. We use temporal mobility indices provided by Descartes labs at both state- and county-level resolutions[65]. These temporal indices are encoded to affect the average contact rates ($\beta^{(d)}$, $\beta^{(u)}$), at both the state- and county-level of geographic resolution.

Non-Pharmaceutical Interventions. We posit that public policy decisions restricting certain classes of population movement or interaction can have a beneficial effect on restricting the progression of the disease[20] at the state-level geographic resolution. The interventions are presented in 6 binary valued time series indicating when an intervention has been activated in one of six categories–school closures, restrictions on bars and restaurants, movement restrictions, mass gathering restrictions, essential businesses declaration, and emergency declaration[66]. This temporal covariate is encoded into the average contact rates ($\beta^{(d)}$, $\beta^{(u)}$).

Demographics. We posit that the age of the individual has a significant outcome on the severity of the disease and the mortality. The Kaiser Family Foundation[8] reports the number of individuals over the age of 60 in different US counties. We encode the effect of this static covariate into the average contact rate ($\beta^{(d)}$, $\beta^{(u)}$), the diagnosis (γ), re-infected (η), recovery ($\rho^{(I,d)}$, $\rho^{(I,u)}$, $\rho^{(H)}$, $\rho^{(C)}$, $\rho^{(V)}$) and death rates ($\kappa^{(I,d)}$, κ^{H}, κ^{C}, κ^{V}), at both the state- and county-level of geographic resolution.

Historical Air Quality. We posit that the historical ambient air quality in a region can have a deleterious effect on COVID-19 morbidity and mortality[67]. We use the BigQuery public dataset that comes from the US Environmental Protection Agency (EPA) that documents historical air quality indices at the county level [9]. This static covariate is encoded into the recovery rates (η), recovery ($\rho^{(I,d)}$, $\rho^{(I,u)}$, $\rho^{(H)}$, $\rho^{(C)}$, $\rho^{(V)}$) and death rates ($\kappa^{(I,d)}$, κ^{H}, κ^{C}, κ^{V}), at both the state- and county-level of geographic resolution.

[8] On BigQuery at c19hcc-info-ext-data:c19hcc_info_public.Kaiser_Health_demographics_by_Counties_States

[9] On Bigquery at bigquery-public-data:epa_historical_air_quality.pm10_daily_summary

Econometrics. We posit that an individual's economic status, as well as the proximity to other individuals in a region has an effect on the rates of infection, hospitalization and recovery. The proximity can be due to high population density in urban areas, or due to economic compulsions. The US census – available from census.gov and on BigQuery Public Datasets[68] – reports state- and county-level static data on population, population density, per capita income, poverty levels, households on public assistance[10]. All of these measures affect transitions into the exposed and infected compartments ($\beta^{(d)}$, $\beta^{(u)}$), as well as the recovery rates ($\rho^{(I,d)}$, $\rho^{(I,u)}$, $\rho^{(H)}$, $\rho^{(C)}$, $\rho^{(V)}$) and death rates ($\kappa^{(I,d)}$, κ^{H}, κ^{C}, κ^{V}), at both the state- and county-level of geographic resolution. In addition, for the state-level model, it also influences the hospitalization rate h, ICU rate c and ventilator rate v.

Hospital Resource Availability. We posit that when an epidemic of like COVID-19 strikes a community with such a rapid progression, local hospital resources can quickly become overwhelmed[38]. We use the BigQuery public dataset that comes from the Center for Medicare and Medicaid Services, a federal agency within the United States Department of Health and Human Services [11]. These static covariates are encoded into the diagnosis rate (γ), recovery rates ($\rho^{(I,d)}$, $\rho^{(I,u)}$, $\rho^{(H)}$, $\rho^{(C)}$, $\rho^{(V)}$), re-infected rate (η) and death rate ($\kappa^{(I,d)}$, κ^{H}, κ^{C}, κ^{V}), at both the state- and county-level of geographic resolution.

Confirmed Cases and Deaths. Past confirmed case counts and deaths can have an effect on the current values of these quantities. We include these as temporal covariates. These have an effect on the average contact rates ($\beta^{(d)}$, $\beta^{(u)}$), the diagnosis rate (γ) and the hospitalization rate h.

11 Graph Model Features

For our county-level models, we use the graph features for the covariates of mobility, mobility index, mobility samples, confirmed cases and deaths. From each covariate, we generate additional covariates using the aggregation operations of mean, median, standard deviation, maximum, and summation, applied on the neighbors on the graphs. Each new feature that is generated as the result of the Cartesian product of the base features and the aggregation operations (Eq. 4) is then added as a covariate input to the appropriate encoder (Table 2).

12 Training Initialization

The start date of the training is set to 1/21/2020. We assume that the compartmental equation regime start when the number of confirmed cases exceed 10 at state level and 3 at county level (before it, to avoid noise, we simple assign the initial values). We initialize the values as follows, where $\psi_{()} \sim U[0,1]$ denote random variables with uniform distribution: $\hat{E}_i[0] = \max(100\psi_{E_{i,1}}, 10\psi_{E_{i,2}}Q_i[0])$, $\hat{I}_i^{(d)}[0] = Q_i[0])$, $\hat{I}_i^{(u)}[0] = \max(100\psi_{E_{i,1}}, 10\psi_{E_{i,2}}Q_i[0])$, $\hat{R}_i^{(d)}[0] = R[0]$, $\hat{R}_i^{(u)}[0] = 5\psi_{R_i}R[0]$, $\hat{H}_i[0] = \mathbb{I}\{H[0]\}H[0] + 0.5\psi_{H_i}(1-\mathbb{I}\{H[0]\})Q[0]$, $\hat{C}_i[0] = \mathbb{I}\{C[0]\}C[0] + 0.2\psi_{C_i}(1-\mathbb{I}\{C[0]\})Q[0]$ and $\hat{V}_i[0] = \mathbb{I}\{V[0]\}V[0] + (1-0.5\psi_{V_i}\mathbb{I}\{V[0]\})Q[0]$. In general, our model is not too sensitive to random initialization of the initial values, and we just define wide ranges to enable exploration.

13 Effective Reproduction Number

The effective reproduction number R_e is the expected number of new infections arising directly from one infected individual in a population where all individuals are susceptible to infection[40]. The R_e for COVID-19 during the early stages of the pandemic in Wuhan, China has been estimated to be around 5.7[69].

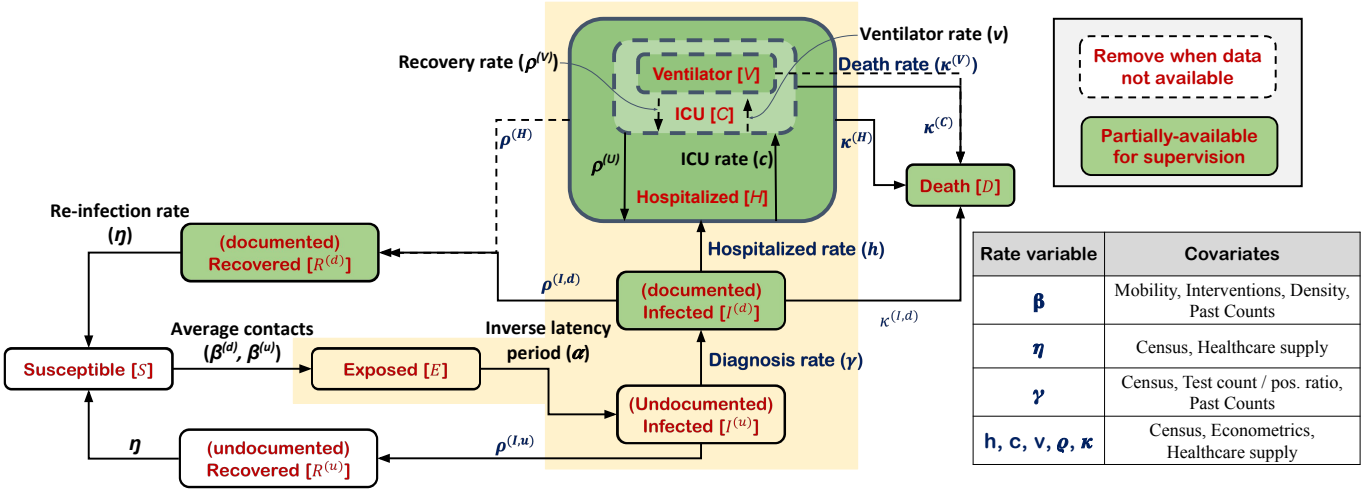
The Next-Generation Matrix[40] is a method to derive expressions for the R_e from a given compartment model. The method involves first finding the disease-free equilibrium (DFE) of the model. The infected sub-system of the compartment model at DFE is identified and its corresponding differential equations are isolated. Then the inflow and outflow terms from each compartment in the sub-system are partitioned between two categories–(i) new infection causing events and (ii) all other flows between compartments.

Two matrices–the new infections matrix **F** and the transitions matrix **V**–are constructed from the inflow and outflow terms.

The DFE for our model is $[S,E,I^{(d)},I^{(u)},R^{(d)},R^{(u)},H,C,V,D] = [N,0,0,0,0,0,0,0,0,0]$. We begin by isolating the infection subsystem as shown in Figure 22. All the individuals in these compartments $\overrightarrow{X} \equiv [E,I^{(d)},I^{(u)},H,C,V]$ are at some stage of the infection. The individuals in $\overrightarrow{Y} \equiv [S,R^{(d)},R^{(u)},D]$ are not infected. From the system of difference equations in Section 3, the

[10] On Bigquery as bigquery-public-data:census_bureau_acs.county_2018_5yr and bigquery-public-data:census_bureau_acs.county_2018_1yr

[11] On Bigquery as bigquery-public-data:cms_medicare.hospital_general_info



Rate variable	Covariates
β	Mobility, Interventions, Density, Past Counts
η	Census, Healthcare supply
γ	Census, Test count / pos. ratio, Past Counts
h, c, v, ϱ, κ	Census, Econometrics, Healthcare supply

Figure 22. Our compartment model with the infection compartments highlighted, and the variables from Table 2 indicated next to each transition.

differential equations for the infection subsystem reduces to:

$$\begin{aligned}
\dot{E} &= (\beta^{(d)} \cdot I^{(d)} + \beta^{(u)} \cdot I^{(u)}) \cdot S/N_i - \alpha \cdot E \\
\dot{I^{(d)}} &= \gamma \cdot I^{(u)} - (\rho^{(I,d)} + \kappa^{(I,d)} + h) \cdot I^{(d)} \\
\dot{I^{(u)}} &= \alpha \cdot E - (\rho^{(I,u)} + \gamma) \cdot I^{(u)} \\
\dot{H} &= h \cdot I^{(d)} - \kappa^{C} \cdot (C - V) - \kappa^{V} \cdot V - (\kappa^{(H)} + \rho^{(H)}) \cdot (H - C) \\
\dot{C} &= c \cdot (H - C) - (\kappa^{(C)} + \rho^{(C)} + \nu) \cdot (C - V) + \kappa^{(V)} \cdot V \\
\dot{V} &= \nu \cdot (C - V) - (\kappa^{(V)} + \rho^{(V)}) \cdot V
\end{aligned} \tag{9}$$

At the DFE, the subsystem is:

$$\begin{aligned}
\dot{E} &= (\beta^{(d)} \cdot I^{(d)} + \beta^{(u)} \cdot I^{(u)}) - \alpha \cdot E \\
\dot{I^{(d)}} &= \gamma \cdot I^{(u)} - (\rho^{(I,d)} + \kappa^{(I,d)} + h) \cdot I^{(d)} \\
\dot{I^{(u)}} &= \alpha \cdot E - (\rho^{(I,u)} + \gamma) \cdot I^{(u)} \\
\dot{H} &= h \cdot I^{(d)} - \kappa^{C} \cdot (C - V) - \kappa^{V} \cdot V - (\kappa^{(H)} + \rho^{(H)}) \cdot (H - C) \\
\dot{C} &= c \cdot (H - C) - (\kappa^{(C)} + \rho^{(C)} + \nu) \cdot (C - V) + \kappa^{(V)} \cdot V \\
\dot{V} &= \nu \cdot (C - V) - (\kappa^{(V)} + \rho^{(V)}) \cdot V
\end{aligned} \tag{10}$$

Examining the right-hand side of the system of equations 10, we see that it is of the form:

$$\dot{\vec{\mathbf{X}}} = \mathbf{M} \times \vec{\mathbf{X}} \tag{11}$$

where $\mathbf{X} \equiv [E, I^{(d)}, I^{(u)}, H, C, V]$ and $\mathbf{M}$ is given by:

$$\begin{bmatrix}
-\alpha & \beta^{(d)} & \beta^{(u)} & 0 & 0 & 0 \\
0 & -h - \kappa^{(I,d)} - \rho^{(I,d)} & \gamma & 0 & 0 & 0 \\
\alpha & 0 & -\gamma - \rho^{(I,u)} & 0 & 0 & 0 \\
0 & h & 0 & -\kappa^{(H)} - \rho^{(H)} & -\kappa^{(C)} + \kappa^{(H)} + \rho^{(H)} & \kappa^{(C)} - \kappa^{(V)} \\
0 & 0 & 0 & c & -c - \nu - \kappa^{(C)} - \rho^{(C)} & \kappa^{(C)} - \kappa^{(V)} + \rho^{(C)} + \nu \\
0 & 0 & 0 & 0 & \nu & -\kappa^{(V)} - \rho^{(V)} - \nu
\end{bmatrix} \tag{12}$$

Upon examination of Figure 22, we see that the only new-infection causing events are described by the rates $\beta^{(d)}$ and $\beta^{(u)}$.

We define the new infections matrix $\mathbf{F}$ as:

$$\begin{bmatrix} 0 & \beta^{(d)} & \beta^{(u)} & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix} \tag{13}$$

We calculate the transitions matrix $\mathbf{V} = -(\mathbf{M} - \mathbf{F})$ to be:

$$\begin{bmatrix} \alpha & 0 & 0 & 0 & 0 & 0 \\ 0 & h+\kappa^{(I,d)}+\rho^{(I,d)} & -\gamma & 0 & 0 & 0 \\ -\alpha & 0 & \gamma+\rho^{(I,u)} & 0 & 0 & 0 \\ 0 & -h & 0 & \kappa^{(H)}+\rho^{(H)} & \kappa^{(C)}-\kappa^{(H)}-\rho^{(H)} & -\kappa^{(C)}+\kappa^{(V)} \\ 0 & 0 & 0 & -c & c+v+\kappa^{(C)}+\rho^{(C)} & -\kappa^{(C)}+\kappa^{(V)}-\rho^{(C)}-v \\ 0 & 0 & 0 & 0 & -v & \kappa^{(V)}+\rho^{(V)}+v \end{bmatrix} \tag{14}$$

From $\mathbf{F}$ and $\mathbf{V}$ we get the Next-Generation Matrix $\mathbf{K} = \mathbf{F} \times \mathbf{V}^{-1}$:

$$\begin{bmatrix} \frac{\beta^{(d)}\gamma}{(\gamma+\rho^{(I,u)})(h+\kappa^{(I,d)}+\rho^{(I,d)})}+\frac{\beta^{(u)}}{\gamma+\rho^{(I,u)}} & \frac{\beta^{(d)}}{h+\kappa^{(I,d)}+\rho^{(I,d)}} & \frac{\beta^{(d)}\gamma}{(\gamma+\rho^{(I,u)})(h+\kappa^{(I,d)}+\rho^{(I,d)})}+\frac{\beta^{(u)}}{\gamma+\rho^{(I,u)}} & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix} \tag{15}$$

Calculating the eigenvalues of $\mathbf{K}$ gives us 5 eigenvalues that are 0, and one non-zero eigenvalue, which is the spectral radius of $\mathbf{K}$. This is the effective reproduction number R_e:

$$R_0 = \frac{\beta^{(d)}\gamma + \beta^{(u)}(h+\kappa^{(I,d)}+\rho^{(I,d)})}{(\gamma+\rho^{(I,u)})(h+\kappa^{(I,d)}+\rho^{(I,d)})}. \tag{16}$$

14 Evaluation Metrics

We use various key metrics to evaluate the quality of the forecasts. For n observations, given the predictions p_i and the ground truth a_i, below are the definiton of the evaluation metrics used:

Mean Absolute Error: $MAE = \frac{1}{n}\sum_{i=1}^{n}|p_i - a_i|$

Root Mean Square Error: $RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(p_i - a_i)^2}$

Mean Absolute Percentage Error: $MAPE = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{p_i - a_i}{a_i}\right|$

Root Mean Squared Logarithmic Error: $RMSLE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\log(p_i+1) - \log(a_i+1))^2}$

We average the performances across multiple locations and timesteps, where n would be proportional to the prediction horizon τ, as well as the number of locations.

15 Results in Other Evaluation Metrics

Tables 10 and 11 show comparisons in Root Mean Squared Logarithmic Error (RMSLE) and Mean Average Percentage Error (MAPE) metrics. We observe consistent outperformance in RMSLE metric, even with a larger margin than RMSE in most cases, showing that our model makes much less large errors on outliers. MAPE is a less reliable metric, especially at the early phases of the disease since the distribution of ground truth deaths was very non-uniform across US states, and many low-count states dominate the MAPE metric. At later phases, as the disease progresses to most locations causing a meaningful number of deaths, MAPE becomes a more meaningful metric. We observe that the MAPE of our model is much lower than the alternatives in most cases. For 7-day forecasting, we observe MAPE values as low as 4.8 %, while the alternative models still remain above 35 %.

Table 10. τ-day average RMSLE for forecasting the number of deaths at state level granularity. Since benchmark models from covid19-forecast-hub repository release forecasts at different dates and horizons, not all models have predictions for all prediction dates/horizons (indicated by "—"). **Bold** indicates the best.

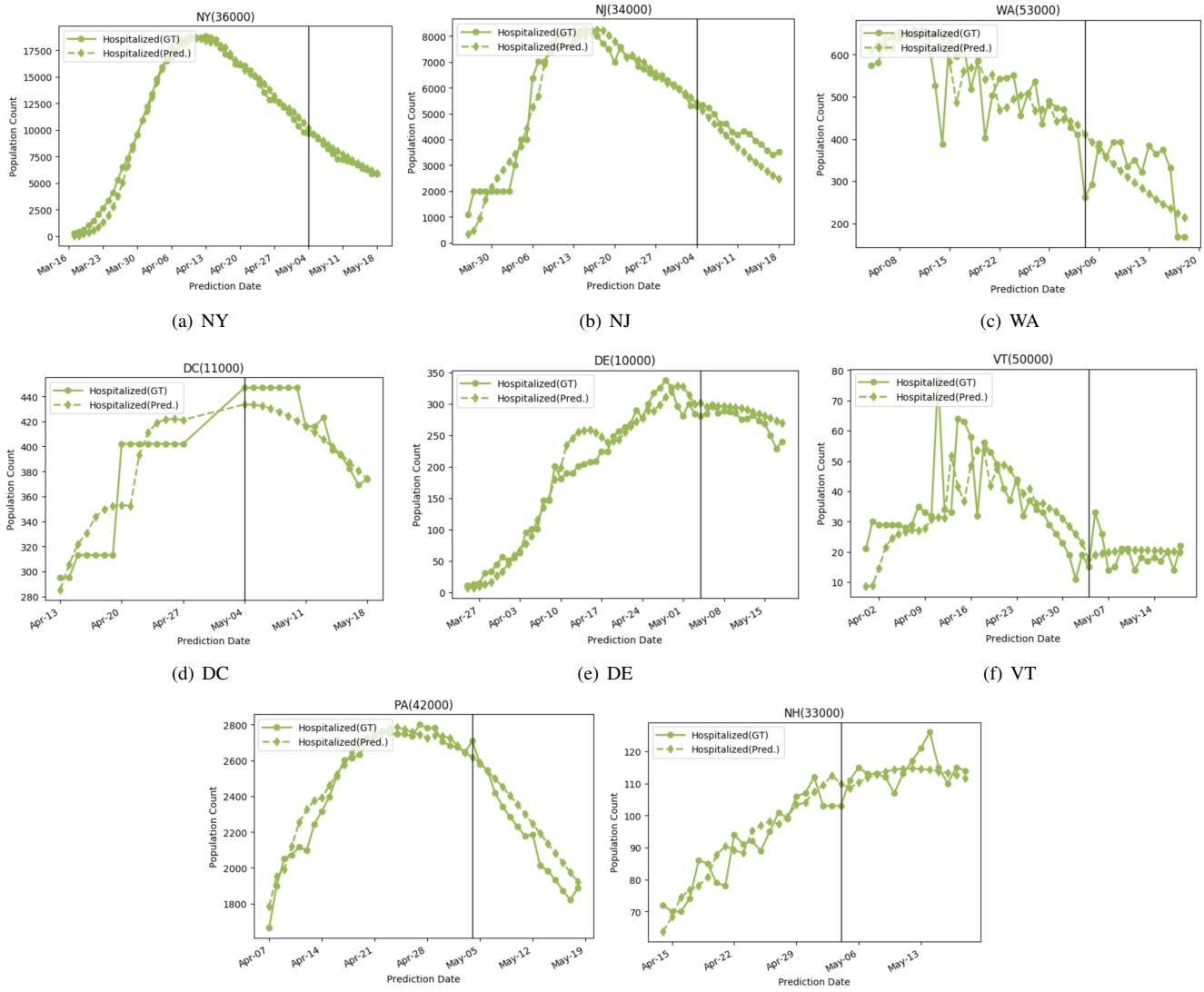
Pred. horizon τ (days)	Pred. date	Ours	IHME	LANL	UT	MIT	YYG	UCLA
5	04/20/2020	**0.404**	0.844	0.778	0.809	—	0.790	—
	04/27/2020	**0.484**	0.903	0.854	0.866	0.867	0.868	—
	05/04/2020	**0.440**	0.849	0.851	0.863	0.876	0.860	—
	05/11/2020	**0.483**	—	0.891	0.891	—	0.893	—
	05/18/2020	**0.421**	—	0.911	0.910	0.916	0.911	—
	05/24/2020	**0.371**	—	—	—	—	0.951	0.945
	05/25/2020	**0.203**	—	0.949	0.953	0.945	0.953	—
	05/31/2020	**0.382**	—	—	—	—	0.985	0.975
	06/01/2020	**0.371**	—	0.979	0.985	0.978	0.984	—
	06/07/2020	**0.315**	—	—	—	—	0.990	0.985
	06/08/2020	**0.256**	—	0.986	0.990	0.991	0.989	—
	06/15/2020	**0.400**	—	—	1.000	1.000	1.000	—
	06/20/2020	**0.156**	—	—	—	—	1.015	—
	06/22/2020	**0.248**	—	—	1.015	1.013	1.014	—
	06/27/2020	**0.486**	—	—	—	—	1.017	—
	06/29/2020	**0.504**	—	—	1.018	1.018	1.016	—
7	04/20/2020	**0.670**	0.893	0.807	0.840	—	0.820	—
	04/27/2020	**0.533**	0.913	0.849	0.866	0.863	0.866	—
	05/04/2020	**0.436**	0.854	0.855	0.869	0.884	0.866	—
	05/11/2020	**0.399**	—	0.895	0.896	—	0.898	—
	05/18/2020	**0.278**	—	—	0.912	0.919	0.914	—
	05/24/2020	**0.360**	—	0.951	—	—	0.954	0.947
	05/25/2020	**0.389**	—	0.952	0.958	0.948	0.956	—
	05/31/2020	**0.482**	—	—	—	—	0.988	0.976
	06/01/2020	**0.496**	—	0.981	0.989	0.980	0.988	–
	06/07/2020	**0.261**	—	—	—	—	0.992	0.97
	06/08/2020	**0.295**	—	0.987	0.993	0.994	0.992	—
	06/15/2020	**0.312**	—	—	1.003	1.003	1.001	—
	06/20/2020	**0.162**	—	—	—	—	1.016	—
	06/22/2020	**0.370**	—	—	1.019	1.015	1.016	—
	06/27/2020	**0.292**	—	—	—	—	1.018	—
	06/29/2020	**0.515**	—	—	1.019	1.019	1.017	—
14	04/20/2020	**0.897**	0.986	—	—	—	—	—
	04/27/2020	**0.739**	0.968	—	—	0.881	—	—
	05/04/2020	**0.508**	0.868	—	0.892	0.91	0.885	—
	05/11/2020	**0.462**	—	0.907	0.915	—	0.912	—
	05/18/2020	**0.514**	—	0.924	0.921	0.929	0.926	—
	05/24/2020	**0.578**	—	0.960	—	—	0.966	0.954
	05/25/2020	**0.502**	—	0.961	0.974	0.953	0.968	—
	05/31/2020	**0.322**	—	—	—	—	1.00	0.982
	06/01/2020	**0.296**	—	0.988	1.000	0.987	0.998	—
	06/07/2020	**0.288**	—	—	—	—	1.000	0.990
	06/08/2020	**0.268**	—	0.991	1.001	1.001	0.999	—
	06/15/2020	**0.433**	—	—	1.011	1.009	1.007	—
	06/22/2020	**0.351**	—	—	1.028	1.020	1.021	—
	06/27/2020	**0.436**	—	—	1.026	1.024	1.021	—

Table 11. τ-day average MAPE for forecasting the number of deaths at state level granularity. For MAPE, we average only non-zero terms to avoid zero terms in the denominator. Since benchmark models from covid19-forecast-hub repository release forecasts at different dates and horizons, not all models have predictions for all prediction dates/horizons (indicated by "—"). **Bold** indicates the best.

Pred. horizon τ (days)	Pred. date	Ours	IHME	LANL	UT	MIT	YYG	UCLA
5	04/20/2020	**20.1**	118.2	89.9	101.3	—	93.0	—
	04/27/2020	**19.6**	56.8	37.1	40.1	34.6	43.0	—
	05/04/2020	**14.4**	20.2	18.4	20.6	25.5	21.5	—
	05/11/2020	21.6	—	**18.3**	19.4	—	18.6	—
	05/18/2020	**11.9**	—	18.7	19.3	24.6	18.9	—
	05/24/2020	**13.1**	—	—	—	—	28.6	25.9
	05/25/2020	**4.9**	—	26.9	29.1	26.8	28.0	—
	05/31/2020	**10.4**	—	—	—	—	37.3	33.0
	06/01/2020	**9.5**	—	34.3	37.9	34.8	36.4	—
	06/07/2020	**4.6**	—	—	—	—	36.2	34.3
	06/08/2020	**6.2**	—	35.3	37.2	37.9	35.8	—
	06/15/2020	**14.5**	—	—	38.8	38.9	37.2	—
	06/20/2020	**4.9**	—	—	—	—	42.5	—
	06/22/2020	**3.9**	—	—	43.9	42.9	42.6	—
	06/27/2020	**14.8**	—	—	—	—	41.4	—
	06/29/2020	**5.4**	—	—	41.7	43.6	41.1	—
7	04/20/2020	**51.6**	122.0	79.0	89.3	—	82.3	—
	04/27/2020	**25.1**	60.5	33.1	36.6	30.7	40.4	—
	05/04/2020	**15.6**	20.9	19.0	21.8	26.7	22.7	—
	05/11/2020	**12.9**	—	19.0	20.7	—	19.3	—
	05/18/2020	**5.5**	—	—	20.0	25.5	19.5	—
	05/24/2020	**8.6**	—	27.5	—	—	29.3	26.2
	05/25/2020	**12.8**	—	27.5	30.8	27.5	28.8	—
	05/31/2020	**6.9**	—	—	—	—	38.5	33.3
	06/01/2020	**12.2**	—	35.0	39.7	35.7	37.6	–
	06/07/2020	**3.4**	—	—	—	—	36.9	34.7
	06/08/2020	**4.8**	—	35.9	38.4	39.3	36.5	—
	06/15/2020	**7.2**	—	—	40.4	40.0	37.8	—
	06/20/2020	**5.3**	—	—	—	—	43.4	—
	06/22/2020	**9.6**	—	—	45.5	43.8	43.4	—
	06/27/2020	**6.9**	—	—	—	—	41.9	—
	06/29/2020	**11.7**	—	—	42.8	44.8	41.6	—
14	04/20/2020	**92.1**	195.4	—	—	—	—	—
	04/27/2020	49.1	84.4	—	—	**26.9**	—	—
	05/04/2020	27.2	**22.3**	—	26.8	31.0	27.9	—
	05/11/2020	**15.7**	—	20.9	25.4	—	21.2	—
	05/18/2020	**18.3**	—	21.1	22.4	28.2	21.8	—
	05/24/2020	28.4	—	28.8	—	—	32.0	**27.3**
	05/25/2020	**26.0**	—	29.4	36.7	29.0	31.4	—
	05/31/2020	**8.0**	—	—	—	—	43.2	34.4
	06/01/2020	**5.3**	—	37.3	47.2	38.2	41.7	—
	06/07/2020	**4.1**	—	—	—	—	39.8	36.0
	06/08/2020	**7.3**	—	37.9	42.7	43.5	39.3	—
	06/15/2020	**7.7**	—	—	45.3	44.3	40.0	—
	06/22/2020	**6.6**	—	—	52.1	47.4	46.2	—
	06/27/2020	**19.5**	—	—	47.0	48.2	43.7	—

16 Hospitalization Prediction

Fig. 23 exemplifies fitted hospitalization predictions for 8 states. Our model can provide robust and accurate forecasts consistently (e.g. in increasing, decreasing or plateauing trends), despite the fluctuations in the past observed data.



(a) NY (b) NJ (c) WA

(d) DC (e) DE (f) VT

(g) PA (h) NH

Figure 23. Fitted hospitalization compartments for 8 states. The vertical line shows the prediction date. Our model can provide robust and accurate forecasts, despite the highly-noisy observed data.

17 State-level 7-Day Forecasts

We show the 7-day forecasts for all 50 US states.

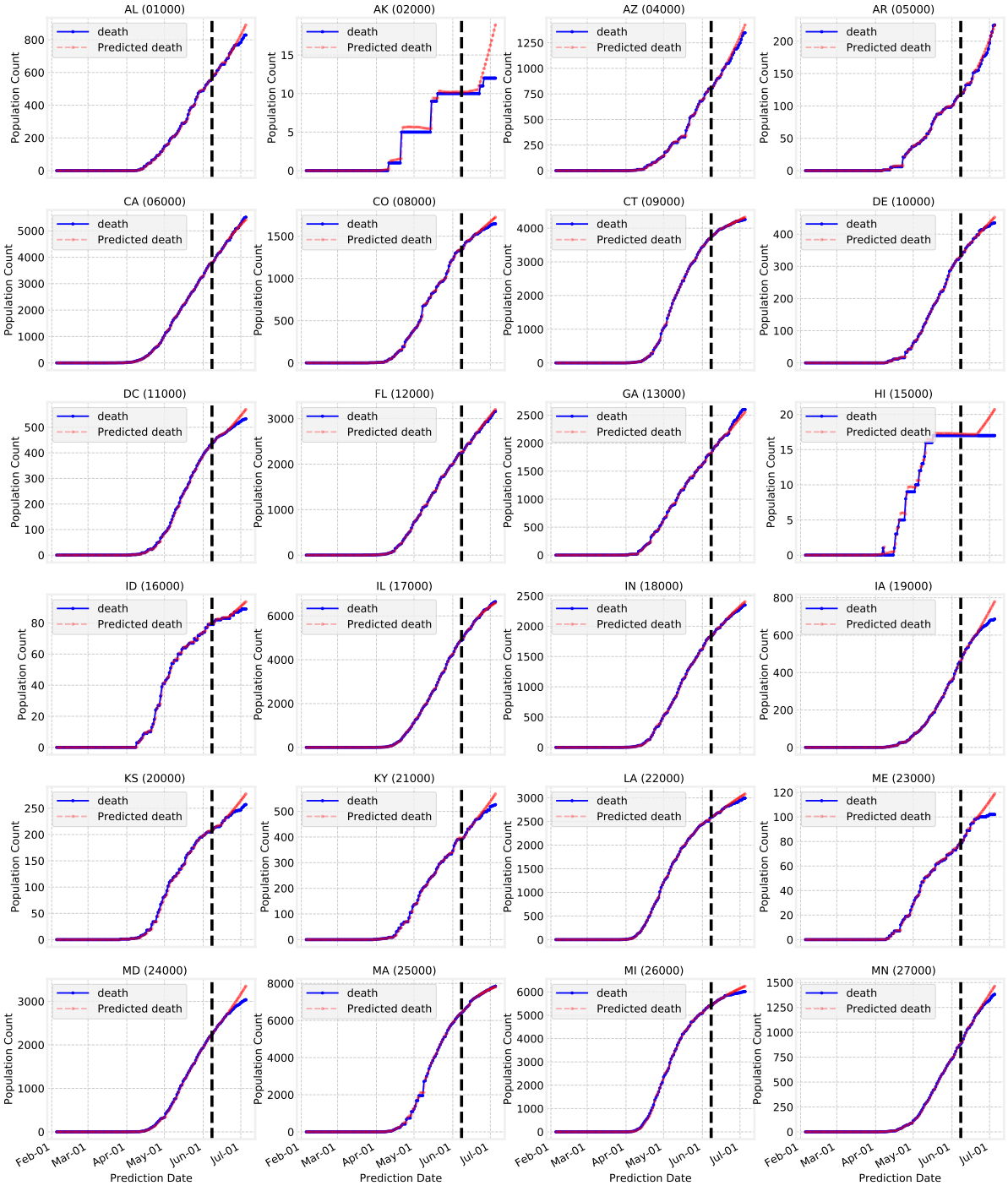


Figure 24. Model performance (deaths) on US states–Alabama to Nebraska. Black dashed line is the training horizon.

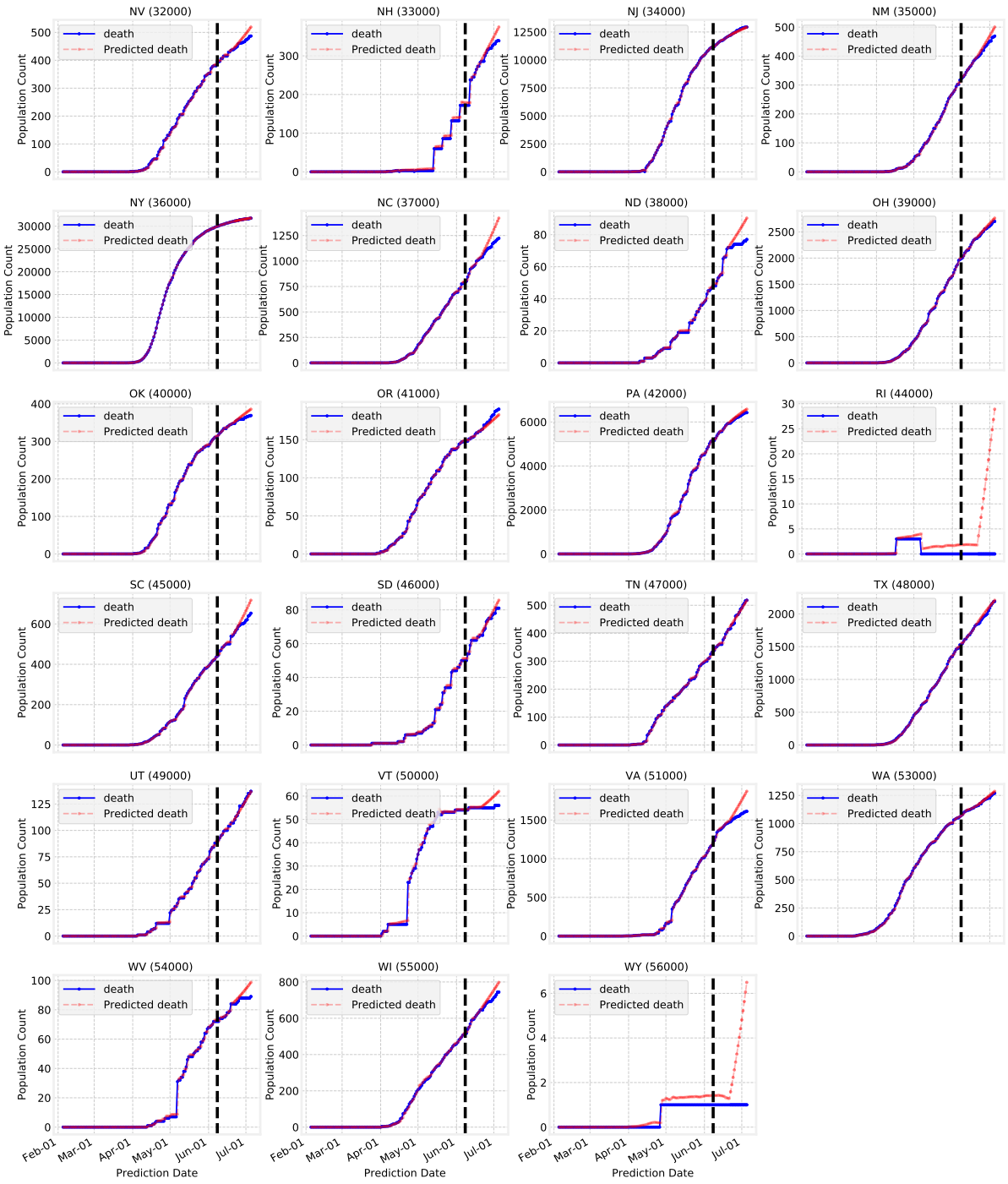


Figure 25. Model performance (deaths) on US states–Nevada to Wyoming. Black dashed line is the training horizon.

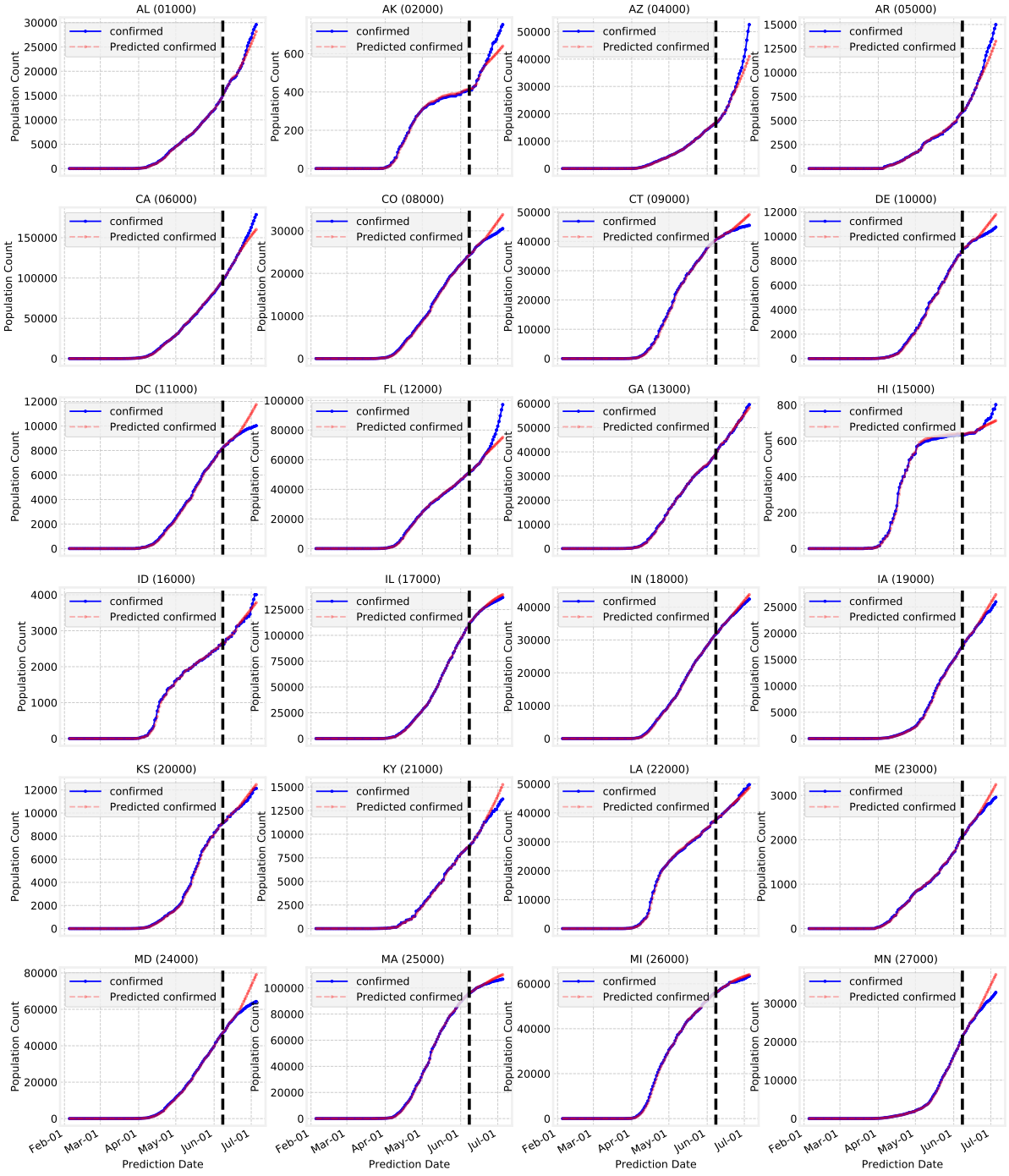


Figure 26. Model performance (confirmed cases) on US states–Alabama to Nebraska. Black dashed line is the training horizon.

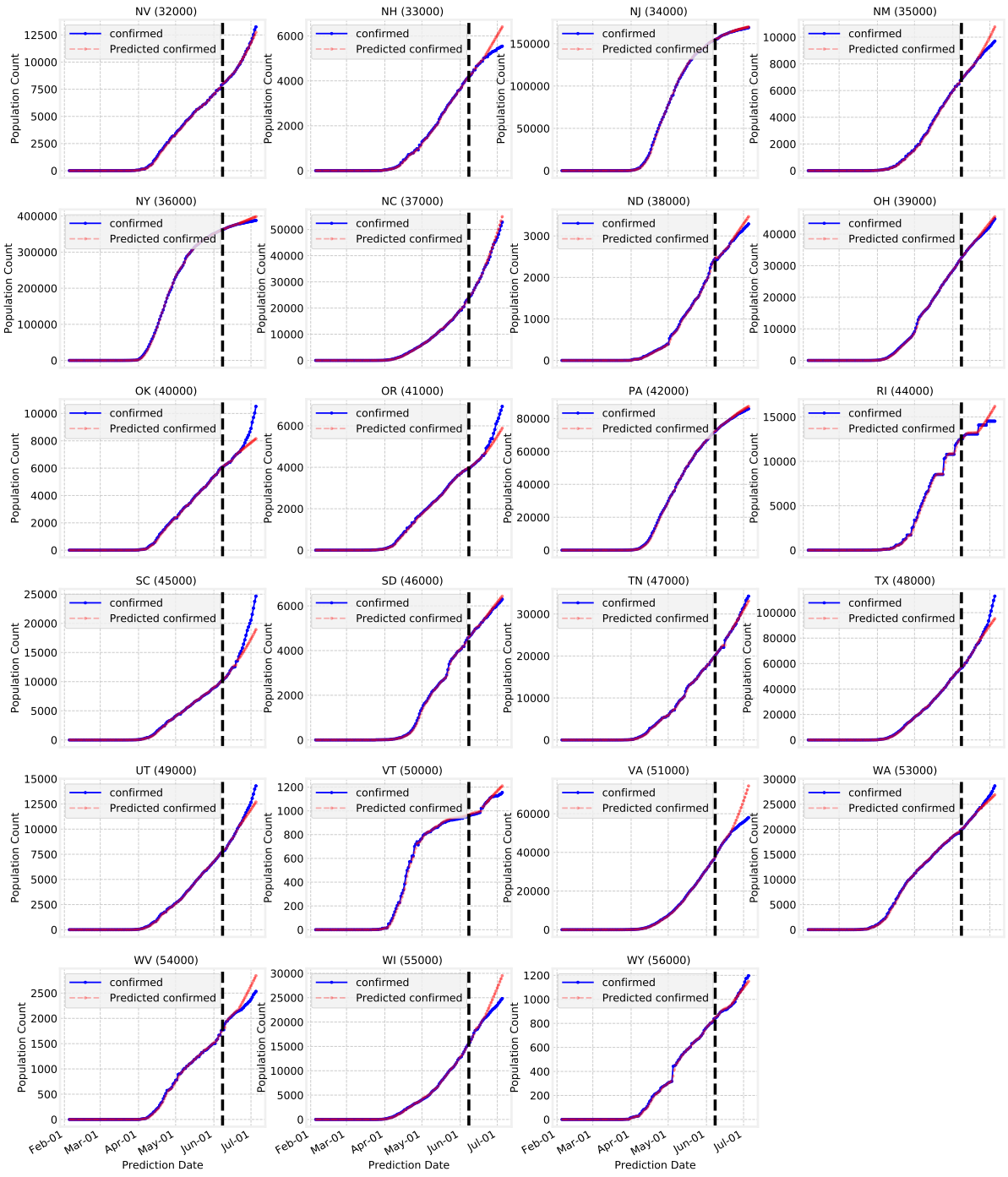


Figure 27. Model performance (confirmed cases) on US states–Nevada to Wyoming. Black dashed line is the training horizon.

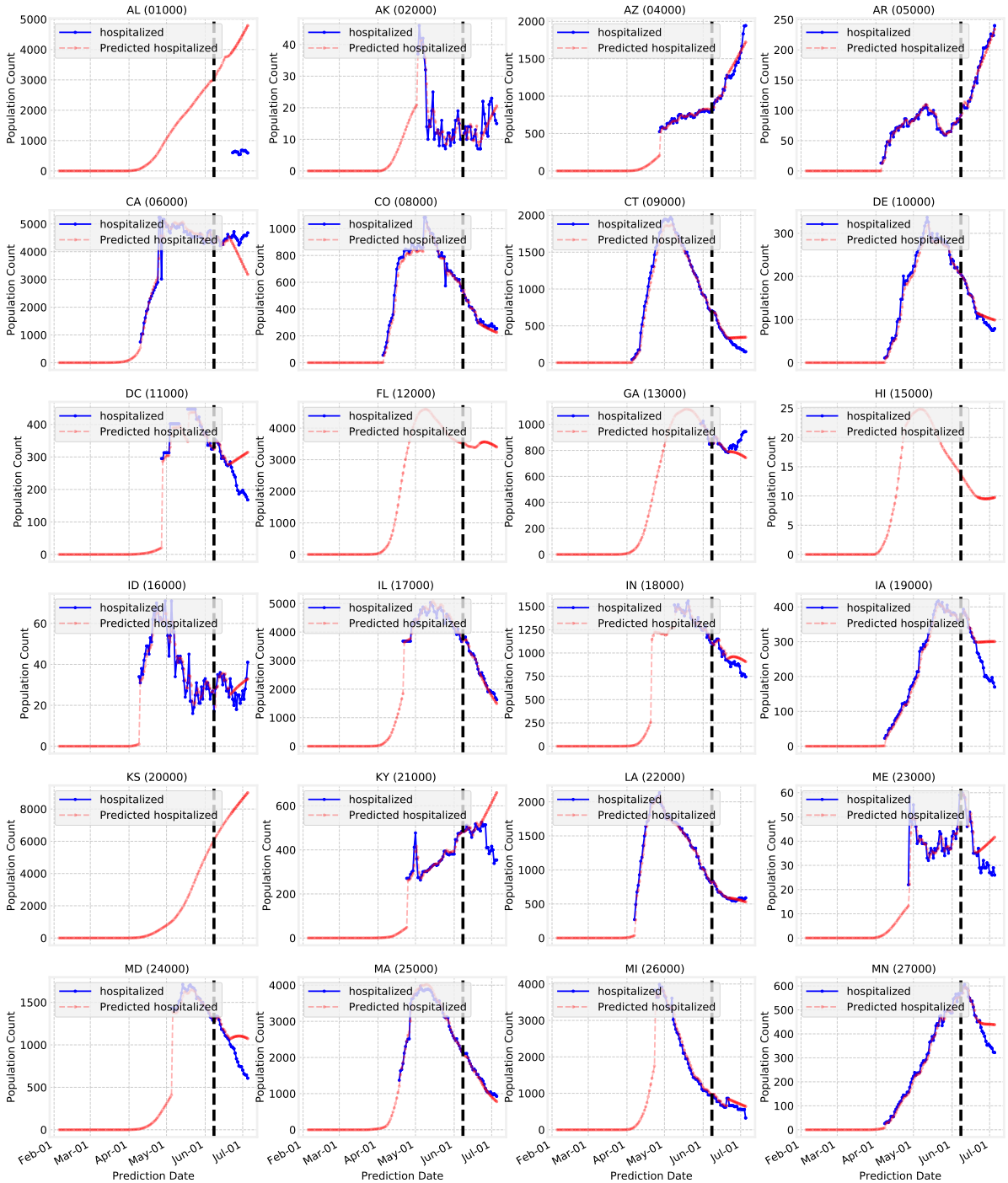


Figure 28. Model performance (hospitalized individuals) on US states–Alabama to Nebraska. Black dashed line is the training horizon. Some states do not have ground truth for hospitalized individuals, hence do not have the corresponding curves in these plots.

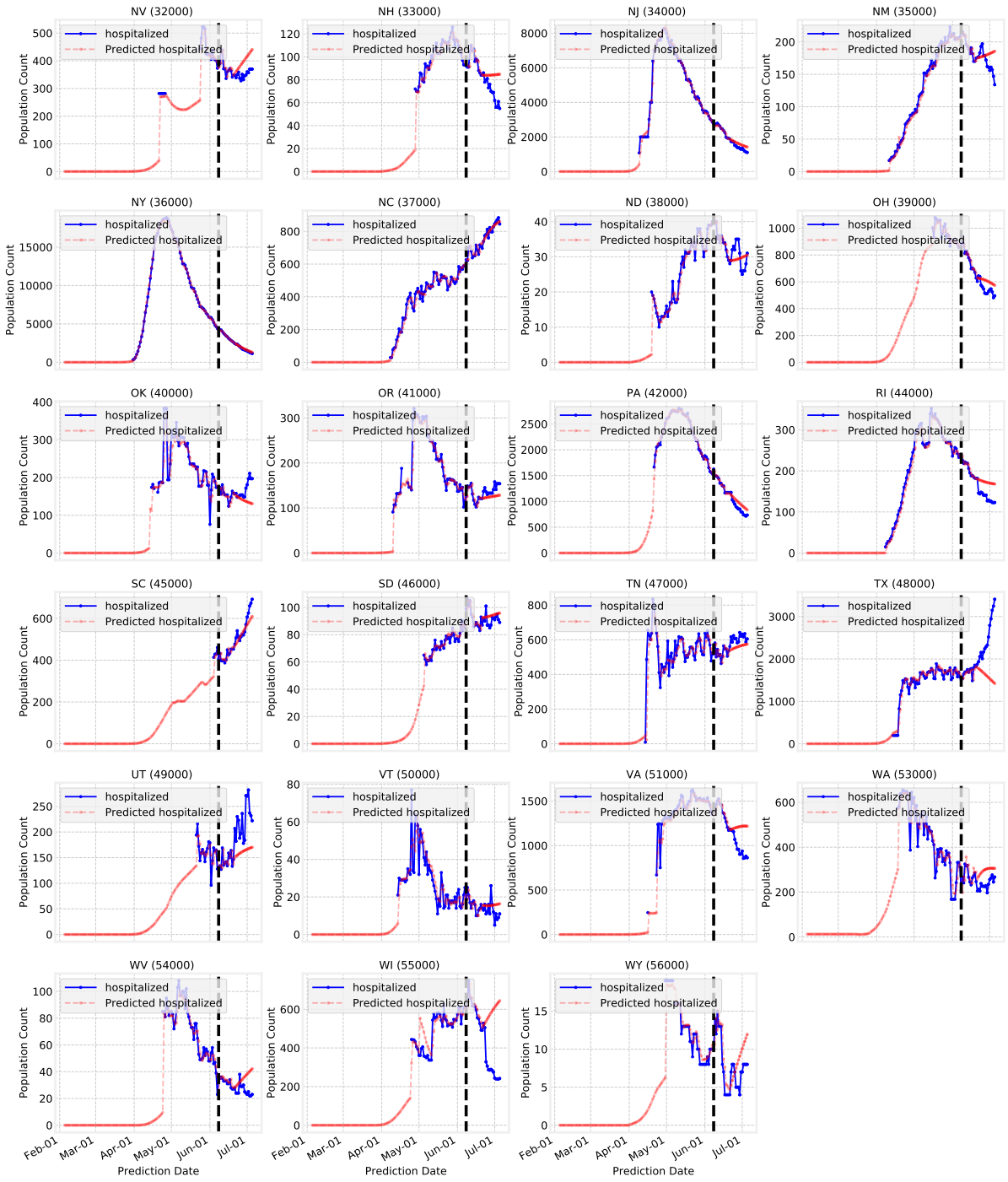


Figure 29. Model performance (hospitalized individuals) on US states–Nevada to Wyoming. Black dashed line is the training horizon.

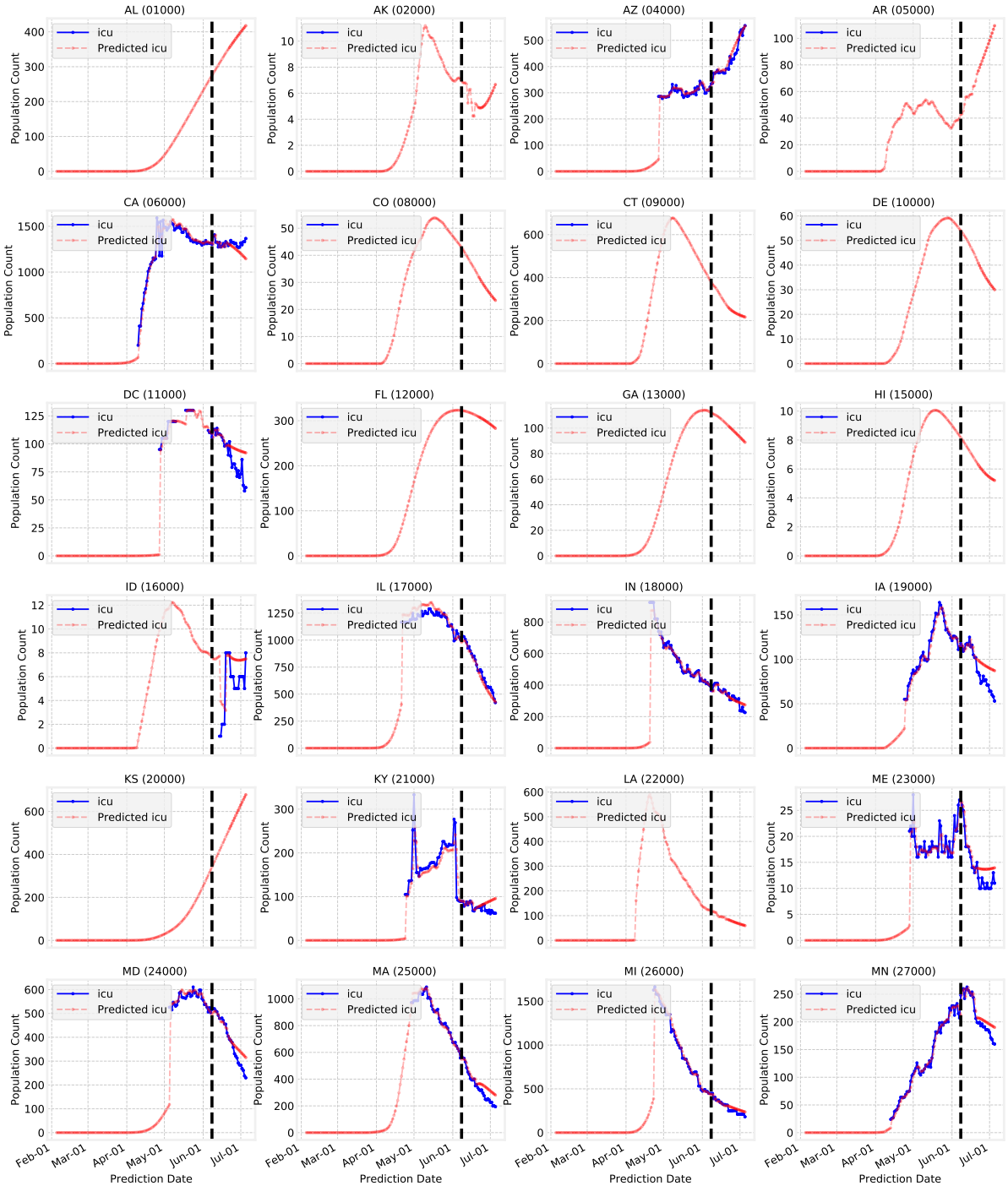


Figure 30. Model performance (Individuals in ICU) on US states–Alabama to Nebraska. Black dashed line is the training horizon. Some states do not have ground truth for hospitalized individuals, hence do not have the corresponding curves in these plots.

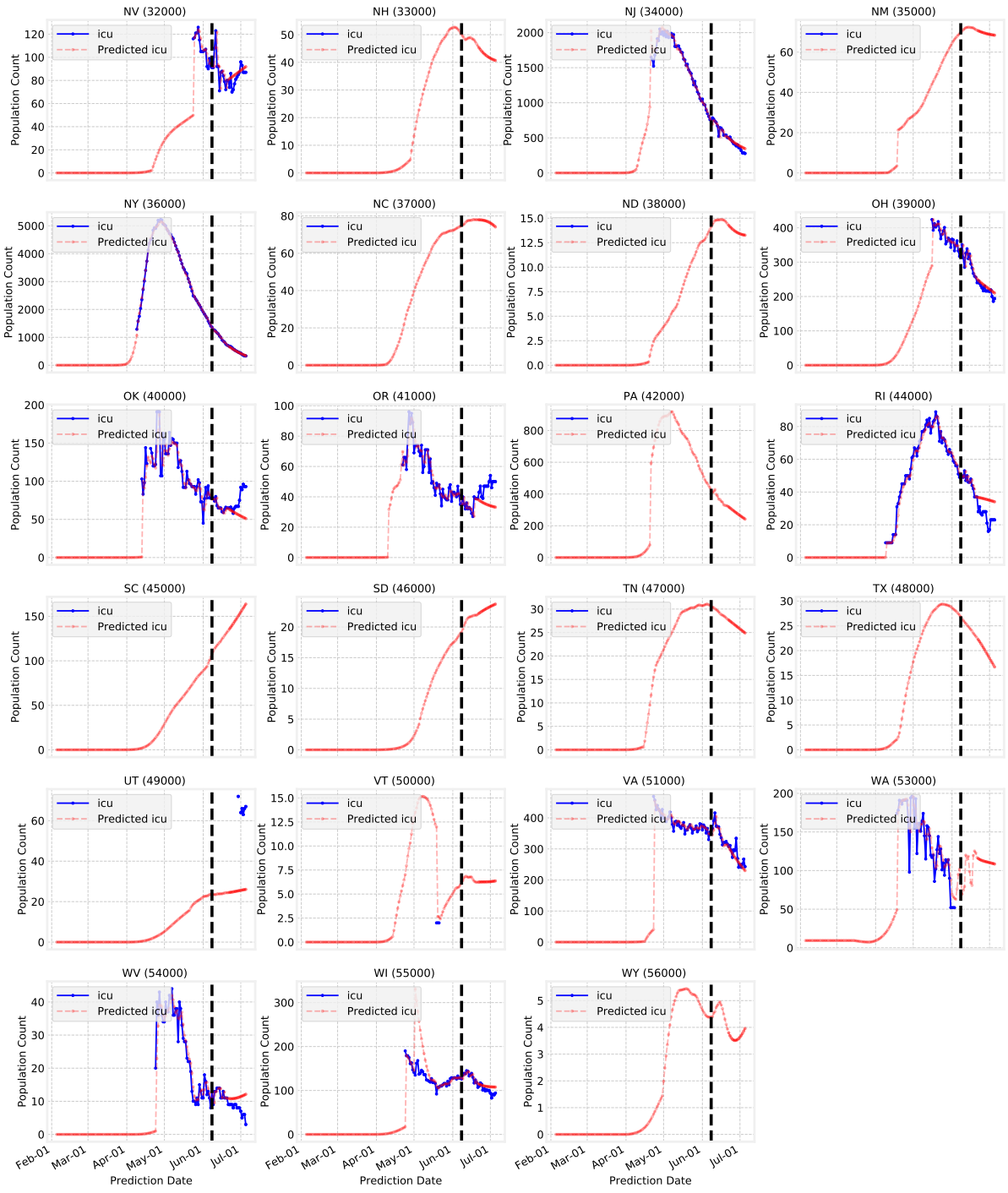


Figure 31. Model performance (Individuals in ICU) on US states–Nevada to Wyoming. Black dashed line is the training horizon. Some states do not have ground truth for hospitalized individuals, hence do not have the corresponding curves in these plots.

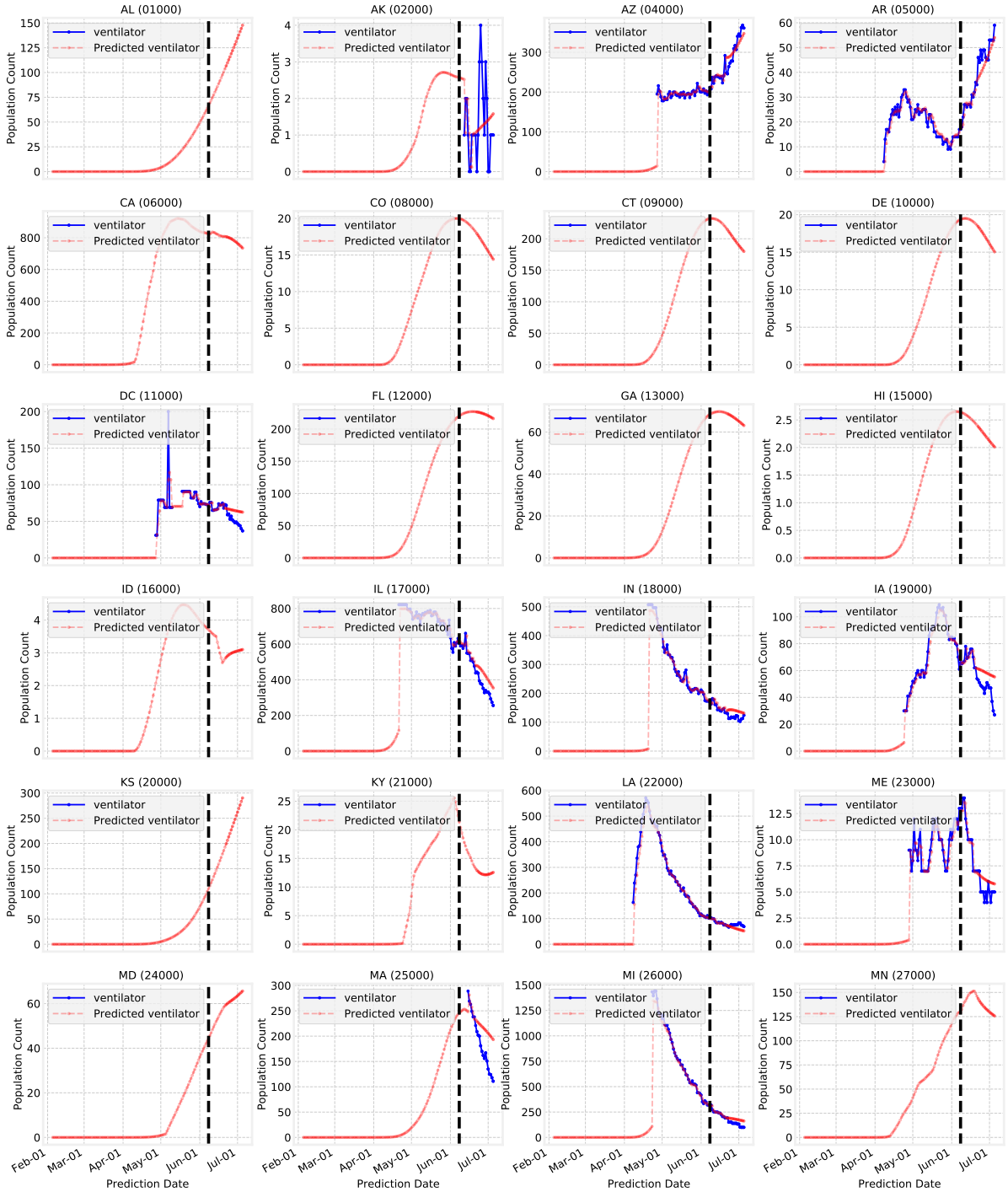


Figure 32. Model performance (Individuals on Ventilator) on US states–Alabama to Nebraska. Black dashed line is the training horizon. Some states do not have ground truth for hospitalized individuals, hence do not have the corresponding curves in these plots.

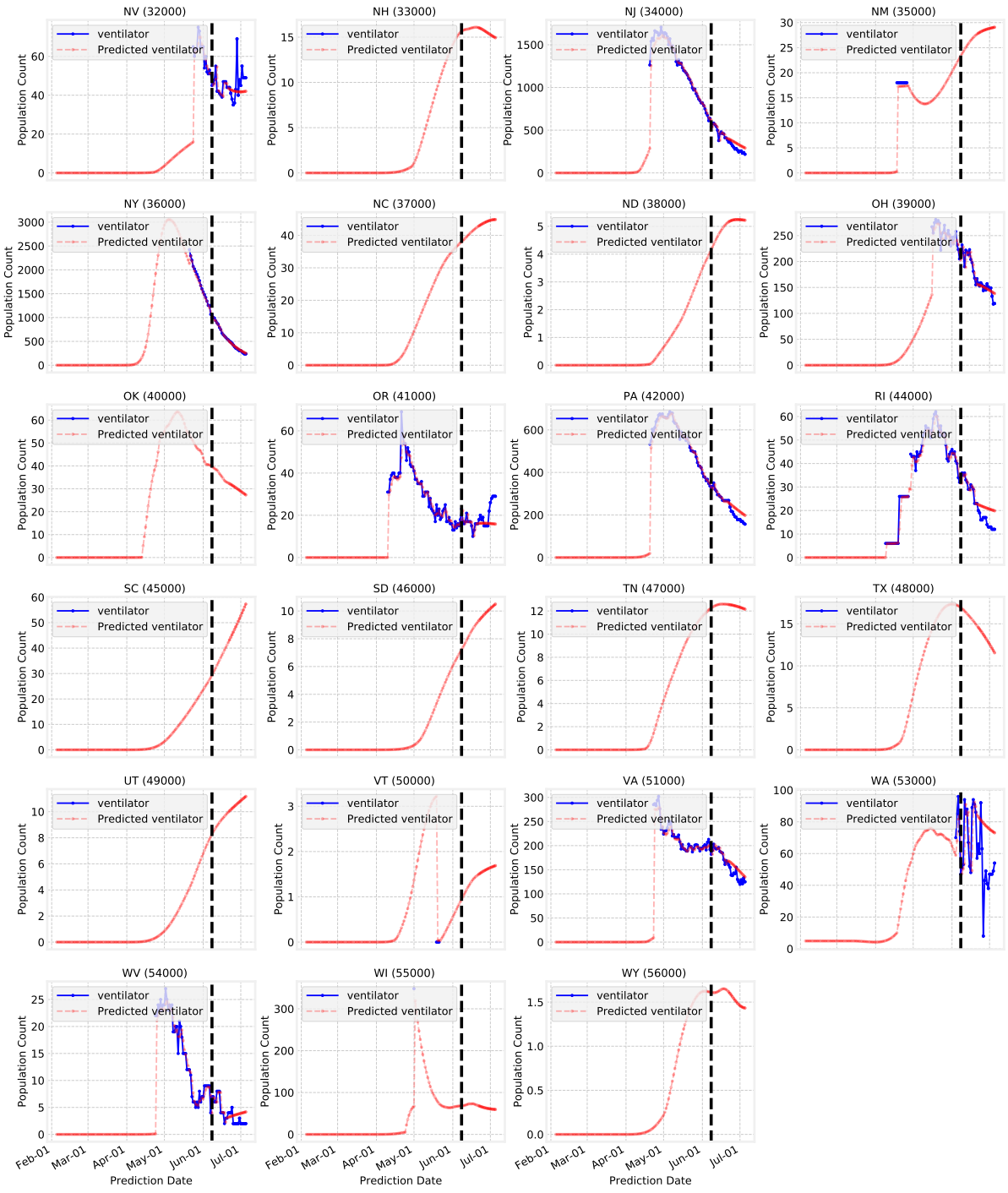


Figure 33. Model performance (Individuals on Ventilator) on US states–Nevada to Wyoming. Black dashed line is the training horizon. Some states do not have ground truth for hospitalized individuals, hence do not have the corresponding curves in these plots.

18 County-Level Forecasts for Top-20 Counties

We present the 7-day deaths forecasts for the top 20 US counties ordered by deaths from COVID-19 in Fig. 34, and confirmed cases forecasts in Fig. 35.

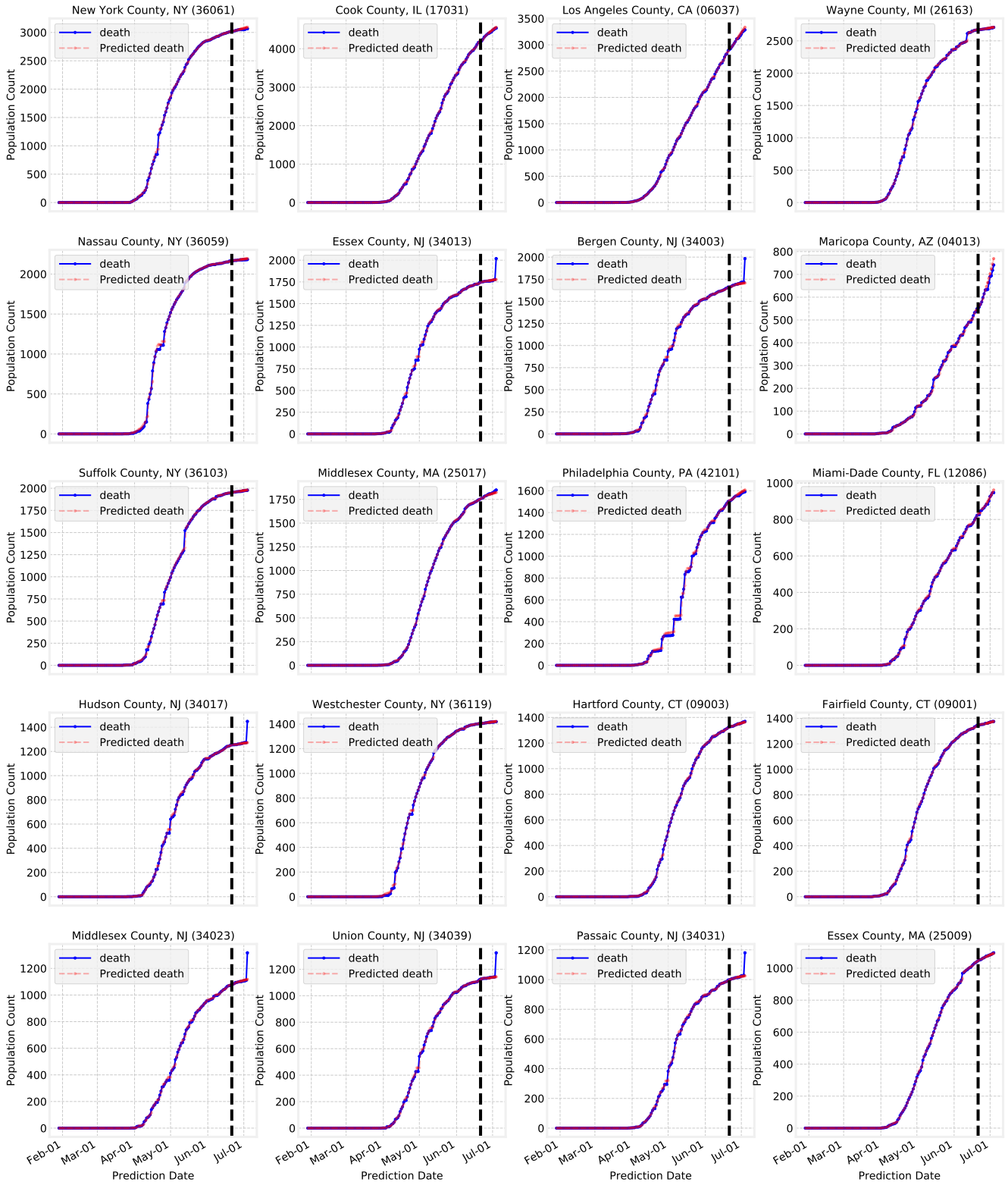


Figure 34. Model performance on deaths in Top-20 US counties ordered by deaths. Black dashed line is the training horizon.

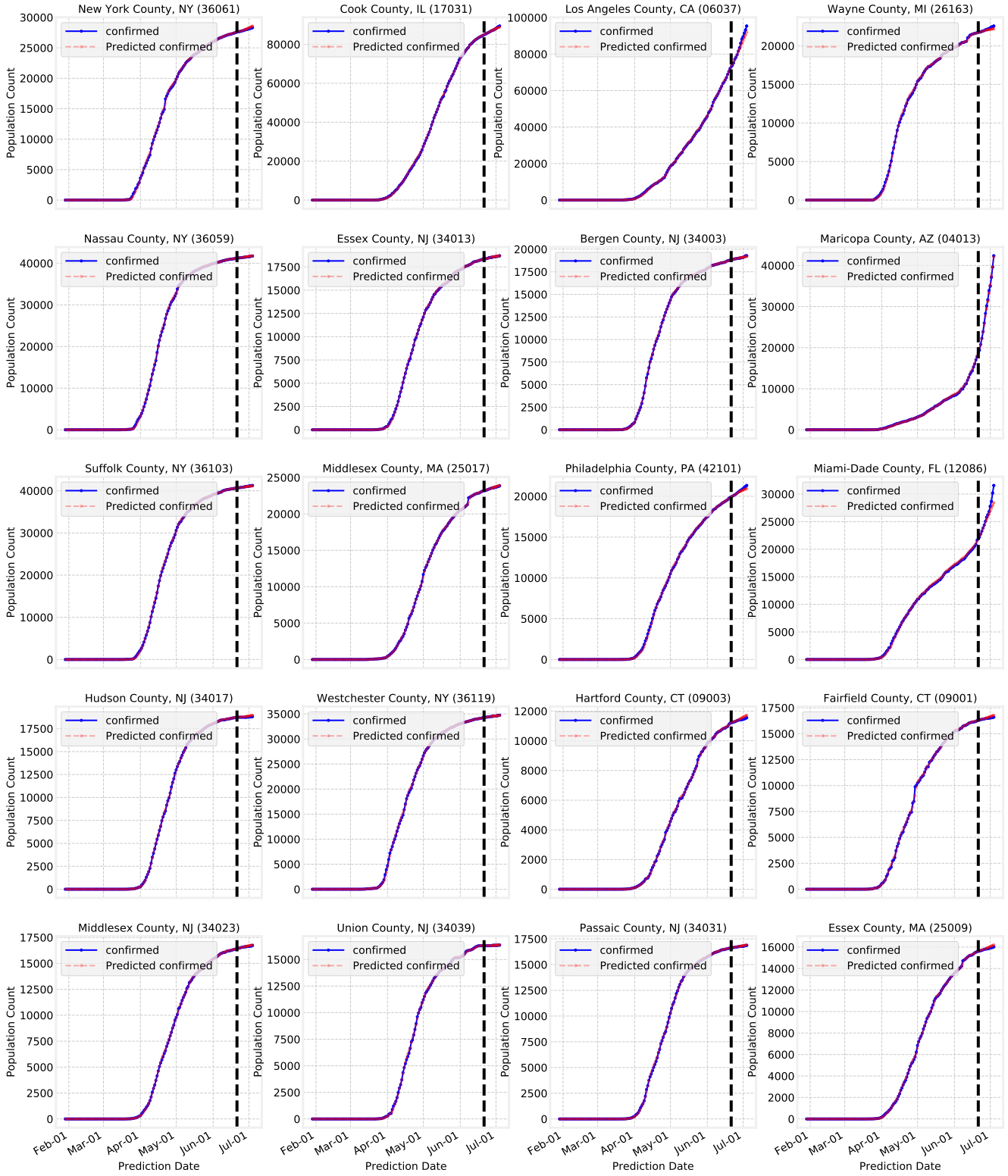


Figure 35. Model performance on confirmed cases in Top-20 US counties ordered by deaths. Black dashed line is the training horizon.

19 Error Distribution Over Counties

Fig. 36 shows the distribution of Mean Absolute Error of deaths over all counties. We see that the mean and median of this distribution are 1.16 and 0.43 respectively. A handful of outlier counties with MAE values ranging from 30 to 50 are present, with the worst performance being observed for Bergen County, NJ (FIPS code 34003).

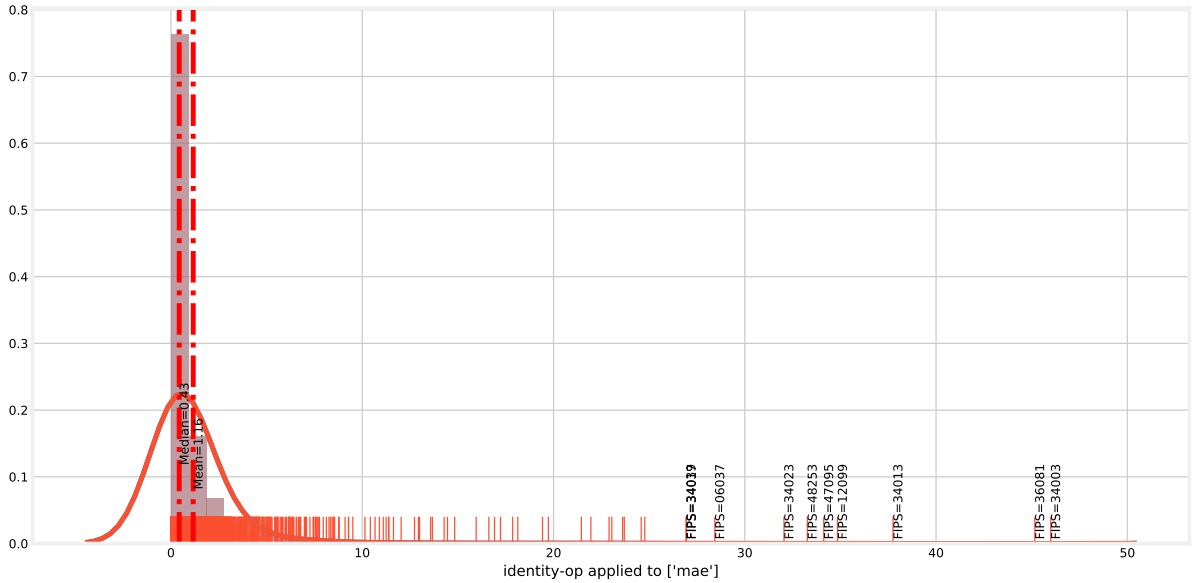


Figure 36. Distribution of MAE of deaths over US counties.

20 Ablation Studies

Table 12. Ablation study: Prediction RMSE for 7-day forecasting of the number of deaths at state-level.

Models	04/27/2020	05/04/2020	05/11/2020
Only SEIR compartments w/o encoders	3841.1	3854.5	3887.2
Our baseline w/o fine-tuning	119.7	121.6	68.2
Our baseline w/o teacher forcing	68.2	103.2	41.0
Our baseline w/o local bias b_i	111.4	141.7	66.3
Our baseline w/o local bias regularization	150.6	113.0	88.0
Our baseline	**53.4**	**48.6**	**25.3**

Table 12 presents the results of ablation cases. We observe the significant benefits of (i) modeling extra compartments and supervision from H, C and V, (ii) partial teacher forcing, (iii) final-fine tuning, adapting to the most recent data after model selection based on validation and (iv) our regularization approaches.

21 Other Things We Have Tried

With the goal of providing guidance to future research, we overview other main ideas we have tried and that we could not get improvements with (while thinking that their modified versions can still be promising):

- **Dropout regularization**: We have considered applying dropout to the inputs or the outputs of the encoders, to improve generalization. We have not observed consistent improvements.
- **Nonlinear processing without feature interactions**: Applying learnable nonlinearity to the covariates can improve the model capacity, while still preserving the explainability somewhat as the feature interactions would be limited to additive aggregation. We have seen that this idea suffers from overfitting.

- **Asymmetric loss functions:** We have tried applying different weights on under- vs. over-prediction, but have not observed a significant benefit that would worth an extra hyperparameter.
- **Normalized loss functions:** We have explored loss terms that are normalized, such as with the ground truth or population, but they seem to affect the training dynamics negatively.
- **Recurrent neural networks for covariate encoding:** We have explored high capacity sequence-to-sequence architectures, particularly variants of LSTMs, which are not explainable, to see whether there is a significant performance gain to obtain if explainability is not considered. But such sequence-to-sequence architectures suffer from optimization difficulties and yield poorer generalization.
- **Second-order optimization:** Since the number of trainable degrees of freedom is small, higher-order optimization is feasible. We have implemented LBFGS for training, but have observed that it yields poorer generalization compared to RMSprop.