INTERFACE

royalsocietypublishing.org/journal/rsif

Research

Cite this article: Qian Y, Zhang K, Marty E, Basu A, O'Dea EB, Wang X, Fox SJ, Rohani P, Drake JM, Li H. 2025 Physics-informed deep learning for infectious disease forecasting. *J. R. Soc. Interface* **22**: 20250379.
https://doi.org/10.1098/rsif.2025.0379

Received: 28 April 2025
Accepted: 25 September 2025

Subject Category:
Life Sciences–Engineering interface

Subject Areas:
biomedical engineering, computational biology

Keywords:
infectious disease forecasting, physics-informed neural networks (PINNs), epidemiological modelling, machine learning

Author for correspondence:
He Li
e-mail: he.li3@uga.edu

Physics-informed deep learning for infectious disease forecasting

Ying Qian[1], Kui Zhang[1], Eric Marty[2], Avranil Basu[2], Eamon B. O'Dea[2], Xianqiao Wang[3], Spencer J. Fox[4], Pejman Rohani[2,5,6], John M. Drake[2] and He Li[1]

[1]School of Chemical, Materials, and Biomedical Engineering, [2]Odum School of Ecology and Center for the Ecology of Infectious Diseases, [3]School of Environmental, Civil Agricultural and Mechanical Engineering, College of Engineering, [4]Department of Epidemiology and Biostatistics, [5]Department of Infectious Diseases, College of Veterinary Medicine, and [6]Center for Influenza Disease and Emergence Research, University of Georgia, Athens, GA, USA

YQ, 0009-0004-9596-8432; KZ, 0009-0007-8243-9234; EBO, 0000-0003-4748-683X; XW, 0000-0003-2461-3015; SJF, 0000-0003-1969-3778; PR, 0000-0002-7221-3801; JMD, 0000-0003-4646-1235

Accurate forecasting of contagious illnesses has become increasingly important to public health policymaking and better prediction could prevent the loss of millions of lives. To better prepare for future pandemics, it is essential to improve forecasting methods and capabilities. In this work, we implement physics-informed neural networks (PINNs), a popular tool in the area of scientific machine learning, to perform infectious disease forecasting. The used PINNs model incorporates dynamical systems representations of disease transmission into the loss function, thereby assimilating epidemiological theory and data using neural networks. Our approach is designed to prevent model overfitting, which often occurs when training deep-learning models with observation data alone. In addition, we use an additional sub-network to account for mobility, cumulative vaccine doses and other covariates that influence the transmission rate, a key parameter in the compartmental model. To demonstrate the capability of the proposed model, we examine the performance of the model using state-level COVID-19 data in California. Our simulation results show that predictions of the PINNs model on the number of cases, deaths and hospitalizations are consistent with existing benchmarks. In particular, the PINNs model outperforms naive baseline forecasts and various sequence deep-learning models, such as recurrent neural networks, long short-term memory networks, gated recurrent units and transformer models. We also show that the performance of the PINNs model is comparable with that of a sophisticated Gaussian infection state forecasting model that combines the compartmental model, a data observation model and a regression model for inferring parameters in the compartmental model. Nonetheless, the PINNs model offers a simpler structure and is easier to implement. In summary, we perform a systematic study of the predictive capability of the PINNs model in forecasting the dynamics of infectious diseases and our results showcase the potential of the proposed model as an efficient computational tool to enhance the current capacity of infectious disease forecasting.

1. Introduction

COVID-19 was the third largest cause of mortality in the United States in 2022 [1] and, in addition to Ebola, Chikungunya, Zika and mPox, was one of five global outbreaks in the past decade caused by emerging pathogens. Climate change is expected to exacerbate disease transmission risks, which before

THE ROYAL SOCIETY PUBLISHING

COVID-19 already accounted for roughly 25% of global mortality [2,3]. In the midst of an exponentially growing outbreak, public health officials must respond rapidly in the face of intense uncertainty [4]. Recent developments in infectious disease forecasting have sought to improve our ability to anticipate future epidemiological trends, such as the number of reported cases, deaths or hospitalizations from a disease and are increasingly used during outbreaks. Forecasting provides a concrete way to integrate data with epidemiological knowledge and can be used to inform the allocation of critical resources, such as antiviral drugs or ventilators, the implementation of non-pharmaceutical interventions and the design of vaccine trials [4–6].

The compartmental epidemiological models, which divide the population into distinct mutually exclusive subsets based on individuals' disease status, have been widely used to represent and forecast the spread of infectious diseases within a population [7–10]. The most basic compartments include individuals who are susceptible to the infection (Susceptible (S)), those who are currently infected and capable of transmitting the disease (Infected (I)) and individuals who have recovered from the infection and are assumed to have acquired immunity or those who have been removed from the susceptible pool due to death (Recovered or Removed (R)) [11,12]. Numerous extensions of the basic compartmental models have introduced more compartments to account for other state variables, such as death or variations in disease staging [13–19] or other flows among compartments (e.g. loss of immunity [8,20]). The transitions between these compartments are governed by a set of ordinary differential equations that describe the rates at which individuals move from one compartment to another. Although compartmental epidemiological models have provided valuable insights into the general dynamics of infectious diseases and inform public health strategies, compartmental models typically assume parameters such as the transmission and recovery rates are fixed. As has been observed in recent high-profile outbreaks, this assumption does not adequately capture the dynamic nature of an unfolding epidemic, especially when factors such as the rollout of new vaccines, public health interventions or virus mutation are involved (see [21–24]).

Over the past decade, alternative approaches to tackling complex social problems have emerged, including artificial intelligence (AI) and machine learning (ML). These approaches have the advantage of efficiently identifying patterns in data and can, in principle, accommodate multiple disparate data streams. However, conventional AI/ML methods require a large amount of training and labelled data to ensure good performance [25–28]. Furthermore, AI/ML models may provide unrealistic predictions, especially for non-stationary systems, as they are not constrained by the mechanisms driving transmission. Furthermore, data-driven models have historically struggled with predicting epidemics because they often fail to differentiate real trends from noise in the data collection process [29]. Recently, scientific ML models for solving partial differential equations (PDEs) and ordinary differential equations (ODEs) have attracted growing attention due to their ability to assimilate the underlying scientific laws represented by ODEs or PDEs with the learning properties of neural networks (NNs). Combining AI with dynamical systems models can help to infer unknown model parameters as constants or functions using limited data. This enables learning from 'small data' as we explicitly utilize the constraints from the physical or biological laws [30–43].

Physics-informed neural networks (PINNs) [44], one of the most popular scientific ML models, provide a powerful framework that integrates data with compartmental models by incorporating both into the loss function. As illustrated in figure 1A, the left side of the architecture represents a fully connected neural network (FCNN), corresponding to the physics-uninformed component, while the right side captures the physics-informed component. The loss function is designed to account for contributions from both observed data and the underlying compartmental models. PINNs have been successfully applied to a range of scientific and engineering challenges, encompassing both forward and inverse problems [45–54]. Previous studies have explored parameter estimation and forecasting using PINNs [55–59]. Specifically, Kharazmi *et al.* [55] demonstrated the effectiveness of PINNs in simultaneously inferring the unknown parameters and the unobserved dynamics using a plurality of epidemiological models. Subsequent studies [60–64] have built upon this framework, using PINNs or their modified variants to estimate time-varying parameters and capture epidemic dynamics using data from multiple spatial locations. Despite these advances, there remains a lack of systematic comparisons between the performance of PINNs and popular data-driven sequence models such as recurrent neural networks (RNNs), long short-term memory (LSTM) networks, gated recurrent units (GRUs) and transformer models. A comprehensive evaluation of these approaches would help elucidate the advantages and limitations of PINNs relative to more conventional ML methods in time-series forecasting.

Here, we introduce a disease forecasting model developed based on PINNs. The compartmental model includes nine state variables (table 1), with the PINNs architecture comprising two sub-networks that both take time t as input. The upper sub-network predicts all nine state variables, contributing to the ODE loss $\mathcal{L}_{\text{ODE}}$, while three observable variables—reported cases (C_r), hospitalizations (A) and reported deaths (D_r)—also contribute to the data loss $\mathcal{L}_{\text{data}}$. The lower sub-network predicts three variables—mobility, cumulative vaccine doses and transmission rate. Since these three outputs share the same network weights, the training of transmission rate is indirectly influenced by the training of mobility and cumulative vaccine doses. The transmission rate feeds into $\mathcal{L}_{\text{ODE}}$ and the other two inform $\mathcal{L}_{\text{data}}$. This design allows for seamless integration of mechanistic modelling with observed data, enhancing forecasting performance. To demonstrate the capability of this approach, we evaluated the performance of the model by comparing one-, two-, three- and four-week-ahead forecasts of COVID-19 cases, deaths and hospitalizations in the state of California using reports of COVID-19, Google's mobility reports and vaccination data available each week. We assessed model accuracy on both point forecasting and quantile forecasting [65]. We compared the score of our model with a naive baseline forecast, various sequence deep-learning models, including RNNs, LSTMs, GRUs, transformer models and a traditional mathematical model.

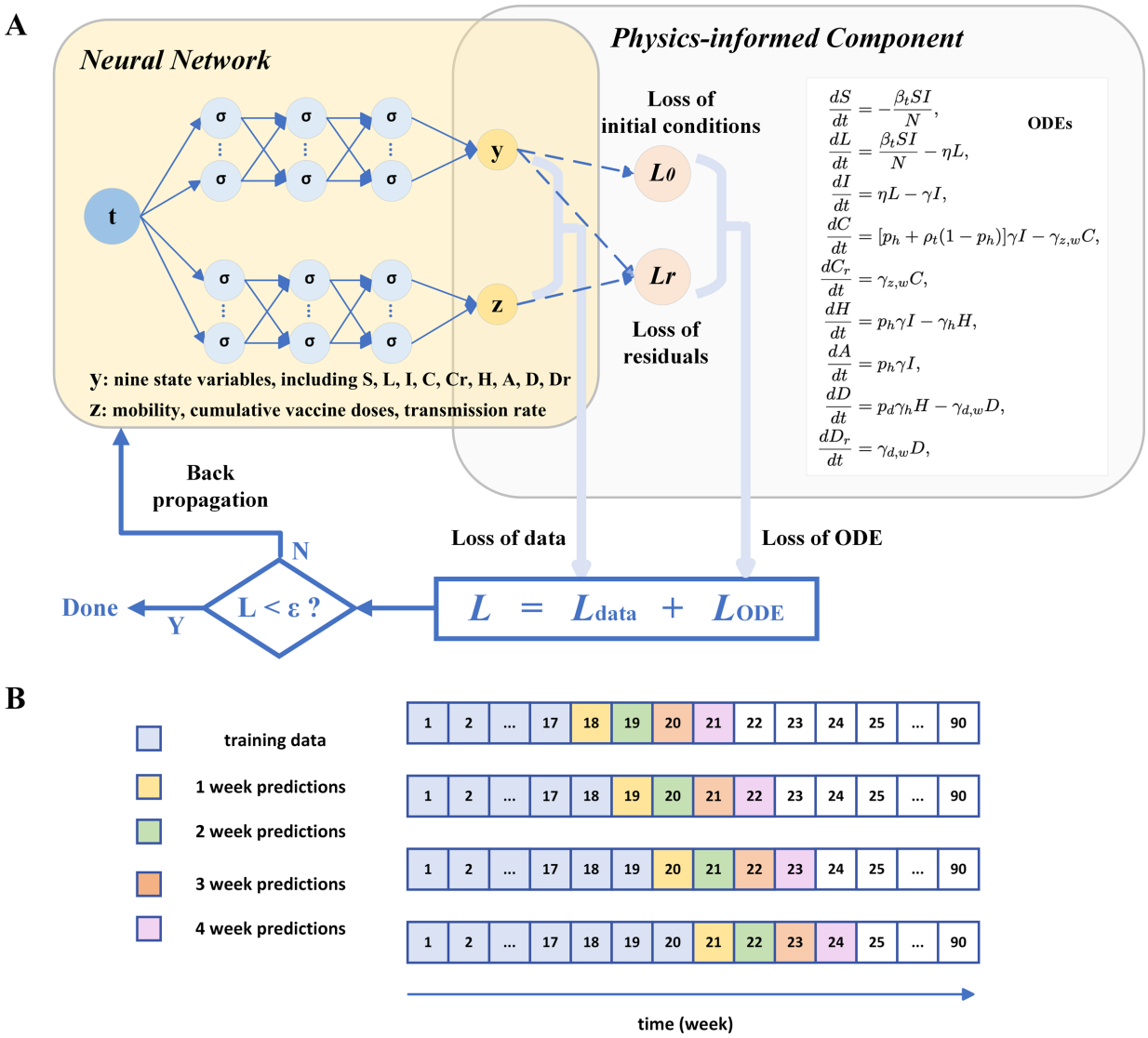

Figure 1. (A) Schematic of the proposed PINNs model for infectious disease forecasting. The model comprises two sub-networks: the upper sub-network predicts the state variables in the compartmental model, while the lower sub-network estimates the time-dependent model parameters. The input of the NN is time t in the form of a vector, and it contains all the residual points where the ODE loss is computed. The output **y** represents the nine compartmental state variables, including S (uninfected and susceptible individuals), L (individuals with latent infections who are not yet infectious), I (individuals who are infectious), C (individuals who have been diagnosed and will be reported as cases but have not yet been reported), C_r (this compartment keeps track of the number of cases reported each day), H (individuals who have been hospitalized), A (individuals who are new hospital admissions), D (individuals who have died from the infection but whose death has not yet been reported), D_r (the number of newly reported deaths each week). The output **z** represents factors including mobility, cumulative vaccine doses and transmission rate. The data loss L_{data} is a weighted sum of observable state variables and factors. (B) Schematic of the rolling window approach. As training data accumulate over time, the PINNs model continuously updates and generates forecasts for the subsequent one to four weeks.

Table 1. State variables in the used compartmental model by O'Dea & Drake [22].

name	definition
S	uninfected and susceptible individuals
L	individuals with latent infections who are not yet infectious
I	individuals who are infectious
C	individuals who have been diagnosed and will be reported as cases but have not yet been reported
C_r	this compartment keeps track of the number of cases reported each day
H	individuals who have been hospitalized
A	individuals who are new hospital admissions
D	individuals who have died from the infection but whose death has not yet been reported
D_r	the number of newly reported deaths each week

2. Data and methods

2.1. Data

2.1.1. Input data

To facilitate a comparison of our model's forecasting performance with that of various other models, we used the forecasting targets defined by the COVID-19 Forecast Hub [22,66], including COVID-19 case and death data sourced from the COVID-19 Data Repository available at https://github.com/CSSEGISandData/COVID-19, which is maintained by the Center for Systems Science and Engineering (CSSE) at Johns Hopkins University, Baltimore, MD, USA (JHU) [67]. Furthermore, we used information on hospitalizations from the COVID-19 Reported Patient Impact and Hospital Capacity by State Time series and COVID-19 Reported Patient Impact and Hospital Capacity by State datasets compiled by the US Department of Health and Human Services (HHS) available on healthdata.gov. Access to these datasets was facilitated through the Delphi Epidata API, as detailed by Reinhart *et al*. [68]. For our analysis, the time series of hospitalizations is the sum of the fields labelled previous_day_admission_adult_covid_confirmed and previous_day_admission_pediatric_covid_confirmed in the tables provided by HHS.

We used data on mobility and cumulative vaccine doses due to their association with the transmission rate, a critical time-dependent parameter in epidemiological forecasting. In particular, mobility can be measured by the amount of time individuals spend in residential areas, as quantified in Google's community mobility reports [69]. For a given forecast date, we obtained the latest snapshot of Google's *Global_Mobility_Report.csv* file on http://web.archive.org that was made before the beginning forecast date in the UTC time zone.

Cumulative vaccine doses represent the total number of vaccine doses administered within a given state. Our modelling assumes that the average susceptibility of individuals who have not been previously infected decreases as the number of administered doses increases. Rather than modelling the number of susceptible individuals directly, we model the effect of vaccination on average susceptibility, acknowledging that a single dose of a COVID-19 vaccine typically does not confer full immunity. We obtained the time series data on administered vaccine doses from the GitHub repository maintained by Johns Hopkins University (https://github.com/govex/COVID). This dataset compiles information from state public dashboards and the Centers for Disease Control and Prevention (CDC) Vaccine Tracker. For each day, the dataset reports the higher value between the two sources to provide the most current estimate. To generate a covariate without missing values, we assumed all values prior to the first reported entry were zero and any missing values within the series were imputed using linear interpolation.

2.1.2. Data pre-processing

The time-series data for cases, deaths, hospitalizations, mobility and cumulative vaccine doses were originally provided at a daily resolution, whereas our prediction targets are weekly totals for cases, deaths and hospitalizations. To ensure consistency with the prediction targets, we pre-processed these datasets by computing a 7-day moving average and aggregating the daily values into a weekly time series. This smoothing strategy not only aligns the temporal resolution with the forecasting targets but also reduces short-term fluctuations in the original data.

The hospitalization data differ from the other time series in several key aspects. First, the hospitalization records were made available later than the case and death data, with reliable reporting beginning in July 2020. Second, the initial portion of the hospitalization time series is affected by an artificial upward trend caused by the gradual increase in the number of hospitals reporting data. To mitigate this bias, we excluded the first 20 weeks of hospitalization data from our analysis.

Towards the end of the time series, a small, incomplete peak appears following two major peaks. To maintain focus on well-characterized dynamics, we chose to model only the two major peaks and discarded data after week 110. These two peaks correspond to different viral variants (Beta and Delta), each associated with distinct transmission dynamics and sufficient for evaluating the performance of our forecasting model.

Figure 2 illustrates both the original (upper panel) and pre-processed (lower panel) time-series data. To address the large differences in scale across variables, we applied linear scaling to normalize each time series to a similar range, approximately between 0 and 1. This normalization is also reflected in the lower panel of figure 2, facilitating better visualization and more stable model training.

2.2. Compartmental model

The compartmental model for COVID-19, adopted from [22], is based on a structured framework that captures the dynamics of COVID-19 transmission, hospitalization and death, incorporating both reported and unreported cases. It accounts for time-varying transmission and progression rates through semi-parametric smoothing splines, allowing for flexible adaptation to real-world data and policy changes. The model introduces nonlinearities primarily through the infection term, which is a product of the susceptible and infectious populations scaled by a time-dependent contact rate. The contact structure implicitly incorporates behavioural and intervention-related effects without explicitly modelling population mixing patterns. A key mathematical property established is the positivity of the solution, ensuring all state variables remain non-negative over time, which is critical for epidemiological interpretability. This well-posedness, along with the model's flexible design, supports robust short- and medium-term forecasts of cases, hospitalizations and deaths.

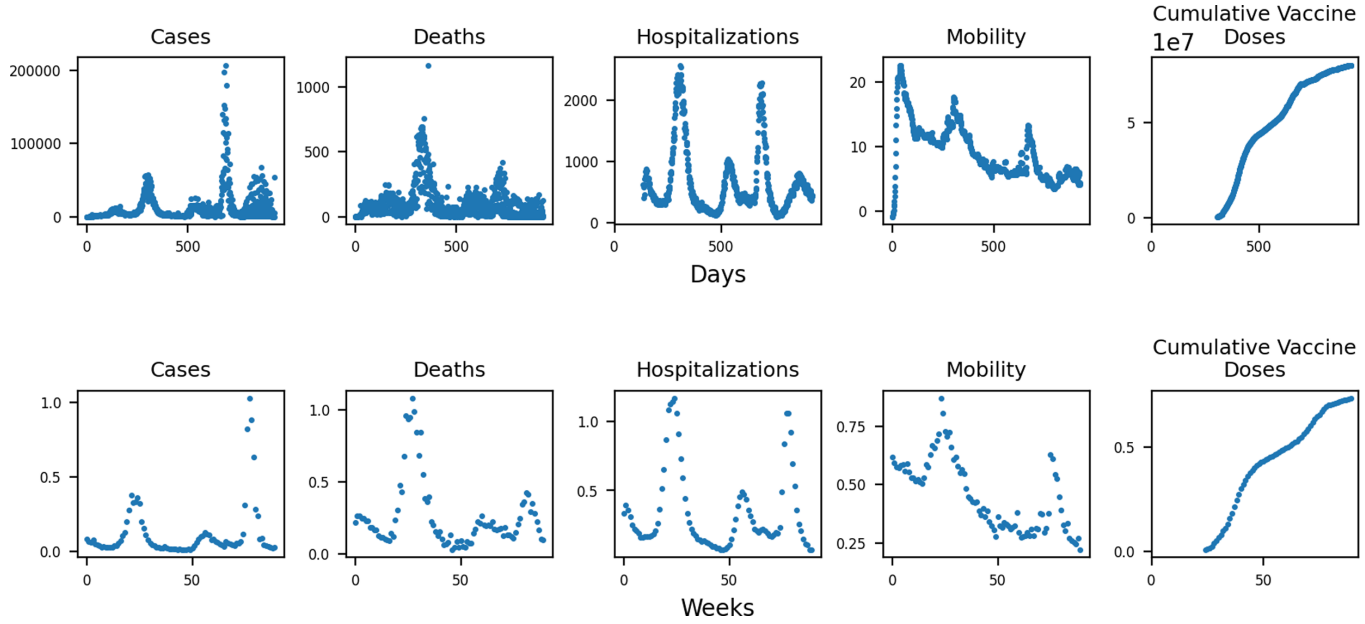

Figure 2. Original dataset (upper panel) and pre-processed (lower panel) dataset used for training and testing the proposed PINNs model.

Table 2. Model parameters used in the compartmental model by O'Dea & Drake [22].

name	definition	value
β_t	transmission rate	estimated
N	population size	39 512 223 persons
η	incubation rate	365.25/4/year
γ	removal rate	365.25/4/year
$\gamma_{d,w}$	rate of reporting deaths on day of week w	365.25/10/year
$\gamma_{z,w}$	rate of reporting cases on day of week w	365.25/1/year
γ_h	rate of exit from hospital	365.25/10/year
p_h	probability of an infection leading to hospitalization	estimated
p_d	probability of a hospitalization leading to death	estimated
ρ_t	probability of a removal on day t being reported as a case	0.5

The compartmental model is governed by a system of nine ODEs involving nine state variables and 11 parameters:

$$\frac{\mathrm{d}S}{\mathrm{d}t} = -\frac{\beta_t SI}{N}, \tag{2.1a}$$

$$\frac{\mathrm{d}L}{\mathrm{d}t} = \frac{\beta_t SI}{N} - \eta L, \tag{2.1b}$$

$$\frac{\mathrm{d}I}{\mathrm{d}t} = \eta L - \gamma I, \tag{2.1c}$$

$$\frac{\mathrm{d}C}{\mathrm{d}t} = [p_h + \rho_t(1-p_h)]\gamma I - \gamma_{z,w} C, \tag{2.1d}$$

$$\frac{\mathrm{d}C_r}{\mathrm{d}t} = \gamma_{z,w} C, \tag{2.1e}$$

$$\frac{\mathrm{d}H}{\mathrm{d}t} = p_h \gamma I - \gamma_h H, \tag{2.1f}$$

$$\frac{\mathrm{d}A}{\mathrm{d}t} = p_h \gamma I, \tag{2.1g}$$

$$\frac{\mathrm{d}D}{\mathrm{d}t} = p_d \gamma_h H - \gamma_{d,w} D, \tag{2.1h}$$

$$\frac{\mathrm{d}D_r}{\mathrm{d}t} = \gamma_{d,w} D. \tag{2.1i}$$

The definitions of the state variables are provided in table 1. Definitions and values for each model parameter are provided in table 2.

2.3. Physics-informed neural networks

PINNs, which combine traditional compartmental models with observational data through NNs, have emerged as one of the most influential approaches in scientific ML. In this work, we incorporate the epidemiological model proposed in [22]—a compartmental model for COVID-19 with time-dependent parameters—into our PINNs framework. This model is governed by a system of nine ODEs involving nine state variables and 11 parameters. Among these, three state variables— C_r, D_r and A—are observable and used as targets for model training and evaluation. The remaining six state variables are unobservable and are treated as latent variables within the PINNs framework.

The system of ODEs is incorporated into the physics-informed component of the model, as illustrated in figure 1A. For consistency and comparability, we generally adopt the parameter settings from [22]: seven parameters are treated as known and constant, while three are estimated during training. Among them, the transmission rate β_t is considered a critical time-dependent parameter and is predicted by the lower sub-network, as depicted in figure 1A. The remaining two parameters, p_h and p_d, are assumed to be constant and are integrated into the model parameters to be estimated during the training process.

As depicted in figure 1A, the PINNs architecture comprises two sub-networks, both taking time t as input. Each sub-network consists of three hidden layers with 50 neurons in each layer. The upper sub-network predicts the values of the nine state variables, all of which are used to compute the ODE loss, denoted as $\mathcal{L}_{\text{ODE}}$. The exponential transformation is applied to the output of the upper and lower sub-networks to enforce positivity. Among these, three variables—C_r, A and D_r—are observable and contribute to the $\mathcal{L}_{\text{data}}$.

The lower sub-network is designed to implicitly capture the relationship between the transmission rate and two closely associated factors: mobility and cumulative vaccine doses. It is trained to predict multiple outputs from the same input t. Each output at inference depends solely on t and does not directly incorporate the others. However, since all outputs share the same network weights, the transmission rate is indirectly affected by the training of mobility and cumulative vaccine doses. This sub-network outputs predictions for mobility, cumulative vaccine doses and the transmission rate. The transmission rate is a key time-dependent parameter in the compartmental model and is involved in the computation of $\mathcal{L}_{\text{ODE}}$, while mobility and cumulative vaccine doses have corresponding reported data and contribute to $\mathcal{L}_{\text{data}}$. This PINNs framework enables the integration of mechanistic modelling with real-world data for improved forecasting accuracy.

The constraint imposed by the compartmental model was integrated into the PINNs training process by combining the ODE loss with the data loss. Specifically, the total loss is defined as a weighted sum of the data loss and the ODE loss:

$$\mathcal{L}(\theta) = \mathcal{L}_{\text{data}}(\theta) + w_{\text{ODE}}\mathcal{L}_{\text{ODE}}(\theta), \tag{2.2}$$

where θ denotes the full set of parameters in PINNs, including the networks' parameters and the physics parameters, i.e. p_h and p_d. $\mathcal{L}_{\text{data}}$ represents data loss, $\mathcal{L}_{\text{ODE}}$ represents ODE loss and w_{ODE} represents the weight for ODE loss.

We denote the upper and lower sub-networks (illustrated in figure 1A) by $\boldsymbol{y}_{\theta_y}(t)$ and $\boldsymbol{z}_{\theta_z}(t)$, respectively, where θ_y and θ_z are the parameters of the upper and lower networks. The variable t denotes time as the input. The output $\boldsymbol{y}_{\theta_y}(t) = [y^1_{\theta_y}(t), y^2_{\theta_y}(t), ..., y^9_{\theta_y}(t)]$ approximates the vector of the nine state variables ordered as in table 1, evaluated at time t, while $\boldsymbol{z}_{\theta_z}(t) = [z^1_{\theta_z}(t), z^2_{\theta_z}(t), z^3_{\theta_z}(t)]$ approximates the vector comprising mobility, cumulative vaccine doses and the transmission rate β_t at t. The PINNs were trained by minimizing the loss function $\mathcal{L}(\theta)$ using the mean squared error as the training metric, with the data loss component defined as

$$\begin{aligned}\mathcal{L}_{\text{data}}(\theta) = &\frac{1}{N_{C_r}}\sum_{i=1}^{N_{C_r}} |y^5_{\theta_y}(t_i^{C_r}) - C_{r,i}|^2 + \frac{1}{N_A}\sum_{i=1}^{N_A} |y^6_{\theta_y}(t_i^A) - A_i|^2 + \frac{1}{N_{D_r}}\sum_{i=1}^{N_{D_r}} |y^9_{\theta_y}(t_i^{D_r}) - D_{r,i}|^2 \\ &+ \frac{1}{N_M}\sum_{i=1}^{N_M} |z^1_{\theta_z}(t_i^M) - M_i|^2 + \frac{1}{N_V}\sum_{i=1}^{N_V} |z^2_{\theta_z}(t_i^V) - V_i|^2,\end{aligned} \tag{2.3}$$

where $\{t_i^{C_r}, C_{r,i}\}_{i=1}^{N_{C_r}}$, $\{t_i^A, A_i\}_{i=1}^{N_A}$, $\{t_i^{D_r}, D_{r,i}\}_{i=1}^{N_{D_r}}$ are measurements for C_r, A, D_r, respectively. Similarly, $\{t_i^M, M_i\}_{i=1}^{N_M}$ and $\{t_i^V, V_i\}_{i=1}^{N_V}$ correspond to the measurements for mobility and cumulative vaccine doses. Here, $N_{C_r}, N_A, N_{D_r}, N_M, N_V$ represent the sizes of these measurements. We emphasize that these measurements may be collected at distinct time points, and the PINNs framework naturally accommodates such asynchronous data sampling [44].

The ODE loss is defined as the sum of the initial condition loss and the residual loss:

$$\mathcal{L}_{\text{ODE}}(\theta) = \mathcal{L}_0(\theta) + \mathcal{L}_r(\theta), \tag{2.4}$$

where $\mathcal{L}_0(\theta)$ denotes the loss associated with the initial conditions and $\mathcal{L}_r(\theta)$ represents the residual loss corresponding to the system dynamics. The initial condition loss quantifies the discrepancy between the model's prediction at the initial time $t_0 = 0$ and the initial values of the nine state variables (denoted by $[I_1, I_2, \ldots, I_9]$) given by [22]. It is defined as

$$\mathcal{L}_0(\theta) = \sum_{j=1}^{9} \left| y^j_{\theta_y}(t_0) - I_j \right|^2. \tag{2.5}$$

The residual loss measures the extent to which the model predictions satisfy the governing compartmental ODE system and is given by

$$\mathcal{L}_r(\theta) = \sum_{j=1}^{9} \frac{w_j}{N_{\text{ODE}}} \sum_{i=1}^{N_{\text{ODE}}} \left| R_j(t_i; \theta) \right|^2, \tag{2.6}$$

where N_{ODE} is the number of collocation points (denoted as $t_i, i = 1, ..., N_{\mathrm{ODE}}$) used to enforce the ODE system defined in equation (2.1), w_j denotes the weighting coefficient for the jth equation and $R_j(t;\theta)$ represents the residual of the jth ODE, computed as the difference between the left- and right-hand sides of equation (2.1), with the state variables and physics parameters approximated by the PINNs. For instance, consider the first equation in the system, $\mathrm{d}S/\mathrm{d}t = -\beta_t SI/N$. The corresponding residual is defined as

$$R_1(t;\theta) := \frac{\mathrm{d}y^1_{\theta_y}}{\mathrm{d}t}(t) + \frac{z^3_{\theta_z}(t)y^1_{\theta_y}(t)y^3_{\theta_y}(t)}{N}, \tag{2.7}$$

where $y^1_{\theta_y}(t)$, $y^3_{\theta_y}(t)$ and $z^3_{\theta_z}(t)$ approximate the variables S, I and the time-dependent transmission rate β_t, respectively.

Overfitting is a common challenge in time series forecasting involving real-world data. To address this issue, we apply L_2 regularization, a widely adopted technique that augments the original loss function with a penalty term proportional to the sum of the squared network parameters. This regularization term is defined as

$$L_2(\theta) = \sum_{j=1}^{n} |\theta_j|^2, \tag{2.8}$$

where θ_j, for $j = 1, \ldots, n$, denotes the full set of parameters in the PINNs, including the parameters of the two sub-networks as well as the physics-related parameters p_h and p_d. The total number of parameters is denoted by n. By penalizing large parameter values, L_2 regularization discourages overly complex models and promotes generalization. In practice, this technique is conveniently implemented in `PyTorch` through the `weight_decay` argument in the optimizer settings.

We used the Adam optimizer [70] with a learning rate of 1×10^{-3} for 50 000 epochs. L_2 regularization with a coefficient of 1×10^{-5} was applied to reduce overfitting. To balance the scales between $\mathcal{L}_{\mathrm{data}}$ and $\mathcal{L}_{\mathrm{ODE}}$, we set the weight $w_{\mathrm{ODE}} = 0.1$. All hyperparameters were manually tuned to optimize predictive accuracy. Specifically, we tuned the learning rate in the range from 1×10^{-5} to 1×10^{-1}, the number of layers in the range of 2–5, the neurons per layer in the range from 20 to 100, the number of training epochs in the range from 10 000 to 200 000, regularization weight in the range from 1×10^{-10} to 1×10^{-1} and the weights assigned to each loss component in the range from 1×10^{-5} to 1. We evaluated different hyperparameter combinations by calculating the corresponding mean absolute scaled error (MASE) for point forecasting and the weighted interval score (WIS) for quantile forecasting, selecting the set that achieved the overall best predictive accuracy. Training and testing were conducted on an NVIDIA GeForce RTX 3090 GPU.

We trained the PINNs model using historical data from July 2020 to April 2022, using a rolling window approach, as illustrated in figure 1B, to enable continuous predictions for the subsequent 1–4 weeks. Initially, PINNs were trained on data from weeks 1 to 17, with predictions made for weeks 18–21. The training window was then extended by one week, retraining the model on data from weeks 1 to 18 and forecasting weeks 19–22. This process was repeated until PINNs were trained using data from weeks 1 to 89 and predictions were made for weeks 90–93. As a result, each week from 21 to 90 was treated as the subsequent 1-, 2-, 3- and 4-week forecast in the rolling window procedure.

2.4. Evaluation metrics

We used two metrics to assess predictive performance: the MASE [71] for point forecasting and the WIS for quantile forecasting. MASE was calculated by dividing the mean absolute error (MAE) of the model's predictions by the MAE of a naive model of the same forecast horizon:

$$\mathrm{MASE} = \frac{\mathrm{MAE}_{\mathrm{PINNs}}}{\mathrm{MAE}_{\mathrm{naive}}}. \tag{2.9}$$

MAE was calculated as the mean absolute difference between model predictions $\hat{y}_i$ and empirical observations y_i:

$$\mathrm{MAE} = \frac{1}{N}\sum_{i=1}^{N} |\hat{y}_i - y_i|, \tag{2.10}$$

where N is the number of data.

The naive model uses data from the previous weeks as predictions for the following weeks. Specifically, when the forecast horizon is one week, the naive model uses data from the past week to predict the following week. Similarly, for a two-week forecast horizon, the naive model uses data from the past two weeks to predict the subsequent two weeks. A MASE value smaller than 1.0 indicates that our PINNs model outperforms the naive model in terms of predictive accuracy. Moreover, a lower MASE value corresponds to more accurate predictions.

In infectious disease forecasting, quantile predictions, namely probabilistic predictions, are typically required to account for error propagation. We assume that the prediction distribution follows a Gaussian distribution, with mean μ_t equal to the point forecast at time t and s.d. σ, which represents the s.d. of all prediction errors and is constant across time. The prediction quantiles at time t are then given by $\mu_t + \sigma\sqrt{2}\,\mathrm{erf}^{-1}(2p-1)$, where erf^{-1} denotes the inverse error function and p represents the desired quantile.

To evaluate probabilistic predictions using the WIS, we used the 'interval-scoring' Python package (https://github.com/adrian-lison/interval-scoring), based on [65]. We calculated the WIS with quantile bins as required by the COVID-19 Forecast Hub, specifically $\alpha_1 = 0.02$, $\alpha_2 = 0.05$, $\alpha_3 = 0.1$, …, $\alpha_{11} = 0.9$ and quantiles including 0.010, 0.025, 0.050, 0.100, 0.150, 0.200, 0.250, 0.300, 0.350, 0.400, 0.450, 0.500, 0.550, 0.600, 0.650, 0.700, 0.750, 0.800, 0.850, 0.900, 0.950, 0.975 and 0.990.

2.5. Ablation study

In contrast to traditional NNs, we incorporated epidemiological knowledge into our model by adding the ODE loss to the overall loss function. To assess the impact of embedding domain-specific epidemiological knowledge into the NN, we also trained a traditional neural network (denoted as NN) and compared its performance with our PINNs model. The NN shares the same architecture as the NN component in the upper left part of figure 1A, but it is purely data-driven and does not integrate the compartmental model.

In contrast to traditional mathematical methods, we used a NN and integrated it with the compartmental model. To evaluate the advantages of PINNs over the compartmental model, we compared the performance of our PINNs with that of the Gaussian infection state space with time dependence (GISST) model [23]. GISST is a conventional mathematical model that solves the same compartmental model.

It is important to note that early historical data may not accurately capture the dynamics of current virus variants due to mutations and may even negatively impact future predictions. To assess this, we investigated the effect of the training data window size on prediction accuracy. While PINNs were originally trained using all available historical data, we also trained models using only recent data and compared their predictive performance. The results are presented in appendix A.1.

Additionally, we conducted an analysis of the impact of L_2 regularization on model performance and evaluated its necessity in our forecasting tasks. This analysis helps to assess whether regularization significantly contributes to improving generalization and preventing overfitting. The results of these experiments, including performance comparisons with and without L_2 regularization, are presented in appendix A.2.

3. Results

3.1. Point forecasts on the number of cases, deaths and hospitalizations for the following one to four weeks

At each time point t, we trained PINNs using historical data from July 2020 up to t and generated forecasts for the subsequent one to four weeks. The predictions, shown in figure 3, demonstrate that the outputs from PINNs are consistently aligned with the observed target values and are more accurate than those from the naive model across all forecasting horizons. Additionally, PINNs outperform most of the corresponding forecasts produced by GISST. The results also indicate that one-week-ahead predictions are the most reliable, with accuracy gradually declining for forecasts on the subsequent two-, three- and four-week horizons. As the forecasting horizon increases, the predicted values tend to diverge more from the ground truth, accompanied by greater fluctuations in the predicted curves.

Quantitative comparisons are reported in table 3. In the column labelled 'method', the name before the colon indicates the forecasting approach, while the number denotes the forecast horizon. Scores highlighted in bold represent the best-performing model (the lowest score) for a given forecast horizon within each metric column. For example, 'PINNs: one week' refers to PINNs predictions for the subsequent first week, while 'PINNs: two weeks' refers to forecasts for the second week ahead. We observe that all of PINNs' MASE scores are no more than 1.0, indicating that PINNs consistently outperform the naive model across all forecast horizons. Among the three state variables, the MASE scores for deaths are generally higher than those for cases and hospitalizations, probably due to greater variability and fluctuations in the original death data.

3.2. Quantile forecasts on the number of cases, deaths and hospitalizations for the following one to four weeks

Similarly, at each time point t, we trained PINNs using historical data from July 2020 up to t and generated predictions for the subsequent one to four weeks. In this case, we visualized the continuous forecasts along with the associated uncertainty in figure 4. For each group of one- to four-week predictions, the points (representing the median predictions at quantile = 0.5) progressively deviate from the ground truth as the forecast horizon extends. Correspondingly, the light blue region, which represents predictive uncertainty, expands with longer horizons. This pattern highlights a common trend: as the forecast horizon increases, prediction accuracy decreases while uncertainty grows. In this study, we assume the s.d. σ of the predictive Gaussian distribution is different for different prediction horizons. σ is computed as the s.d. of the errors of the 70 predictions for the one-, two-, three- and four-week predictions, respectively. The values for σ are provided in table 4.

Quantitative comparisons of WIS are presented in table 3. The scaled WIS was computed by dividing the WIS of each model by the WIS of the naive model for the corresponding forecast horizon. We observe that PINNs achieve scaled WIS scores that are generally below 1.0, indicating superior performance over the naive baseline across most forecasting horizons. The only exception occurs in the one-week-ahead quantile predictions for death, where the PINNs' WIS score slightly exceeds 1.0. However, this deviation is minor and still significantly lower than the scores of other models.

3.3. Comparison with sequence deep-learning models and traditional mathematical model

We first compared our PINNs' predictions with results obtained from the NN without incorporating the compartmental model, as well as with the results of a traditional mathematical model named GISST [22]. As shown in table 3, both MASE and WIS scores for PINNs are lower than the corresponding scores of NN for each forecast horizon, which demonstrates the absolute advantage of PINNs over NN for both point forecasting and quantile forecasting. This is probably due to the regularizing effect of the compartmental model.

Table 3. MASE and WIS comparison of different methods for point forecasting and quantile forecasting, respectively. Column 'hosp' denotes hospitalizations. Scaled WIS is computed by dividing the WIS by the WIS of naive models of the same forecast horizon. Scores in bold denote the lowest score of the same forecast horizon in that column.

method	MASE			WIS			scaled WIS		
	cases	deaths	hosp	cases	deaths	hosp	cases	deaths	hosp
PINNs: one week	**0.75**	1.00	**0.58**	16 247	**103**	**250**	0.70	**1.04**	**0.57**
PINNs: two week	0.77	0.92	**0.62**	32 954	146	**556**	0.79	1.00	**0.66**
PINNs: three week	0.81	0.86	**0.68**	49 272	**176**	**882**	0.87	**0.91**	**0.72**
PINNs: four week	0.85	0.89	**0.74**	63 311	217	**1232**	0.94	0.89	**0.79**
NN: one week	0.91	1.32	0.65	25 152	158	298	1.09	1.60	0.68
NN: two week	1.27	1.76	0.79	52 761	298	729	1.27	2.04	0.86
NN: three week	1.56	1.86	0.89	88 920	442	1281	1.57	2.29	1.05
NN: four week	1.65	1.84	0.97	1 18 515	558	1885	1.76	2.28	1.22
GISST: one week	0.85	**0.88**	0.91	**15 346**	163	401	**0.66**	1.65	0.92
GISST: two week	**0.75**	**0.64**	0.78	**26 233**	185	798	**0.63**	1.27	0.94
GISST: three week	**0.71**	**0.45**	0.77	**37 041**	187	1300	**0.65**	0.97	1.06
GISST: four week	**0.76**	**0.34**	0.89	58 148	**181**	2159	0.86	**0.74**	1.39
RNN: one week	0.91	1.28	0.70	17 653	126	278	0.76	1.27	0.64
RNN: two week	0.92	1.17	0.82	33 544	168	642	0.81	1.15	0.76
RNN: three week	0.94	1.09	0.94	46 641	205	1105	0.82	1.06	0.90
RNN: four week	0.92	0.99	1.03	**53 964**	242	1554	**0.80**	0.99	1.00
LSTM: one week	1.00	1.29	0.79	19 622	124	325	0.85	1.25	0.74
LSTM: two week	1.00	1.39	0.93	35 249	194	786	0.85	1.33	0.93
LSTM: three week	0.98	1.16	1.03	49 661	226	1249	0.87	1.17	1.02
LSTM: four week	0.94	1.09	1.04	56 474	268	1624	0.84	1.09	1.05
GRU: one week	0.89	1.22	0.77	16 952	115	324	0.73	1.16	0.74
GRU: two week	0.90	1.24	0.88	33 573	178	749	0.81	1.22	0.88
GRU: three week	0.93	1.14	0.98	47 267	225	1198	0.83	1.17	0.98
GRU: four week	0.90	1.03	1.01	55 414	262	1608	0.82	1.07	1.04
transformer: one week	0.96	1.14	0.81	22 155	111	356	0.96	1.12	0.81
transformer: two week	0.99	0.98	0.86	40 494	**145**	730	0.97	**0.99**	0.86
transformer: three week	1.05	0.97	0.94	56 804	193	1179	1.00	1.00	0.96
transformer: four week	1.08	0.91	1.04	69 297	235	1629	1.03	0.96	1.05

Table 4. Standard deviations (σ) of forecast errors by week for cases, deaths and hospitalizations.

week	cases	deaths	hosp
1	0.3715	0.3303	0.1739
2	0.5525	0.4217	0.3134
3	0.7732	0.4997	0.4598
4	0.9658	0.6248	0.6126

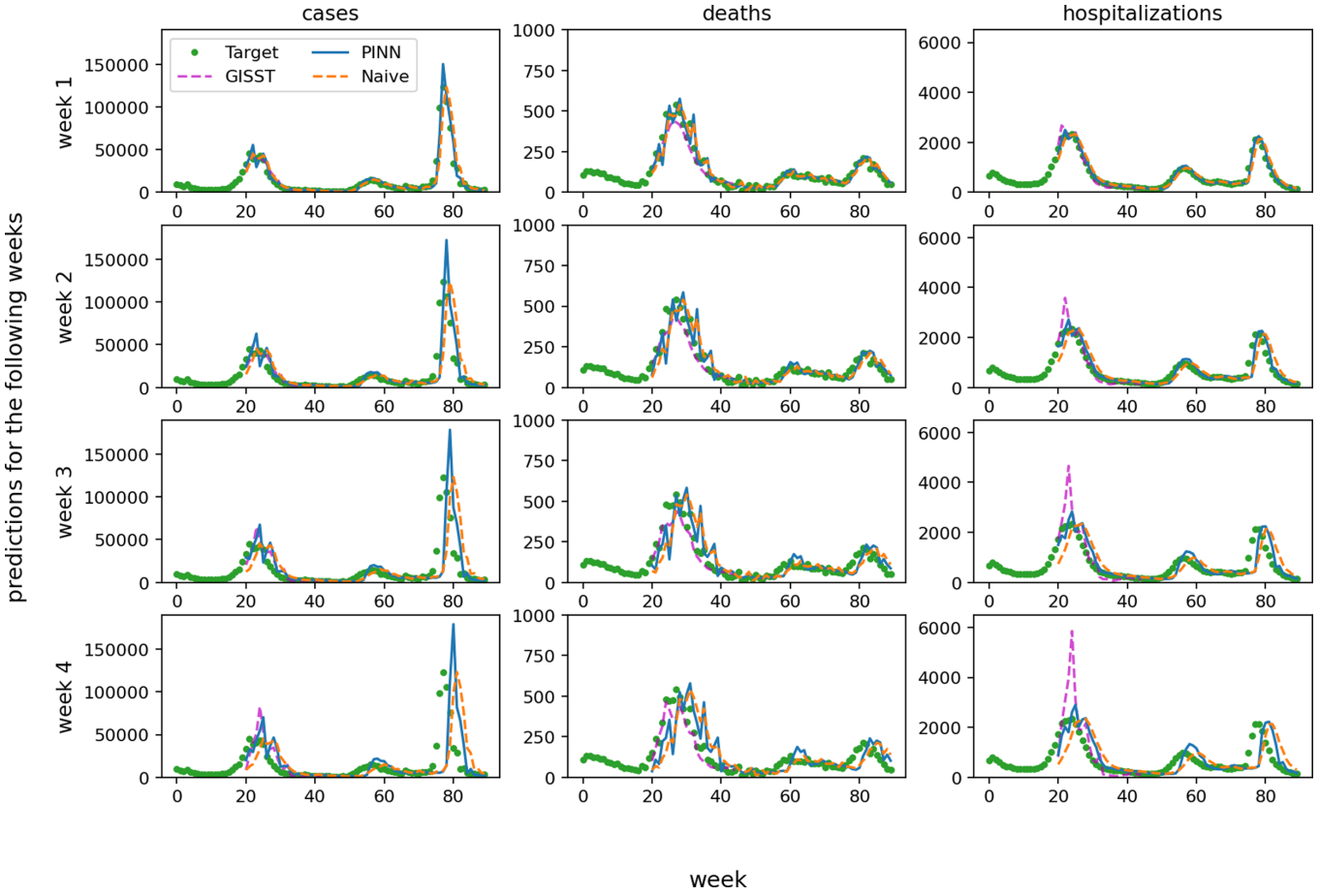

Figure 3. PINNs' point predictions on the number of cases, deaths and hospitalizations for the following one to four weeks. GISST represents a mathematical model named the Gaussian infection state space with time dependence [22]. The naive model uses data from the previous weeks as predictions for the following weeks.

Additionally, PINNs typically outperformed GISST when forecasting hospitalizations. GISST outperformed PINNs for cases and forecasting performance for deaths, which depended on whether the evaluation was for point forecasts (GISST superior) or quantile forecasts (PINNs superior). These results demonstrate the important role of epidemiological domain knowledge as a constraint to avoid unreasonable predictions that may arise in purely data-driven modelling.

To further evaluate the performance of PINNs, we compared them against several state-of-the-art NN architectures commonly used for sequential data processing. Specifically, we implemented RNNs [72], LSTMs [73], GRUs [74] and transformer models [75] to address the same forecasting task. The RNN, LSTM and GRU models share a similar architecture configuration. Each model consists of an input layer with three neurons, two hidden layers with 50 neurons each and an output layer with three neurons. The transformer model is composed of an input layer with three neurons, followed by two encoder layers, each with a hidden dimensionality of 64 and four attention heads and an output layer with three neurons. These models take weekly data on cases, deaths and hospitalizations as input features and generate corresponding predictions for each of the three variables. A window size of seven weeks was used for input. Training was conducted using the Adam optimizer with a learning rate of 1×10^{-3} over 10 000 epochs.

The window size is a crucial hyperparameter for these models. Thus, we evaluated the models using window sizes of 3, 5, 7, 10, 12 and 15 weeks and the results are presented in figure 5. The findings align with our expectations: a window that is too short does not adequately capture the temporal dependencies of the data, whereas an excessively long window introduces redundancy. A window size of seven weeks strikes an optimal balance and is well suited for this forecasting task.

The predictions for one- two-, three- and four-week forecasts are shown in figure 6A. In comparison with PINNs, the forecasts generated by RNNs, LSTMs, GRUs and transformers exhibit noticeable fluctuations, indicating a higher degree of overfitting. These fluctuations are more clearly observed in figure 6B, which provides a zoomed-in view of the four-week predictions for deaths. This contrast underscores the advantage of incorporating physical information into PINNs, which contributes to more stable and reliable predictions. A quantitative comparison is provided in table 3, where PINNs consistently outperform all other prominent NN models commonly used for time-series forecasting, in terms of both MASE for point forecasting and WIS for quantile forecasting. Despite architectural advances in RNNs, LSTMs, GRUs and transformers, none of these models demonstrates substantial performance gains on this task. These findings emphasize the importance of integrating physics-based domain knowledge into NN models for improved generalization and predictive accuracy.

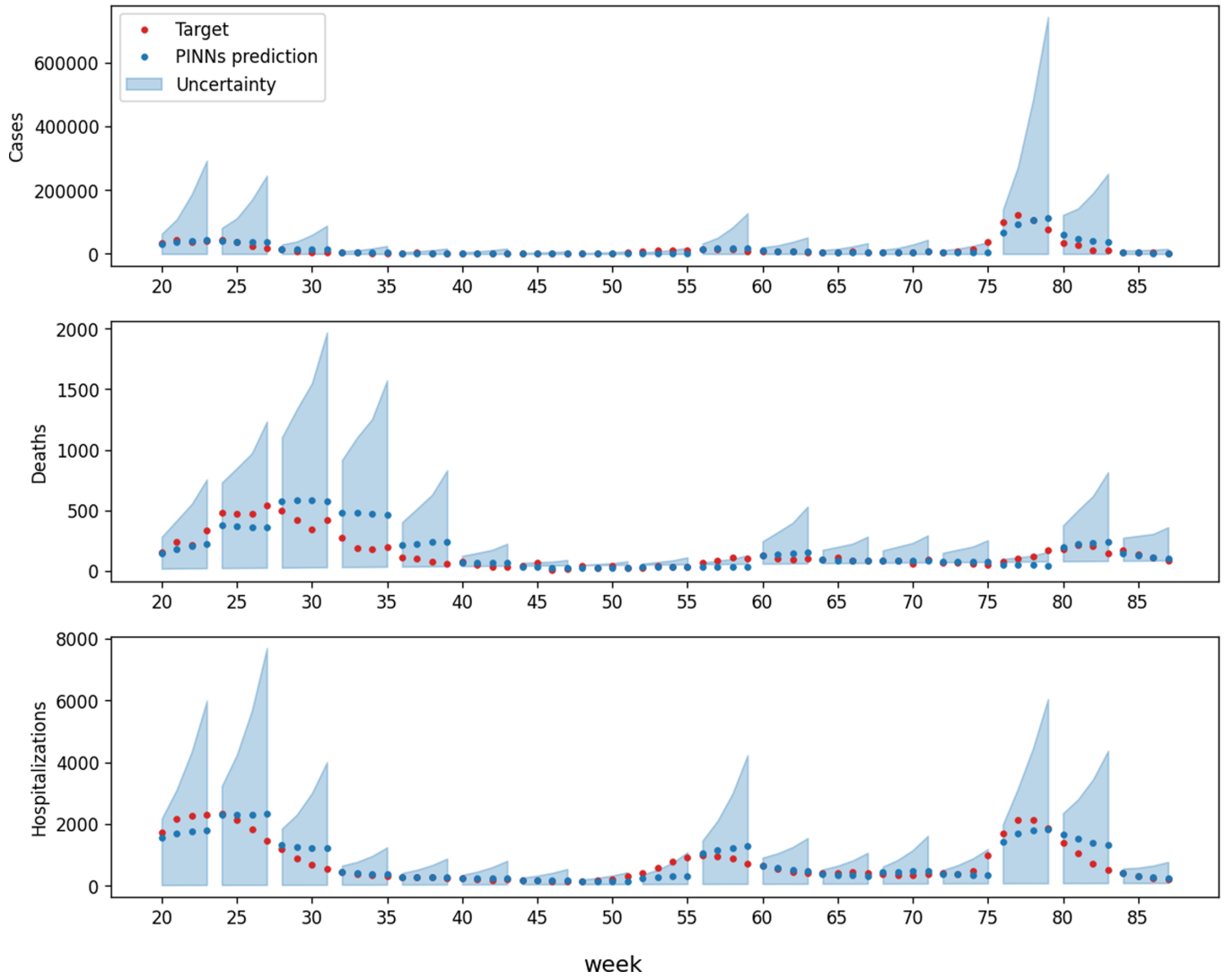

Figure 4. PINNs' quantile predictions on the number of cases, deaths and hospitalizations for the following one to four weeks. The red dots represent the ground truth. The blue points indicate PINNs' predictions and the light blue region represents the associated uncertainty in the forecasts. Each polygon spans a continuous four-week prediction interval, encompassing the one-, two-, three- and four-week forecasts made from a time point one week prior to the start of the polygon.

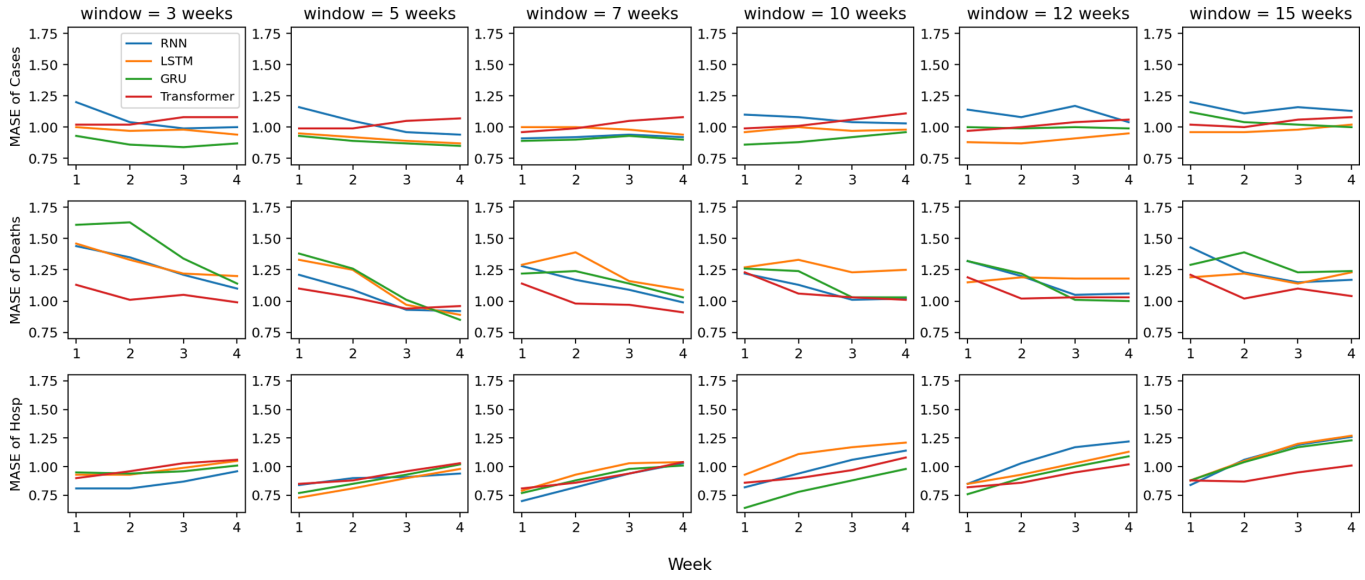

Figure 5. A MASE comparison of RNNs, LSTM networks, GRUs and transformer models, with training window sizes ranging from 3 to 15 weeks.

4. Discussion and summary

Epidemic forecasting can traditionally be framed as an inverse-forward problem, wherein model parameters must first be inferred before generating future predictions. In many scenarios, the mechanisms underlying disease transmission are partially understood

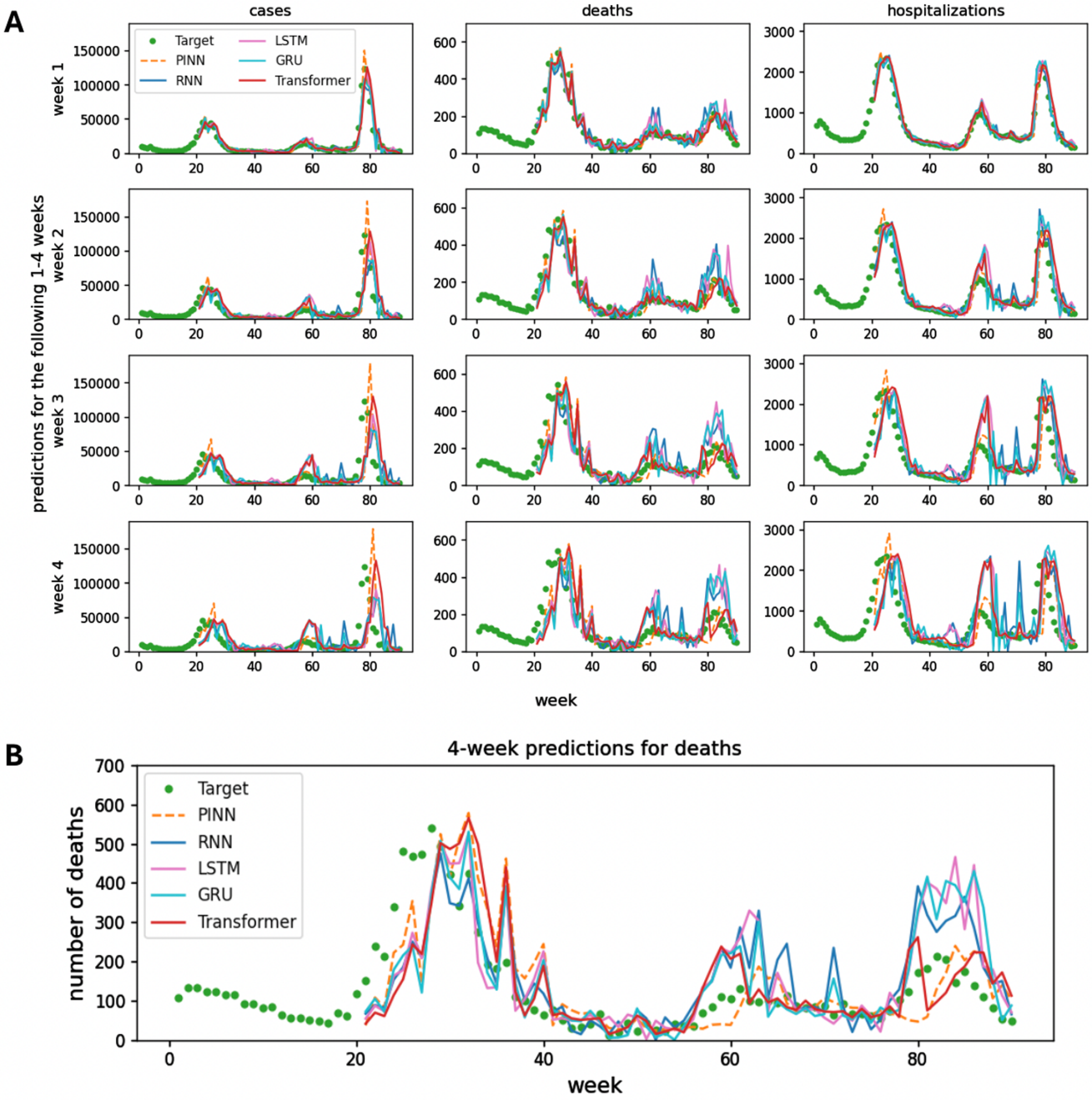

Figure 6. MASE comparison with RNNs, LSTM networks, GRUs and transformers. (A) An overview comparison of the number of cases, deaths and hospitalizations for the following one to four weeks. (B) A focused comparison of deaths for the four-week predictions.

and represented using compartmental models in the form of ODEs. However, several model parameters remain unknown and must be estimated by fitting the model to observational data. Once the parameters are identified, these compartmental models can be used to forecast future trends in disease spread and evaluate the impact of public health interventions or the emergence of new variants.

Despite their utility, traditional mathematical models suffer from inherent limitations due to model misspecification. For example, they often assume fixed transition rates, which makes it challenging to adapt to evolving conditions, such as changes in transmission dynamics caused by new variants. Moreover, forecasting is not feasible unless all relevant parameters are accurately estimated. In contrast, data-driven NN models have emerged as powerful tools for addressing complex social and scientific problems in recent years. However, conventional NNs typically require large volumes of training data and labelled examples to perform effectively, and they often disregard prior domain knowledge. As a result, ML models may produce unrealistic or unreliable forecasts, particularly in extrapolative scenarios, due to their lack of grounding in the underlying physical or epidemiological processes.

In this work, we used PINNs to integrate epidemiological knowledge, as described by compartmental models, with observed data to enhance the accuracy of infectious disease forecasts. Our results demonstrate that the proposed approach offers several key advantages that contribute to its superior performance. First, the PINNs framework leverages NNs to model complex and nonlinear patterns inherent in infectious disease dynamics, enabling the system to capture intricate relationships that traditional

models may miss. Second, PINNs incorporate domain-specific knowledge by embedding the compartmental models into the loss function. This integration ensures that the learnt representations conform to the underlying principles and constraints of disease transmission. Moreover, the inclusion of mechanistic models in the training process helps mitigate overfitting, a common issue when relying solely on data-driven ML approaches. By constraining the learning process with known epidemiological laws, PINNs achieve a balanced trade-off between flexibility and generalizability.

Additionally, PINNs can be used to estimate key epidemiological parameters—such as transmission and recovery rates—directly from data, thereby enhancing the model's adaptability to evolving disease dynamics. In particular, PINNs treat these parameters as outputs of the NN, allowing them to vary over time. This enables the model to capture temporal changes in transmission behaviour, such as those driven by viral mutations or shifts in public health interventions. Furthermore, PINNs are capable of continuously incorporating newly acquired data, allowing for real-time model updates and progressively improved forecasts as more information becomes available during an ongoing outbreak. This adaptability makes PINNs particularly well suited for real-world epidemic surveillance and response. In addition, PINNs can integrate heterogeneous data sources—including social, environmental and demographic factors—offering a more comprehensive and holistic framework for infectious disease forecasting.

Recently, several PINNs-based models have been developed for infectious disease forecasting, including applications to COVID-19 [62,76,77] and measles [78]. Compared with these existing approaches, our proposed model incorporates a more complex compartmental framework with nine state variables embedded into the NN architecture. This significantly increases the challenge of ensuring consistency with the underlying physics described by the ODE system, thereby demonstrating the flexibility and robustness of our model. Notably, our model achieved better performance metrics—normalized root mean squared error (NR1 and NR2) and normal deviation (ND), as defined in [76]—compared with the EINNs [76] in forecasting COVID-19 deaths, which is the only predicted state variable in EINNs. Specifically, our model achieved 0.53 (EINNs: 0.54) on NR1, 0.15 (EINNs: 0.24) on NR2 and 0.33 (EINNs: 0.38) on ND, while simultaneously predicting two additional state variables. Furthermore, while prior models primarily offer point forecasts, we extend our approach to quantile forecasts, which allows us to quantify uncertainty in the predictions.

Despite the strengths of the PINNs framework in handling noisy and limited data, we note that while the estimated transmission rate β_t falls within the reasonable scale as in observation data, it did not correlate well with the observed variables, namely cases, deaths and hospitalizations. This is probably due to the highly underdetermined nature of the problem: our ODE system includes nine state variables, yet only three of them have associated observational data. These three observation data do not directly impact the ODEs that contain β_t. As a result, the PINNs model struggles to infer transmission rates that align with the reference values, as illustrated in blue dots in figure 8. To further understand this limitation, we performed additional experiments in which an increased number of sparse ground-truth data points of the transmission rate were provided to guide the PINNs model. Figure 9 shows that with even a small number of such data points, the model was able to capture the dynamic changes in the transmission rate. These results suggest that when applying PINNs to high-dimensional ODE systems with limited observation data, the model may still produce reliable predictions for the variables with observation data, even if it cannot accurately infer all the unknown parameters in the ODE system.

The current model also lacks a built-in mechanism for quantifying the uncertainty of its predictions in the presence of input noise. A well-established method for regression with uncertainty estimation is Gaussian process regression (GPR), which, rooted in the Bayesian framework, provides a principled way to quantify uncertainty in the solution of differential equations. However, GPR is typically limited to small, linear problems due to its high computational complexity, which scales poorly with large datasets. A promising direction to address this limitation involves the use of Bayesian PINNs, as proposed by [38]. These models combine the physics-informed learning capability of PINNs with the probabilistic nature of Bayesian inference, allowing for uncertainty-aware solutions even in ill-posed settings, such as those with incomplete or under-specified boundary conditions. Future research should investigate the applicability and effectiveness of Bayesian PINNs in the context of infectious disease forecasting, particularly in scenarios involving high uncertainty, evolving dynamics or limited data availability.

In summary, the application of PINNs in infectious disease forecasting offers a synergistic approach that combines the power of data-driven deep learning with the interpretability and constraints provided by underlying epidemiological principles, leading to more accurate, interpretable and adaptable models for predicting the spread of infectious diseases.

Ethics. This work did not require ethical approval from a human subject or animal welfare committee.

Data accessibility. All relevant code and data are available from Zenodo [79].

Declaration of AI use. We have not used AI-assisted technologies in creating this article.

Authors' contributions. Y.Q.: conceptualization, data curation, formal analysis, investigation, methodology, validation, visualization, writing—original draft, writing—review and editing; K.Z.: investigation, methodology, validation; E.M.: conceptualization, project administration, resources, supervision; A.B.: data curation, resources; E.B.O.: data curation, resources; X.W.: resources, supervision; S.J.F.: conceptualization, resources, supervision, writing—review and editing; P.R.: conceptualization, resources, supervision, writing—review and editing; J.M.D.: conceptualization, project administration, resources, supervision, writing—review and editing; H.L.: conceptualization, funding acquisition, methodology, project administration, resources, supervision, writing—review and editing.

All authors gave final approval for publication and agreed to be held accountable for the work performed therein.

Conflict of interest declaration. We declare we have no competing interests.

Funding. This work was supported by National Institutes of Health grants R21HL168507 and NSF SCH Award Number 2406212. High-performance computing resources were provided by the Georgia Advanced Computing Resource Center (GACRC) at the University of Georgia.

Acknowledgements. H.L. would like to thank Shan-Ho Tsai from the Georgia Advanced Computing Resource Center (GACRC) for providing technical support.

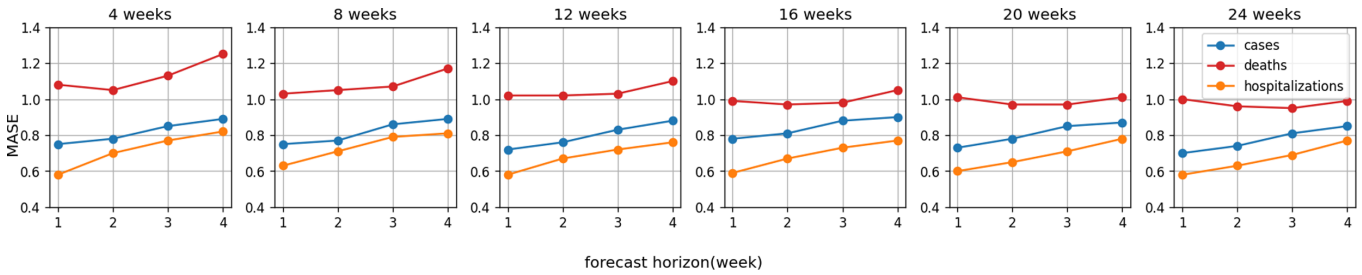

Figure 7. MASE for the predictions of PINNs models on cases, deaths and hospitalizations when trained with varied lengths of historical data ranging from 4 to 24 weeks.

Appendix A.

A.1. Impact of training data length

We trained PINNs using datasets comprising the most recent 4, 8, 12, 16, 20 and 24 weeks of observations and evaluated their forecasting performance over the period spanning weeks 48–110. The results, shown in figure 7, illustrate the MASE scores for predictions of cases, deaths and hospitalizations across one-, two-, and four-week horizons using different training data lengths.

For all three targets, we observe a consistent trend: as the length of training data increases, the MASE scores generally decrease and then stabilize. Models trained on only four or eight weeks of data tend to underperform, probably due to insufficient information to accurately capture the underlying disease dynamics. In contrast, models trained on 12 or 16 weeks of data achieve predictive accuracy comparable to those trained on the full historical dataset. Notably, further increasing the training window to 20 or 24 weeks yields only marginal improvements in accuracy. This suggests that data beyond 16 weeks—spanning nearly half a year—may be less relevant to current epidemiological conditions and contribute less to near-term forecasting performance.

A.2. Impact of L_2 regularization

We further demonstrate that the absence of L_2 regularization leads to a substantial decline in model accuracy. Specifically, removing L_2 regularization results in average performance drops of 22.48% for cases, 9.16% for deaths and 42.54% for hospitalizations over the one- to four-week forecasting horizons. These results highlight the importance of incorporating L_2 regularization in the PINNs framework, as it contributes to improved generalization and prediction stability across multiple epidemiological targets.

A.3. Estimates of transmission rate

In the current study, we treat the transmission rate as a time-dependent variable of the ODE system, and it is associated with mobility and cumulative vaccine dose information. We adapt original statistical data from https://data.cdc.gov/Public-Health-Surveillance/United-States-COVID-19-County-Level-of-Community-T/nra9-vzzn/data and get the transmission rate after processing the original data. According to [80], we mapped community transmission level to specific values and calculated average transmission rate among communities in California. These data, shown as blue dots in figures 8 and 9, will act as additional information in the training process as well as references for inferred β_t from PINNs. As illustrated in the upper panel of figure 8, when no real transmission data are used to inform the PINNs model, the inferred β_t did not capture the trend of the real transmission rate. On the other hand, figure 9 shows that when training data are gradually added from one to six data points, PINNs can roughly capture the tendency of transmission rate.

Recall that in the main text, the transmission rate β_t is modelled as the third output of the lower sub-network. To ensure positivity, an exponential transformation is applied, i.e. $z^3_{\theta_z} = \exp(\alpha)$, where α denotes the third output of the fully connected network within the lower sub-network. In this section, we explore two alternative output transformations for modelling β_t within the PINNs framework:

(1) $z^3_{\theta_z} = (\arctan(\alpha) + \frac{\pi}{2}) \times 10^{-3}$,
(2) $z^3_{\theta_z} = \exp(\min(\log(1000), \alpha + \beta_{\text{res}} z^1_{\theta_z} + \beta_{\text{doses}} z^2_{\theta_z}))$, according to eqn (2.4) of [22].

In the second transformation, β_{res} and β_{doses} are additional trainable physical parameters determined during the PINNs training process, while $z^1_{\theta_z}$ and $z^2_{\theta_z}$ approximate mobility and cumulative vaccine doses, respectively. The prediction performance of the model based on these two ways of involving the auxiliary variables is summarized in table 5 and figure 8, which shows no significant difference in model performance compared with the exponential transformation used in the main text. This is probably due to the highly underdetermined nature of the problem: our ODE system includes nine state variables, yet only three of them have associated observational data. These three observation data do not directly impact the ODEs that contain β_t. As a result, the PINNs model struggles to infer transmission rates that align with the reference values.

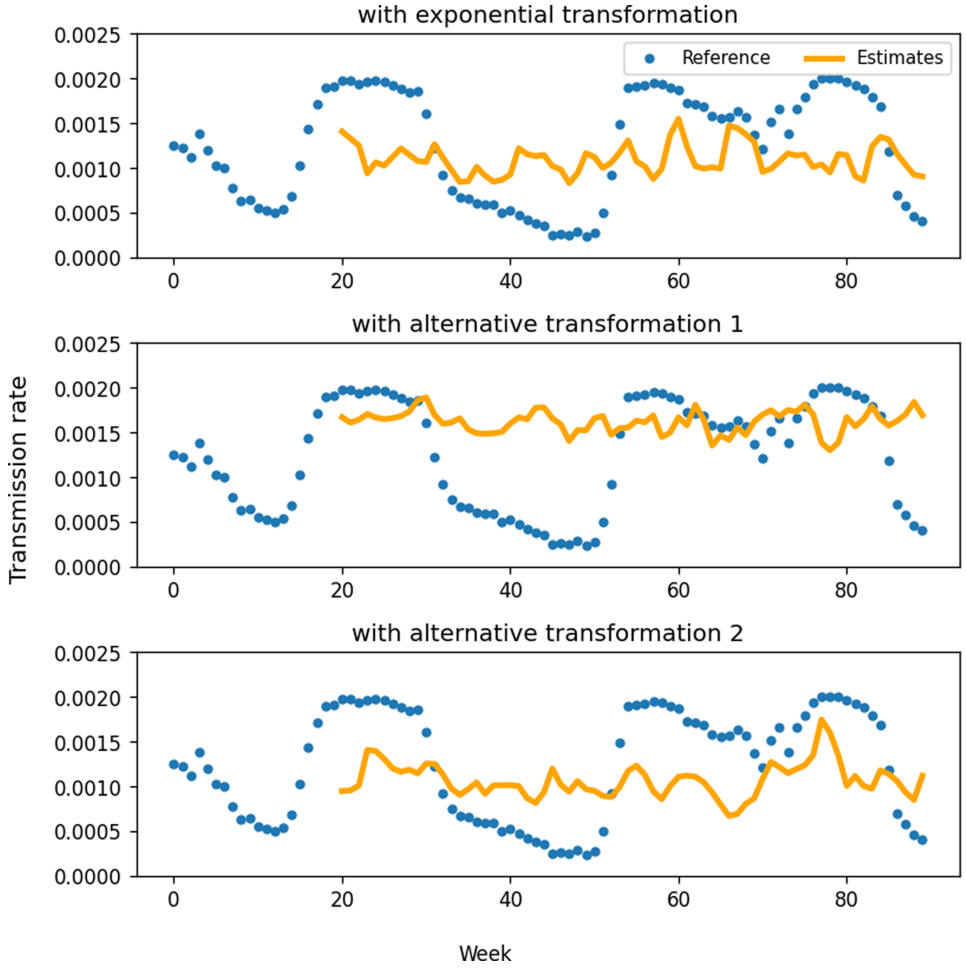

Figure 8. The transmission rate is estimated by the proposed PINNs model without observation points. Blue dots represent data collected from CDC United States COVID-19 County Level of Community Transmission Historical Changes. The upper panel corresponds to the PINNs model in the main text, whereas the middle and lower panels correspond to PINNs with the two alternative output transformations as defined in §3.3.

Table 5. MASE and WIS comparison of PINNs methods with different formulation of transmission rate β_t. In PINNs$_1$, β_t is modelled by $z^3_{\theta_z} = (\arctan(\alpha) + \frac{\pi}{2}) \times 10^{-3}$, while in PINNs$_2$, β_t is modelled by $z^3_{\theta_z} = \exp(\min(\log(1000), \alpha + \beta_{\text{res}} z^1_{\theta_z} + \beta_{\text{doses}} z^2_{\theta_z}))$ according to [22], where α denotes the third output of the fully connected network within the lower sub-network. Column 'hosp' denotes hospitalizations. Scaled WIS is computed by dividing the WIS by the WIS of naive models of the same forecast horizon.

method	MASE			WIS			scaled WIS		
	cases	deaths	hosp	cases	deaths	hosp	cases	deaths	hosp
PINNs$_1$: one week	0.78	1.03	0.59	16 716	104	254	0.72	1.05	0.58
PINNs$_1$: two week	0.79	0.95	0.64	33 465	150	564	0.80	1.03	0.67
PINNs$_1$: three week	0.81	0.88	0.70	49 666	179	894	0.87	0.93	0.73
PINNs$_1$: four week	0.86	0.90	0.75	63 474	219	1246	0.94	0.89	0.80
PINNs$_2$: one week	0.76	1.00	0.60	16 336	103	254	0.71	1.04	0.58
PINNs$_2$: two week	0.78	0.95	0.64	33 155	148	562	0.80	1.01	0.66
PINNs$_2$: three week	0.82	0.88	0.69	49 621	179	889	0.87	0.93	0.73
PINNs$_2$: four week	0.86	0.89	0.75	63 894	218	1241	0.95	0.89	0.80

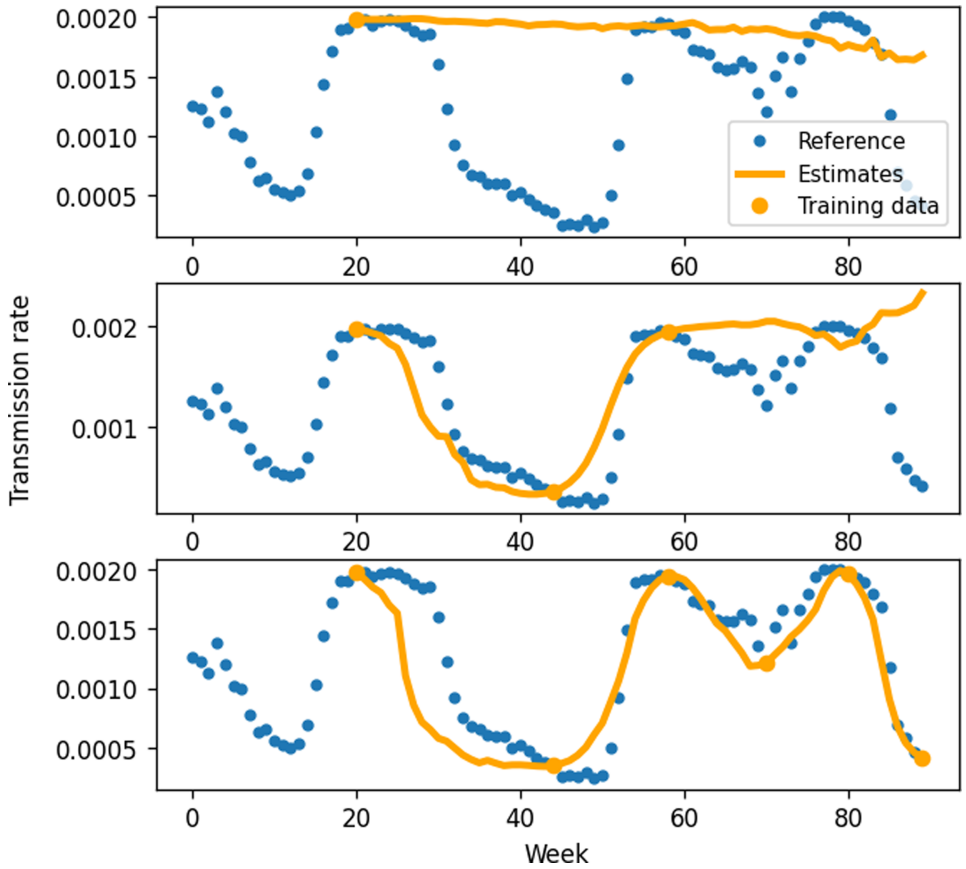

Figure 9. The transmission rate is estimated by the proposed PINNs model with a varied number of observation points. One, three and five data points are examined, respectively, and the results are plotted from top to bottom. Blue dots represent data collected from CDC United States COVID-19 County Level of Community Transmission Historical Changes.

References

1. Ahmad FB, Cisewski JA, Anderson RN. 2022 Provisional mortality data—United States, 2021. *MMWR Morb. Mortal. Wkly Rep*. **71**, 597–600. (doi:10.15585/mmwr.mm7117e1)
2. Morens DM, Folkers GK, Fauci AS. 2004 The challenge of emerging and re-emerging infectious diseases. *Nature* **430**, 242–249. (doi:10.1038/nature02759)
3. Carlson CJ, Albery GF, Merow C, Trisos CH, Zipfel CM, Eskew EA, Olival KJ, Ross N, Bansal S. 2022 Climate change increases cross-species viral transmission risk. *Nature* **607**, 555–562. (doi:10.1038/s41586-022-04788-w)
4. Biggerstaff M, Slayton RB, Johansson MA, Butler JC. 2022 Improving pandemic response: employing mathematical modeling to confront coronavirus disease 2019. *Clin. Infect. Dis*. **74**, 913–917. (doi:10.1093/cid/ciab673)
5. Lutz CS *et al*. 2019 Applying infectious disease forecasting to public health: a path forward using influenza forecasting examples. *BMC Public Health* **19**, 1659. (doi:10.1186/s12889-019-7966-8)
6. Case BKM, Salcedo MV, Fox SJ. 2025 An accurate hierarchical model to forecast diverse seasonal infectious diseases. *medRxiv*. (doi:10.1101/2025.03.03.25323259)
7. Anderson RM, May RM. 1991 *Infectious diseases of humans*. Oxford, UK: Oxford University Press.
8. Keeling MJ, Rohani P. 2008 *Modelling infectious diseases: in humans and animals*. Princeton, NJ: Princeton University Press. (doi:10.1515/9781400841035)
9. Brauer F. 2008 Compartmental models in epidemiology. In *Mathematical epidemiology* (eds F Brauer, P van den Driessche, J Wu), pp. 19–79. Berlin, Germany: Springer. (doi:10.1007/978-3-540-78911-6_2)
10. Brauer F, Castillo-Chavez C, Feng Z. 2019 *Mathematical models in epidemiology*. New York, NY: Springer. (doi:10.1007/978-1-4939-9828-9)
11. Martcheva M. 2015 *An introduction to mathematical epidemiology*. New York, NY: Springer. (doi:10.1007/978-1-4899-7612-3)
12. Huppert A, Katriel G. 2013 Mathematical modelling and prediction in infectious disease epidemiology. *Clin. Microbiol. Infect*. **19**, 999–1005. (doi:10.1111/1469-0691.12308)
13. Wearing HJ, Rohani P, Keeling MJ. 2005 Appropriate models for the management of infectious diseases. *PLoS Med*. **2**, e174. (doi:10.1371/journal.pmed.0020174)
14. Massonis G, Banga JR, Villaverde AF. 2021 Structural identifiability and observability of compartmental models of the COVID-19 pandemic. *Annu. Rev. Control* **51**, 441–459. (doi:10.1016/j.arcontrol.2020.12.001)
15. Biala TA, Khaliq AQM. 2021 A fractional-order compartmental model for the spread of the COVID-19 pandemic. *Commun. Nonlinear Sci. Numer. Simul*. **98**, 105764. (doi:10.1016/j.cnsns.2021.105764)
16. Ndaïrou F, Area I, Bader G, Nieto JJ, Torres DFM. 2020 Mathematical modeling of COVID-19 transmission dynamics with a case study of Wuhan. *Chaos Solitons Fractals* **135**, 109846. (doi:10.1016/j.chaos.2020.110311)
17. Abou-Ismail A. 2020 Compartmental models of the COVID-19 pandemic for physicians and physician-scientists. *SN Compr. Clin. Med*. **2**, 852–858. (doi:10.1007/s42399-020-00330-z)
18. Batistela CM, Correa DPF, Bueno ÁM, Piqueira JRC. 2021 SIRSi compartmental model for COVID-19 pandemic with immunity loss. *Chaos Solitons Fractals* **142**, 110388. (doi:10.1016/j.chaos.2020.110388)
19. Odagaki T. 2023 New compartment model for COVID-19. *Sci. Rep*. **13**, 5409. (doi:10.1038/s41598-023-32159-6)

20. Domenech de Cellès M, Magpantay FMG, King AA, Rohani P. 2018 The impact of past vaccination coverage and immunity on pertussis resurgence. *Sci. Transl. Med*. **10**, j1748. (doi:10.1126/scitranslmed.aaj1748)
21. Kucharski AJ, Camacho A, Flasche S, Glover RE, Edmunds WJ, Funk S. 2015 Measuring the impact of Ebola control measures in Sierra Leone. *Proc. Natl Acad. Sci. USA* **112**, 14366–14371. (doi:10.1073/pnas.1508814112)
22. O'Dea EB, Drake JM. 2022 A semi-parametric, state-space compartmental model with time-dependent parameters for forecasting COVID-19 cases, hospitalizations and deaths. *J. R. Soc. Interface* **19**, 20210702. (doi:10.1098/rsif.2021.0702)
23. Drake JM, Handel A, Marty É, O'Dea EB, O'Sullivan T, Righi G, Tredennick AT. 2023 A data-driven semi-parametric model of SARS-CoV-2 transmission in the United States. *PLOS Comput. Biol*. **19**, e1011610. (doi:10.1371/journal.pcbi.1011610)
24. Gibson GC, Reich NG, Sheldon D. 2020 Real-time mechanistic bayesian forecasts of COVID-19 mortality. medRxiv. (doi:10.1101/2020.12.22.20248736)
25. Wu Y, Yang Y, Nishiura H, Saitoh M. 2018 Deep learning for epidemiological predictions. In *41st Int. ACM SIGIR Conf. on Research & Development in Information Retrieval*, pp. 1085–1088. New York, NY: ACM. (doi:10.1145/3209978.3210077)
26. Deng Y, Li H. 2023 Deep learning for few-shot white blood cell image classification and feature learning. *Comput. Methods Biomech. Biomed. Eng*. **11**, 2081–2091. (doi:10.1080/21681163.2023.2219341)
27. Lu L, Qian Y, Dong Y, Su H, Deng Y, Zeng Q, Li H. 2023 A systematic study of the performance of machine learning models on analyzing the association between semen quality and environmental pollutants. *Front. Phys*. **11**, 1259273. (doi:10.3389/fphy.2023.1259273)
28. Gao Q, Lin H, Qian J, Liu X, Cai S, Li H, Fan H, Zheng Z. 2023 A deep learning model for efficient end-to-end stratification of thrombotic risk in left atrial appendage. *Eng. Appl. Artif. Intell*. **126**, 107187. (doi:10.1016/j.engappai.2023.107187)
29. Vytla V, Ramakuri SK, Peddi A, Kalyan Srinivas K, Nithish Ragav N. 2021 Mathematical models for predicting COVID-19 pandemic: a review. *J. Phys*. **1797**, 012009. (doi:10.1088/1742-6596/1797/1/012009)
30. Karniadakis GE, Kevrekidis IG, Lu L, Perdikaris P, Wang S, Yang L. 2021 Physics-informed machine learning. *Nat. Rev. Phys*. **3**, 422–440. (doi:10.1038/s42254-021-00314-5)
31. Raissi M, Yazdani A, Karniadakis GE. 2020 Hidden fluid mechanics: learning velocity and pressure fields from flow visualizations. *Science* **367**, 1026–1030. (doi:10.1126/science.aaw4741)
32. Lu L, Dao M, Kumar P, Ramamurty U, Karniadakis GE, Suresh S. 2020 Extraction of mechanical properties of materials through deep learning from instrumented indentation. *Proc. Natl Acad. Sci. USA* **117**, 7052–7062. (doi:10.1073/pnas.1922210117)
33. Zou Z, Karniadakis GE. 2025 Multi-head physics-informed neural networks for learning functional priors and uncertainty quantification. *J. Comput. Phys*. **531**, 113947. (doi:10.1016/j.jcp.2025.113947)
34. Zou Z, Meng X, Karniadakis GE. 2024 Correcting model misspecification in physics-informed neural networks (PINNs). *J. Comput. Phys*. **505**, 112918. (doi:10.1016/j.jcp.2024.112918)
35. Zhang Z, Zou Z, Kuhl E, Karniadakis GE. 2024 Discovering a reaction–diffusion model for Alzheimer's disease by combining PINNs with symbolic regression. *Comput. Methods Appl. Mech. Eng*. **419**, 116647. (doi:10.1016/j.cma.2023.116647)
36. Zou Z, Meng X, Karniadakis GE. 2025 Uncertainty quantification for noisy inputs–outputs in physics-informed neural networks and neural operators. *Comput. Methods Appl. Mech. Eng*. **433**, 117479. (doi:10.1016/j.cma.2024.117479)
37. Zou Z, Meng X, Psaros AF, Karniadakis GE. 2024 NeuralUQ: a comprehensive library for uncertainty quantification in neural differential equations and operators. *SIAM Rev*. **66**, 161–190. (doi:10.1137/22m1518189)
38. Yang L, Meng X, Karniadakis GE. 2021 B-PINNs: Bayesian physics-informed neural networks for forward and inverse PDE problems with noisy data. *J. Comput. Phys*. **425**, 109913. (doi:10.1016/j.jcp.2020.109913)
39. Meng X, Li Z, Zhang D, Karniadakis GE. 2020 PPINN: parareal physics-informed neural network for time-dependent PDEs. *Comput. Methods Appl. Mech. Eng*. **370**, 113250. (doi:10.1016/j.cma.2020.113250)
40. Yu J, Lu L, Meng X, Karniadakis GE. 2022 Gradient-enhanced physics-informed neural networks for forward and inverse PDE problems. *Comput. Methods Appl. Mech. Eng*. **393**, 114823. (doi:10.1016/j.cma.2022.114823)
41. Lou Q, Meng X, Karniadakis GE. 2021 Physics-informed neural networks for solving forward and inverse flow problems via the Boltzmann-BGK formulation. *J. Comput. Phys*. **447**, 110676. (doi:10.1016/j.jcp.2021.110676)
42. Lu L, Meng X, Mao Z, Karniadakis GE. 2021 DeepXDE: a deep learning library for solving differential equations. *SIAM Rev*. **63**, 208–228. (doi:10.1137/19m1274067)
43. Toscano JD, Oommen V, Varghese AJ, Zou Z, Ahmadi Daryakenari N, Wu C, Karniadakis GE. 2025 From PINNs to PIKANs: recent advances in physics-informed machine learning. *Mach. Learn. Comput. Sci. Eng*. **1**, 15. (doi:10.1007/s44379-025-00015-1)
44. Raissi M, Perdikaris P, Karniadakis GE. 2019 Physics-informed neural networks: a deep learning framework for solving forward and inverse problems involving nonlinear partial differential equations. *J. Comput. Phys*. **378**, 686–707. (doi:10.1016/j.jcp.2018.10.045)
45. Zhang Q, Sampani K, Xu M, Cai S, Deng Y, Li H, Sun JK, Karniadakis GE. 2022 AOSLO-net: a deep learning-based method for automatic segmentation of retinal microaneurysms from adaptive optics scanning laser ophthalmoscopy images. *Transl. Vis. Sci. Technol*. **11**, 7. (doi:10.1167/tvst.11.8.7)
46. Chen Q, Ye Q, Zhang W, Li H, Zheng X. 2023 TGM-Nets: a deep learning framework for enhanced forecasting of tumor growth by integrating imaging and modeling. *Eng. Appl. Artif. Intell*. **126**, 106867. (doi:10.1016/j.engappai.2023.106867)
47. Daneker M, Cai S, Qian Y, Myzelev E, Kumbhat A, Li H, Lu L. 2024 Transfer learning on physics-informed neural networks for tracking the hemodynamics in the evolving false lumen of dissected aorta. *Nexus* **1**, 100016. (doi:10.1016/j.ynexs.2024.100016)
48. Chen Q, Li H, Zheng X. 2025 A deep neural network for operator learning enhanced by attention and gating mechanisms for long-time forecasting of tumor growth. *Eng. Comput*. **41**, 423–533. (doi:10.1007/s00366-024-02003-0)
49. Qian Y, Zhu G, Zhang Z, Modepalli S, Zheng Y, Zheng X, Frydman G, Li H. 2024 Coagulo-Net: enhancing the mathematical modeling of blood coagulation using physics-informed neural networks. *Neural Networks* **180**, 106732. (doi:10.1016/j.neunet.2024.106732)
50. Cai S, Li H, Zheng F, Kong F, Dao M, Karniadakis GE, Suresh S. 2021 Artificial intelligence velocimetry and microaneurysm-on-a-chip for three-dimensional analysis of blood flow in physiology and disease. *Proc. Natl Acad. Sci. USA* **118**, e2100697118. (doi:10.1073/pnas.2100697118)
51. Linka K, Schäfer A, Meng X, Zou Z, Karniadakis GE, Kuhl E. 2022 Bayesian physics informed neural networks for real-world nonlinear dynamical systems. *Comput. Methods Appl. Mech. Eng*. **402**, 115346. (doi:10.1016/j.cma.2022.115346)
52. Patel RG, Manickam I, Trask NA, Wood MA, Lee M, Tomas I, Cyr EC. 2022 Thermodynamically consistent physics-informed neural networks for hyperbolic systems. *J. Comput. Phys*. **449**, 110754. (doi:10.1016/j.jcp.2021.110754)

53. Patel RG, Trask NA, Wood MA, Cyr EC. 2021 A physics-informed operator regression framework for extracting data-driven continuum models. *Comput. Methods Appl. Mech. Eng*. **373**, 113500. (doi:10.1016/j.cma.2020.113500)
54. Zou Z, Wang Z, Karniadakis GE. 2025 Learning and discovering multiple solutions using physics-informed neural networks with random initialization and deep ensemble. (https://arxiv.org/abs/2503.06320)
55. Kharazmi E, Zhang Z, Karniadakis GEM. 2020 hp-VPINNs: variational physics-informed neural networks with domain decomposition. *Comput. Methods Appl. Mech. Eng*. **374**, 113547. (doi:10.1016/j.cma.2020.113547)
56. Malinzi J, Gwebu S, Motsa S. 2022 Determining COVID-19 dynamics using physics informed neural networks. *Axioms* **11**, 121. (doi:10.3390/axioms11030121)
57. Millevoi C. 2024 Physics-informed neural network in porous media and epidemiological applications. Doctoral thesis, Università degli Studi di Padova, Italy.
58. Olumoyin KD, Khaliq AQM, Furati KM. 2021 Data-driven deep-learning algorithm for asymptomatic COVID-19 model with varying mitigation measures and transmission rate. *Epidemiologia* **2**, 471–489. (doi:10.3390/epidemiologia2040033)
59. Torku T, Khaliq A, Rihan F. 2023 SEINN: a deep learning algorithm for the stochastic epidemic model. *Math. Biosci. Eng*. **20**, 16330–16361. (doi:10.3934/mbe.2023729)
60. Ning X, Guan J, Li XA, Wei Y, Chen F. 2023 Physics-informed neural networks integrating compartmental model for analyzing COVID-19 transmission dynamics. *Viruses* **15**, 1749. (doi:10.3390/v15081749)
61. Nguyen L, Raissi M, Seshaiyer P. 2022 Modeling, analysis and physics informed neural network approaches for studying the dynamics of COVID-19 involving human-human and human-pathogen interaction. *Comput. Math. Biophys*. **10**, 1–17. (doi:10.1515/cmb-2022-0001)
62. Berkhahn S, Ehrhardt M. 2022 A physics-informed neural network to model COVID-19 infection and hospitalization scenarios. *Adv. Contin. Discret. Model*. **2022**, 61. (doi:10.1186/s13662-022-03733-5)
63. Shaier S, Raissi M, Seshaiyer P. 2021 Data-driven approaches for predicting spread of infectious diseases through DINNs: disease informed neural networks. (https://arxiv.org/abs/2110.05445)
64. Ogueda-Oliva AG, Martínez-Salinas EJ, Arunachalam V, Seshaiyer P. 2023 Machine learning for predicting the dynamics of infectious diseases during travel through physics informed neural networks. *J. Mach. Learn. Model. Comput*. **4**, 17–35. (doi:10.1615/jmachlearnmodelcomput.2023047213)
65. Bracher J, Ray EL, Gneiting T, Reich NG. 2021 Evaluating epidemic forecasts in an interval format. *PLoS Comput. Biol*. **17**, e1008618. (doi:10.1371/journal.pcbi.1008618)
66. Ray EL *et al*. 2020 Ensemble forecasts of coronavirus disease 2019 (COVID-19) in the US. *medRxiv.* (doi:10.1101/2020.08.19.20177493)
67. Dong E, Du H, Gardner L. 2020 An interactive web-based dashboard to track COVID-19 in real time. *Lancet Infect. Dis*. **20**, 533–534. (doi:10.1016/s1473-3099(20)30120-1)
68. Reinhart A *et al*. 2021 An open repository of real-time COVID-19 indicators. *Proc. Natl Acad. Sci. USA* **118**, e2111452118. (doi:10.1073/pnas.2111452118)
69. Google LLC. 2020 *Google COVID-19 community mobility reports*. See https://www.google.com/covid19/mobility/.
70. Kingma DP, Ba J. 2014 Adam: a method for stochastic optimization. (https://arxiv.org/abs/1412.6980)
71. Hyndman RJ, Koehler AB. 2006 Another look at measures of forecast accuracy. *Int. J. Forecast*. **22**, 679–688. (doi:10.1016/j.ijforecast.2006.03.001)
72. Grossberg S. 2013 Recurrent neural networks. *Scholarpedia* **8**, 1888. (doi:10.4249/scholarpedia.1888)
73. Hochreiter S, Schmidhuber J. 1997 Long short-term memory. *Neural Comput*. **9**, 1735–1780. (doi:10.1162/neco.1997.9.8.1735)
74. Dey R, Salem FM. 2017 Gate-variants of gated recurrent unit (GRU) neural networks. In *2017 IEEE 60th Int. Midwest Symp. on Circuits and Systems (MWSCAS)*, *Boston, MA, USA, 6–9 August 2017*, pp. 1597–1600. (doi:10.1109/MWSCAS.2017.8053243)
75. Ghojogh B, Ghodsi A. 2020 Attention mechanism, transformers, BERT, and GPT: tutorial and survey. OSF. (doi:10.31219/osf.io/m6gcn)
76. Rodríguez A, Cui J, Ramakrishnan N, Adhikari B, Prakash BA. 2023 EINNs: epidemiologically-informed neural networks. *Proc. AAAI Conf. Artif. Intell*. **37**, 14453–14460. (doi:10.1609/aaai.v37i12.26690)
77. Barmparis GD, Tsironis GP. 2022 Physics-informed machine learning for the COVID-19 pandemic: adherence to social distancing and short-term predictions for eight countries. *Quant. Biol*. **10**, 139–149. (doi:10.15302/j-qb-022-0281)
78. Madden WG, Jin W, Lopman B, Zufle A, Dalziel B, Metcalf CJE, Grenfell BT, Lau MSY. 2024 Neural networks for endemic measles dynamics: comparative analysis and integration with mechanistic models. *medRxiv.* (doi:10.1101/2024.05.28.24307979)
79. Ying Q. 2025 Data & code. Physics-informed deep learning for infectious disease forecasting. Zenodo. (doi:10.5281/zenodo.17196508)
80. Christie A, Brooks JT, Hicks LA, Sauber-Schatz EK, Yoder JS, Honein MA, CDC COVID-19 and Response Team. 2021 Guidance for implementing COVID-19 prevention strategies in the context of varying community transmission levels and vaccination coverage. *MMWR Morb. Mortal. Wkly Rep*. **70**, 1044–1047. (doi:10.15585/mmwr.mm7030e2)