International Journal of Data Science and Analytics (2023) 15:247–266
https://doi.org/10.1007/s41060-021-00295-9

REGULAR PAPER

A spatiotemporal machine learning approach to forecasting COVID-19 incidence at the county level in the USA

Benjamin Lucas[1] · **Behzad Vahedi**[1] · **Morteza Karimzadeh**[1]

Received: 31 August 2021 / Accepted: 19 November 2021 / Published online: 15 January 2022

Abstract
With COVID-19 affecting every country globally and changing everyday life, the ability to forecast the spread of the disease is more important than any previous epidemic. The conventional methods of disease-spread modeling, compartmental models, are based on the assumption of spatiotemporal homogeneity of the spread of the virus, which may cause forecasting to underperform, especially at high spatial resolutions. In this paper, we approach the forecasting task with an alternative technique—spatiotemporal machine learning. We present COVID-LSTM, a data-driven model based on a long short-term memory deep learning architecture for forecasting COVID-19 incidence at the county level in the USA. We use the weekly number of new positive cases as temporal input, and hand-engineered spatial features from Facebook movement and connectedness datasets to capture the spread of the disease in time and space. COVID-LSTM outperforms the COVID-19 Forecast Hub's Ensemble model (COVIDhub-ensemble) on our 17-week evaluation period, making it the first model to be more accurate than the COVIDhub-ensemble over one or more forecast periods. Over the 4-week forecast horizon, our model is on average 50 cases per county more accurate than the COVIDhub-ensemble. We highlight that the underutilization of data-driven forecasting of disease spread prior to COVID-19 is likely due to the lack of sufficient data available for previous diseases, in addition to the recent advances in machine learning methods for spatiotemporal forecasting. We discuss the impediments to the wider uptake of data-driven forecasting, and whether it is likely that more deep learning-based models will be used in the future.

Keywords COVID-19 · LSTM · Spatiotemporal machine learning · Geospatial artificial intelligence

1 Introduction

Since the first known case of COVID-19 in December 2019, the disease has grown into a pandemic of unprecedented scale, significantly impacting modern life in the twenty-first century. There have been over 216 million confirmed COVID-19 infections globally as of August 2021, with the USA being particularly hard hit, accounting for over one-third of all infections and 636,000 deaths [25]. As a result, forecasting the spread of the disease within the USA has been a significant focus of the Centers for Disease Control and Prevention (CDC) and National Institutes of Health (NIH).

The nationwide spread of COVID-19 has necessitated continual adaptations in planning and response decisions. While substantial uncertainty exists surrounding the continuing spread of COVID-19, a robust forecast can be used to inform policy, targeted interventions, and mitigation strategies. During the pandemic, forecasts have been used to allocate medical resources that are in short supply (e.g., ventilators, personal protective equipment, gowns, sanitizer) to the areas with high COVID risk [17,18,33]. Moreover, they have influenced the assignment of travel nurses, a group of qualified nurses who are not employed at specific locations, but rather, have multiple short-term appointments at hospitals located anywhere in the USA based on demand [3]. Forecasts have also been shown to inform response and mitigation strategies [63,96] and help identify preferred locations for vaccine efficacy trials [22]. The decision of many government authorities to 'lockdown' the population has been

✉ Morteza Karimzadeh
karimzadeh@colorado.edu

Benjamin Lucas
benjamin.lucas@colorado.edu

Behzad Vahedi
behzad@colorado.edu

[1] Department of Geography, University of Colorado Boulder, Boulder, CO, USA

based upon the combination of the current incidence rate and short-term forecasting of COVID-19 incidence [55,68]. The importance of accurate forecasting is further underlined by the fact that tens of teams from a variety of academic areas and industries have shifted their research focus to COVID-19 forecasting since March 2020 [19].

In March 2020, the University of Massachusetts Amherst created the COVID-19 Forecast Hub [19,78], and since then, has been publishing weekly forecasts of COVID-19 incidence and mortalities at the scales of national, state, and county level. In this work, we focus on forecasting incidence at the highest spatial resolution for which validation data is widely available, i.e., county level, as the most important scale for planning, resource allocation, and medical equipment distribution. Additionally, adopting preventative or mitigation strategies such as business restrictions or mask mandates are largely implemented locally, based on county-level incidence and prevalence.

The majority of the forecasting surrounding the spread of COVID-19 has been produced using compartmental models from epidemiology, in particular SEIR models [31,41,84,107]. These models divide the population into compartments—such as Susceptible to the virus (S), Exposed to the virus (E), Infected with the virus (I), or Recovered from the virus (R)—and then, use the characteristics of the virus and population to estimate the flow of proportions between these categories in order to forecast the spread and/or duration of an epidemic. These models have produced reasonable forecasts for many decades and different epidemics [48,79,94], but their main strength lies in providing a framework for characterizing the reproduction rate of a disease, i.e., the expected number of secondary cases produced by a single (typical) infection in a completely susceptible population. When it comes to forecasting, their performance is undermined by the underpinning assumption of the spatiotemporal homogeneity of the spread of the virus [1,37].

In this paper, we propose an alternate approach to incidence forecasting by implementing a data-driven framework based on a spatiotemporal deep learning architecture, which we call COVID-LSTM, to reflect our adoption of long short-term memory (LSTM) network architecture. Our method utilizes human movement and county connectedness data, published by Facebook and acquired from mobile devices carrying the Facebook app, to derive features quantifying the spread of the virus between counties. We integrate these hand-engineered spatiotemporal features into a LSTM deep learning model for multivariate time series. Our results demonstrate that this method is the first to be more accurate on average than the COVID-19 Forecast Hub's Ensemble model (COVIDhub-ensemble) at predicting COVID-19 incidence over multiple forecast horizons.

The main contributions of this paper can be summarized as follows:

1. COVID-LSTM: A novel framework for integrating spatial features and temporal incidence data using spatiotemporal deep learning for disease-spread forecasting;
2. The first model to produce more accurate forecasts, on average, than the COVIDhub-ensemble at multiple forecast horizons;
3. A novel use of Facebook's Social Connectedness Index and Movement Range datasets to define the strength of spatial connections between counties in the USA, and the amount of inter-county and intra-county population movement; and,
4. Openly providing code and data for reproducibility, wider implementation, and future research.

The remainder of this paper is set out as follows: in Sect. 2, we discuss the COVID-19 Forecast Hub and present both compartmental models and other existing data-driven approaches; in Sect. 3, we discuss the case data and explain our use of Facebook's Social Connectedness Index and Movement Range datasets to derive spatial features; in Sect. 4, we present COVID-LSTM in detail; in Sect. 5, we present our experiments and comparisons with the leading models currently used by the CDC and NIH, and discuss some of the model parameters; in Sect. 6, we comment on our forecasts, the methods, and the future of COVID-19 forecasting; and finally, we draw conclusions in Sect. 7.

2 COVID-19 forecasting and related work

Severe Acute Respiratory Syndrome Coronavirus 2 (SARS-CoV-2) is the highly contagious virus causing the COVID-19 respiratory illness, and responsible for the COVID-19 pandemic. It is a relative of SARS-CoV-1, the disease that was responsible for the SARS epidemic in 2003–2004, and which initially provided valuable insight into the potential spread of this disease [97].

This section gives an overview of COVID-19 disease-spread modeling, including the role of both compartmental models and machine learning models. We discuss some of the highest-performing models used to forecast the spread of COVID-19, with an emphasis on models addressing the same problem we are—forecasting county-level incidence in the USA. We also discuss the COVID-19 Forecast Hub and the resulting COVIDhub-baseline and COVIDhub-ensemble models, as used by CDC, which we later use for comparison in our experiments.

2.1 COVID-19 modeling

The international impact of COVID-19 has led to significant research efforts being invested into modeling various aspects of the disease and policy responses. In a compre-

hensive review of COVID-19 modeling [13], the authors identify over 200k published articles on COVID-19, with approximately 22k of them specifically related to modeling. COVID-19 modeling publications cover a large range of aspects of the pandemic including disease spread [6,35,73,83,102,107], transmission dynamics [45,57,98,106], diagnosis [20,70], contact tracing [46,49], medical treatment [8,57], non-pharmaceutical interventions [9,23,28,30], and socioeconomic influence and impact [24,62,76,95].

The work dedicated to modeling the spread of the disease can be categorized into two strategies—compartmental models and data-driven models—which approach the problem from different directions. Compartmental models are based on characteristics of the disease while data-driven models learn the pattern and rate of spread through previously observed data.

2.2 Compartmental models

Compartmental models are based on a conventional mathematical modeling technique for predicting the spread of infectious diseases. They stratify the population into compartments depending on their relationship with the disease in question. The basis for each of these models is the SIR model, which was developed in the early twentieth century, and assigns members of the population into one of three categories: *S*usceptible to the virus, *I*nfected with the virus, or *R*ecovered from the virus [50]. The flow of members of the population from one state to the next is modeled by the following set of differential equations:

$$\begin{aligned} \frac{\mathrm{d}S}{\mathrm{d}t} &= -\frac{\beta I S}{N} \\ \frac{\mathrm{d}I}{\mathrm{d}t} &= \frac{\beta I S}{N} - \gamma I \\ \frac{\mathrm{d}R}{\mathrm{d}t} &= \gamma I \end{aligned} \tag{1}$$

where S is the number of subjects in the population susceptible, I is the number of subjects currently infected, R is the number of subjects recovered, and N is the total population size. Parameters β and γ are based on characteristics of the disease in the population, and calculate the proportion of the susceptible population that are becoming infected with the disease and the rate at which people recover from the disease, respectively. The ratio of β to γ, referred to as the reproduction rate or R_0, is the expected number of new infections for each individual infection in a completely susceptible population. An R_0 of greater than 1 represents a disease that is growing in the community, while an R_0 of less than 1 represents a disease that is declining in incidence.

The classical SIR model can be extended by including additional compartments such as exposed (resulting in an S*E*IR model) and/or deceased (SEIR*D*), or other compartments specific to the disease [43]. For COVID-19 modeling, an exposed category is useful to represent the population who are in proximity to someone with the disease, but are not yet showing symptoms due to the incubation period of the virus [41,47]. In [102], the authors categorize the community into 6 classes: Susceptible, Exposed, Infected, Quarantined, Insusceptible, and Recovered, resulting in a SEIQPR model. The categories of this model do not follow a direct linear sequence either, i.e., an exposed individual may progress to being infected, or they may pass to being quarantined, allowing for differing regulations between locations. In [107], the authors propose a SuEIR model, including a category for unreported or unconfirmed COVID-19 cases, as the number of cases reported in the USA is believed to have largely been under-reported throughout the pandemic. Other methods have added free parameters to compartmental models to account for government policy such as social distancing [71] and travel restrictions [16].

In [38], the authors build upon the traditional SEIRD model to facilitate real-time COVID-19 forecasting. Their model, named the Mechanistic Bayesian Model (UMass-MechBayes), uses a nonparametric model of the transmission rate (β_t) against time. This allows for the transmission rate to increase or decrease for each measurement period. A similar approach is presented in [2], where researchers use machine learning models to accurately and dynamically quantify the transitions between model compartments. Both this model, named the COVID-19 Public Forecast model (Google_Harvard-CPF), and the UMass-MechBayes model have been identified as producing highly accurate county-level forecasts in the USA [19], and therefore, we use them as comparison models in Sect. 5.

2.3 Data-driven approaches

As the COVID-19 epidemic developed into a pandemic during 2020, it also provided data at a scale unprecedented in epidemiology. Within the USA, incidence and death data has been recorded at the county level, and made freely available on a daily basis. Data of this magnitude has provided an opportunity for re-examining the conventional forecasting methods, and devising data-driven forecasting.

The wealth of data collected during the COVID-19 pandemic (as a result of its importance as well as the disease prevalence) provides an opportunity for researchers to use autoregressive processes and machine learning as alternative approaches to compartmental models. Machine learning is data-driven, meaning that the models identify, and learn from, underlying spatiotemporal trends in the data. Alternately, compartmental modeling is based on the assumptions of the spatiotemporal homogeneity and the homogeneity of the population [37]—assumptions that may be incorrect in

the case of COVID-19. While models have been extended to include free parameters to account for demographic factors [14,61], dependence of transmission rates on time [53], and metapopulation structure [65,100], this often ends up with a large number of parameters that must be calibrated for a given disease and population, which can introduce errors in incidence forecasting.

One potential reason that data-driven methods are not common–relative to compartmental models–in the history of disease forecasting, is the potential sparsity of the data. For example, the World Health Organization states that around 8,000 people globally were infected with the SARS-CoV-1 outbreak in 2003, and only 8 of these were in the USA [74]. It is reasonable to assume that machine learning methods, which traditionally improve in performance proportionally to the quantity of data available [40,90], might have underperformed in forecasting this outbreak. However, with millions of reported infections worldwide, data scarcity is not an issue for researchers using data-driven approaches to forecasting the COVID-19 pandemic.

With the global-scale disrupting effects of the COVID-19 pandemic, several research groups leveraging applied Artificial Intelligence have diverted their attention to COVID-19 forecasting. Additionally, recent research in machine learning for sparse data and transfer learning should further facilitate its adoption for disease forecasting, even for smaller epidemics [88].

In [81], researchers present an autoregressive time series model (CMU-TimeSeries) that uses time series of both incidence and death data to make forecasts at high spatial resolutions, such as the county level.

DeepCOVID [83] is recognized as the first deep learning-based COVID-19 forecasting model published for US data. It uses a multi-layer perceptron to produce state-level forecasts using cases, deaths, and hospitalizations as a multivariate input. As this was the first data-driven model published, it is notable that the resulting forecasts are highly comparable with the state-of-the-art compartmental models at the state level [19].

Subsequent to DeepCOVID, various studies have treated COVID-19 forecasting as a time series forecasting problem, and addressed it with deep learning. The most popular is LSTM networks, which is also the architecture that we also utilize in this work, as they are the state of the art in learning patterns in temporal data. LSTM networks are implemented for national-level COVID-19 forecasting in [6,15,35,73,104,107]. It is likely due to data sparsity, irregularity, and the required pre-processing, that many researchers have not used LSTM networks at smaller scales [83]. One notable model, however, is the Neural Relational Autoregressive Model by Facebook AI Research (FAIR-NRAR) [60]. This model uses temporal series of incidence, deaths, and other covariates quantifying the mobility of the population as input to a recurrent neural network. This model is the most similar in structure to ours and is therefore included as a comparison model in our experiments in Sect. 5.

2.4 COVID-19 Forecast Hub

The COVID-19 Forecast Hub [19,78], created by the University of Massachusetts Amherst, is a repository of COVID-19 forecasts from various research groups across the USA. Each week, groups submit forecasts for the numbers of new cases, hospitalizations, and deaths in future days, weeks, and months at the national, state, and county level in the USA. The repository hosts over 100 million rows of data that are openly accessible.

The Forecast Hub works closely with the CDC and passes on forecasts for use in government communications. The two common forecast models published by the CDC are the COVIDhub-baseline model and the COVIDhub-ensemble model; these are also the two models that we use for comparison in our experiments in Sect. 5. The COVIDhub-baseline model's forecast is a neutral reference model with a predictive median equal to that observed over the same time period immediately prior. In our experiments, this means that the predicted number of COVID-19 incidence for a given county during any future week will be equal to the number of reported infections in that county during the current week, i.e., persistence. The COVIDhub-ensemble model's forecast is a collaboration between the CDC, 21 research groups, five private industry groups, and two other government groups. The forecast value is the median prediction out of all eligible models that are submitted through the COVID-19 Forecast Hub for a given forecast date, hence the size of the ensemble varies by week and location. The number of individual models in the ensemble ranges from single figures during April 2020 to 49 during December 2020, however the forecasts for some locations include fewer models as not all model submissions contain predictions for all locations. The COVIDhub-ensemble's forecast is rarely the most accurate for an individual point prediction, but is shown to be significantly better on average than any single model across all forecast horizons [19].

Forecast Hub predictions are published for various time horizons. The focus of this work, and the majority of the model submissions, is on the more reliable 'short-term' forecasts, which include forecasts for 1, 2, 3, and 4 weeks into the future. One important note is that the forecasts are made for epi-weeks, meaning that the forecast value for the 2-week horizon is equal to the numbers anticipated to occur during the week (7 days) that is 2 weeks in the future, rather than a cumulative forecast for the coming 2 weeks (14 days). Epi-weeks run Sunday through Saturday, as defined by the CDC, and are common practice in epidemiology.

Both a point prediction and a set of quantiles are necessary for each Forecast Hub submission to enable the creation of prediction intervals. For forecasting incidence at the county level, the published quantiles are 0.025, 0.1, 0.25, 0.5, 0.75, 0.9, and 0.975, and therefore, we have included these quantiles in our results (Sect. 5.4).

3 Data

The task we aim to address is to predict COVID-19 incidence at the county level in the contiguous US over 1-, 2-, 3-, and 4-week forecast horizons. Our data has 10 temporal input features per instance: 2 derived from the raw number of cases; 6 features derived from Facebook datasets representing human movement and inter-county connectedness; and 2 weather features. Each feature x is included with a temporal lag of n weeks: $[x_t, x_{t-1}, x_{t-2}, \ldots, x_{t-n}]$.

The input features are shown in Table 1 and discussed further in the following subsections.

3.1 Reported cases

The raw data are downloaded from Johns Hopkins University's Center for Systems Science and Engineering [25] as the cumulative number of confirmed COVID-19 infections per county per day, i.e., incidence. We download data starting April 1, 2020, through February 20, 2021. The period used for our evaluation begins on Saturday October 31, 2020, through February 20, 2021. This evaluation period was chosen as it covers three different phases of the pandemic—the sharp increase in late 2020, the peak around the December–January holiday period, and the general decline beginning in the new year. This is reflected in the national case numbers shown in Fig. 1. All data between the start date and the given evaluation (i.e., forecasting) date is used for training the forecast model, and all evaluation is performed on unseen (held out) data.

Table 1 The features used in COVID-LSTM

Source	Feature
Johns Hopkins University CSSE [25]	New weekly COVID-19 incidence
	Monthly mean cumulative COVID-19 incidence
Facebook [4,42]	Stay put index
	Rate of weekly change of stay put index
	Change in movement index
	Rate of weekly change in change in movement index
	Weekly change in social proximity to cases
	Monthly mean social proximity to cases
Weather	Average minimum temperature
	Average maximum temperature

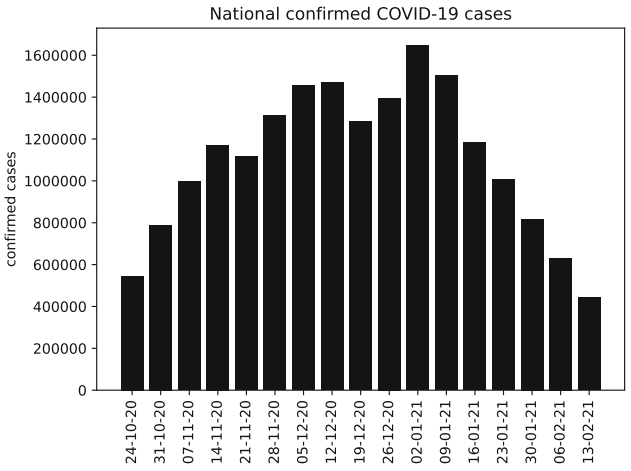

Fig. 1 The observed COVID-19 incidence for our 17-week evaluation period at the national level. This period covers three periods of differential growth of the virus, a rapid increase in November and Early December, steady growth across the new year period, and a decline in growth from mid-January to February

To process this data into a weekly dataframe, we first take a 7-day rolling average to smooth the irregularities caused by inconsistent reporting, especially on weekends and holidays. From here, we calculate 2 input variables aligned with the epi-week-based forecast dates: (1) the mean number of cumulative cases during a given week, and (2) the increase in new cases reported during a given week, i.e., incidence. This second variable is also the target variable for our model.

Although we have applied smoothing to the raw data, there are still a number of irregularities present. To combat this, if a county reports a weekly increase of 0 cases, any input instance containing that value is excluded from the training data. This assists the model to learn underlying temporal trends without the noise of inconsistent reporting.

We note that while our model is trained on a target of rolling average values, the final evaluation is compared to the raw number of reported cases in a given week (we discuss this further in Sect. 5.5).

Our data contains the counties of 48 states in the contiguous US. However, as we are using a time lag of length l, each county contributes multiple instances to the training set. That is, for a single county's case data: $x_1, x_2, x_3, \ldots, x_t$, where we are trying to forecast x_{t+1}, the training data will include as individual instances:

$$x_1, x_2, \ldots, x_{1+l} \rightarrow x_{2+l}$$
$$x_2, x_3, \ldots, x_{2+l} \rightarrow x_{3+l}$$
$$x_3, x_4, \ldots, x_{3+l} \rightarrow x_{4+l}$$
$$\ldots$$

$$x_{t-1-l}, x_{t-l}, \ldots, x_{t-1} \rightarrow x_t$$

where the values to the right of the arrows are the target of the respective training instances. The test instance will be:

$$x_{t-l}, x_{t-l+1}, \ldots, x_t \rightarrow x_{t+1}.$$

This means that the model is essentially learning patterns in multivariate temporal series of length $l + 1$, for each county and time step.

The results presented in Sect. 5 use input data with a temporal lag of 9. We investigate the use of other temporal lags in Sect. 5.7.

3.2 Facebook-derived spatial features

In order to account for *intra-county* human movement patterns, we included movement variables derived from the Facebook Movement Range dataset [42]. This anonymized, privacy-protecting dataset is generated by Facebook, and derived from mobile devices carrying the Facebook app, i.e., by tracking the location of users' log-ins over time to measure the flow of the population. Within this dataset, there are two metrics, called (1) Change in Movement and (2) Stay Put. Change in Movement is a measure of the relative change in aggregated movement within a county compared to a baseline of the month of February 2020, which is the month prior to the first cases of COVID-19 being recorded in the USA. Stay Put is a measure of the proportion of a county's population that has stayed within a small radius for a 24 h period. The Movement Range dataset is published daily, and we have calculated the change in movement and the rate of this change for the appropriate epi-week.

In order to account for *inter-county* spread of the disease, we have also incorporated an index called Social Proximity to Cases (SPC) [58], which is a COVID-19-specific metric incorporating Facebook's Social Connectedness Index (SCI) [4]. The SCI is a dataset published by Facebook that uses the general home location of Facebook friends to quantify the connectedness between those two administrative units (in our case, the locations are US counties, however SCI is not available at this scale in all locations internationally). This means that two counties can be connected to one another without being spatially adjacent. The Social Connectedness (SC) between two counties is calculated as the ratio of Facebook-friendships between users in those counties to the total number of possible Facebook-friendships between those counties, i.e., SC represents the probability that any two Facebook users in different US counties are friends on Facebook, given their respective locations. The SC between counties i and j would be:

$$\mathrm{SC}_{i,j} = \frac{\mathrm{FB\_friendships}_{i,j}}{\mathrm{FB\_users}_i \times \mathrm{FB\_users}_j}. \tag{2}$$

The published value for SCI equals the value for SC scaled to a range of between 1 and 1,000,000,000 and rounded to the nearest integer.

In [58], the authors found that SCI was highly correlated with the early spread of COVID-19 cases at the county level in the USA. Their work defined the SPC metric to quantify the likely exposure of the population of a given county to positive cases from connected counties. It is calculated as the weighted sum of the positive cases in the connected counties, where the weights are the social connectedness between counties. Specifically, the SPC for county i at time t is:

$$\mathrm{SPC}_{i,t} = \sum_{j \in C} \mathrm{Incidence\_rate}_{j,t} \times \frac{\mathrm{SC}_{i,j}}{\sum_{h \in C} \mathrm{SC}_{i,h}} \tag{3}$$

where $\mathrm{Incidence\_rate}_{j,t}$ is the number of positive COVID-19 cases per 10,000 people in county j at time t, and C is the set of all counties socially connected to county i (in our case, C is the set of all other counties in the USA). SCI is a static index produced annually; however, as SPC is weighted by the weekly COVID-19 incidence rate, SPC is a dynamic measure of proximity to cases.

Together, these features help capture the heterogeneous spatial spread of COVID-19, both intra- and inter-county, for a given week [93]. In order for our model to learn the spatiotemporal spread, we create temporal series of the weekly averages of each of these variables. We also include a series of the rate of change of the variables (as the slope of a linear regression model) over that week, to account for average values that have high or low variance.

3.3 Weather features

There is evidence to suggest a correlation between climate and the spread of COVID-19 [7,67,92], and as such, we have included temperature as a feature in COVID-LSTM. The features included are the weekly average minimum temperature and weekly average maximum temperature, per county. The averaging is performed only for the populated areas of each county based on the US Census's Populated Places. As above, these variables are used as a temporal sequence of l+1 weeks in length.

4 COVID-LSTM

Our COVID-LSTM model is an ensemble regression model based on a stacked long short-term memory (LSTM) deep

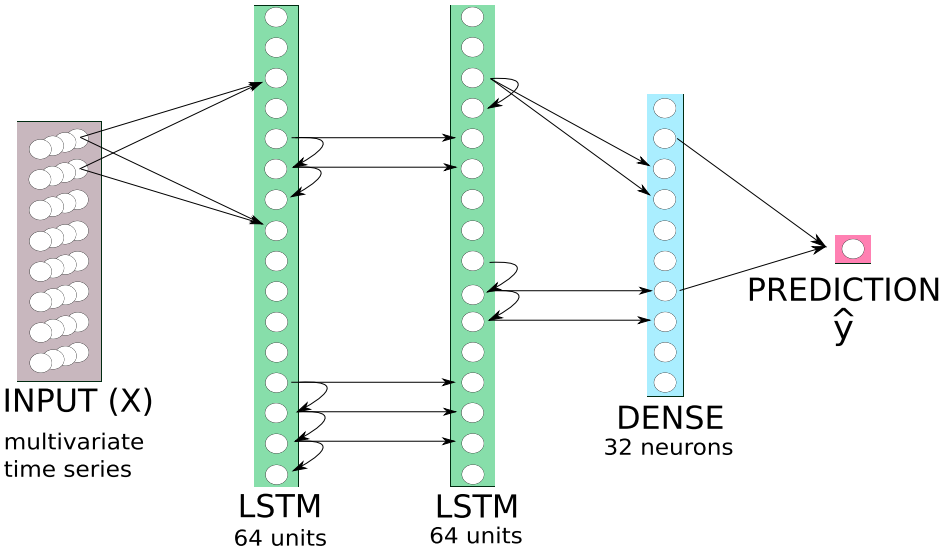

Fig. 2 The architecture of one constituent models of COVID-LSTM, which is an ensemble of 10 identical architectures

learning architecture [44], which is a type of recurrent neural network (RNN). LSTM networks incorporate multiple loop connections, which enable information to be retained and flow from one point of the network to the next. As a result, they have found much success in sequential and list-based data, such as speech recognition, image captioning, and time series forecasting, as in our application [36,39,85]. In this section, we will outline the COVID-LSTM architecture in depth, and discuss its hyperparameters.

4.1 Architecture

COVID-LSTM is an ensemble of 10 LSTM-based networks, all identical in architecture, but each initialized randomly and trained separately. Each network takes a 10-channel multivariate time series as input, has 3 hidden layers, and a single node as output. The first two hidden layers are LSTM layers, each with 64 units. The final hidden layer is a dense (fully connected) layer of size 32. The architecture is shown in Fig. 2.

A single LSTM unit is comprised of three 'gates,' which regulate the information flow through the cell and retain past information learnt from the sequence of input data. The input gate determines new information to store from the current state, the forget gate determines what information to discard from the previous state, while the output gate determines the next hidden state or the output value of the unit. The functioning of these gates is the reason that LSTM networks outperform 'standard' RNNs in almost all tasks [89].

4.2 Hyperparameters

Each individual model was trained for 15 epochs with the final model chosen as the one with the lowest training error out of the 15 epochs. We used an Adam optimizer [51] with a learning rate of 0.001 to minimize the mean squared error loss. The experiments were performed in Tensorflow 2.5.0.

4.3 COVID-LSTM ensemble model

We designed COVID-LSTM as an ensemble model to reduce the variability of the results of a single run. As the training of deep learning methods is a stochastic process, it may lead to different fitted models for each training; and therefore, models fit separately on the same data may generate variable results. Ensembling generally reduces this variability, while also improving overall performance [29,66,72,103]. Furthermore, the uncertainty in epidemiological modeling and data has resulted in ensemble models consistently outperforming individual models [48,79,94]. We define our ensemble prediction as the median predicted value of all of the constituent model predictions, providing enhanced stability and consistency of our forecasts. The curators of the COVIDhub-ensemble model investigated more sophisticated methods of combining models to form an ensemble, such as using a trained or untrained weighted mean of the constituent models, and ultimately found that taking the simple median generated equally competitive results. [10].

5 Experiments

In this section, we present the results of forecasting county-level reported COVID-19 incidence using COVID-LSTM. We compare our forecasts to 6 comparison models, including the one used by the CDC to inform decisions at a federal level, and present our results in terms of both mean absolute error and mean absolute percentage error. All models are assessed across 1-, 2-, 3- and 4-week forecast horizons.

After providing comparison with other leading models, we investigate the uncertainty in our forecasted values using quantile regression, and turn to look at our forecast errors and the outliers still present in the data. Finally, we aggregate our output to investigate trends in our forecast errors at a national level.

All experiments were run on a machine with an Intel Xenon processor and Ubuntu version 20.04. Our models were trained using a NVIDIA GeForce RTX 3080 graphics card with 10GB of RAM.

5.1 Comparison metrics

We compare our forecasts to those of different models by mean average error (MAE) and mean average percentage error (MAPE). MAE is the metric used by the COVID-19 Forecast Hub to compare the performance of forecasts, and therefore, we have chosen to align with this decision, as the Hub is the primary reference point for COVID-19 forecasting in the USA.

It is calculated as:

$$\text{MAE}_t = \frac{\sum_{j \in C} |\widehat{y_{j,t}} - y_{j,t}|}{|C|} \tag{4}$$

where C is all of the counties included in the forecast, $\widehat{y_{j,t}}$ is the forecasted value for county j in week t and $y_{j,t}$ is the true value.

MAPE calculates the error in the forecast as a percentage of the ground truth value. In our application, this means that the error in the forecast of each county's incidence contributes equally to the evaluation metric. It is calculated as:

$$\text{MAPE}_t = \frac{100}{|C|} \sum_{j \in C} \frac{|\widehat{y_{j,t}} - y_{j,t}|}{y_{j,t}} \tag{5}$$

where C is all of the counties included in the forecast, $\widehat{y_{j,t}}$ is the forecasted value for county j in week t and $y_{j,t}$ is the true value.

As a percentage error, there are certain instances where MAPE is undefined or does not make sense [56]. For example, when forecasting COVID-19 at the county level in the USA, many small counties record zero weekly incidence, meaning that any forecast greater than zero will have an infinite MAPE. For this reason, we have reported MAPE for only the 50 most populous counties in the USA.

5.2 Comparison models

Our experiments presented in Sect. 5.3 compare the forecast of COVID-LSTM to 6 other leading forecasting models: 2 models that have been identified by the Forecast Hub as consistently high performing at the county level [19]; 2 models that share notable similarities to COVID-LSTM; and, 2 models produced by the Forecast Hub, one of which has been the basis of reporting and projection by the US CDC.

High-performing individual models
The University of Massachusetts Mechanistic Bayesian model (UMass-MechBayes) [38][1] and Google and Harvard University's COVID-19 Public Forecast model (Google_Harvard-CPF) [2] are identified as consistently high-performing forecasts of COVID-19 incidence at the county level in the USA [19,78]. UMass-MechBayes is a SEIRD model modified to include nonparametric estimates of varying transmission rates and nonparametric modeling of case discrepancies to account for testing and reporting issues. Google_Harvard-CPF is a machine learning-infused SEIR model that emphasizes explainability. It uses an encoder model to extract information from spatial and temporal covariates to update the transitions between compartments in the model.

Similar models The Carnegie Mellon Delphi Group's Time Series model (CMU-TimeSeries) [81][2] and Facebook Artificial Intelligence Research's Neural Relational Autoregressive model (FAIR-NRAR) [60] share similarities with our technique as they frame the problem as a data-driven forecasting problem. CMU-TimeSeries uses incidence and deaths as inputs to an autoregressive time series model for forecasting at the county level. FAIR-NRAR is based on a recurrent neural network architecture and adds covariates to represent regional sociodemographics, the population mobility, and local policies.

Forecast Hub models The final two models are those produced by the Forecast Hub—COVIDhub-baseline and COVIDhub-ensemble, as discussed in Sect. 2.4. The COVIDhub-baseline represents persistence, i.e., the following week will have the same number of incidence as the previous week, and is developed as a universal benchmark in the USA, while the COVIDhub-ensemble is the best county-level forecasting model (through ensembling multiple models created by leading universities and tech companies), and is used by the US CDC and other government departments in decision-making.

[1] We note that the UMass-MechBayes model only publishes forecasts for 485–490 US counties per week and the results shown in the experiments section are based on these predictions only.

[2] We note that the CMU-TimeSeries model only publishes forecasts for 199 US counties per week and the results shown in the experiments section are based on these predictions only.

Table 2 Average weekly mean absolute error of COVID-19 cases per county over the whole evaluation period

	Forecast horizon			
	1 week	2 weeks	3 weeks	4 weeks
CMU-TimeSeries	810.90	1235.14	1531.42	1706.18
UMass-MechBayes	457.82	711.11	962.01	1307.32
Google_Harvard-CPF	136.07	200.91	259.87	321.32
FAIR-NRAR	98.31	156.96	192.08	213.10
COVIDhub-baseline	91.13	139.55	180.14	214.62
COVIDhub-ensemble	**78.77**	121.87	155.40	183.32
COVID-LSTM	87.29	**110.97**	**121.46**	**133.22**

Model with lowest error shown in bold

5.3 Forecasting COVID-19 incidence

The average weekly MAE of the forecasts produced by COVID-LSTM and the six comparison models over the four different forecast horizons are listed in Table 2.

The results show that four models—FAIR-NRAR, COVIDhub-baseline, COVIDhub-ensemble, and COVID-LSTM—have substantially lower average weekly MAE than the remaining three models. Figure 3 shows the MAE of COVID-LSTM against these 3 closest competitor models for each forecast date in our test period. For the 1-week ahead forecast horizon, different models generate lower errors, depending on the specific week. However, it is evident that COVID-LSTM has a lower error in many weeks across the 2-, 3-, and 4-week forecast horizons. Over the 1-week ahead forecast, the COVIDhub-ensemble model is on average 9 cases per county more accurate than COVID-LSTM, which itself is 4 cases per county more accurate than the COVIDhub-baseline. Over the 2-, 3-, and 4-week horizons however, COVID-LSTM considerably outperforms FAIR-NRAR, the COVIDhub-baseline, and COVIDhub-ensemble models on average. Specifically, COVID-LSTM is on average 11, 34, and 50 cases per county more accurate than the COVIDhub-ensemble for the 2-, 3-, and 4-week horizons, respectively. COVID-LSTM is also 11, 45, 71, and 80 cases per county more accurate on average than FAIR-NRAR for the 1-, 2-, 3-, and 4-week forecast horizons, respectively.

Figure 3 shows a common trend in which our model error peaks around the national holidays in late November and late December, which may be indicative of lags and inconsistencies in reporting. We will explore this further in Sect. 5.5.

The average weekly MAPE of the forecasts of each model is shown in Table 3, and the MAPE of the 4 highest-

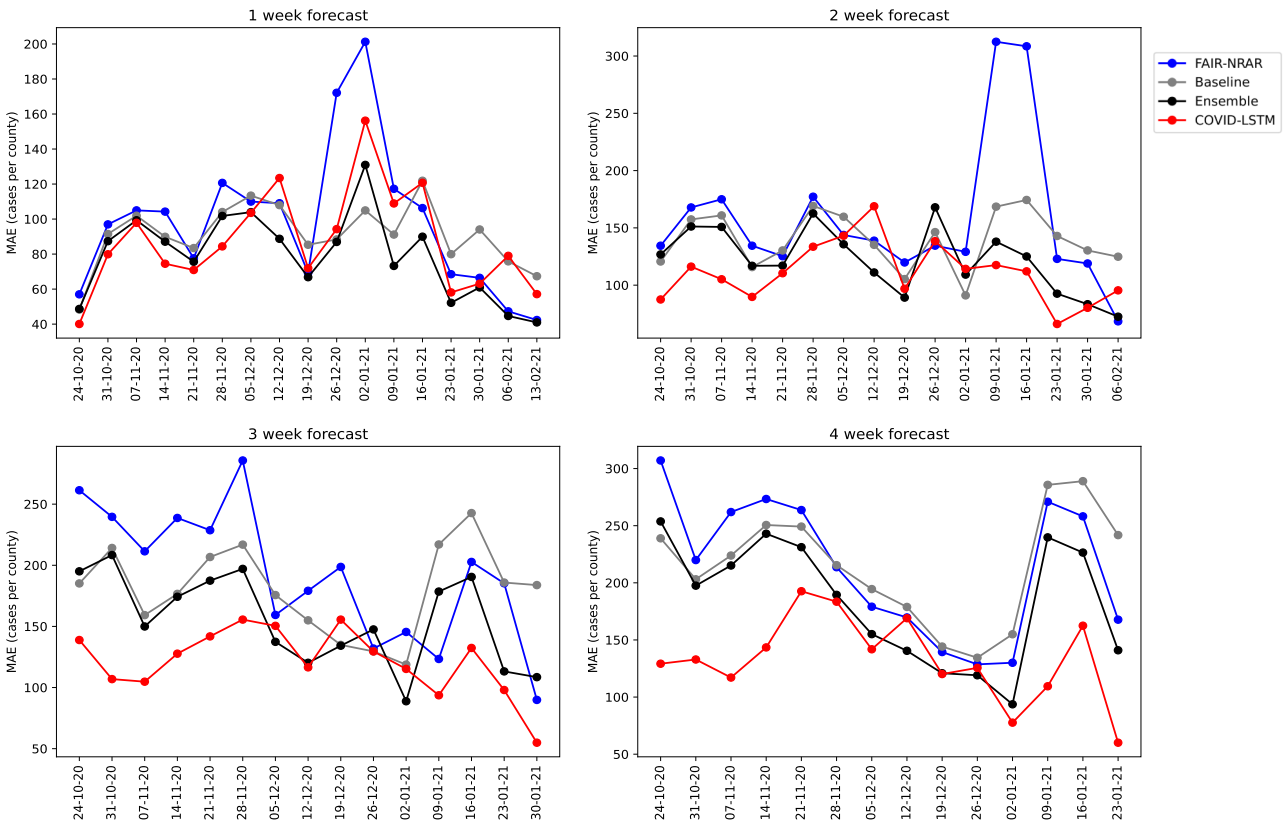

Fig. 3 The mean absolute error for COVID-LSTM against the best published comparison models, for forecast horizons of 1–4 weeks in advance

Table 3 Average weekly mean absolute percentage error (MAPE) of COVID-19 cases per county over our evaluation period in the 50 most populous counties

	Forecast horizon			
	1 week	2 weeks	3 weeks	4 weeks
CMU-TimeSeries	24.26	38.01	49.39	47.31
UMass-MechBayes	24.94	38.21	53.92	62.72
Google_Harvard-CPF	29.04	49.13	66.79	71.41
FAIR-NRAR	21.91	42.39	57.12	54.16
COVIDhub-baseline	23.30	37.64	51.09	51.39
COVIDhub-ensemble	**19.72**	32.33	42.95	43.31
COVID-LSTM	22.06	**29.29**	**35.60**	**38.30**

Model with lowest error shown in bold

performing models is shown in Fig. 4 for the whole evaluation period.

The MAPE results, similar to the MAE results previously reported, show that the COVIDhub-ensemble performs best at the 1-week forecast horizon while COVID-LSTM is best across all of the 2-, 3-, and 4-week forecast horizons.

To further contextualize the significance of these results, in [78], the authors stated that approximately half of the models submitted to the Forecast Hub had errors larger than the COVIDhub-baseline model. Furthermore, the COVIDhub-ensemble model, which is used by the CDC in reporting forecasts, is an ensemble of tens of individually calculated models each week, and the best COVID-19 Forecasting model in the USA [19]. COVID-LSTM is the first COVID-19 forecasting model to outperform the COVIDhub-ensemble model on average over any forecast horizon [19], and it does so across two evaluation metrics and three forecast horizons.

5.4 Prediction intervals

To align with the COVID-19 Forecast Hub predictions and submission requirements, COVID-LSTM is also capable of producing prediction intervals for each forecast. These are produced using quantile regression adapted for deep learning [52,54,101]. To generate quantiles, we adapt our output layer to be of the same size as the number of quantile predictions required and modify the loss function to minimize cumulative loss over all quantiles. Figure 5 shows our predictions for a 1-week forecast horizon for six individual counties. The corresponding 95% prediction interval is shown in gray. We note that these counties—Los Angeles, California; Milwaukee, Wisconsin; Pulaski, Arkansas; Boulder, Colorado; Duval, Florida; and Philadelphia, Pennsylvania—are for illustrative purposes and have not been chosen for any specific purpose, with the exception of Los Angeles County, which has the highest number of cumulative reported COVID-19 cases in the USA.

Figure 5 shows that the large majority of our 95% prediction intervals contain the true value. Where the true value falls outside of the value, it is often because of a significant change in the number of reported cases that week, either an increase or a decrease. For example, the week starting January 16, 2021, saw a decrease of approximately 40% in recorded COVID-19 infections in Los Angeles County compared to the previous week—the week of January 9, when most likely, case numbers piled up (and under-reported) during the New Year's holidays were added to reports, and thus, was followed by a week with much fewer cases. Likewise, and most likely for the same reason, the week starting January 2, 2021, saw an increase of approximately 40% in Duval county compared to the previous week, and then a similar decline the following week. In each of these examples, the reported case numbers fell outside of our model's prediction interval, but as we elaborate in the next section, this is largely due to noise in the reported data as opposed to a poor model performance.

5.5 Outliers and rolling average incidence

The reporting of confirmed COVID-19 cases has been noisy and inconsistent in most countries worldwide; however, the inconsistencies are more pronounced at smaller scales such as US counties. For instance, according to Johns Hopkins data [25], 23 counties in Utah have recorded zero cases of COVID-19 since April 2020 through July 2021, including the counties of Cache, Washington, and Weber, each of which have populations exceeding 100,000 people. Similarly, in all counties of Nebraska, zero positive cases were recorded in the months of June and July 2021.

Inconsistencies in county incidence reporting can also be observed at a national level, as shown in Fig. 6. During the holiday period in December 2020, the national number of new daily cases dropped from 240,000 on December 23 to below 100,000 on December 25, and then increased again to 240,000 on December 31, only to drop by 90,000 cases the following day. There are also observable weekly cycles in the daily data with lower cases reported on the weekends.

The variation in case data can be attributed to various causes including: the availability of tests and testers, lags in reporting, false positives, the number of asymptomatic cases, the incubation time of the virus, political motivations, and public policy [5,12,75,91,99,105].

We address the noise in the data to some extent by using a 7-day rolling average of incidence in training, instead of using the raw numbers. This form of data smoothing is standard practice in time series analysis [11]. Although we used the smoothed values in training data, we calculated MAE presented in Fig. 3 against the raw data (as ground truth) without smoothing. It can also be argued that the smoothed data should be used as the ground truth in eval-

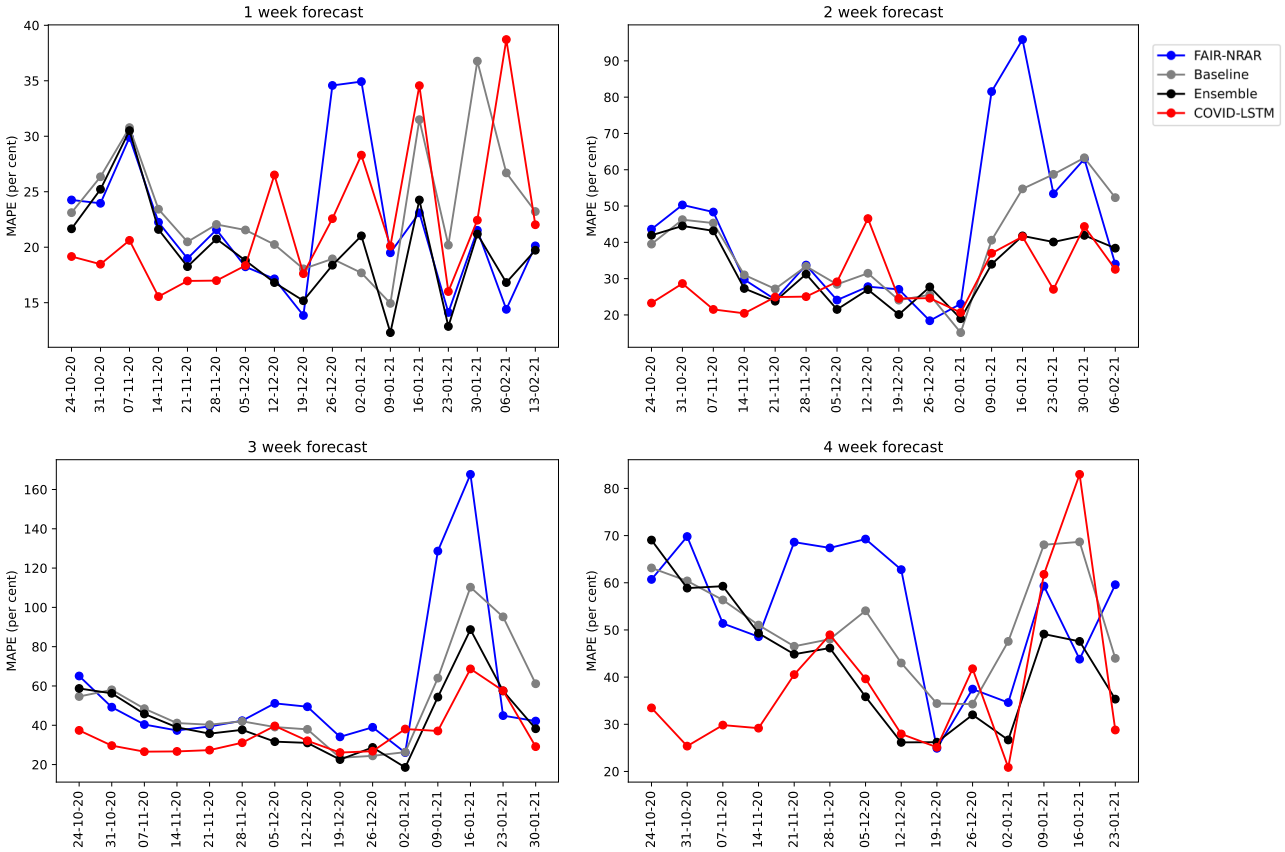

Fig. 4 The mean absolute percentage error for COVID-LSTM against the best published comparison models, for forecast horizons of 1–4 weeks ahead in the 50 most populous counties in the USA

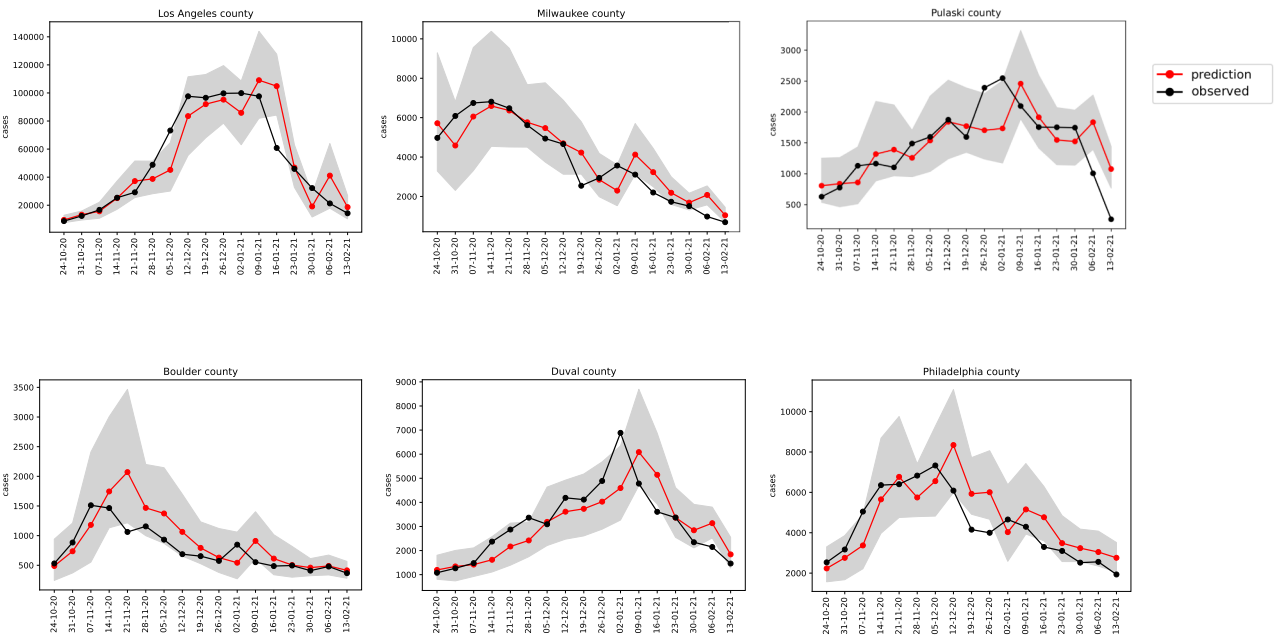

Fig. 5 COVID-19 forecasts and 95% prediction intervals from COVID-LSTM for 6 counties in the USA

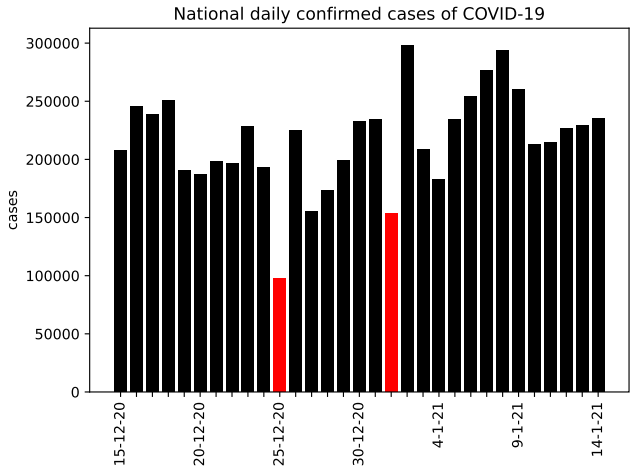

Fig. 6 US daily reported cases of COVID-19 highlighting the reporting variability around the 2020–21 holiday period

Table 4 Comparison of the performance of COVID-LSTM using average weekly MAE of COVID-19 cases per county against raw case numbers and a 7-day rolling average of case numbers

	Forecast horizon			
	1 week	2 weeks	3 weeks	4 weeks
Raw	87.29	110.97	121.46	133.22
Rolling ave	62.58	91.33	103.79	121.76

uation (i.e., in calculating MAE), as otherwise the models may be penalized heavily and unfairly in weeks during which data reporting was inconsistent (e.g., due to holidays or other reasons discussed above). When we compare the forecasts of COVID-LSTM to the rolling average case numbers, as opposed to the raw numbers, the average MAE of our model decreases at every forecast horizon (as shown in Table 4).

A comparison of MAE calculated against the raw versus smoothed case numbers for the evaluation period is shown in Fig. 7. These plots show that when evaluating against the smoothed case numbers, COVID-LSTM performs better in almost every prediction. While the Forecast Hub submissions are compared based on the prediction of raw case numbers, we would recommend that any decision-making or data-driven policy is based on the smoothed values of confirmed COVID-19 cases to account for inconsistencies in reporting.

5.6 Error analysis using aggregated county-level incidence

Our county-level evaluations above are presented as MAE, which provides no indication if COVID-LSTM's results are consistently over- or under-estimating the true totals. To investigate this, we have aggregated our county-level forecasts to the national level for comparison with the observed national incidence at both 1- and 4-week forecast horizons. We have also aggregated the county-level forecast for the COVIDhub-ensemble for further comparison. A plot of these is shown in Fig. 8. We note that these are not national-level forecasts as they are aimed at minimizing the per-county error, as opposed to minimizing the error in the national incidence.

The 1-week forecast horizon shows that both COVID-LSTM and the COVIDhub-ensemble underestimate weekly incidence when cases are rising nationally, and overestimate it when cases are declining, i.e., they appear to have a bias toward persistence. When considering the 4-week horizon, this effect is less pronounced for COVID-LSTM but still very evident for the COVIDhub-ensemble. There is potential for future work to investigate this further in an attempt to reduce bias and improve the forecasts even further.

We note that on one forecast date—January 2, 2021—both models substantially underestimate the true number of cases, which is likely due to the previous week's low data reporting during the holidays (as discussed in Sect. 5.5) and the fact that this date represents the global maximum of new COVID-19 cases.

5.7 Temporal lag

The results presented in Sect. 5.3 use input data with a temporal lag of 9 weeks, i.e., series with a length of 10—the current week and the 9 weeks prior. We chose this value based on an initial cross-validation performed on a smaller sample of the training data, as this is common for hyperparameter tuning in machine learning [69,87]. However, there is no guarantee that the chosen value is optimal. Figure 9 shows the average MAE over the evaluation period using data with longer or shorter temporal series for a 1-week forecast horizon. We can see that our choice to use a temporal lag of 9 weeks produces a low MAE but not conclusively the lowest.

Table 5 shows the average MAE over the whole evaluation period. These averages suggest that our results may have been marginally stronger if we used a lag of 10 weeks rather than 9, however this is obviously based on a comparison of evaluations on the test data, which would not be possible to see prior to performing the experiments, which are simulations of real-world model deployment.

5.8 Socioeconomic variables

Many models of disease spread incorporate socioeconomic and demographic variables such as median household income; proportion of the population over 65; proportion of black or Hispanic residents; and the political leanings of county residents [26,34,77]. Socioeconomic variables are latent variables representing factors that may cause the disease to spread faster or slower in a given county. For example, having a lower median income may imply reduced access to

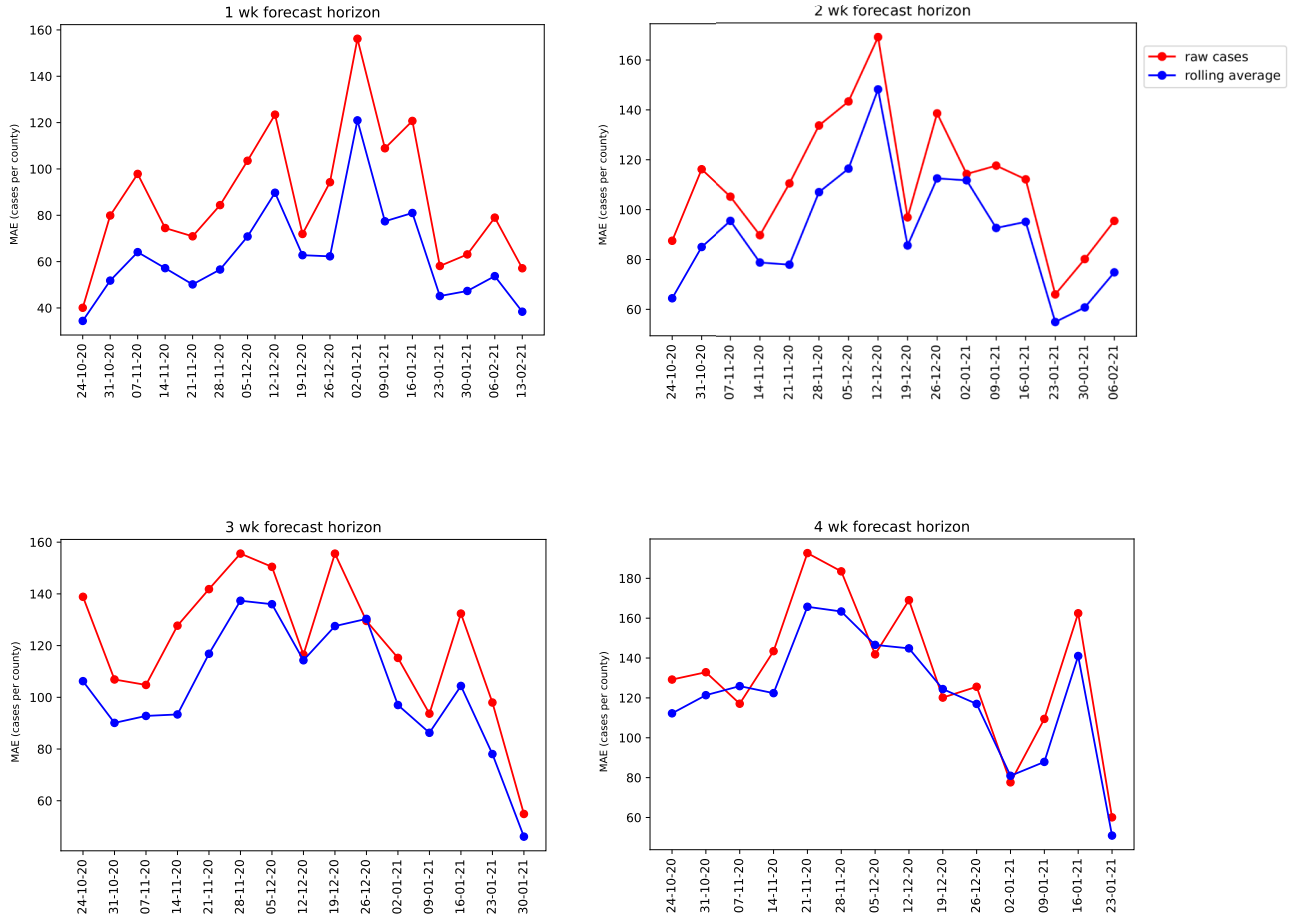

Fig. 7 The mean absolute error for COVID-LSTM forecasts when compared to the raw confirmed cases and a 7-day rolling average of these values

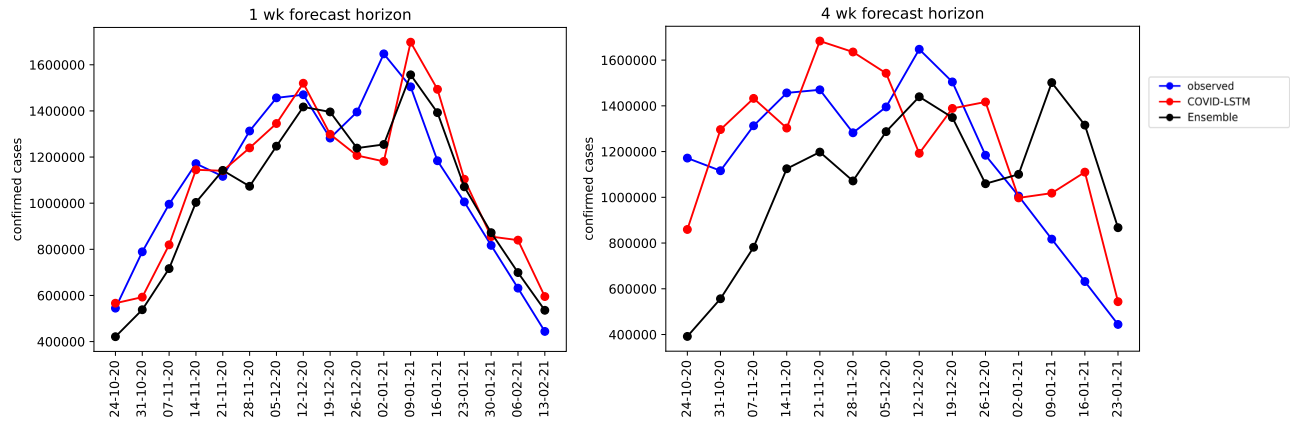

Fig. 8 US weekly total reported cases of COVID-19 shown against the aggregated county-level numbers forecasted by COVID-LSTM and the COVIDhub-ensemble at 1- and 4-week horizons

healthcare or fewer white-collar workers who can work from home.

In order to test whether socioeconomic and demographic variables would improve our forecasts, we defined a hybrid-LSTM model to incorporate atemporal variables alongside our existing temporal ones. The hybrid-LSTM architecture is shown in Fig. 10. The variables we considered in this model were: population density, proportion of black residents, proportion of Hispanic residents, proportion of indigenous residents, proportion of residents aged over 65 years, rural land as a proportion of the county area, proportion of residents who voted for Donald Trump in the 2016 US presidential election, and median household income, all at the county level.

A comparison of the MAE for the hybrid-LSTM model against COVID-LSTM and the COVIDhub-ensemble is shown in Fig. 11. The results show that the additional variables do not significantly improve the forecast over COVID-LSTM on all dates over a 1-week forecast horizon. As listed in Table 6, the average MAE across the evaluation period for the hybrid-LSTM is 88.05, which is comparable with COVID-LSTM (87.29). The similarity of these errors indicates that the majority of the predictive power results from the temporal features and the Facebook-derived connectedness and movement features. These findings are also consistent with those of the Forecast Hub researchers who

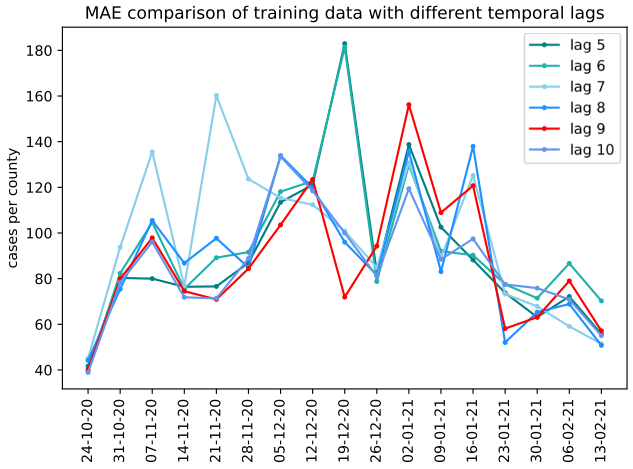

Fig. 9 The mean absolute error for COVID-LSTM trained on input data with different temporal lags for the 17-week evaluation period. The 9-week lag (red) was chosen for use in the final model through cross-validation on the training data

Table 5 The average weekly mean absolute error for COVID-LSTM trained on input data with different temporal lags for the whole evaluation period

Temporal lag (weeks)	Average MAE
5	90.38
6	94.21
7	96.78
8	89.38
9	87.29
10	86.05

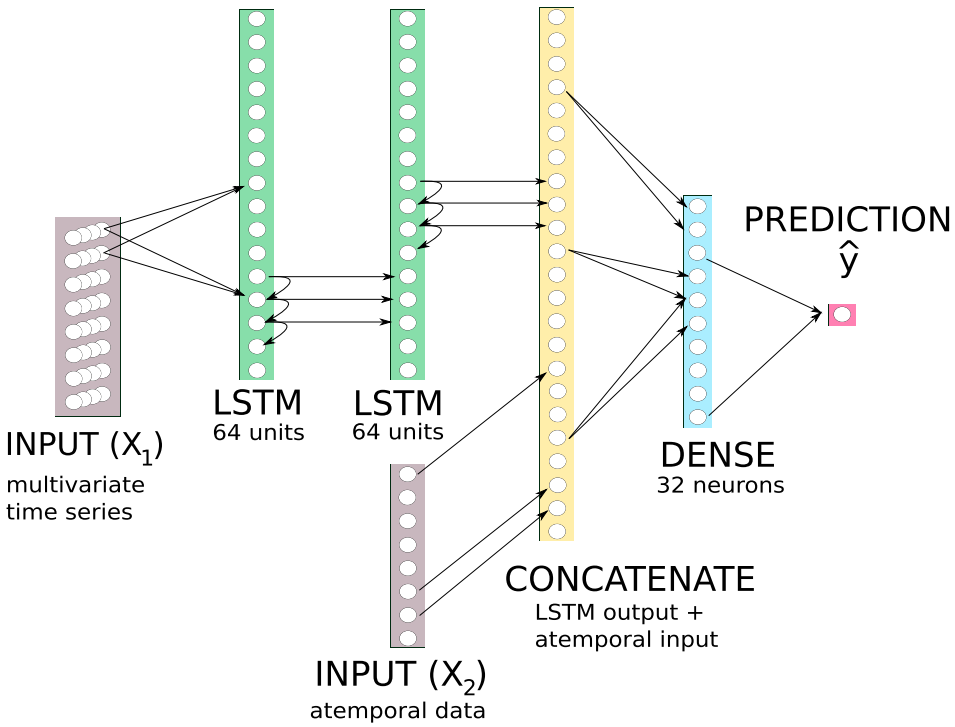

Fig. 10 The architecture of one constituent model of our Hybrid-LSTM architecture, which is an ensemble of 10 identical architectures

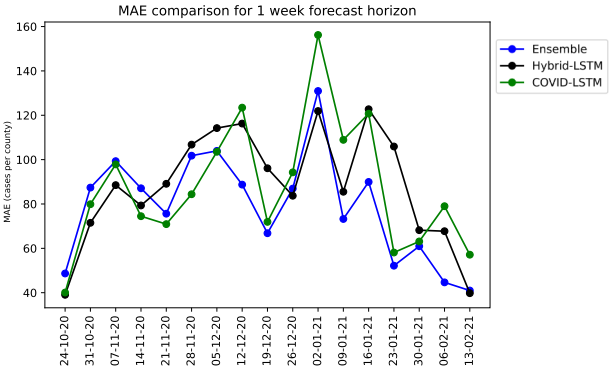

Fig. 11 The mean absolute error for the hybrid-LSTM shown against COVID-LSTM and the COVIDhub-ensemble

Table 6 The mean absolute error for the hybrid-LSTM shown against COVID-LSTM, COVIDhub-ensemble, and COVIDhub-baseline

Model	Average MAE
COVIDhub-baseline	91.13
COVIDhub-ensemble	78.77
COVID-LSTM	87.29
Hybrid-LSTM	88.05

observed that models with fewer variables tended to perform very well in COVID-19 short-term forecasting [19].

6 Discussion

Our evaluation results indicate that our data-driven, spatiotemporal forecasting using deep learning is more accurate than the COVID-19 Hub Ensemble of multiple models at predicting county-level incidence over 2-, 3-, and 4-week forecast horizons. Further, the performance metrics of our data-driven algorithm (including on the 1-week horizon) would further improve if a smoothed 7-day rolling average of incidence is used as criteria for calculating MAE. The major benefit of a data-driven disease-spread model is that the external factors that drive the spread across time and space, which are represented in the recorded case numbers for each county, are directly utilized in forecasting. This means that we can produce highly accurate forecasts without the need to quantify variables such as the reproduction rate (R_0), which currently, are calibrated based on estimates that may not generalize to the larger population well. We have demonstrated this by using an evaluation period that covers a time of high spread, a plateauing of the spread, and a decline in the spread of COVID-19. However, circumventing parameters such as R_0 is a double-edged sword, as many of these variables are key indicators of disease spread, in a universally agreed-upon framework to epidemiologists. While we have demonstrated that our data-driven method is superior to the state-of-the-art epidemiological models across most forecast horizons, the lack of these indicators may prevent its uptake in the field. Nevertheless, we here emphasize the merit of each approach for its strength: compartmental models for characterizing parameters of a disease, and machine learning for forecasting incidence in the general population.

Another advantage of a data-driven, autoregressive, spatiotemporal model, such as COVID-LSTM, is that many variables that are difficult to measure or model are automatically captured using proxy variables. For instance, in the USA, with many local governments and states leading their own policy interventions (e.g., school closures, business restrictions, masking or social distancing), keeping track of these policies and codifying them to data is next to impossible for approximately 36,000 municipalities and townships, or 13,506 school districts. While we attempted to capture social distancing through the inclusion of daily movement ranges, we are not explicitly including any dataset listing policy mandates or business restrictions. Further, we are not tracking mask mandates (since temporal data on masking at the county level does not exist). However, policy intervention or individual behavior change reflects itself in the number of observed cases in the current and prior weeks, which we use in the models as an input feature.

The same can be said about the seasonality of the disease spread. COVID-19 cases have peaked in the USA both during the Fall of 2020 followed by a winter surge, as well as summer of 2021 in many parts of the country. Our autoregressive, data-driven model captures potential seasonality by following the trend of cases in each county, based on its specific climate.

Similarly, vaccination can be a confounder in many states in late 2021. To this date (September 2021), several states are not making county-level data available for vaccination, and the rate of vaccination has been spatially heterogeneous across the country. While our model evaluation period presented in this paper does not overlap with the public availability of vaccines, in general, more vaccinations in a county would lead to a drop to new infections, which would be captured by our autoregressive model on-the-fly, as opposed to the traditional compartmental models that may require updating of the model parameters and incorporating new assumptions. We leave the incorporation of vaccination statistics and investigating its potential effect on forecasting performance for future research.

It is worth noting that while Facebook-derived datasets may not be perfectly representative of different demographic layers in the USA, our models generate county-level forecasts (in a similar fashion used by the Forecast Hub and the CDC), and not age-, race-, or gender-specific forecasts. Nevertheless, in terms of representation, Facebook has more than 200 million users in the USA, resulting in one of the

largest datasets available on human movement and county-connectivity. Furthermore, Facebook, which has 2.89 billion users worldwide, releases similar datasets for other countries as well, including locations where other sources of data are scarce.

We also note that our data-driven models benefit from the wealth of data present for COVID-19 cases when compared to previous pandemics or epidemics. If the data were more sparse—i.e., more values of zero recorded—a LSTM-based model would have more difficulty learning patterns [27,64].

Based on the information presented in this section and the overall accuracy of COVID-LSTM, we argue that disease-spread modeling should move to a realm of coexistence of compartmental models and data-driven models.

6.1 Future opportunities

While our model produces highly accurate forecasts, we acknowledge that there are outstanding challenges, which if addressed, may further improve its performance, and the performance of COVID-19 forecasting overall.

First, while we have accounted for some of the noise in the data through smoothing, a systematic algorithm for identifying and handling irregularities in the county-level data would likely improve model performance. Researchers at the Forecast Hub in the U.S. have invested significant effort for building automated anomaly detection methods; however, they report that their experiments did not yield consistently satisfactory results that improve over human judgment [80].

Secondly, while our model can account for vaccination rates through weekly incidence, an explicit input variable representing the proportion of vaccinated residents in the county may improve the model further. We note that this is not publicly available at the county level in the USA at the time of writing (September 2021). There is also scope to explore the relationship between movement variables and vaccination status, i.e., whether the people moving about are or are not vaccinated.

Thirdly, forecasts will improve from a deeper understanding of the disease characteristics. Studies have found different values for R_0 for different geographical areas and different stages of the pandemic, ranging from 3.1 in Brazil [21] to 5.7 in Wuhan, China [86]. Similarly, the incubation period is widely reported as 5 days; however, it has been measured as potentially being up to 14 days and likely depends on the variant [59]. An increase in the genomic sequencing of confirmed COVID-19 cases, and further understanding of the transmissibility of different variants of the disease, may also help improve forecasting performance. However, at present, very little genomic sequencing of confirmed cases in the USA is being conducted compared to other countries [32,82].

While the above represent notable challenges and opportunities for improving future COVID-19 forecasts, there are many other complexities of the data and disease (as identified in [13]), which a deeper understanding of will lead to improved forecasting performance.

7 Conclusion

In this paper, we presented COVID-LSTM, a data-driven approach to county-level forecasting of COVID-19. We approached the task as a spatiotemporal machine learning problem by using a temporal series of cases and hand-engineered spatial features derived from Facebook movement and connectedness datasets. COVID-LSTM outperforms the COVIDhub-ensemble on our 17-week evaluation period, making it the first model to be more accurate than the COVIDhub-ensemble over one or more forecast periods. Specifically, over the 2-, 3-, and 4-week forecast horizons, COVID-LSTM is 11, 34, and 50 cases per county more accurate than the COVIDhub-ensemble.

The high predictive power of our deep learning-based approach for forecasting COVID-19 incidence at high spatial resolutions is notable, especially given its incorporation of spatial features derived from Facebook. Facebook has over 2.89 users worldwide, and releases similar datasets for many other countries. This means that a similar forecasting approach could be useful in data-poor regions or where other sources of data may be scarce.

We have demonstrated that a data-driven spatiotemporal approach to forecasting would be greatly informative and beneficial for decision-making and planning purposes. We also acknowledge that (1) our conclusions are valid at the spatial scale and for the study area/period that we have used, not necessarily for any specific region, gender, or age group, and (2) forecasts are not the only output of compartmental epidemiological models, and that the uptake of data-driven approaches in the field may require future work to integrate and utilize both data-driven and traditional epidemiological models.

Future research should investigate methods for calibrating compartmental models in the framework of deep learning, to offer the benefits of both models: more precise forecasting and better characterization of the contagion. Another worthwhile research direction includes designing deep architectures for direct incorporation of spatial features, to potentially improve upon our current approach of hand-engineered spatial features. Lastly, future research should investigate the incorporation of vaccination variables for forecasting at high spatial resolutions, especially that annual booster shots and seasonal COVID-19 epidemics may define the world's new norm.

Supplementary material

To aid replication, the code for our method and the raw results of all experiments are available at https://github.com/geohai/covid-lstm.

Acknowledgements This work was supported by the Population Council, and the University of Colorado Population Center (CUPC) funded by Eunice Kennedy Shriver National Institute of Child Health & Human Development of the National Institutes of Health (P2CHD066613). The content is solely the responsibility of the authors and does not reflect the views of the Population Council, or official views of the NIH, CUPC, or the University of Colorado.

Declaration

Conflict of interest The authors have no relevant financial or non-financial interests to disclose.

References

1. Ansumali, S., Prakash, M.K.: A very flat peak: why standard SEIR models miss the plateau of COVID-19 infections and how it can be corrected. medRxiv (2020)
2. Arik, S.O., Li, C.L., Yoon, J., Sinha, R., Epshteyn, A., Le, L.T., Menon, V., Singh, S., Zhang, L., Yoder, N., Nikoltchev, M., Sonthalia, Y., Nakhost, H., Kanal, E., Pfister, T.: Interpretable sequence learning for COVID-19 forecasting. **2008.00646** (2021). http://arxiv.org/abs/2008.00646
3. Astin Cole, H.A., Ahmed, A., Hamasha, M., Jordan, S.: Identifying patterns of turnover intention among alabama frontline nurses in hospital settings during the COVID-19 pandemic. J. Multidiscip. Healthc. **14**, 1783 (2021)
4. Bailey, M., Cao, R., Kuchler, T., Stroebel, J., Wong, A.: Social connectedness: measurement, determinants, and effects. J. Econ. Perspect. **32**(3), 259–80 (2018)
5. Balmford, B., Annan, J.D., Hargreaves, J.C., Altoè, M., Bateman, I.J.: Cross-country comparisons of Covid-19: policy, politics and the price of life. Environ. Resour. Econ. **76**(4), 525–551 (2020)
6. Bandyopadhyay, S.K., Dutta, S.: Machine learning approach for confirmation of COVID-19 cases: positive, negative, death and release. MedRxiv (2020)
7. Bashir, M.F., Ma, B., Komal, B., Bashir, M.A., Tan, D., Bashir, M.: Correlation between climate indicators and COVID-19 pandemic in New York, USA. Sci. Total Environ. **728**, 138835 (2020). https://doi.org/10.1016/j.scitotenv.2020.138835
8. Beck, B.R., Shin, B., Choi, Y., Park, S., Kang, K.: Predicting commercially available antiviral drugs that may act on the novel coronavirus (SARS-CoV-2) through a drug-target interaction deep learning model. Comput. Struct. Biotechnol. J. **18**, 784–790 (2020)
9. Brinati, D., Campagner, A., Ferrari, D., Locatelli, M., Banfi, G., Cabitza, F.: Detection of COVID-19 infection from routine blood exams with machine learning: a feasibility study. J. Med. Syst. **44**(8), 1–12 (2020)
10. Brooks, L.C., Ray, E.L., Bien, J., Bracher, J., Rumack, A., Tibshirani, R.J., Reich, N.G.: Comparing ensemble approaches for short-term probabilistic COVID-19 forecasts in the US. International Institute of Forecasters (2020)
11. Brown, R.G.: Smoothing, Forecasting and Prediction of Discrete Time Series. Courier Corporation, North Chelmsford (2004)
12. Campolieti, M.: COVID-19 deaths in the USA: Benford's law and under-reporting. J. Public Health (Oxford, England) (2021)
13. Cao, L., Liu, Q.: Covid-19 modeling: a review. **2104.12556** (2021). http://arxiv.org/abs/2104.12556
14. Castillo-Chavez, C., Hethcote, H.W., Andreasen, V., Levin, S.A., Liu, W.M.: Epidemiological models with age structure, proportionate mixing, and cross-immunity. J. Math. Biol. **27**(3), 233–258 (1989)
15. Chimmula, V.K.R., Zhang, L.: Time series forecasting of COVID-19 transmission in Canada using LSTM networks. Chaos Solitons Fractals **135**, 109864 (2020)
16. Chinazzi, M., Davis, J.T., Ajelli, M., Gioannini, C., Litvinova, M., Merler, S., Piontti, A.P., Mu, K., Rossi, L., Sun, K., et al.: The effect of travel restrictions on the spread of the 2019 novel coronavirus (Covid-19) outbreak. Science **368**(6489), 395–400 (2020)
17. Chu, J., Ghenand, O., Collins, J., Byrne, J., Wentworth, A., Chai, P.R., Dadabhoy, F., Hur, C., Traverso, G.: Thinking green: modelling respirator reuse strategies to reduce cost and waste. BMJ Open (2021). https://doi.org/10.1136/bmjopen-2021-048687
18. Cohen, J., van der Meulen Rodgers, Y.: Contributing factors to personal protective equipment shortages during the COVID-19 pandemic. Prev. Med. **141**, 106263 (2020)
19. Cramer, E.Y., Ray, E.L., Lopez, V.K., Bracher, J., Brennen, A., Rivadeneira, A.J.C., Gerding, A., Gneiting, T., House, K.H., Huang, Y., Jayawardena, D., Kanji, A.H., Khandelwal, A., Le, K., Mühlemann, A., Niemi, J., Shah, A., Stark, A., Wang, Y., Wattanachit, N., Zorn, M.W., Gu, Y., Jain, S., Bannur, N., Deva, A., Kulkarni, M., Merugu, S., Raval, A., Shingi, S., Tiwari, A., White, J., Woody, S., Dahan, M., Fox, S., Gaither, K., Lachmann, M., Meyers, L.A., Scott, J.G., Tec, M., Srivastava, A., George, G.E., Cegan, J.C., Dettwiller, I.D., England, W.P., Farthing, M.W., Hunter, R.H., Lafferty, B., Linkov, I., Mayo, M.L., Parno, M.D., Rowland, M.A., Trump, B.D., Corsetti, S.M., Baer, T.M., Eisenberg, M.C., Falb, K., Huang, Y., Martin, E.T., McCauley, E., Myers, R.L., Schwarz, T., Sheldon, D., Gibson, G.C., Yu, R., Gao, L., Ma, Y., Wu, D., Yan, X., Jin, X., Wang, Y.X., Chen, Y., Guo, L., Zhao, Y., Gu, Q., Chen, J., Wang, L., Xu, P., Zhang, W., Zou, D., Biegel, H., Lega, J., Snyder, T.L., Wilson, D.D., McConnell, S., Walraven, R., Shi, Y., Ban, X., Hong, Q.J., Kong, S., Turtle, J.A., Ben-Nun, M., Riley, P., Riley, S., Koyluoglu, U., DesRoches, D., Hamory, B., Kyriakides, C., Leis, H., Milliken, J., Moloney, M., Morgan, J., Ozcan, G., Schrader, C., Shakhnovich, E., Siegel, D., Spatz, R., Stiefeling, C., Wilkinson, B., Wong, A., Gao, Z., Bian, J., Cao, W., Ferres, J.L., Li, C., Liu, T.Y., Xie, X., Zhang, S., Zheng, S., Vespignani, A., Chinazzi, M., Davis, J.T., Mu, K., Piontti, A.P.y., Xiong, X., Zheng, A., Baek, J., Farias, V., Georgescu, A., Levi, R., Sinha, D., Wilde, J., Penna, N.D., Celi, L.A., Sundar, S., Cavany, S., España, G., Moore, S., Oidtman, R., Perkins, A., Osthus, D., Castro, L., Fairchild, G., Michaud, I., Karlen, D., Lee, E.C., Dent, J., Grantz, K.H., Kaminsky, J., Kaminsky, K., Keegan, L.T., Lauer, S.A., Lemaitre, J.C., Lessler, J., Meredith, H.R., Perez-Saez, J., Shah, S., Smith, C.P., Truelove, S.A., Wills, J., Kinsey, M., Obrecht, R., Tallaksen, K., Burant, J.C., Wang, L., Gao, L., Gu, Z., Kim, M., Li, X., Wang, G., Wang, Y., Yu, S., Reiner, R.C., Barber, R., Gaikedu, E., Hay, S., Lim, S., Murray, C., Pigott, D., Prakash, B.A., Adhikari, B., Cui, J., Rodríguez, A., Tabassum, A., Xie, J., Keskinocak, P., Asplund, J., Baxter, A., Oruc, B.E., Serban, N., Arik, S.O., Dusenberry, M., Epshteyn, A., Kanal, E., Le, L.T., Li, C.L., Pfister, T., Sava, D., Sinha, R., Tsai, T., Yoder, N., Yoon, J., Zhang, L., Abbott, S., Bosse, N.I., Funk, S., Hellewel, J., Meakin, S.R., Munday, J.D., Sherratt, K., Zhou, M., Kalantari, R., Yamana, T.K., Pei, S., Shaman, J., Ayer, T., Adee, M., Chhatwal, J., Dalgic, O.O., Ladd, M.A., Linas, B.P., Mueller, P., Xiao, J., Li, M.L., Bertsimas, D., Lami, O.S., Soni, S., Bouardi, H.T., Wang, Y., Wang, Q., Xie, S., Zeng, D., Green,

A., Bien, J., Hu, A.J., Jahja, M., Narasimhan, B., Rajanala, S., Rumack, A., Simon, N., Tibshirani, R., Tibshirani, R., Ventura, V., Wasserman, L., O'Dea, E.B., Drake, J.M., Pagano, R., Walker, J.W., Slayton, R.B., Johansson, M., Biggerstaff, M., Reich, N.G.: Evaluation of individual and ensemble probabilistic forecasts of COVID-19 mortality in the U.S. medRxiv (2021). https://doi.org/10.1101/2021.02.03.21250974

20. Das, D., Santosh, K., Pal, U.: Truncated inception net: Covid-19 outbreak screening using chest x-rays. Phys. Eng. Sci. Med. **43**(3), 915–925 (2020)
21. de Souza, W.M., Buss, L.F., da Silva Candido, D., Carrera, J.P., Li, S., Zarebski, A.E., Pereira, R.H.M., Prete, C.A., de Souza-Santos, A.A., Parag, K.V., et al.: Epidemiological and clinical characteristics of the COVID-19 epidemic in Brazil. Nat. Hum. Behav. **4**(8), 856–865 (2020)
22. Dean, N.E., Piontti, A.P., Madewell, Z.J., Cummings, D.A., Hitchings, M.D., Joshi, K., Kahn, R., Vespignani, A., Halloran, M.E., Longini, I.M.: Ensemble forecast modeling for the design of COVID-19 vaccine efficacy trials. Vaccine **38**(46), 7213–7216 (2020)
23. Dehning, J., Zierenberg, J., Spitzner, F.P., Wibral, M., Neto, J.P., Wilczek, M., Priesemann, V.: Inferring change points in the spread of COVID-19 reveals the effectiveness of interventions. Science **369**(6500), 160 (2020)
24. del Rio-Chanona, R.M., Mealy, P., Pichler, A., Lafond, F., Farmer, J.D.: Supply and demand shocks in the COVID-19 pandemic: an industry and occupation perspective. Oxford Rev. Econ. Policy **36**(Supplement 1), S94–S137 (2020). https://doi.org/10.1093/oxrep/graa033
25. Dong, E., Du, H., Gardner, L.: An interactive web-based dashboard to track COVID-19 in real time. Lancet Infect. Dis. **20**(5), 533–534 (2020)
26. Doti, J.L.: Examining the impact of socioeconomic variables on COVID-19 death rates at the state level. J. Bioecon. **23**(1), 15–53 (2021)
27. Drumond, R.R., Marques, B.A.D., Vasconcelos, C.N., Clua, E.: An LSTM recurrent network for motion classification from sparse data. In: Proceedings of the 13th international joint conference on computer vision, imaging and computer graphics theory and applications, vol. 1, pp. 215–22 (2018)
28. Fang, Y., Nie, Y., Penny, M.: Transmission dynamics of the Covid-19 outbreak and effectiveness of government interventions: a data-driven analysis. J. Med. Virol. **92**(6), 645–659 (2020)
29. Fawaz, H.I., Lucas, B., Forestier, G., Pelletier, C., Schmidt, D.F., Weber, J., Webb, G.I., Idoumghar, L., Muller, P.A., Petitjean, F.: Inceptiontime: finding alexnet for time series classification. Data Min. Knowl. Discov. **34**(6), 1936–1962 (2020)
30. Flaxman, S., Mishra, S., Gandy, A., Unwin, H.J.T., Mellan, T.A., Coupland, H., Whittaker, C., Zhu, H., Berah, T., Eaton, J.W., et al.: Estimating the effects of non-pharmaceutical interventions on COVID-19 in Europe. Nature **584**(7820), 257–261 (2020)
31. Friedman, J., Liu, P., Troeger, C.E., Carter, A., Reiner, R.C., Barber, R.M., Collins, J., Lim, S.S., Pigott, D.M., Vos, T., et al.: Predictive performance of international COVID-19 mortality forecasting models. Nat. Commun. **12**(1), 1–13 (2021)
32. Furuse, Y.: Genomic sequencing effort for SARS-CoV-2 by country during the pandemic. Int. J. Infect. Dis. **103**, 305–307 (2021)
33. Galanis, P., Vraka, I., Fragkou, D., Bilali, A., Kaitelidou, D.: Nurses' burnout and associated risk factors during the COVID-19 pandemic: a systematic review and meta-analysis. J. Adv. Nurs. (2021)
34. Garnier, R., Benetka, J.R., Kraemer, J., Bansal, S.: Socioeconomic disparities in social distancing during the COVID-19 pandemic in the United States: observational study. J. Med. Internet Res. **23**(1), e24591 (2021)
35. Gautam, Y.: Transfer learning for COVID-19 cases and deaths forecast using LSTM network. ISA Trans (2021)
36. Gers, F.A., Schmidhuber, J., Cummins, F.: Learning to forget: continual prediction with LSTM. Neural Comput. **12**(10), 2451–2471 (2000). https://doi.org/10.1162/089976600300015015
37. Getz, W.M., Salter, R., Mgbara, W.: Adequacy of SEIR models when epidemics have spatial structure: Ebola in Sierra Leone. Philos. Trans. R. Soc. **374**(1775), 20180282 (2019)
38. Gibson, G.C., Reich, N.G., Sheldon, D.: Real-time mechanistic Bayesian forecasts of covid-19 mortality. medRxiv (2020). https://www.medrxiv.org/content/early/2020/12/24/2020.12.22.20248736
39. Graves, A., Liwicki, M., Fernández, S., Bertolami, R., Bunke, H., Schmidhuber, J.: A novel connectionist system for unconstrained handwriting recognition. IEEE Trans. Pattern Anal. Mach. Intell. **31**(5), 855–868 (2009). https://doi.org/10.1109/TPAMI.2008.137
40. Halevy, A., Norvig, P., Pereira, F.: The unreasonable effectiveness of data. IEEE Intell. Syst. **24**(2), 8–12 (2009)
41. He, S., Peng, Y., Sun, K.: SEIR modeling of the COVID-19 and its dynamics. Nonlinear Dyn. **101**(3), 1667–1680 (2020)
42. Herdağdelen, A., Dow, A.: Protecting privacy in facebook mobility data during the COVID-19 response (2020). https://research.fb.com/blog/2020/06/protecting-privacy-in-facebook-mobility-data-during-the-covid-19-response/
43. Hethcote, H.W.: The mathematics of infectious diseases. SIAM Rev. **42**(4), 599–653 (2000)
44. Hochreiter, S., Schmidhuber, J.: Long short-term memory. Neural Comput. **9**(8), 1735–1780 (1997). https://doi.org/10.1162/neco.1997.9.8.1735
45. Hong, H.G., Li, Y.: Estimation of time-varying reproduction numbers underlying epidemiological processes: a new statistical tool for the Covid-19 pandemic. PLoS ONE **15**(7), 1–15 (2020)
46. Ibrahim, M.R., Haworth, J., Lipani, A., Aslam, N., Cheng, T., Christie, N.: Variational-LSTM autoencoder to forecast the spread of coronavirus across the globe. PLoS ONE **16**(1), E0246120 (2021)
47. IHME COVID-19 forecasting team: Modeling COVID-19 scenarios for the United States. Nat. Med. (2020)
48. Johansson, M.A., Apfeldorf, K.M., Dobson, S., Devita, J., Buczak, A.L., Baugher, B., Moniz, L.J., Bagley, T., Babin, S.M., Guven, E., et al.: An open challenge to advance probabilistic forecasting for dengue epidemics. Proc. Natl. Acad. Sci. **116**(48), 24268–24274 (2019)
49. Keeling, M.J., Hollingsworth, T.D., Read, J.M.: Efficacy of contact tracing for the containment of the 2019 novel coronavirus (COVID-19). J. Epidemiol. Community Health **74**(10), 861–866 (2020)
50. Kendall, D.G.: Deterministic and stochastic epidemics in closed populations, pp. 149–166. University of California Press (1956). https://doi.org/10.1525/9780520350717-011
51. Kingma, D.P., Ba, J.: Adam: a method for stochastic optimization. **1412.6980** (2014). http://arxiv.org/abs/1412.6980
52. Kocherginsky, M., He, X., Mu, Y.: Practical confidence intervals for regression quantiles. J. Comput. Graph. Stat. **14**(1), 41–55 (2005). https://doi.org/10.1198/106186005X27563
53. Koelle, K., Cobey, S., Grenfell, B., Pascual, M.: Epochal evolution shapes the phylodynamics of interpandemic influenza a (h3n2) in humans. Science **314**(5807), 1898–1903 (2006)
54. Koenker, R., Bassett, G.: Regression quantiles. Econometrica **46**(1), 33–50 (1978)
55. Koh, D.: COVID-19 lockdowns throughout the world. Occup. Med. **70**(5), 322 (2020)
56. Kolassa, S., Schütz, W., et al.: Advantages of the MAD/MEAN ratio over the MAPE. Foresight Int. J. Appl. Forecast. **6**, 40–43 (2007)

57. Kontis, V., Bennett, J.E., Rashid, T., Parks, R.M., Pearson-Stuttard, J., Guillot, M., Asaria, P., Zhou, B., Battaglini, M., Corsetti, G., et al.: Magnitude, demographics and dynamics of the effect of the first wave of the Covid-19 pandemic on all-cause mortality in 21 industrialized countries. Nat. Med. **26**(12), 1919–1928 (2020)
58. Kuchler, T., Russel, D., Stroebel, J.: The geographic spread of COVID-19 correlates with the structure of social networks as measured by facebook. J. Urban Econ. 103314 (2021)
59. Lauer, S.A., Grantz, K.H., Bi, Q., Jones, F.K., Zheng, Q., Meredith, H.R., Azman, A.S., Reich, N.G., Lessler, J.: The incubation period of coronavirus disease 2019 (COVID-19) from publicly reported confirmed cases: estimation and application. Ann. Intern. Med. **172**(9), 577–582 (2020)
60. Le, M., Ibrahim, M., Sagun, L., Lacroix, T., Nickel, M.: Neural relational autoregression for high-resolution COVID-19 forecasting. Facebook AI Research (2020). https://ai.facebook.com/research/publications/neural-relational-autoregression-for-high-resolution-covid-19-forecasting
61. Leclerc, P.M., Matthews, A.P., Garenne, M.L.: Fitting the HIV epidemic in Zambia: a two-sex micro-simulation model. PLoS ONE **4**(5), e5439 (2009)
62. Li, J., Vidyattama, Y., La, H.A., Miranti, R., Sologon, D.M.: The impact of COVID-19 and policy responses on australian income distribution and poverty. **2009.04037** (2020). http://arxiv.org/abs/2009.04037
63. Lipsitch, M., Finelli, L., Heffernan, R.T., Leung, G.M., Redd; for the 2009 H1N1 Surveillance Group, S.C.: Improving the evidence base for decision making during a pandemic: the example of 2009 influenza a/h1n1. Biosecurity Bioterrorism Biodefense Strategy Pract. Sci. **9**(2), 89–115 (2011)
64. Liu, S., Ni'mah, I., Menkovski, V., Mocanu, D.C., Pechenizkiy, M.: Efficient and effective training of sparse recurrent neural networks. Neural Comput. Appl. 1–12 (2021)
65. Lloyd, A.L., Jansen, V.A.: Spatiotemporal dynamics of epidemics: synchrony in metapopulation models. Math. Biosci. **188**(1–2), 1–16 (2004)
66. Lucas, B., Shifaz, A., Pelletier, C., O'Neill, L., Zaidi, N., Goethals, B., Petitjean, F., Webb, G.I.: Proximity forest: an effective and scalable distance-based classifier for time series. Data Min. Knowl. Discov. **33**(3), 607–635 (2019). https://doi.org/10.1007/s10618-019-00617-3
67. Malki, Z., Atlam, E.S., Hassanien, A.E., Dagnew, G., Elhosseini, M.A., Gad, I.: Association between weather data and COVID-19 pandemic predicting mortality rate: machine learning approaches. Chaos Solitons Fractals **138**, 110137 (2020)
68. Melnick, E.R., Ioannidis, J.P.: Should governments continue lockdown to slow the spread of COVID-19? BMJ **369**, m1924 (2020)
69. Moore, A.W., Lee, M.S.: Efficient algorithms for minimizing cross validation error. In: Machine Learning Proceedings 1994, pp. 190–198. Elsevier (1994)
70. Mukherjee, H., Ghosh, S., Dhar, A., Obaidullah, S.M., Santosh, K., Roy, K.: Shallow convolutional neural network for covid-19 outbreak screening using chest x-rays. Cogn. Comput. 1–14 (2021)
71. Mwalili, S., Kimathi, M., Ojiambo, V., Gathungu, D., Mbogo, R.: SEIR model for COVID-19 dynamics incorporating the environment and social distancing. BMC Res. Notes **13**(1), 1–5 (2020)
72. Opitz, D.W., Shavlik, J.W.: Actively searching for an effective neural network ensemble. Connect. Sci. **8**(3–4), 337–354 (1996)
73. Pal, R., Sekh, A.A., Kar, S., Prasad, D.K.: Neural network based country wise risk prediction of COVID-19. Appl. Sci. **10**(18), 6448 (2020)
74. Peiris, J.S., Guan, Y., Yuen, K.Y.: Severe acute respiratory syndrome. Nat. Med. **10**(12), S88–S97 (2004)
75. Pettengill, M.A., McAdam, A.J., Miller, M.B.: Can we test our way out of the COVID-19 pandemic? J. Clin. Microbiol. **58**(11), e02225-20 (2020). https://doi.org/10.1128/JCM.02225-20
76. Pichler, A., Pangallo, M., del Rio-Chanona, R.M., Lafond, F., Farmer, J.D.: Production networks and epidemic spreading: How to restart the UK economy? **2005.10585** (2020). http://arxiv.org/abs/2005.10585
77. Quan, D., Wong, L.L., Shallal, A., Madan, R., Hamdan, A., Ahdi, H., Daneshvar, A., Mahajan, M., Nasereldin, M., Van Harn, M., et al.: Impact of race and socioeconomic status on outcomes in patients hospitalized with COVID-19. J. Gen. Intern. Med. **36**(5), 1302–1309 (2021)
78. Ray, E.L., Wattanachit, N., Niemi, J., Kanji, A.H., House, K., Cramer, E.Y., Bracher, J., Zheng, A., Yamana, T.K., Xiong, X., Woody, S., Wang, Y., Wang, L., Walraven, R.L., Tomar, V., Sherratt, K., Sheldon, D., Reiner, R.C., Prakash, B.A., Osthus, D., Li, M.L., Lee, E.C., Koyluoglu, U., Keskinocak, P., Gu, Y., Gu, Q., George, G.E., España, G., Corsetti, S., Chhatwal, J., Cavany, S., Biegel, H., Ben-Nun, M., Walker, J., Slayton, R., Lopez, V., Biggerstaff, M., Johansson, M.A., Reich, N.G., on behalf of the COVID-19 Forecast Hub Consortium: Ensemble forecasts of coronavirus disease 2019 (COVID-19) in the U.S. medRxiv (2020). https://doi.org/10.1101/2020.08.19.20177493
79. Reich, N.G., McGowan, C.J., Yamana, T.K., Tushar, A., Ray, E.L., Osthus, D., Kandula, S., Brooks, L.C., Crawford-Crudell, W., Gibson, G.C., et al.: Accuracy of real-time multi-model ensemble forecasts for seasonal influenza in the US. PLoS Comput. Biol. **15**(11), e1007486 (2019)
80. Reich Lab - University of Massachusetts Amherst: Data anomalies (2020). https://github.com/reichlab/covid19-forecast-hub/tree/master/data-anomalies
81. Reinhart, A., Brooks, L., Jahja, M., Rumack, A., Tang, J., Saeed, W.A., Arnold, T., Basu, A., Bien, J., Cabrera, Á.A., Chin, A., Chua, E.J., Clark, B., DeFries, N., Forlizzi, J., Gratzl, S., Green, A., Haff, G., Han, R., Hu, A.J., Hyun, S., Joshi, A., Kim, J., Kuznetsov, A., Motte-Kerr, W.L., Lee, Y.J., Lee, K., Lipton, Z.C., Liu, M.X., Mackey, L., Mazaitis, K., McDonald, D.J., Narasimhan, B., Oliveira, N.L., Patil, P., Perer, A., Politsch, C.A., Rajanala, S., Rucker, D., Shah, N.H., Shankar, V., Sharpnack, J., Shemetov, D., Simon, N., Srivastava, V., Tan, S., Tibshirani, R., Tuzhilina, E., Van Nortwick, A.K., Ventura, V., Wasserman, L., Weiss, J.C., Williams, K., Rosenfeld, R., Tibshirani, R.J.: An open repository of real-time covid-19 indicators. medRxiv (2021). https://www.medrxiv.org/content/early/2021/07/16/2021.07.12.21259660
82. Robishaw, J.D., Alter, S.M., Solano, J.J., Shih, R.D., DeMets, D.L., Maki, D.G., Hennekens, C.H.: Genomic surveillance to combat Covid-19: challenges and opportunities. Lancet Microbe (2021)
83. Rodríguez, A., Tabassum, A., Cui, J., Xie, J., Ho, J., Agarwal, P., Adhikari, B., Prakash, B.A.: Deepcovid: an operational deep learning-driven framework for explainable real-time COVID-19 forecasting. medRxiv (2020). https://doi.org/10.1101/2020.09.28.20203109
84. Rădulescu, A., Williams, C., Cavanagh, K.: Management strategies in a SEIR-type model of COVID-19 community spread. Sci. Rep. **10**(1), 1–16 (2020)
85. Sak, H., Senior, A., Beaufays, F.: Long short-term memory based recurrent neural network architectures for large vocabulary speech recognition (2014). arXiv:1402.1128
86. Sanche, S., Lin, Y.T., Xu, C., Romero-Severson, E., Hengartner, N., Ke, R.: High contagiousness and rapid spread of severe acute respiratory syndrome coronavirus 2. Emerg. Infect. Dis. **26**(7), 1470 (2020)
87. Schaffer, C.: Selecting a classification method by cross-validation. Mach. Learn. **13**(1), 135–143 (1993)

88. Shastri, S., Singh, K., Kumar, S., Kour, P., Mansotra, V.: Time series forecasting of COVID-19 using deep learning models: India-USA comparative case study. Chaos Solitons Fractals **140**, 110227 (2020). https://doi.org/10.1016/j.chaos.2020.110227
89. Sherstinsky, A.: Fundamentals of recurrent neural network (RNN) and long short-term memory (LSTM) network. Phys. D Nonlinear Phenom. **404**, 132306 (2020)
90. Sun, C., Shrivastava, A., Singh, S., Gupta, A.: Revisiting unreasonable effectiveness of data in deep learning era. In: 2017 IEEE International Conference on Computer Vision (ICCV), vol. 2017-, pp. 843–852. IEEE (2017)
91. Surkova, E., Nikolayevskyy, V., Drobniewski, F.: False-positive COVID-19 results: hidden problems and costs. Lancet Respir. Med. **8**(12), 1167–1168 (2020)
92. Tosepu, R., Gunawan, J., Effendy, D.S., Lestari, H., Bahar, H., Asfian, P., et al.: Correlation between weather and COVID-19 pandemic in Jakarta, Indonesia. Sci. Total Environ. **725**, 138436 (2020)
93. Vahedi, B., Karimzadeh, M., Zoraghein, H.: Spatiotemporal prediction of COVID-19 cases using inter- and intra-county proxies of human interactions. Nat. Commun. **12**, 6440 (2021). https://doi.org/10.1038/s41467-021-26742-6
94. Viboud, C., Sun, K., Gaffey, R., Ajelli, M., Fumanelli, L., Merler, S., Zhang, Q., Chowell, G., Simonsen, L., Vespignani, A., et al.: The RAPIDD EBOLA forecasting challenge: synthesis and lessons learnt. Epidemics **22**, 13–21 (2018)
95. Walker, P.G.T., Whittaker, C., Watson, O.J., Baguelin, M., Winskill, P., Hamlet, A., Djafaara, B.A., Cucunubá, Z., Mesa, D.O., Green, W., Thompson, H., Nayagam, S., Ainslie, K.E.C., Bhatia, S., Bhatt, S., Boonyasiri, A., Boyd, O., Brazeau, N.F., Cattarino, L., Cuomo-Dannenburg, G., Dighe, A., Donnelly, C.A., Dorigatti, I., van Elsland, S.L., FitzJohn, R., Fu, H., Gaythorpe, K.A.M., Geidelberg, L., Grassly, N., Haw, D., Hayes, S., Hinsley, W., Imai, N., Jorgensen, D., Knock, E., Laydon, D., Mishra, S., Nedjati-Gilani, G., Okell, L.C., Unwin, H.J., Verity, R., Vollmer, M., Walters, C.E., Wang, H., Wang, Y., Xi, X., Lalloo, D.G., Ferguson, N.M., Ghani, A.C.: The impact of COVID-19 and strategies for mitigation and suppression in low- and middle-income countries. Science **369**(6502), 413–422 (2020). https://doi.org/10.1126/science.abc0035
96. Wallinga, J., van Boven, M., Lipsitch, M.: Optimizing infectious disease interventions during an emerging epidemic. Proc. Natl. Acad. Sci. **107**(2), 923–928 (2010)
97. Wang, C., Liu, Z., Chen, Z., Huang, X., Xu, M., He, T., Zhang, Z.: The establishment of reference sequence for SARS-CoV-2 and variation analysis. J. Med. Virol. **92**(6), 667–674 (2020)
98. Wang, L., Didelot, X., Yang, J., Wong, G., Shi, Y., Liu, W., Gao, G.F., Bi, Y.: Inference of person-to-person transmission of Covid-19 reveals hidden super-spreading events during the early outbreak phase. Nat. Commun. **11**(1), 1–6 (2020)
99. Watson, J., Whiting, P.F., Brush, J.E.: Interpreting a COVID-19 test result. Br. Med. J. (2020). https://doi.org/10.1136/bmj.m1808
100. Watts, D.J., Muhamad, R., Medina, D.C., Dodds, P.S.: Multiscale, resurgent epidemics in a hierarchical metapopulation model. Proc. Natl. Acad. Sci. **102**(32), 11157–11162 (2005)
101. Wei, Y., Pere, A., Koenker, R., He, X.: Quantile regression methods for reference growth charts. Stat. Med. **25**(8), 1369–1382 (2006). https://doi.org/10.1002/sim.2271
102. Xu, C., Yu, Y., Chen, Y., Lu, Z.: Forecast analysis of the epidemics trend of COVID-19 in the USA by a generalized fractional-order SEIR model. Nonlinear Dyn. **101**(3), 1621–1634 (2020)
103. Yang, J., Zeng, X., Zhong, S., Wu, S.: Effective neural network ensemble approach for improving generalization performance. IEEE Trans. Neural Netw. Learn. Syst. **24**(6), 878–887 (2013)
104. Zeroual, A., Harrou, F., Dairi, A., Sun, Y.: Deep learning methods for forecasting COVID-19 time-series data: a comparative study. Chaos Solitons Fractals **140**, 110121 (2020)
105. Zhao, H., Lu, X., Deng, Y., Tang, Y., Lu, J.: COVID-19: asymptomatic carrier transmission is an underestimated problem. Epidemiol. Infect. **148**, e116 (2020)
106. Zhou, T., Ji, Y.: Semiparametric Bayesian inference for the transmission dynamics of Covid-19 with a state-space model. Contemp. Clin. Trials **97**, 106146 (2020)
107. Zou, D., Wang, L., Xu, P., Chen, J., Zhang, W., Gu, Q.: Epidemic model guided machine learning for COVID-19 forecasts in the United States. medRxiv (2020). https://doi.org/10.1101/2020.05.24.20111989