Contents lists available at ScienceDirect

Applied Soft Computing Journal

journal homepage: www.elsevier.com/locate/asoc

Forecasting the dynamics of cumulative COVID-19 cases (confirmed, recovered and deaths) for top-16 countries using statistical machine learning models: Auto-Regressive Integrated Moving Average (ARIMA) and Seasonal Auto-Regressive Integrated Moving Average (SARIMA)

K.E. ArunKumar [a,*], Dinesh V. Kalaga [b,*], Ch. Mohan Sai Kumar [c], Govinda Chilkoor [d], Masahiro Kawaji [b], Timothy M. Brenza [a,e,**]

[a] *Department of Chemical and Biological Engineering, South Dakota School of Mines and Technology, Rapid City, SD 57701, USA*
[b] *Mechanical Engineering Department, City College of New York, New York, NY 10031, USA*
[c] *Process Chemistry and Technology, CSIR-Central Institute of Medicinal and Aromatic Plants, Lucknow, UP, 226015, India*
[d] *Department of Civil and Environmental Engineering, South Dakota School of Mines and Technology, Rapid City, SD, 57701, USA*
[e] *Biomedical Engineering program, South Dakota School of Mines and Technology, Rapid City, SD 57701, USA*

ARTICLE INFO

Article history:
Received 29 September 2020
Received in revised form 8 January 2021
Accepted 28 January 2021
Available online 8 February 2021

Keywords:
COVID-19
Statistical modeling
Pandemic
ARIMA
SARIMA
Time-series forecast
Auto-Correlation Function (ACF)
Akaike Information Criteria (AIC)
Bayesian Information Criterion (BIC)

ABSTRACT

Most countries are reopening or considering lifting the stringent prevention policies such as lockdowns, consequently, daily coronavirus disease (COVID-19) cases (confirmed, recovered and deaths) are increasing significantly. As of July 25th, there are 16.5 million global cumulative confirmed cases, 9.4 million cumulative recovered cases and 0.65 million deaths. There is a tremendous necessity of supervising and estimating future COVID-19 cases to control the spread and help countries prepare their healthcare systems. In this study, time-series models — Auto-Regressive Integrated Moving Average (ARIMA) and Seasonal Auto-Regressive Integrated Moving Average (SARIMA) are used to forecast the epidemiological trends of the COVID-19 pandemic for top-16 countries where 70%–80% of global cumulative cases are located. Initial combinations of the model parameters were selected using the auto-ARIMA model followed by finding the optimized model parameters based on the best fit between the predictions and test data. Analytical tools Auto-Correlation function (ACF), Partial Auto-Correlation Function (PACF), Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC) were used to assess the reliability of the models. Evaluation metrics Mean Absolute Error (MAE), Mean Square Error (MSE), Root Mean Square Error (RMSE) and Mean Absolute Percent Error (MAPE) were used as criteria for selecting the best model. A case study was presented where the statistical methodology was discussed in detail for model selection and the procedure for forecasting the COVID-19 cases of the USA. Best model parameters of ARIMA and SARIMA for each country are selected manually and the optimized parameters are then used to forecast the COVID-19 cases. Forecasted trends for confirmed and recovered cases showed an exponential rise for countries such as the United States, Brazil, South Africa, Colombia, Bangladesh, India, Mexico and Pakistan. Similarly, trends for cumulative deaths showed an exponential rise for countries Brazil, South Africa, Chile, Colombia, Bangladesh, India, Mexico, Iran, Peru, and Russia. SARIMA model predictions are more realistic than that of the ARIMA model predictions confirming the existence of seasonality in COVID-19 data. The results of this study not only shed light on the future trends of the COVID-19 outbreak in top-16 countries but also guide these countries to prepare their health care policies for the ongoing pandemic. The data used in this work is obtained from publicly available John Hopkins University's COVID-19 database.

1. Introduction

In the last week of December 2019, a group of patients at local hospitals in Wuhan, China demonstrated novel a form of viral pneumonia [1]. All the patients shared a common history of visiting a wet market in Wuhan, China. The patients were not

[*] Corresponding authors.
[**] Correspondence to: 501 E Saint Joseph St.
E-mail addresses: eswararunkumar.kalaga@mines.sdsmt.edu (K.E. ArunKumar), dkalaga@ccny.cuny.edu (D.V. Kalaga), Timothy.Brenza@sdsmt.edu (T.M. Brenza).

https://doi.org/10.1016/j.asoc.2021.107161
1568-4946/

Nomenclature

y_t	Non-stationary time-series.
y'_t	Time-series after first order differencing.
y''_t	Time-series after second order differencing.
y_{t-1}	Observation at time $t-1$ in the past, one-time step away from current time stamp t.
y'_s	Seasonal time-series data.
y_{t-m}	Observation at $t-m$ in the past, m time steps away from the current time-stamp t.
m	Number of time steps for a single seasonal period.
y_t	Time-series data at time t.
C	Intercept or constant.
$\phi_{i\ or\ p}$	Auto-regressive parameter at ith or pth time-stamp.
ϕ_1	Auto-regressive parameter at $t-1$ time-stamp.
ϕ_2	Auto-regressive parameter at $t-2$ time-stamp.
θ_1	Moving average parameter at $t-1$ time-stamp.
θ_2	Moving average parameter at $t-2$ time-stamp.
$\mathcal{E}_t$	Random error or residual term for the tth day.
p	Auto-regressive term of ARIMA model.
d	Ordinary differencing term of ARIMA model.
q	Moving average term of ARIMA model.
P	Auto-regressive term of SARIMA model.
D	Seasonal differencing term of SARIMA model.
Q	Moving average term of SARIMA model.
Φ_P	Auto-regressive parameter of SARIMA model at Pth time-stamp.
Θ_Q	Moving average parameter of SARIMA model at Qth time-stamp.
B	Backshift operator.
$\varphi(\mathrm{B})$	Non-seasonal auto-regressive polynomial.
$\theta(\mathrm{B})$	Non-seasonal moving average polynomial.
$\Phi_\mathrm{P}(\mathrm{B}^m)$	Seasonal auto-regressive polynomial.
$(1-\mathrm{B}^\mathrm{m})^\mathrm{D}$	Seasonal differencing.
$\Theta_\mathrm{Q}(\mathrm{B}^m)$	Seasonal moving average polynomial.
$\mathrm{L}\left(\hat{\theta}\right)$	Likelihood of the candidate model given the data evaluated at $\hat{\theta}$.
$\hat{\theta}$	The set of model parameters.
k	The number of estimated parameters in the candidate model.
n	Sample size or number of observations.

responding to medicine and the agent causing the disease was identified as Severe Acute Respiratory Syndrome Corona Virus-2 (SARS-CoV-2) which is a strain of the coronaviruses family [2]. Sooner the outbreak was declared as a pandemic, the disease is named as COVID-19, by World Health Organization (WHO) on the 11th of March 2020 [3]. Ever since the outbreak was declared as pandemic, many countries over the world were affected severely by novel coronavirus disease (COVID-19) and took various measures to control the spread. For example, countries such as the USA, Australia, India took various preventive measures such as using facemasks, implementing stay-at-home order, social-distancing, and lockdowns [4–7]. As a result of the control measures the daily confirmed cases decreased drastically. For example, due to the implementation of stay-at-home orders in the USA, the daily confirmed cases on June 7h, 2020 was 19,370 which was decreased from 32,074 on April 9th, 2020 [8].

Currently, most of the countries are at the breaking point in terms of health services, following the stay-at-home, mandatory face masks, and social-distancing orders. This can be evident from the surge in the reported 63000 daily confirmed new cases (5-day average) on 15th July 2020, which was 3.2 folds of the cases reported on 7th June 2020 [8]. Similarly, Italy's health care system has been pushed beyond the limits. The exponential rise in confirmed cases required exponential rise in health care supplies and the deployment of healthcare personal [9]. Currently, there are 16 countries where 80% of the global COVID-19 confirmed cases are concentrated. As there is no specific treatment for the COVID-19 illness, the preparation of the health care system and prevention is of utmost urgency [10]. The healthcare system can be prepared to control the outbreak, by accurately predicting the forecast of the COVID-19 dynamics using statistical modeling tools. These models can be used for making short-term and long-term forecast of the disease spread thereby providing an idea on the amount of additional healthcare resources will be needed.

Various statistical models are used to predict the upcoming number of cases and forecast the spread of infectious disease in the near future [11]. Zhang et al. [12] have used the SARIMA model to forecast Typhoid fever. In another study Chen et al. [13] have forecasted the influenza incidence in urban and rural areas of Shenyang, China using SARIMA model. Similarly, ARIMA models were used to forecast infectious diseases such as tuberculosis [14], Dengue fever [15] and Brucellosis [16]. Recently, ARIMA models were used to predict the prevalence, growth rate, the life cycle of COVID-19 pandemic. Ceylan. Z [17] has used ARIMA models to predict the epidemiological trend in Italy, Spain, and France. Leila et al. [18] have used the ARIMA model to predict and forecast the number of COVID-19 patients for the next 30 days in Iran. They reported the number of daily cases would be 3,574 by April 20. Marbaniang S. P. [19] has reported the use of ARIMA models and predicted and forecasted the total confirmed cases for the next 20 days from May 18th, 2020. He reported that the cases in India will increase to 2,45,000 in the first week of June 2020. Perone [20] has used ARIMA models to forecast the cumulative cases in Italy for more than 40 days. Their results showed that the number of COVID-19 cases in Tuscany (Italy) will reach plateau on 55th day of the forecast.

Further, several researchers have reported the short-term forecast of COVID-19 pandemic using the machine learning models other than ARIMA and SARIMA Ghosal et al. [21] have used the linear and multiple linear regression techniques to forecast the number of fatalities in India for a short period for six weeks. Authors have reported that the fatalities in India will be doubled if the COVID-19 preventive measures are unchanged or not implemented. Parbat and Chakraborty [22] have employed the Support Vector Regression (SVR) for predicting the COVID-19 cases in India for 60 days based on the time-series data reported for the period of 1st March 2020 to 30th April 2020. Their results indicate that the SVR model has an accuracy of ∼97% in predicting the cumulative fatalities cases, cumulative recovered cases, cumulative confirmed cases. Their model also able to predict the daily new COVID cases with an accuracy of 87%. Maleki et al. [23] have used Auto-Regressive (AR) models based on two-piece scale mixture normal distributions to forecast the confirmed and recovered COVID-19 cases. Their model performed well in forecasting confirmed and recovered global COVID-19 cases. Ribeiro et al. [4,24] have used Cubist Regression, Random Forest, Ridge Regression, SVR, and ARIMA models for short-term forecasting of COVID-19 confirmed cases in Brazil. Their findings reveal that the best performing models are SVR, ARIMA. Salgotra et al. [25] have used models based on genetic programming for predicting the cumulative confirmed cases and cumulative fatalities in India. Authors have found that their model is less sensitive to the variables and highly reliable in predicting the cumulative confirmed cases and cumulative deaths. Chimmula and Zhang [26] have employed a deep learning Long Short-Term Memory (LSTM) network to predict the COVID-19 trends in Canada. It is reported that the pandemic in Canada will be ending in about three months. Mehdi et al. [27] have employed LSTM network, SARIMA and Holt winter's exponential smoothing and moving average methods to forecast COVID-19 cases in Iran. Their comparative study reported that the LSTM model outperformed other models. Ardabili et al. [28] have implemented a multi-layer perceptron model and adaptive network-based fuzzy interface system for predicting the COVID-19 outbreak. Their research work has recommended developing individual machine learning models for each country due to the existence of fundamental differences among different countries.

In this study, we made an attempt to forecast the cumulative COVID-19 confirmed cases, recovered cases, and confirmed deaths for the top-16 countries, where 70%–80% of global COVID-19 cases concentrated. The top-16 countries were chosen based on the total accumulative confirmed cases. The pie chart for the percentage distribution of COVID-19 cases per each country is depicted in Fig. 1. The present study uses the COVID-19 cases are reported for the period of Jan 22nd, 2020 to July 24th, 2020 and the data was obtained from Johns Hopkins coronavirus resource center [29]. The rest of the paper is organized as follows: Section 2 describes the statistical models, their underlying mathematics along with the analytical tools, evaluation metrics. The computational framework of the model parameter selection procedure is discussed in Section 3. In Section 4, the model parameter selection and parameter optimization procedure are discussed in great detail by taking the time-series analysis of cumulative confirmed cases of the USA as a case study. Further, forecasted trends of the cumulative confirmed cases, recovered and deaths, based on ARIMA and SARIMA models, are given in the results and discussion section (Section 5). Finally, Section 6 provides the conclusions drawn from the present work.

2. Statistical models and description

We have used ARIMA and SARIMA statistical models to generate a 60-day forecast of cumulative COVID-19 cases for top-16 countries, the proposed models are country-specific and were optimized by selecting the best model parameters. For each country, we have considered the date on which the first case was reported as the starting day of the time-series, hence, the date of the first case reported varies from country to country. To have a statistically meaningful forecast of time-series data, the minimum sample size of 30 observations is required [30]. The number of observations (i.e. sample size) used in the present work is much greater than the minimum size required to carry out the meaningful time-series forecasting, as the data collected for the duration of seven months (22nd January 2020 to 3rd August 2020).

Time-series data is a sequence of numerical values that has a time-stamp associated with each value [31]. Time-series data can be classified into two categories namely stationary data and non-stationary data. A stationary time-series data has no patterns with respect to the time whereas a non-stationary time-series data has patterns, also known as seasonality. Therefore, the mean and variance of the non-stationary data are not constant over time. The non-stationary time-series data can be converted into stationary by calculating the difference between two successive observations. This technique is called differencing, it removes the changes in the level of the time-series thereby eliminating the trends and seasonality. There are two widely used differencing techniques, known as ordinary differencing and seasonal differencing. The ordinary first-order differencing, second-order differencing are mathematically represented as Eqs. (1) and (2), respectively.

$$y`_t = y_t - y_{t-1} \tag{1}$$

$$y"_t = y_t - 2y_{t-1} + y_{t-2} \tag{2}$$

Where y_t is non-stationary time-series data, y'_t is the time-series after first-order differencing, y''_t is the time-series after second-order differencing, y_{t-1} is the observation at time-stamp $t-1$, y_{t-2} is the observation at time-stamp $t-2$. Second order differencing is needed when the data is not stationary after first order differencing. In seasonal differencing, the difference is equal to the difference between an observation and the previous observation from the same season. The first order of seasonal differencing can be written as follows.

$$y'_s = y_t - y_{t-m} \tag{3}$$

where y'_s is the seasonal time-series after first order differencing, y_{t-m} is the observation at time-stamp $t-m$, m is the number of time step corresponding to a single seasonal period. The time-series data was first subjected to differencing for removing the seasonality and then the resulted data frame is used for forecasting. For developing the statistical models based on the time-series data the following assumptions were made:

1. Time-series data does not contain anomalies/outliers.
2. Data is univariate meaning the time-series data is comprised of only one variable, as both the ARIMA and SARIMA model regresses a variable with its past values.
3. The model assumes that the data is stationary requiring the mean and variance are constant over time.
4. Model parameters and error terms are assumed to be constant with respect to time.

2.1. Auto-Regressive Integrated Moving Average (ARIMA(p,d,q))

ARIMA(p,d,q) model was first introduced by Box and Jenkin in 1976 [32], it can be used for forecasting the non-seasonal stationary time-series data. An ARIMA model is characterized by 3 terms: p, d, q where p is the order of the Auto-Regression (AR) term, q is the order of the Moving Average (MA) term, d is the order of differencing required to make the time-series stationary. Auto-Regression is nothing but the regression of the variable against itself to forecast the variable of interest. It correlates the pattern of the one-time period to its previous time periods. MA is a regression-like model that uses the errors associated with the forecast at a previous time-step to forecast a variable at a later time-step. The following are the generalized equations of pth order AR model (Eq. (4)) and qth order MA model (Eq. (5)).

$$y_t = C + \phi_1 y_{t-1} + \phi_2 y_{t-2} + \cdots\cdots + \phi_p + y_{t-p} + \mathcal{E}_t \tag{4}$$

$$y_t = C + \mathcal{E}_t + \theta_1 \mathcal{E}_{t-1} + \theta_2 \mathcal{E}_{t-2} + \cdots\cdots + \theta_q \mathcal{E}_{t-q} \tag{5}$$

ARIMA models are built upon incorporating the AR model (Eq. (4)), integration (I) and the MA model (Eq. (5)). The integration (I) is the reverse process of differencing to generate the forecast. The generalized ARIMA model is mathematically represented as in Eq. (6).

$$y_t = C + \phi_1 y + \phi_p y_{t-p} + \cdots\cdots + \phi_n y_{t-n} + \theta_1 \mathcal{E}_{t-1} + \theta_q \mathcal{E}_{t-q} + \mathcal{E}_t \tag{6}$$

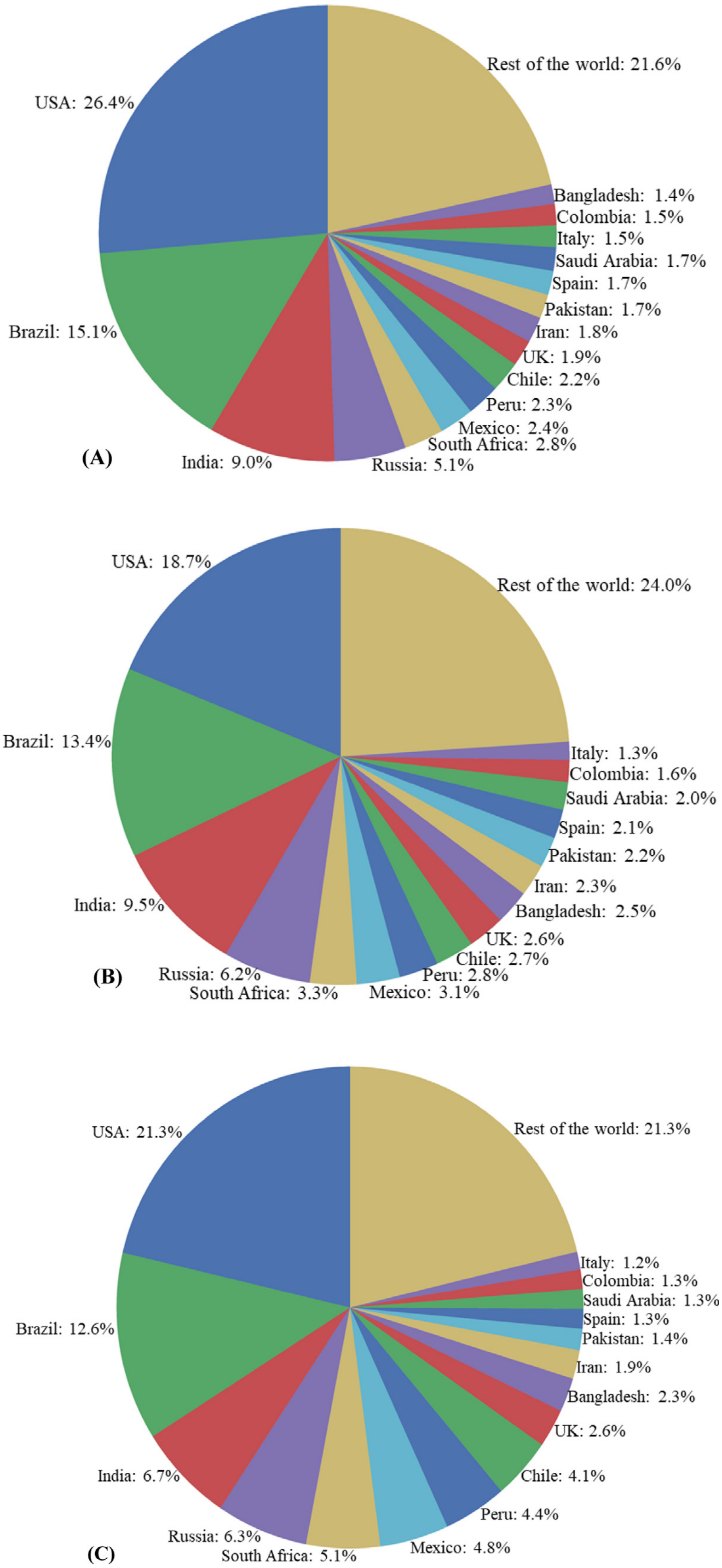

Fig. 1. Pie-chart showing the percentage distribution of global COVID-19 data (A) confirmed cases (B) recovered cases (C) deaths.

Where C is intercept, $\phi_i(i = 1, 2 \ldots \text{p})$ is auto-regressive model parameters, $\theta_i(i = 1, 2 \ldots \text{p})$ is moving average model parameters, y_t is current time-series value, y_{t-1}, y_{t-2}...y_{t-p} is past values and $\mathcal{E}_t$ is random error or residual term for the tth day and it is given by the following equation:

$$\mathcal{E}_t = y_t - y_{t-1} \tag{7}$$

2.2. Seasonal Auto-Regressive Integrated Moving Average (SARIMA (p, d, q)(P, D, Q))

Seasonal-ARIMA (SARIMA) model includes non-seasonal ARIMA(p, d, q) and additional seasonal terms $(\text{P, D, Q})_\text{m}$ to account for the seasonality of the time-series data for m number of time steps corresponding to a single seasonal period. The terms P, Q and D are the order of seasonal AR term, seasonal moving average term, seasonal differencing term, respectively. The general SARIMA model is mathematically represented as follows:

$$\Phi_P\left(B^m\right)\phi_p\left(B\right)\left(1-B^m\right)^D\left(1-B\right)^d y_t = \Theta_\text{Q}(\text{B}^m)\theta_q(B)w_t \tag{8}$$

Where y_t is the non-stationary time-series, w_t is the Gaussian white noise process, $\varphi(\text{B})$ is non-seasonal auto-regressive polynomial and $\theta(\text{B})$ is non-seasonal moving average polynomial, D is seasonal differencing term is equal to 1 or 2 etc. However, the value of $\text{D} = 1$ is sufficient to enforce stationarity into the data, $\Phi_\text{P}(\text{B}^m)$ is seasonal auto-regressive polynomial, and $\Theta_\text{Q}(\text{B}^m)$ is seasonal moving average polynomial. Where, B is defined as the backshift operator which is expressed as follows:

$$B^k y_t = y_{t-k} \tag{9}$$

The expressions for the non-seasonal auto-regressive model (Eq. (10)), moving-average (Eq. (11)) model, seasonal terms for seasonal AR model (Eq. (12)) and seasonal MA (Eq. (13)) model are given below.

$$\varphi\ (\text{B}) = 1 - \varphi_1 B - \varphi_2 B^2 - \cdots - \varphi_p B^p \tag{10}$$

$$\theta\ (\text{B}) = 1 + \theta_1 B + \theta_2 B^2 + \cdots + \theta_q B^q \tag{11}$$

$$\Phi_\text{P}(\text{B}^m) = 1 - \Phi_1 \text{B}^m - \Phi_2 \text{B}^{2m} + \cdots\cdots + \Phi_P \text{B}^{Pm} \tag{12}$$

$$\Theta_\text{Q}(\text{B}^m) = 1 + \Theta_1 B^m + \Theta_2 B^{2m} + \cdots\cdots + \Theta_\text{Q} B^{Qs} \tag{13}$$

2.3. Analytical tools and model evaluation

The following analytical tools are used for assessing the reliability of time-series analysis: Auto-Correlation Function (ACF), Partial Auto-Correlation Function (PACF), Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC). These measures indicate the relation between the observations within the time-series. ACF gives the correlation of time-series data with its previous time-series data, whereas PACF correlates the time-series with its own lagged values separated by certain time units. AIC and BIC are both penalized-likelihood criteria, the lower the AIC and BIC values mean that the model is more likely to be considered as a true model. The evaluation metrics used in this study are Mean Absolute Error (MAE), Mean Square Error (MSE), and Root Mean Square Error (RMSE).

2.3.1. Auto-Correlation Function (ACF) and Partial Auto-Correlation Function (PACF)

The correlation between current observation with the observations from previous time-steps (lags) in a time-series data is called auto-correlation. The plot of auto-correlation vs lags in the time-series is called an auto-correlation plot, and the ACF describes the linear relationship between observation at time t and observation at a previous time (t-k). To illustrate, the ACF for time-series y_t is given by:

$$ACF(y_t, y_{t-k}) = \frac{Covariance(y_t, y_{t-k})}{variance(y_t)} \tag{14}$$

where k is lag, and it is defined as the difference between y_t and y_{t-k}. Lag k auto-correlation means the correlation between the observations that are k time periods apart. On the other hand, in partial auto-correlation, the intermediate observations are considered while calculating the correlation between two observations at different times. For instance, consider that a time-series y_t. The PACF between two observations y_t and y_{t-2} (assuming $k = 2$) can be written as shown in the equation.

$$\text{PACF}(y_t, y_{t-2}) = \frac{covariance(y_t, y_{t-2}|y_{t-1})}{\sqrt{variance(y_t|y_{t-1})}\sqrt{variance(y_{t-2}|y_{t-1})}} \tag{15}$$

2.3.2. Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC)

The generated models need to be tested for the goodness of the model performance in terms of explaining the relationships between the variables. We have used the information criteria to determine how well a model explains the relationships. Two popular criteria are AIC and BIC, these information criteria access the quality of the models by giving credit to models which has less error while applying penalty for models with too many parameters. AIC is mathematically represented as follows.

$$AIC = -2logL\left(\hat{\theta}\right) + 2K \tag{16}$$

$\log L\left(\hat{\theta}\right)$ represents the likelihood function and K is the total number of model parameters. Similarly, BIC is another model selection criterion. BIC imposes a lesser penalty on the number of parameters when compared to AIC. In both AIC and BIC settings the lower value represents the best model which has a higher likelihood value. Thus, assisting time-series analysts in choosing the best model amongst the finite number of potential models generated. BIC is mathematically represented as follows.

$$BIC = -2logL\left(\hat{\theta}\right) + KlogN \tag{17}$$

Where N is the number of observations.

2.3.3. Evaluation metrics

MAE, MSE, RMSE and MAPE are used often to evaluate the accuracy of the proposed model, which are given by the following equations:

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \tag{18}$$

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \tag{19}$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \tag{20}$$

$$MAPE = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{y_i - \hat{y}_i}{y_i}\right| \tag{21}$$

Where, $\hat{y}_i$ is model predicted value, y_i is actual value.

3. Computational framework for model development

In the first step, each time-series was checked for the presence of non-stationarity using ACF and PACF plots. If the auto-correlation reduces very marginally as the number of lags

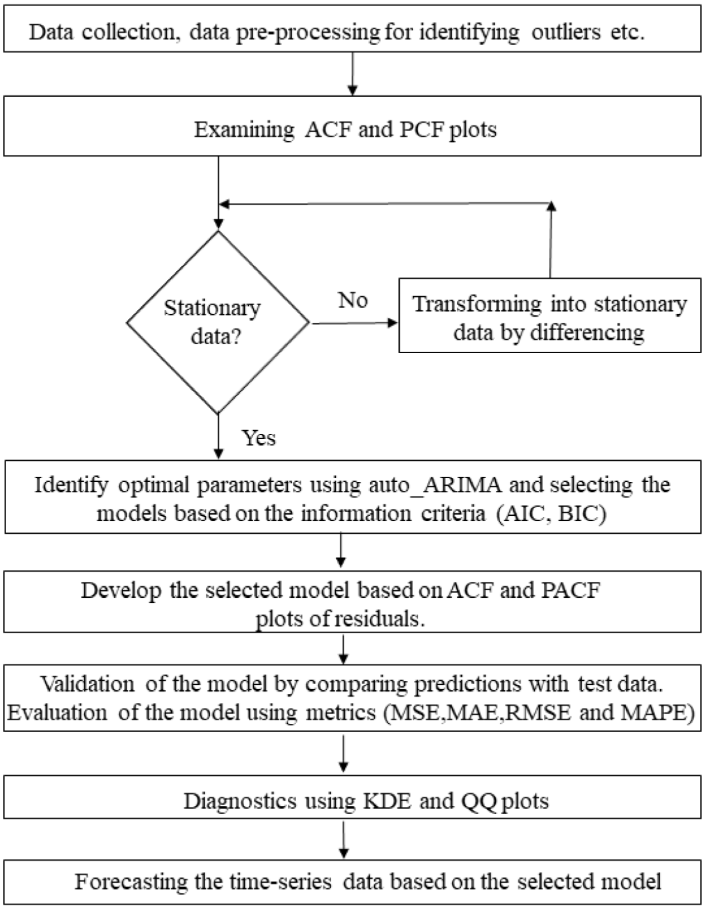
Data collection, data pre-processing for identifying outliers etc.
Examining ACF and PCF plots
Stationary data?
No
Transforming into stationary data by differencing
Yes
Identify optimal parameters using auto_ARIMA and selecting the models based on the information criteria (AIC, BIC)
Develop the selected model based on ACF and PACF plots of residuals.
Validation of the model by comparing predictions with test data. Evaluation of the model using metrics (MSE,MAE,RMSE and MAPE)
Diagnostics using KDE and QQ plots
Forecasting the time-series data based on the selected model

Fig. 2. Algorithm showing the methodology for developing ARIMA and SARIMA models.

increase, it indicates that the time-series is non-stationary. Such time-series with evidence of non-stationary was differenced before performing the ARIMA or SARIMA modeling. The raw time-series was used for modeling without any differencing if ACF and PACF plots indicate the presence of stationarity. The non-stationary time-series was subjected to first-order differencing to stabilize the mean of the time-series before performing the forecast. However, in some cases the second-order differencing was performed if the first order differenced time-series has a trend or seasonality. The algorithm showing the stepwise procedure for developing the ARIMA and SARIMA models were given in Fig. 2. The scripts were written in Python (Ver. 3.7.) programming language installed in the Anaconda environment. The numerical simulations were performed in the cloud computing platform (Google COLAB) and on a local computer (OS: Windows 10, Processor: Intel-I7). The snapshots of the Python script depicting the important steps of the data analysis can be found in Appendix A. The average computational time taken for each simulation on the local computer is about 3 secs for the ARIMA model and 6 secs for the SARIMA models.

Most of the countries' cumulative cases required second-order differencing except for confirmed cases of South Africa and Spain (Table 1), for recovered cases of UK (Table 3), for recovered cases Spain and UK (Table 4). ACF plot of stationary time-series was used to get a basic idea on whether AR terms or MA terms will fit to the data to deliver a superior model. If the ACF plot has negative auto-correlation at the first lag, it suggests using MA terms. If the PACF plot of the differenced time-series showed a sharp cutoff which is positive, we consider adding AR terms to the model. Selecting the best parameter (p, d, q) manually using ACF and PACF plots for ARIMA can be time-consuming as the number of models to assess is a permutation of the number of model order parameters, and it can be even more expensive for selecting the parameters of SARIMA(p, d, q)(P, D, Q)$_m$. To select the proper combination of the model parameter values we performed a grid search using pmdarima (Pyramid ARIMA) library available in statsmodels (a python module). The pmdarima uses AIC as an evaluation metric to choose the best model from various ARIMA and SARIMA models. The seasonality of the data was checked using the seasonal_decompose function that is available in statsmodels. Then the stepwise parameter selection was performed to identify the best combination by setting the seasonality to "True" during the grid search. Since the cumulative COVID-19 cases are of only few months, the parameter that represents seasonality (m) was assigned to 3, 7, 12. Our data analysis showed that seasonality terms varied from country to country. The model with the best seasonal term was identified using information criteria (AIC and BIC). Zohair Malki et al. [33] have used a similar approach for identifying the seasonality term (m) for COVID -19 data by assigning it 3,7 and 12.

The time-series data of all the selected top-16 countries were split into 80% training and 20% testing/validating datasets. The model development and parameter selection were done using the training dataset and the performance of the developed model was tested with the validation dataset. The ACF and PACF plots of the residuals were used to further determine the model's goodness of fit. If the ACF and PACF plots of the residuals displayed correlation coefficients that are significantly different from

Table 1
Selected ARIMA models for forecasting cumulative confirmed cases.

Country	ARIMA (p,d,q)	AIC	BIC	MSE	MAE	RMSE	MAPE
South Africa	7,2,1	1.93E+03	1.96E+03	1.33E+05	1.91E+02	3.64E+02	5.42E−02
Bangladesh	6,2,2	1.72E+03	1.75E+03	38,97,458	1.35E+03	1.97E+03	6.00E−01
Brazil	6,2,1	2.63E+03	2.65E+03	1.29E+09	2.99E+04	3.59E+04	1.39E+00
Chile	7,2.4	2.20E+03	2.24E+03	3,62,491	4.87E+02	6.02E+02	1.50E−01
Columbia	2,2,2	1.77E+03	1.77E+03	1.11E+07	3.30E+03	3.33E+03	1.82E+00
India	5,2,2	2.48E+03	2.50E+03	2.99E+07	3.52E+03	5.47E+03	3.28E−01
Iran	0,2,0	2.10E+03	2.11E+03	7.43E+04	2.22E+02	2.73E+02	8.06E−02
Italy	7,2,4	2.28E+03	2.32E+03	1.97E+04	1.23E+02	1.40E+02	4.50E−01
Mexico	6,2,2	1.96E+03	1.98E+03	8.08E+06	2.65E+03	2.84E+03	7.50E−01
Pakistan	4,2,1	2.26E+03	2.28E+03	4.25E+05	4.25E+02	6.52E+02	1.63E−01
Peru	0,2,1	2.14E+03	2.14E+03	9.02E+06	2.45E+03	3.00E+03	6.77E−01
Russia	3,2,3	2.36E+03	2.38E+03	8.66E+06	1.88E+03	2.94E+03	2.37E−01
Saudi Arabia	3,2,1	1.81E+03	1.83E+03	8.10E+05	6.42E+02	9.00E+02	2.57E−01
Spain	3,2,4	2.75E+03	2.77E+03	3.06E+08	3.97E+03	5.53E+03	1.55E+00
UK	7,2,1	2.26E+03	2.29E+03	8.71E+06	2.33E+03	2.95E+03	9.00E−02
USA	7,2,1	3.03E+03	3.06E+03	2.04E+08	1.17E+04	1.43E+04	9.91E−02

Table 2
Selected SARIMA models for forecasting cumulative confirmed cases.

Country	SARIMA (p,d,q)(P,D,Q,m)	AIC	BIC	MSE	MAE	RMSE	MAPE
South Africa	(2,1,2)(1,1,1,7)	1.87E+03	1.89E+03	2.12E+05	3.29E+02	4.60E+02	9.00E−02
Bangladesh	(5,2,2)(2,1,2,7)	1.39E+03	1.43E+03	6.63E+04	2.13E+02	2.57E+02	1.84E−01
Brazil	(3,2,2)(2,0,1,7)	2.30E+03	2.32E+03	2.69E+09	4.45E+04	5.18E+04	2.04E+00
Chile	(0,2,1)(1,0,1,7)	2.04E+03	2.05E+03	3.59E+05	4.97E+02	6.00E+02	1.58E−01
Colombia	(4,2,1)(2,1,1,7)	1.40E+03	1.43E+03	1.42E+08	9.71E+03	1.19E+04	4.83E+00
India	(2,2,1)(1,1,1,7)	2.21E+03	2.23E+03	2.97E+09	3.87E+04	5.45E+04	3.10E+00
Iran	(3,2,0)(2,0,1,7)	1.69E+03	1.71E+03	1.33E+07	3.06E+03	3.64E+03	1.09E+00
Italy	(1,2,2)(2,0,1,7)	1.99E+03	2.02E+03	1.76E+06	1.13E+03	1.33E+03	4.55E−01
Mexico	(1,2,1)(1,0,2,7)	1.73E+03	1.75E+03	4.35E+07	5.78E+03	6.59E+03	1.61E+00
Pakistan	(2,2,2)(0,0,1,3)	2.50E+03	2.52E+03	1.35E+08	9.53E+03	1.16E+04	3.56E+00
Peru	(0,2,1)(1,0,0,12)	1.95E+03	1.95E+03	2.73E+07	4.37E+03	5.22E+03	1.21E+00
Russia	(4,2,4)(4,1,4,3)	2.09E+03	2.14E+03	5.73E+06	1.54E+03	2.39E+03	2.13E−01
Saudi Arabia	(0,2,0)(1,0,0,3)	1.77E+03	1.77E+03	1.38E+07	3.01E+03	3.72E+03	1.17E+00
Spain	(3,1,1)(2,1,1,3)	2.55E+03	2.57E+03	6.59E+06	1.80E+03	2.57E+03	6.71E−01
UK	(1,2,1)(1,0,2,7)	2.32E+03	2.33E+03	9.05E+05	8.11E+02	9.51E+02	2.73E−01
USA	(3,2,4)(2,0,4,7)	2.46E+03	2.50E+03	3.62E+08	1.60E+04	1.90E+04	4.00E+00

zero at higher lags, then we developed higher-order ARIMA or SARIMA models, otherwise, the simple models suggested by auto-ARIMA were used. The evaluation of the model was done using the evaluation metrics: MAE, MSE, RMSE and MAPE (Fig. 2). The actual vs predicted values were plotted to visually understand the error. Once the finest model was identified by training on the training dataset, the model was used to predict values of the test data followed by forecasting for the next 60 days of cumulative COVID-19 cases for top-16 countries.

In the first step, we checked for stationarity of the raw data followed by differenced data of all the countries using ACF and PACF plots as mentioned in the case study. As mentioned before, both seasonal and non-seasonal ARIMA models for all the top-16 countries' COVID-19 cases. The SARIMA models capture both trend and seasonality using non-seasonal differencing (d) and seasonal differencing (D) respectively. For COVID-19 cases, we have considered seasonality in this time-series which is in between 3 to 12 days pattern. There are various factors that control and contribute to the seasonal pattern of the pandemics such as influenza and COVID-19. Some of those factors include social distancing on weekdays vs weekends [34], climatic conditions [35]. For example, the seasonality of the confirmed cases of the USA has an oscillating pattern on every 7 days as discussed in Section 4. For instance, in the case of the confirmed cases of Peru a simple ARIMA model (0,2,1) was selected as the best model with the lowest AIC and BIC values of 2,139.9 and 2,143.3 as shown in Table 1. The selected ARIMA(0,2,1) model was used to forecast the cumulative confirmed cases in Peru because the ACF and PACF plot of the residuals did not show any correlation coefficients that are significantly different from zero at least until 10 lags as shown in Figure S1B (supplementary document). For any model such as ARIMA(0,2,1), with second-order differencing (I or d = 2), implies that the forecast and the trend of the time-series was adapted over time, hence the trend is equal to the exponentially smoothed values of the previous slopes (change in the process). Similarly, ARIMA(0,2,0) was the best model to forecast cumulative confirmed cases of Iran, in the forecast process of ARIMA(0,2,0), new observation is predicted based on the most recent value and the trend is the most recent change in the process. The predicted observation and trend determine the value of the next period in the forecast [36]. Moreover, higher order models such as ARIMA(6,2,2) are developed to fit the data of countries such as Italy as shown in Table 1. ARIMA(6,2,2) for Italy means that the response variable (y) is a combination of 6th (p) order auto-regression model, 2nd (q) order moving average model and the d value of 2 represents the integrative part of the model. Similarly, SARIMA models were developed based on ACF plots of differenced data as described in the case study (Fig. 3). For example, SARIMA(2,2,2)(2,2,1,7) for Peru presented in Table 4, has both second-order seasonal (D) and second-order ordinary differences (d) as indicated by the 2's in the second place of each part of the model. It also has a 2nd order auto-regressive model and 2nd order moving average model along with 2nd order seasonal Auto-Regressive model and 1 seasonally lagged error.

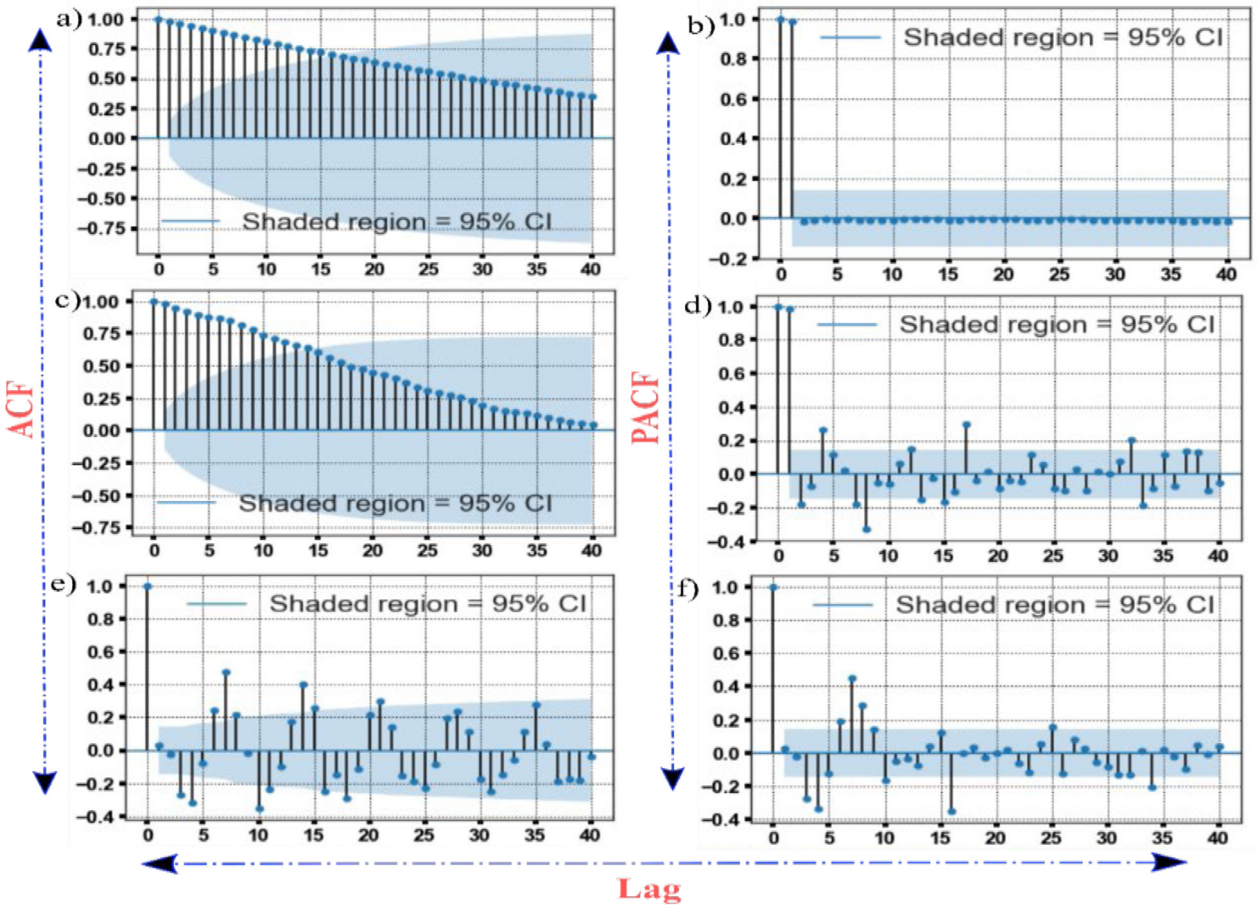

Fig. 3. The Auto-Correlation Function (ACF) and Partial Auto-Correlation Function (PACF) plots of the cumulative confirmed cases of the USA as function of the Lag. The shaded region represents the 95% confidence interval (CI). (A) ACF plot of actual data, (B) PACF plot of the actual data, (C) ACF plot of actual data after first differencing, (D) PACF of data after first differencing (E) ACF of data after second differencing, (F) PACF of the data after second differencing.

In this study, we have reported the final optimized ARIMA and SARIMA models used for forecasting COVID-19 cases of the top-16 countries. The complete details of these models are presented in Tables 1–6 which has the information criteria (AIC, BIC) values, and evaluation metrics (MSE, MAE, RMSE and MAPE) values. Tables 1 and 2 have the details of ARIMA and SARIMA models used to forecast cumulative confirmed cases of top-16 countries. The details of models used to forecast cumulative recovered cases are described in Table 3 (ARIMA models) and Table 4 (SARIMA models). Similarly, Tables 5 and 6 provide details of the selected models for forecasting cumulative death cases for 60 days using ARIMA and SARIMA models, respectively. Such selected ARIMA and SARIMA models were used to forecast the next 60 days from the recent reported date of the COVID-19 cases. The following section presents a case study based on the USA data, and the 60-day forecast of COVID-19 cases for top-16 countries.

4. Forecasting cumulative confirmed cases of USA: A case study

This section describes the detailed forecasting procedure for the USA's confirmed cases. Fig. 3 displays the ACF and PACF plots of the actual time-series data, first order and second order differenced cumulative confirmed cases of the USA. Fig. 3A and 3B are ACF and PACF plots of the actual data, respectively, the auto-correlation coefficients gradually decrease as the number of lags increase (Fig. 3A). This suggests that the data is non-stationary, hence, there is a need to apply the differencing technique to convert the data to stationary. The ACF plot (Fig. 3C) of the time-series after first order differencing, shows the correlation coefficients decreased gradually representing the existence of non-stationarity in the time-series. So, the time-series was differenced for the second time to introduce stationarity. Fig. 3D is the PACF plot of the time-series data after first differencing, it displays a sharp cutoff after lag 1. Moreover, on inspecting Fig. 3E, the ACF plot of the second time differenced time-series shows an oscillation indicating a seasonal series, the sharp significant peak (greater correlation) occurs at lags of 7 days because the data at 22nd January correlates with 29th January and so on. This pattern strongly supports the existence of seasonality in the time-series. This could be because of a greater number of social distancing violations on weekends than on the weekdays. The Fig. 3F is a PACF plot of second-order differenced data displayed a sharp cutoff after lag 0. The second-order differencing indicated an integrated order (I) of 2 must be used in developing the model because taking the second-order differencing made the USA data stationary. Similarly, we did second-order differencing for recovered and death cases to stabilize the datasets whenever required. The ARIMA(0, 2, 0) was the best ARIMA model with the lowest AIC and BIC values 3,113.8 and 3,120.1, respectively. However, while determining the goodness of the fit, the auto-correlation plots of residuals displayed coefficients at higher lags that are significantly different from zero. So, we developed a higher-order ARIMA model (7,2,1) with AIC and BIC values of

Table 3
Selected ARIMA models for forecasting cumulative recovered cases.

Country	ARIMA (p,d,q)	AIC	BIC	MSE	MAE	RMSE	MAPE
South Africa	(3,2,1)	1.77E+03	1.78E+03	2.15E+08	1.10E+04	1.47E+04	5.03E+00
Bangladesh	(6,2,1)	2.08E+03	2.11E+03	2.33E+07	4.00E+03	4.83E+03	3.47E+00
Brazil	(0,2,1)	2.58E+03	2.58E+03	9.00E+08	2.55E+04	3.00E+04	1.64E+00
Chile	(0,2,1)	2.25E+03	2.16E+03	1.90E+08	1.18E+04	1.38E+04	3.81E+00
Columbia	(6,2,1)	1.76E+03	1.79E+03	2.20E+08	1.16E+04	1.48E+04	1.14E+01
India	(10,2,2)	2.49E+03	2.53E+03	3.25E+08	1.01E+04	1.80E+04	1.21E+00
Iran	(1,2,6)	2.08E+03	2.10E+03	2.88E+07	4.48E+03	5.37E+03	1.84E+00
Italy	(5,2,5)	2.22E+03	2.26E+03	2.27E+06	1.22E+03	1.51E+03	6.00E−01
Mexico	(1,2,1)	2.06E+03	2.07E+03	1.66E+08	9.99E+03	1.29E+04	3.65E+00
Pakistan	(6,2,2)	1.25E+03	2.19E+03	2.95E+07	3.97E+03	5.43E+03	1.92E+00
Peru	(3,2,3)	1.98E+03	2.00E+03	9.42E+06	2.50E+03	3.07E+03	9.98E−01
Russia	(6,2,2)	2.51E+03	2.54E+03	1.95E+08	1.23E+04	1.40E+04	2.19E+00
Saudi Arabia	(5,2,2)	1.96E+03	1.98E+03	1.25E+08	1.05E+04	1.12E+04	5.18E+00
Spain	(2,2,2)	2.35E+03	2.37E+03	3.00E−04	1.50E−02	1.70E−02	8.62E+00
UK	(2,1,2)	1.49E+03	1.51E+03	3.85E+03	5.50E+01	6.20E+01	4.06E+00
USA	(2,2,1)	3.18E+03	3.20E+03	2.54E+08	1.16E+04	1.59E+04	9.74E−01

Table 4
Selected SARIMA models for forecasting cumulative recovered cases.

Country	SARIMA (p,d,q)(P,D,Q,m)	AIC	BIC	MSE	MAE	RMSE	MAPE
South Africa	(4,2,2)(3,2,2,7)	1.49E+03	1.52E+03	2.20E+08	1.17E+04	1.48E+04	5.46E+00
Bangladesh	(0,2,1)(1,0,0,7)	1.96E+03	1.96E+03	1.15E+07	2.92E+03	3.40E+03	2.57E+00
Brazil	(2,2,1)(1,1,1,12)	2.26E+03	2.28E+03	5.46E+07	2.13E+03	2.34E+03	3.12E+00
Chile	(1,2,1)(1,0,1,7)	2.28E+03	2.28E+03	8.23E+07	7.80E+03	9.07E+03	2.53E+00
Colombia	(5,2,2)(4,2,2,7)	1.12E+03	1.15E+03	5.71E+04	1.16E+02	2.38E+02	1.30E−01
India	(7,2,6)(3,2,6,3)	1.98E+03	2.05E+03	1.46E+08	7.15E+03	1.21E+04	8.70E−01
Iran	(4,2,4)(3,1,2,3)	1.86E+03	1.87E+03	2.16E+07	3.80E+03	4.65E+03	9.50E−02
Italy	(4,2,2)(1,1,1,7)	1.96E+03	1.98E+03	3.78E+06	1.79E+03	1.94E+03	9.23E−01
Mexico	(1,2,1)(1,0,1,7)	1.93E+03	1.93E+03	1.19E+08	8.82E+03	1.09E+04	3.24E+00
Pakistan	(4,2,2)(2,1,2,7)	1.97E+03	1.98E+03	9.44E+04	1.58E+02	3.07E+02	7.80E−02
Peru	(2,2,2)(2,2,1,7)	1.53E+03	1.55E+03	5.82E+06	1.99E+03	2.41E+03	7.70E−01
Russia	(5,2,0)(1,0,1,7)	2.51E+03	2.54E+03	1.95E+08	1.23E+04	1.40E+04	9.70E+01
Saudi Arabia	(5,2,2)(4,0,2,7)	1.47E+03	1.51E+03	1.32E+09	1.08E+04	1.15E+04	8.00E−02
Spain	(1,1,3)(2,0,1,7)	2.35E+03	2.37E+03	5.00E−04	2.00E−02	2.00E−03	9.88E−01
UK	(4,1,2)(2,0,1,3)	1.41E+03	1.44E+03	2.10E+02	1.25E+01	1.45E+01	8.00E−01
USA	(2,2,2)(1,0,1,7)	2.99E+03	3.01E+03	6.03E+08	1.72E+04	2.46E+04	1.41E+00

3,025 and 3,056 respectively (Table 1) for prediction and forecasting the cumulative confirmed cases in the USA. An ARIMA(7,2,1), the auto-correlation plots of the residuals did not display lags that are significantly different from zero as shown in Figure S1B, S2B.

To further investigate the ARIMA(7,2,1) model, we have used the Quantile–Quantile (Q–Q) plot and the probability density Q–Q plot was constructed using the residuals (Fig. 4). The residual errors have a normal distribution as shown in Figs. 4B and 4D the linear plot of residuals with respect to quantiles follow a linear relationship except few blue dots at the ends but all other dots lie close to the straight line. This bell-shaped distribution of residuals suggests that the data came from a normal distribution. Higher-order model ARIMA(7,2,1) was selected and used to predict the cumulative cases and forecast to the near future. However, when we considered seasonality in the model i.e. SARIMA(3,2,4)(2,1,4,7), the Q–Q plot displayed lesser outliers at the tails when compared to ARIMA(7,2,1). The Kernel Density Estimate Plot (KDE) of the residuals of ARIMA(7,2,1) and SARIMA(3,2,4)(2,1,4,7) has a gaussian-like distribution but it is sharper suggesting an asymmetric exponential distribution as shown in Fig. 4(B & D). Moreover, the KDE plot of SARIMA(3,2,4)(2,1,4,7) (Fig. 4D) shows that the distribution of residuals is more normal/gaussian than that of ARIMA(7,2,1) (Fig. 4B). The results of diagnostic plots (Q–Q plot and KDE plot) of the residuals are in strong support of choosing SARIMA(3,2,4)(2,1,4,7) as the better model to fit with zero-auto correlated errors as shown in ACF and PACF plots of the residuals in Figures S3B & S4B (supporting document).

The Fig. 5 displays the comparison between the test data (20% of the actual data) and the predictions of the test data obtained by ARIMA, SARIMA models, along with the LSTM (Long–Short Term Memory) and GRU (Gated Recurrent Unit) models, developed in our recent work [37]. The model evaluation was carried out by calculating the MAE, MSE, RMSE and MAPE using the Eqs. (18), (19), (20) and (21), respectively. The calculated errors are reported in Tables 1 and 2.

From Fig. 5, it is evident that both the ARIMA and SARIMA models predicted the test data reasonably well. Further, SARIMA model outperformed the complex deep learning models such as LSTM and GRU models confirming that the simple machine learning models are sufficient to accurately predict the test data. The predictions of SARIMA(3,2,4)(2,1,4,7) matched the test data better than the ARIMA(7,2,1) predictions. Therefore, the prediction for test data and forecast for the next 60 days of cumulative confirmed cases of the USA was done using the ARIMA(7,2,1) and SARIMA(3,2,4)(2,1,4,7) models. Fig. 6(B) and Fig. 7(B) shows the 60-day forecast with 95% (CI) using ARIMA(7,2,1) and SARIMA(3,2,4)(2,1,4,7), respectively. Both models' forecast suggests that the USA's actual cumulative confirmed cases might continue to increase exponentially in 60 days. Our best forecast ARIMA and SARIMA models for the USA projects the number of cumulative confirmed cases might reach 7.5 million by the end of September. According to ARIMA(7,2,1) the cumulative

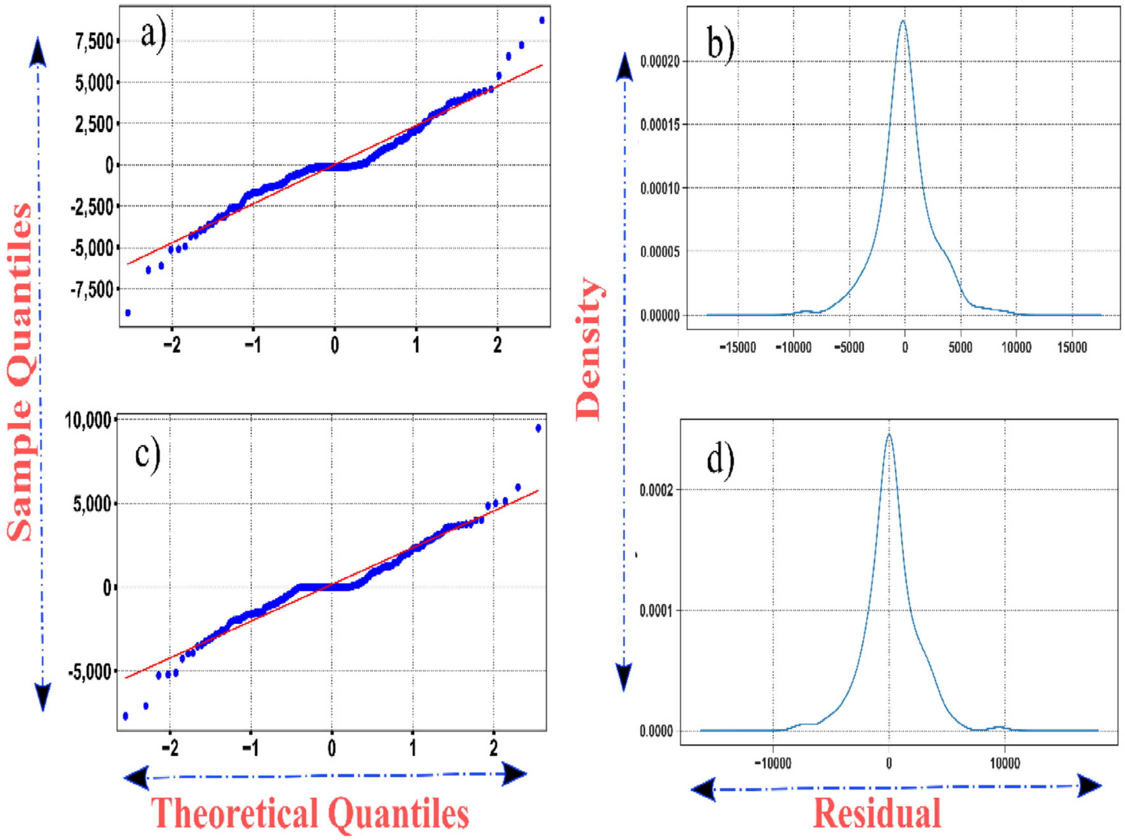

Fig. 4. Diagnostics for the models used in the case study — prediction and forecast of cumulative confirmed cases of the USA. (A) Normal Q–Q plot of residuals of ARIMA(7,2,1), (B) KDE of residuals of ARIMA(7,2,1), (C) Normal Q–Q plot of residuals of SARIMA(3,2,4)(2,0,1,7), (D) KDE of residuals of SARIMA(3,2,4)(2,0,1,7).

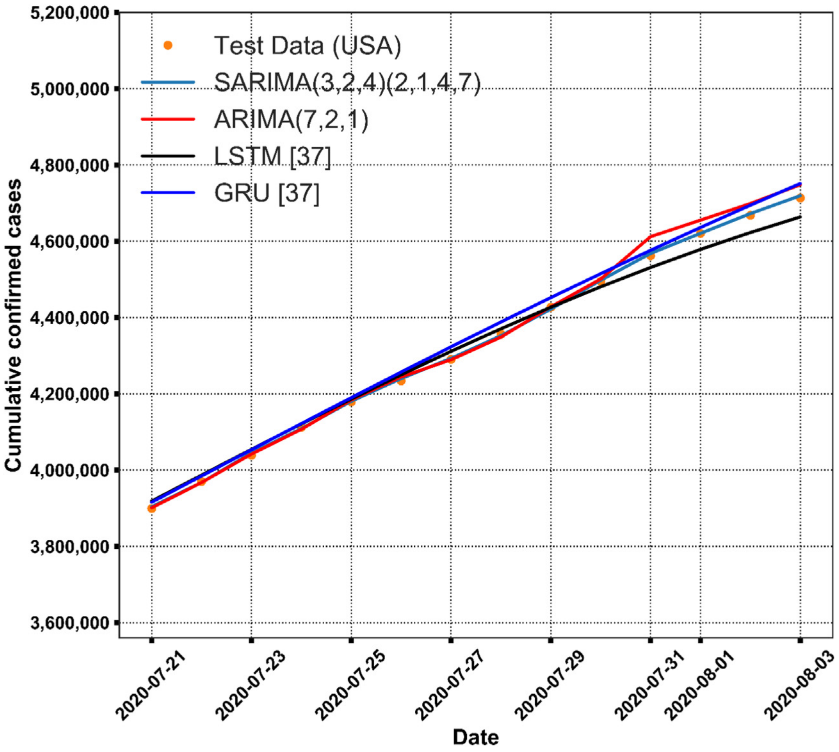

Fig. 5. Comparison of ARIMA and SARIMA models' predictions with test data of cumulative confirmed cases in the USA.

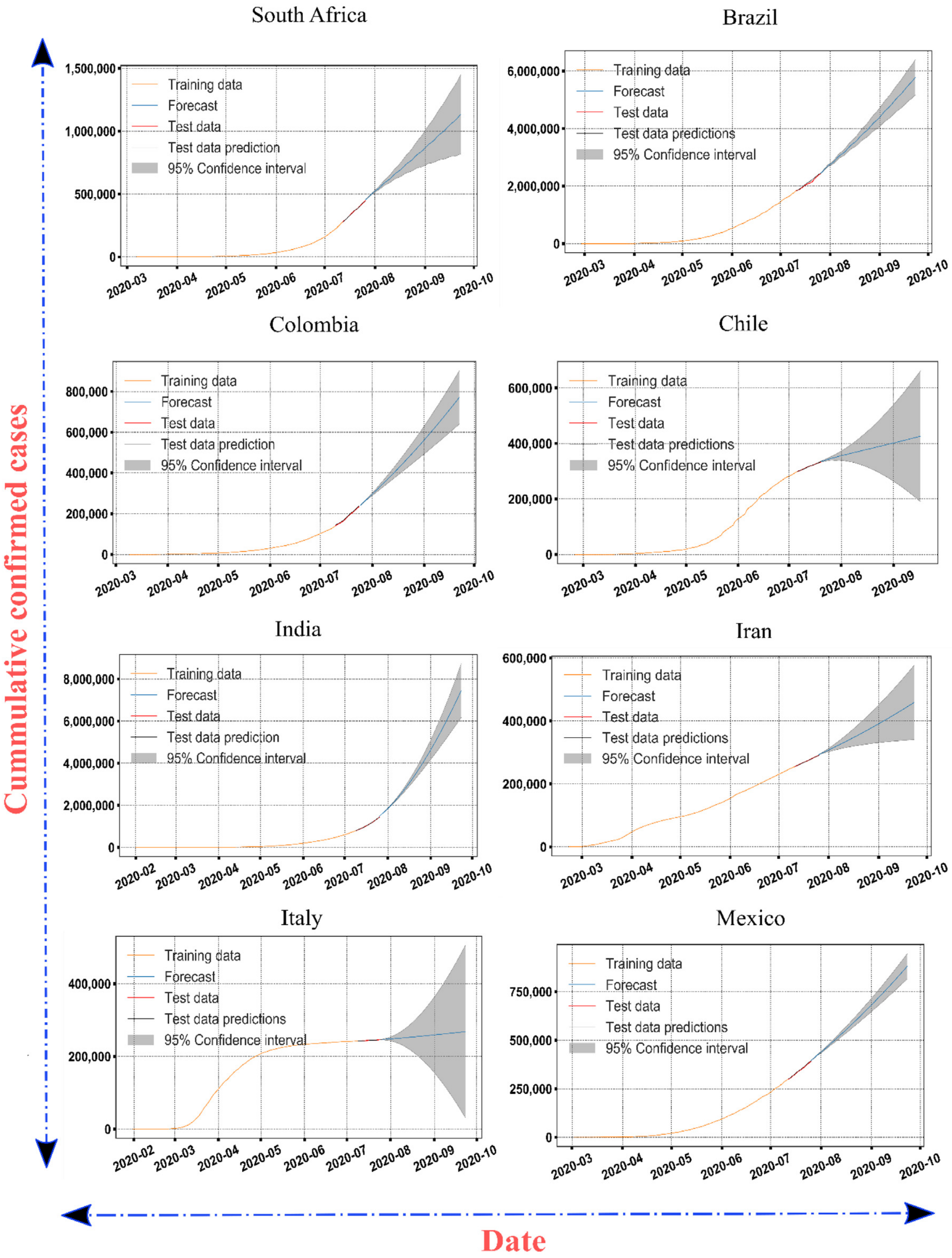

Fig. 6(A). 60-day ahead forecast of the cumulative confirmed cases in the top-16 countries (1–8) generated on July 26th of 2020 based on the best ARIMA models selected for each country.

confirmed cases will increase to 6,478,221 on September 1st, 2020. Whereas SARIMA(3,2,4)(2,1,4,7) indicates that the cases will be 2,677 lesser than what ARIMA(7,2,1) predicted on 1st September. Before forecasting 60 days into the future, a similar robust analysis was done for all three (confirmed, recovered, deaths) cumulative COVID-19 cases for proposing an optimized model for each country in top-16 countries.

5. Results and discussion

The percentage distribution of cumulative COVID-19 cases (confirmed cases, recovered cases and deaths) of the top-16 countries are present as shown in Fig. 1. We selected top-16 countries based on the number of cumulative confirmed cases, the top-16 countries include the USA, Brazil, India, Russia, Peru, Chile,

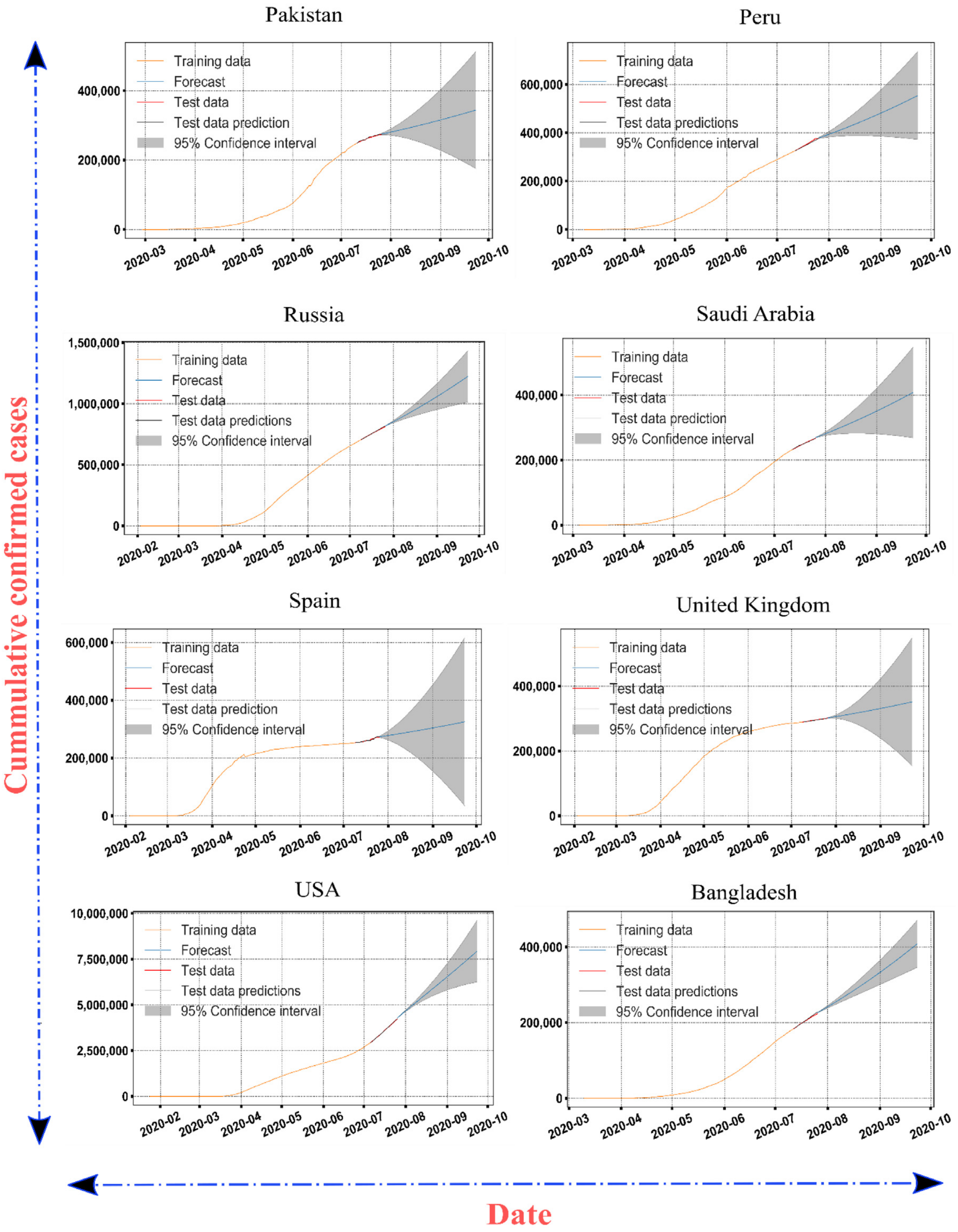

Fig. 6(B). 60-day ahead forecast of the cumulative confirmed cases in the top-16 countries (9–16) generated on July 26th of 2020 based on the best ARIMA model selected for each country.

Mexico, the UK, South Africa, Iran, Spain, Pakistan, Italy, Saudi Arabia, and Turkey as of 25th July. From Fig. 1A, it is evident that out of 16 countries the USA had 26.4% of global cumulative confirmed cases followed by Brazil (15.1%), India (9.0%), Russia (5.1%) and South Africa (2.8%). This work is accounted for 78.4% of the total confirmed cases (≈16.5 M) which are reported in top-16 countries but not accounted for those reported in the rest of the world. The country-based percentage distribution of the cumulative recovered cases is given in Fig. 1B. The order of countries with high to low recovered cases is as follows: USA (18.7%), Brazil (13.4%), India (9.5%), Russia (6.2%), South Africa (3.3%), Mexico (3.1%), Peru (2.8%), Chile (2.7%), UK (2.6%), Bangladesh (2.5%), Iran (2.3%), Pakistan (2.2%), Spain (2.1%), Saudi Arabia (2.0%), Colombia (1.6%) and Italy (1.3%). The total percentage of recovered cases recorded by these countries is 76% of the global cases (≈9.4

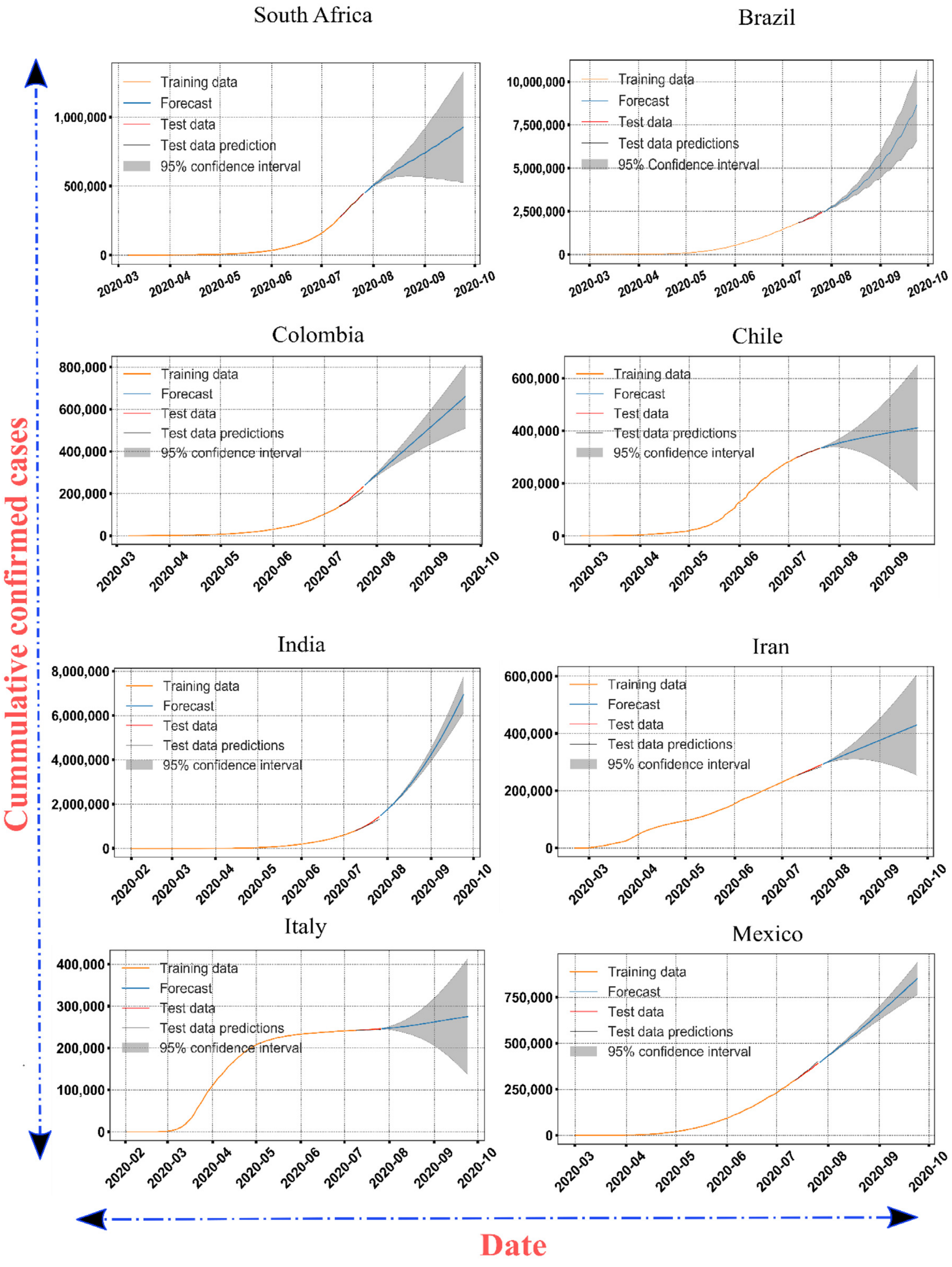

Fig. 7(A). 60-day ahead forecast of the cumulative confirmed cases in the top-16 countries (1–8) generated on July 26th of 2020 based on the best SARIMA models selected for each country.

M). Similarly, total deaths reported by the top-16 countries are account for 78.7% of the global deaths (≈650,000). Countries have such as the USA, Brazil, India, Russia, South Africa have reported a high number of deaths. The present work has accounted for approximately 80% of the global confirmed cases, recovered cases and deaths for developing a reliable statistical model. Hence, the results from these models can be used for predicting the COVID-19 trends in other countries, which are not considered in this work, as well as for forecasting the global COVID-19 cases.

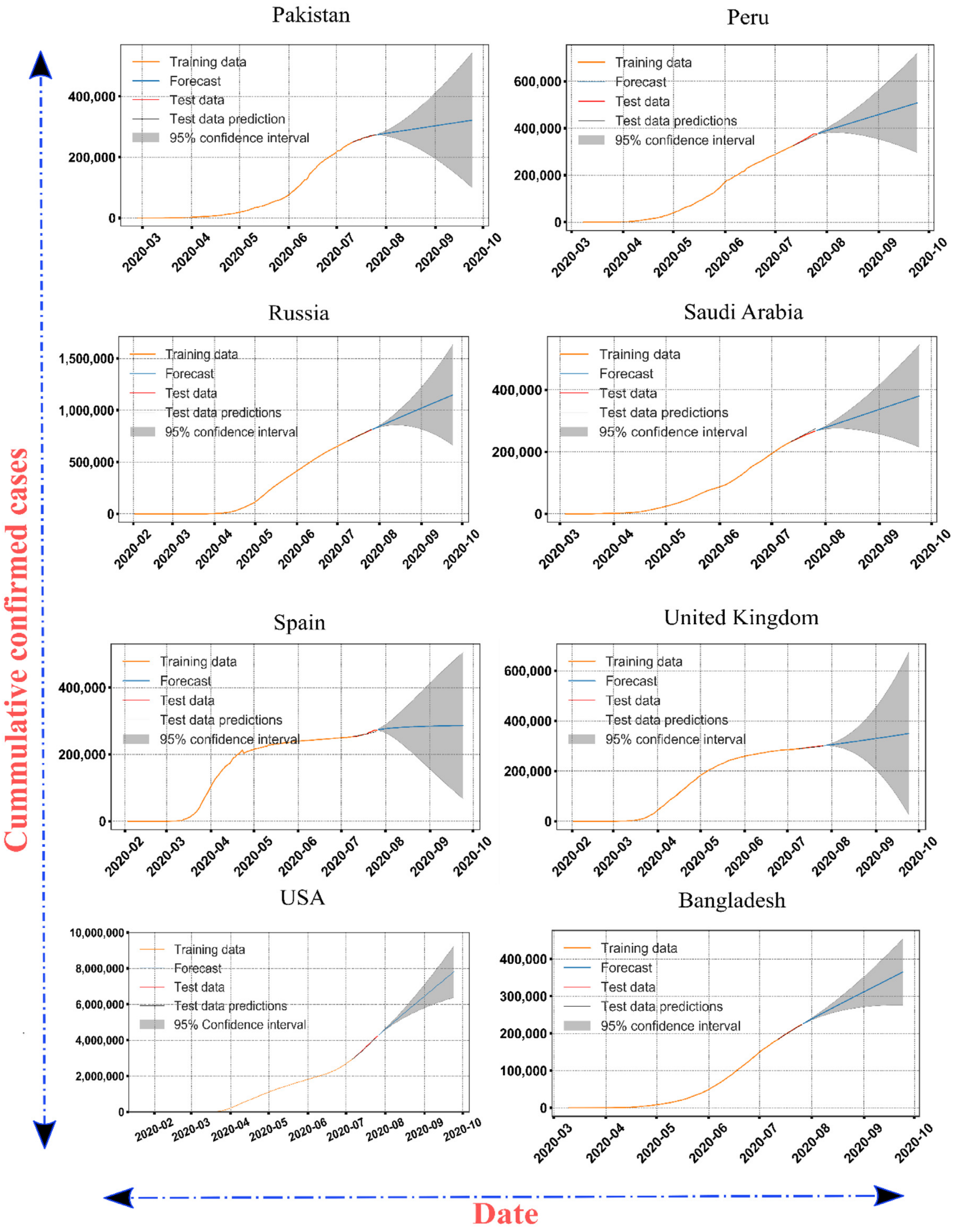

Fig. 7(B). 60-day ahead forecast of the cumulative confirmed cases in the top-16 countries (9–16) generated on July 26th of 2020 based on the best SARIMA models selected for each country.

5.1. Cumulative confirmed cases

The 60-day forecast of confirmed cases for top-16 countries are shown in Figs. 6 & 7. It is important to mention that the yellow line represents the reported data, the blue line represents the forecasted data and the shaded region is the 95% Confidence Interval (CI) of the forecasted data. Fig. 6 displays the 60-day forecast of the top-16 countries based on ARIMA modes whereas Fig. 7 shows the SARIMA based models' forecasts of cumulative confirmed cases of top-16 countries. From Fig. 6(A), it is evident that South Africa will have a cumulative confirmed case of ≈1,100,000 by September 22nd of 2020. The forecast for Brazil has an exponential trend with a narrow 95% CI, the ARIMA model predicted that the cumulative confirmed cases will be ≈5,900,000 by the end of the 2nd week of September. Similarly, a 60-day forecast of Colombia reveals that the number

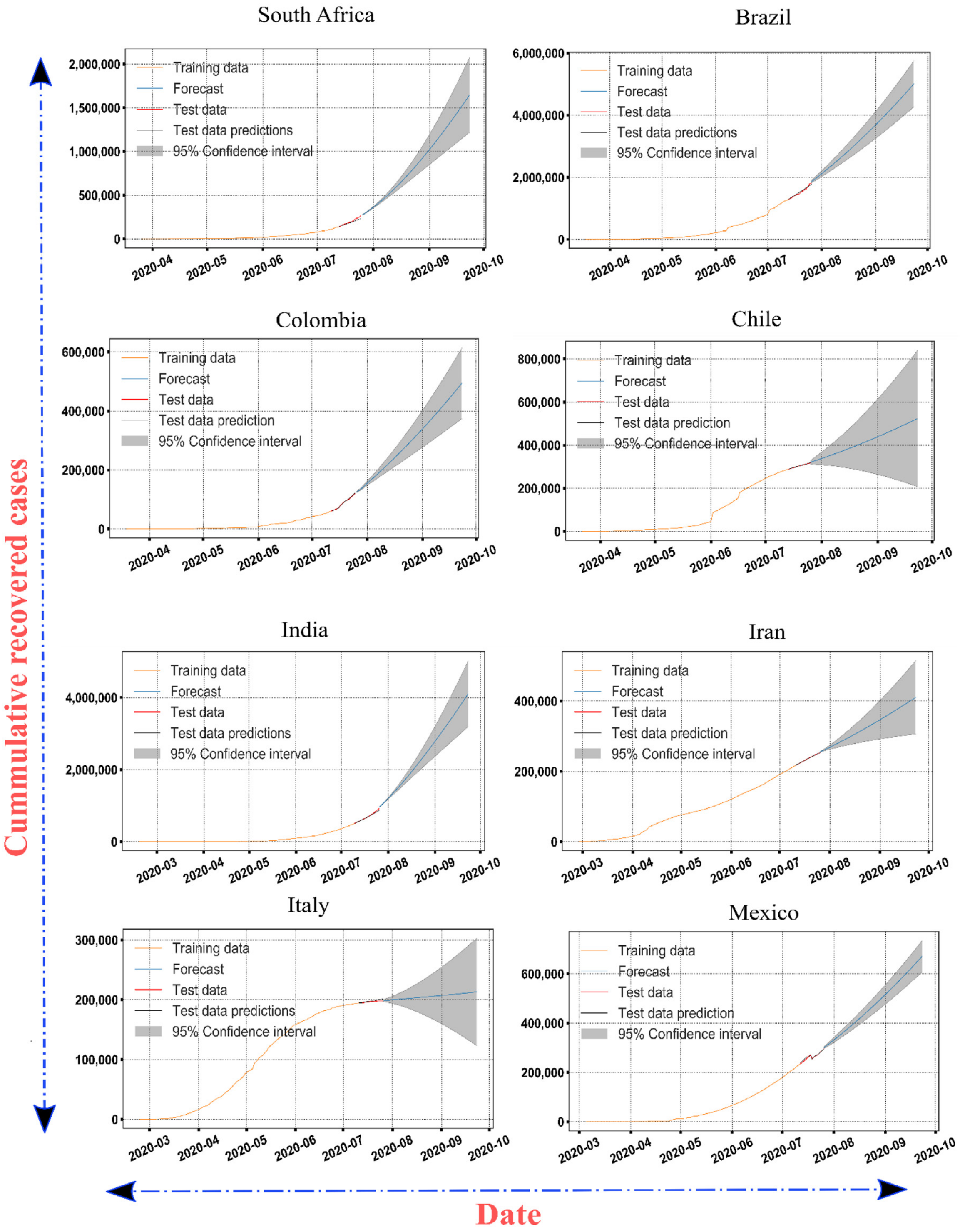

Fig. 8(A). 60-day ahead forecast of the cumulative recovered cases in the top-16 countries (1–8) generated on July 26th of 2020 based on the best ARIMA models selected for each country.

of cumulative confirmed cases will reach ≈799,000 by the 3rd week of September, upper and lower limits of the 95% CI of the forecast strongly follows the exponential growth of confirmed cases as seen in Fig. 6(A). A similar exponential trend of the number of confirmed cases is observed in the case of the USA, South Africa, Colombia, Brazil, India, Mexico, and Bangladesh (Figs. 6(A) and 6(B)). However, the forecast of Saudi Arabia, Pakistan, Chile, Russia, Peru, Iran shows a steep linear increment in the number of cumulative confirmed cases at a steady pace. Italy, the UK and Spain's forecast showed very steady linear increment in the number of cumulative confirmed cases. The selected ARIMA models projected the number of cases in South Africa, Brazil, Colombia, Chile, India, Iran, Italy, Mexico, Pakistan, Peru, Russia, Saudi Arabia, Spain, UK, USA, and Bangladesh will be ≈1,750,000, ≈5,800,000, ≈799,000, ≈401,000, ≈6,900,000, ≈425,000, ≈290,000, ≈820,000, ≈325,000, ≈505,000, ≈1200000,

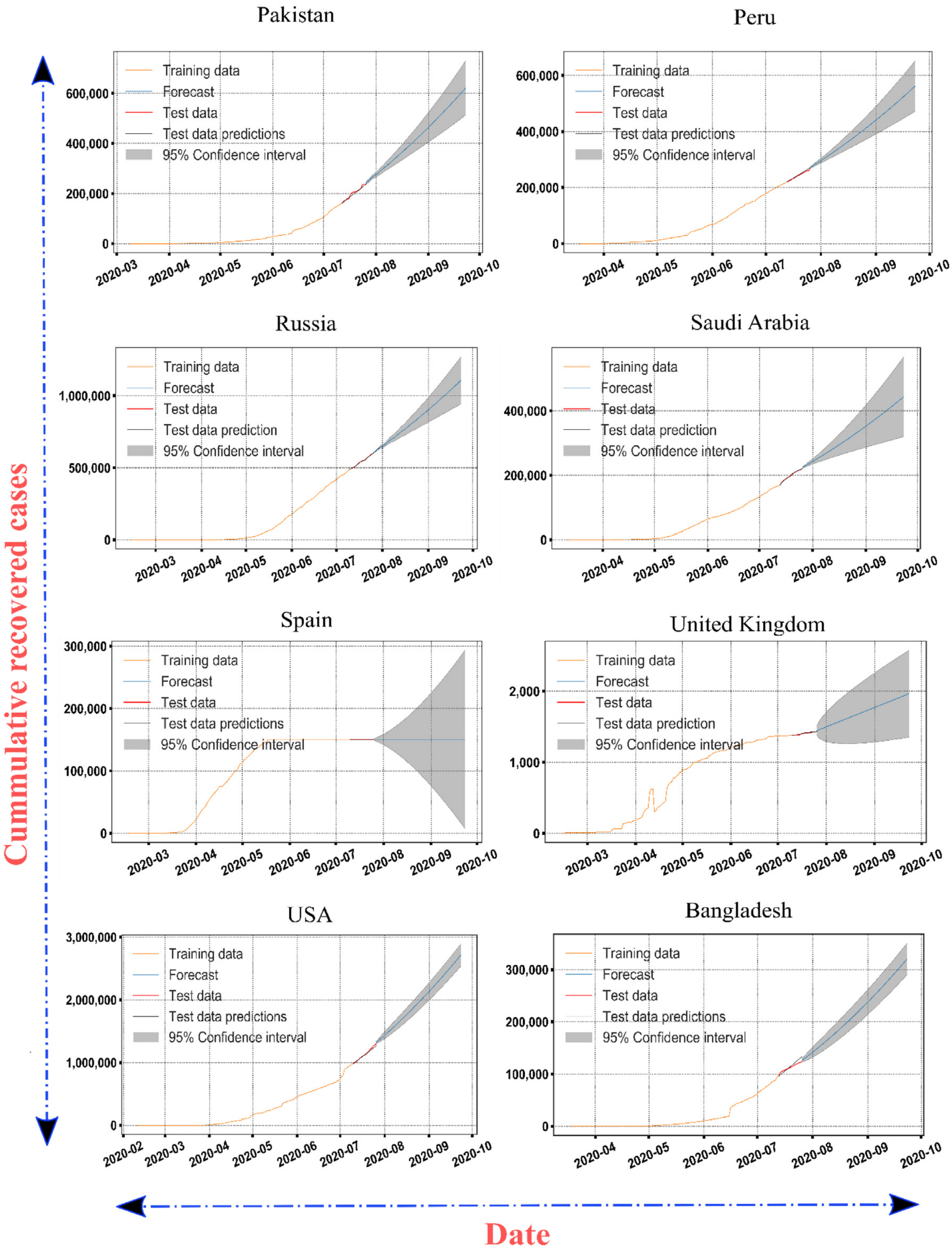

Fig. 8(B). 60-day ahead forecast of the cumulative recovered cases in the top-16 countries (9–16) generated on July 26th of 2020 based on the best ARIMA models selected for each country.

≈401,000, ≈301,000, ≈330,000, ≈7,850,000 and ≈410,000 respectively according to Figs. 6(A) and 6(B).

When seasonality is considered for SARIMA models, the seasonal forecast of the confirmed cases has captured the variance and seasonality in the time-series and projected well into the forecast. This is more evident from the forecast of Brazil as shown in Fig. 7(A). The SARIMA(3,2,2)(2,1,17) of Brazil has better captured the seasonality when compared to non-seasonal forecast of Brazil as shown in Fig. 6(A). The forecasted data is capable of recognizing the continuous seasonal patterns of the reported data. The number of cumulative confirmed cases predicted by SARIMA(3,2,2)(2,1,1,7) of Brazil is 2,000,000 greater than the ARIMA predicted cumulative confirmed cases by the end of 2nd

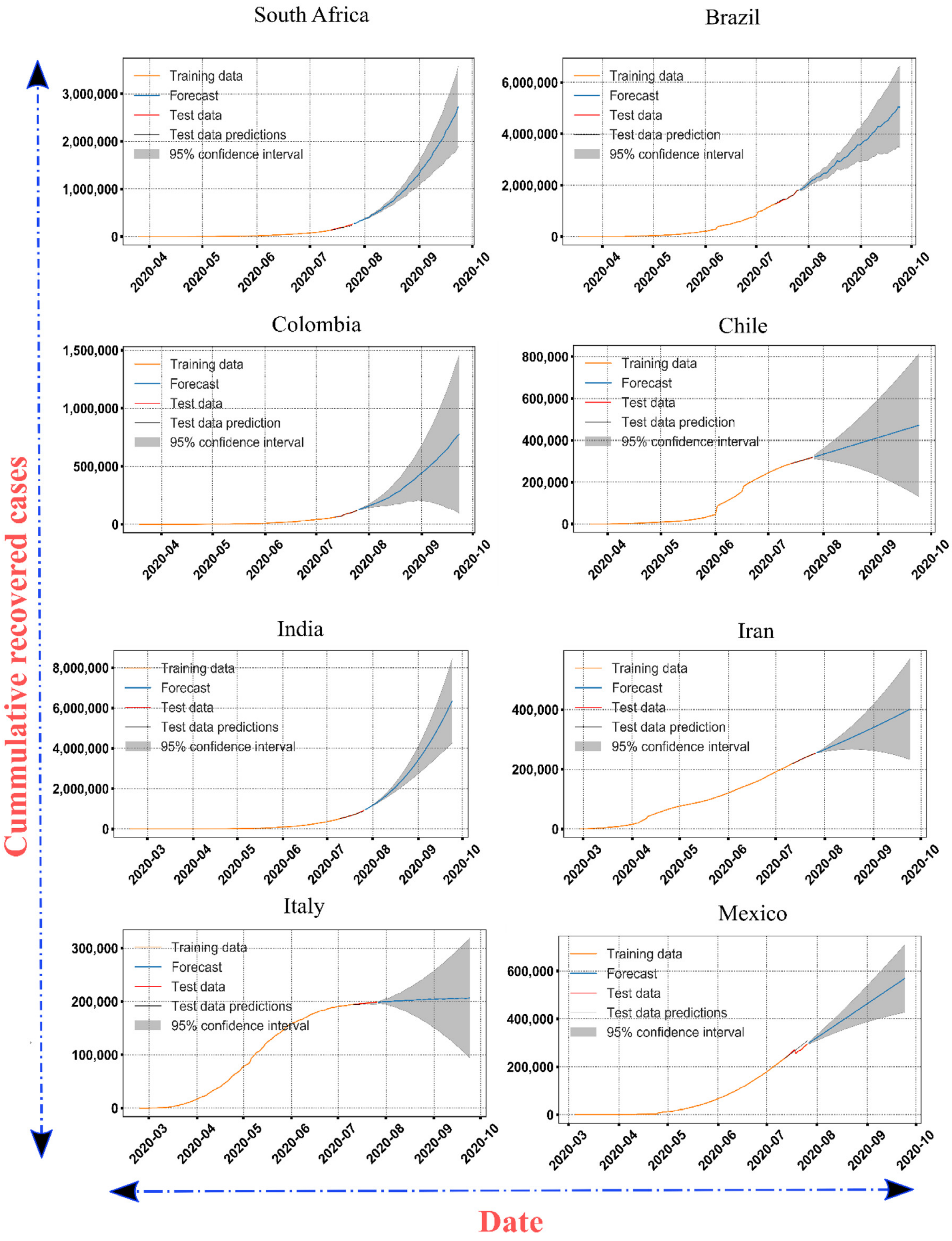

Fig. 9(A). 60-day ahead forecast of the cumulative recovered cases in the top-16 countries (1–8) generated on July 26th of 2020 based on the best SARIMA models selected for each country.

week of September. The SARIMA models' predicted number of cumulative confirmed cases of most of the countries are lesser than that of the ARIMA models' predictions. Such countries include South Africa with ≈109,000, Colombia with ≈169,000, India with ≈200,000, Iran with 30,000, Mexico with 45,000, Russia with ≈100,000, Saudi Arabia with ≈20,000, UK with ≈35,000, USA with ≈350,000 and Bangladesh with ≈30,000 lesser cumulative confirmed cases when compared to cumulative confirmed cases predicted by their respective ARIMA models. Further, the countries including USA, Peru, Pakistan, Iran, Italy and Chile has broad 95% CI even after 3 weeks of forecast (Fig. 7(B)). The lower limit of forecast's 95% CI of these countries indicates the decline in the number of confirmed cases, whereas the upper limit indicates the rapid exponential raise in the number of confirmed cases.

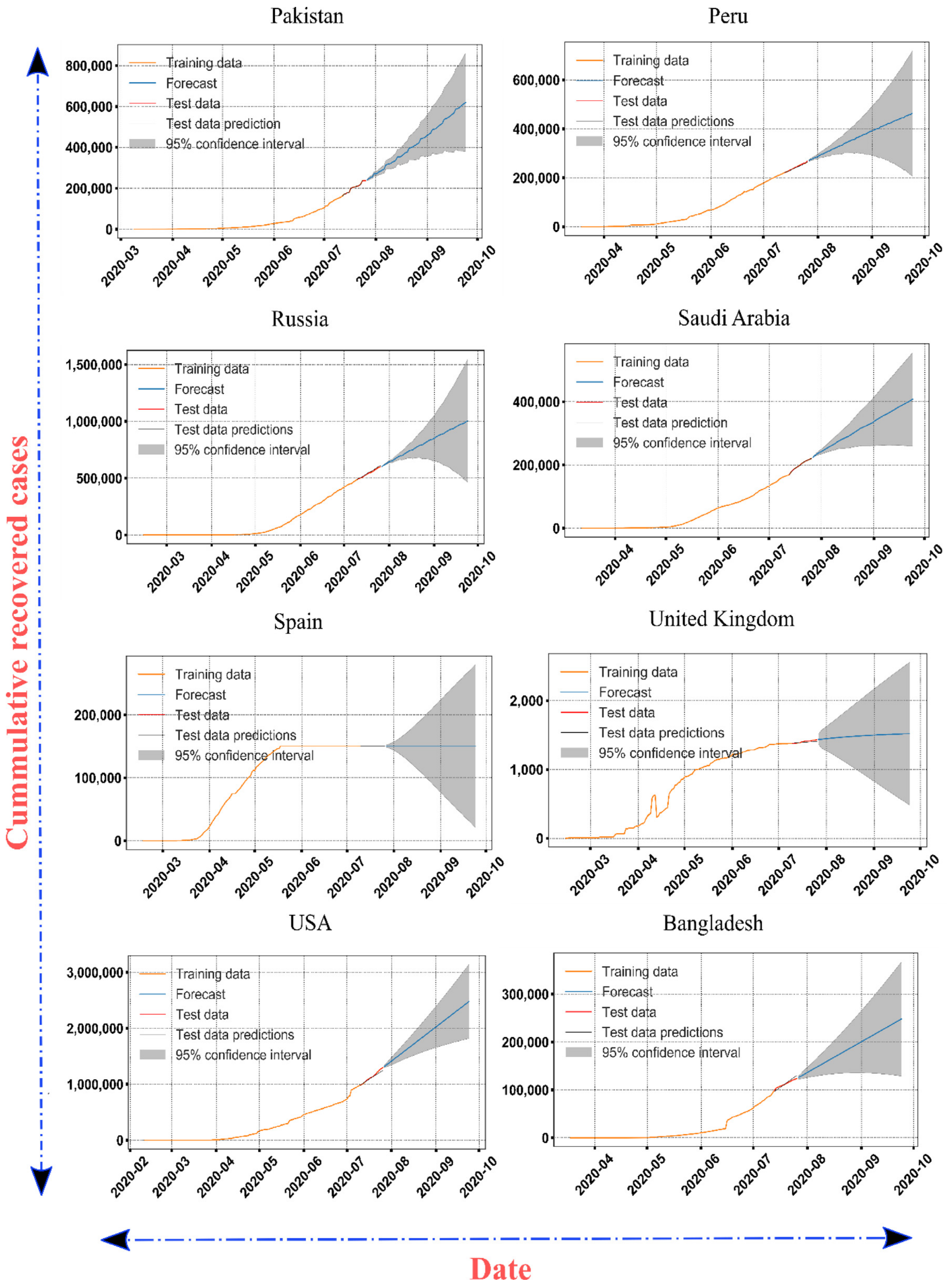

Fig. 9(B). 60-day ahead forecast of the cumulative recovered cases in the top-16 countries (9–16) generated on July 26th of 2020 based on the best SARIMA models selected for each country.

Certainly, most of the countries are reopening and loosening the COVID-19 restrictions, we can see the rapid rise in the confirmed cases in the next 60 days as suggested by the upper limit and not the significant declining trend predicted by the models. Due to the relaxation of preventive measures such as lockdowns, social distancing and reopening of restaurants, and other local businesses, the fast-rising infection rates may lead to an exponential growth of COVID-19 victims in these countries, the effect of the reopening of the economy is clearly visible in various countries.

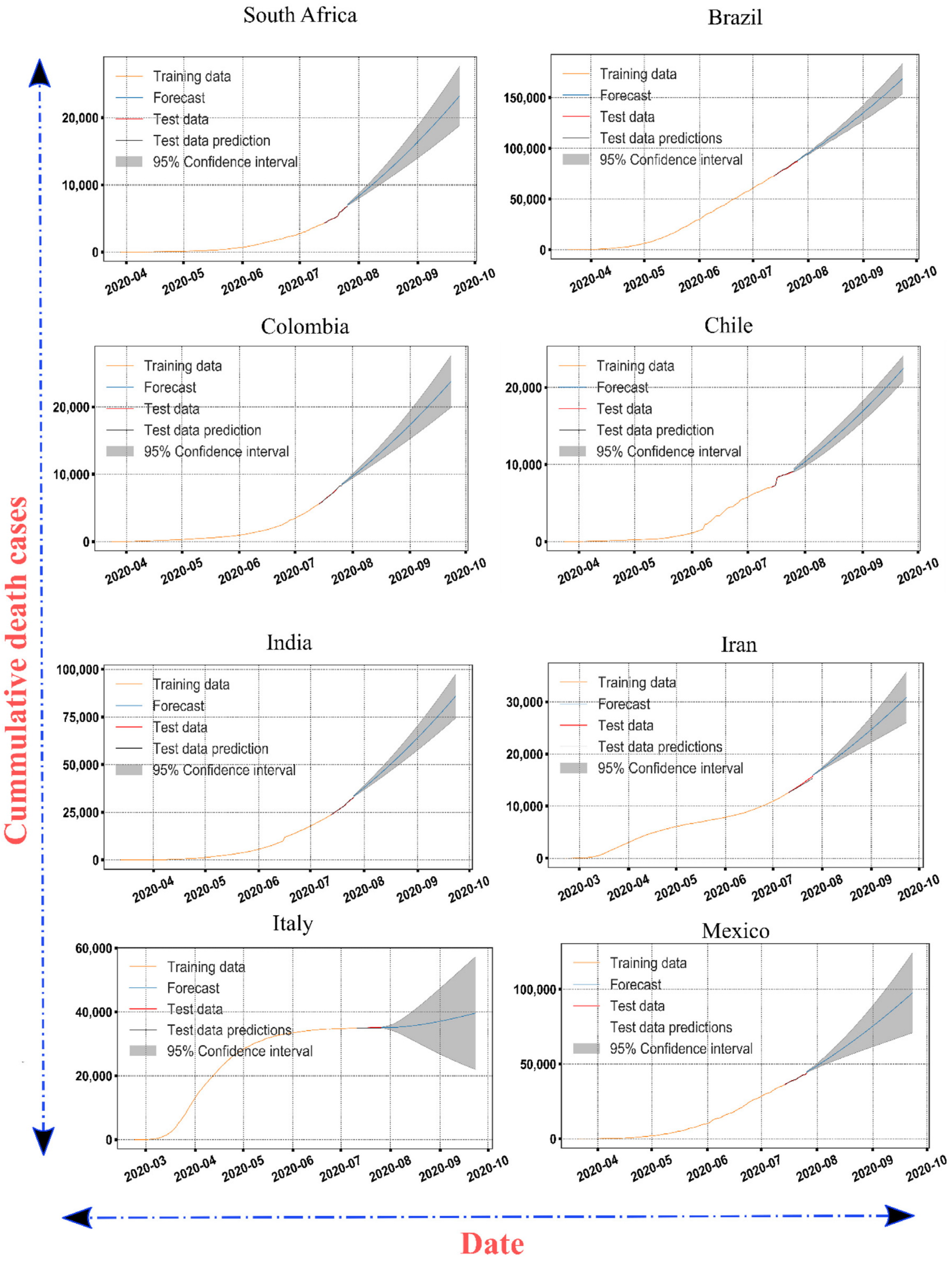

Fig. 10(A). 60-day ahead forecast of the cumulative death cases in the top-16 countries (1–8) generated on July 26th of 2020 based on the best ARIMA models selected for each country.

For example, the cumulative cases in India were less at the beginning of the pandemic, which was due to the implementation of the lockdown (April–May 2020). However, the cases rise as soon as the lockdown was removed (June to August 2020). Whereas in the USA, the reaction was relatively slow toward COVID-19, leading to a continuous raise in COVID-19 cases. Similarly, Iran has noticed a significant drop in new cases after implementing stringent lockdown policies, Iran reopened in April 2020, due to which the number of COVID-19 cases in Iran skyrocketed again in May 2020 [18]. The forecast for cumulative cases in Iran suggests that the cases might reach ≈450,000 by the second week of September. Similarly, the forecasts of confirmed cases

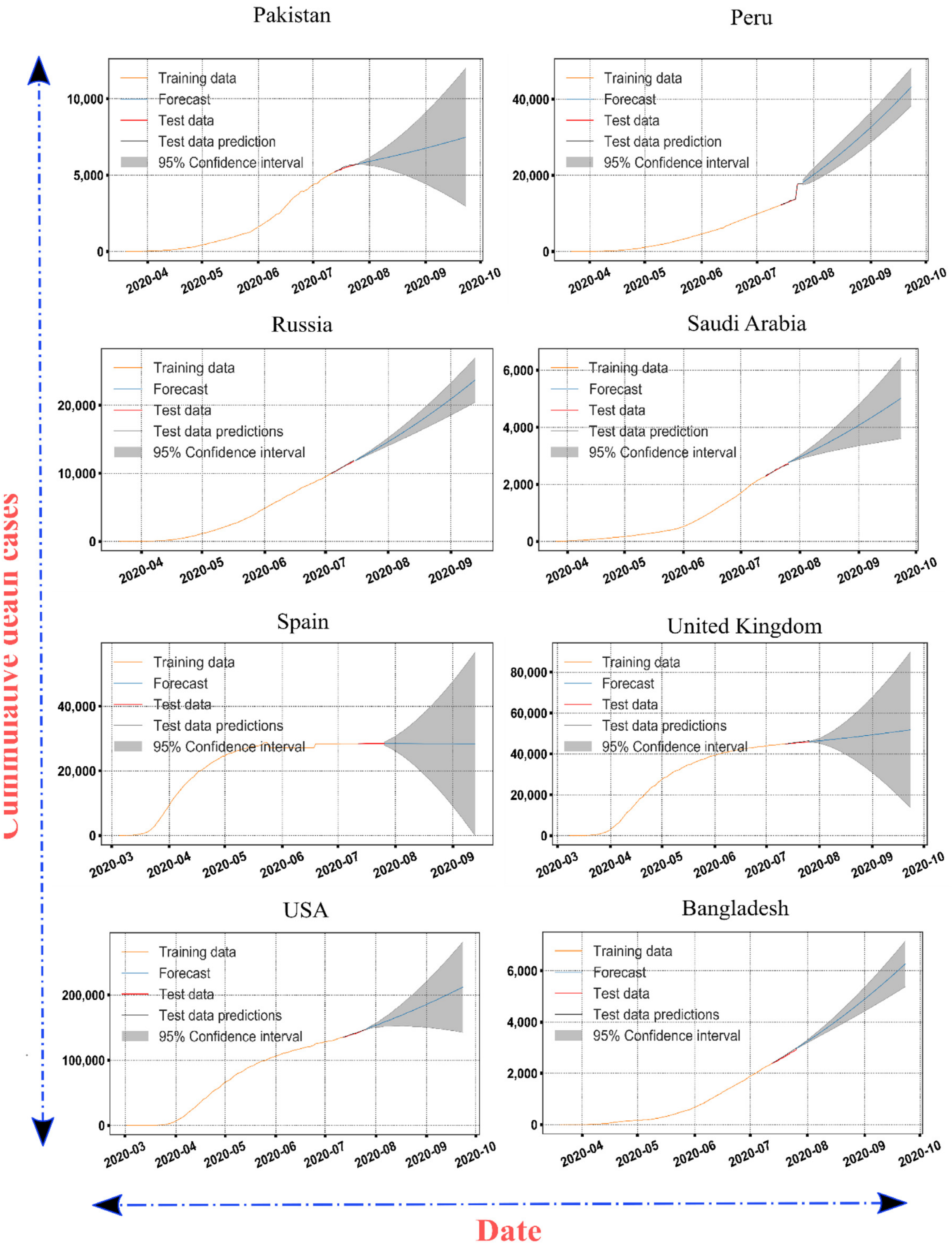

Fig. 10(B). 60-day ahead forecast of the cumulative death cases in the top-16 countries (9–16) generated on July 26th of 2020 based on the best ARIMA models selected for each country.

in India suggests that the cumulative confirmed cases will reach ≈65,00,000 according to ARIMA(5,2,2) but when we considered the seasonality, SARIMA(2,2,1)(1,1,1,7), the projected cumulative confirmed cases will be ≈70,00,000 which is ≈500,000 greater than the ARIMA(5,2,2) model's prediction by 3rd week of September 2020 (Figs. 6(A) & 7(A)). Similar to ARIMA models, three different trends in the forecasted profiles such as exponential rise (USA, South Africa, Colombia, Brazil, India, Mexico, and Bangladesh), steep linear increment (Saudi Arabia, Pakistan, Chile, Russia, Peru, Iran) and gradual linear increment (Italy, UK and Spain) are observed. As shown in Figs. 7(A) and 7(B), the selected SARIMA models projected number of cases of South

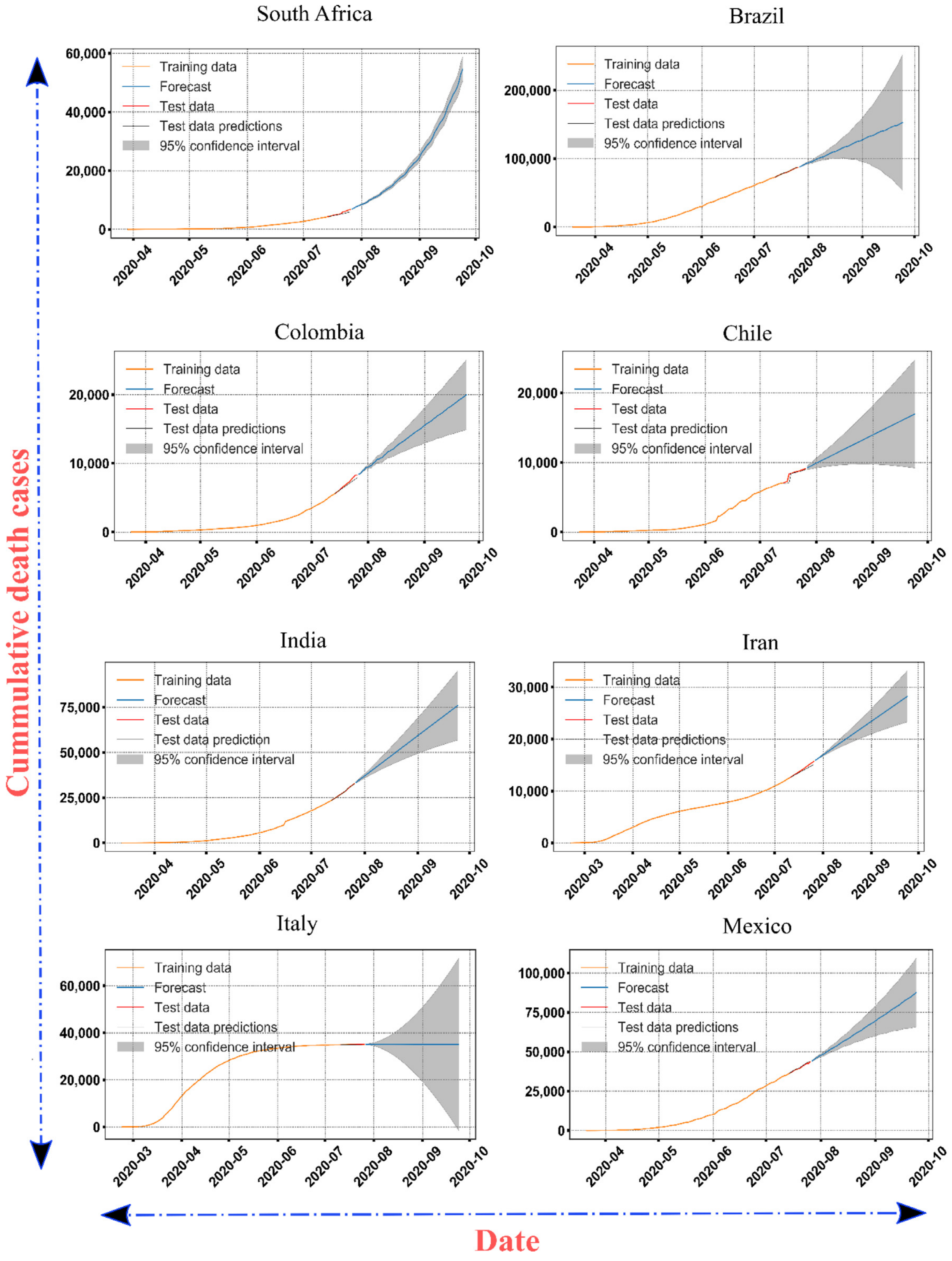

Fig. 11(A). 60-day ahead forecast of the cumulative death cases in the top-16 countries (1–8) generated on July 26th of 2020 based on the best SARIMA models selected for each country.

Africa, Brazil, Colombia, Chile, India, Iran, Italy, Mexico, Pakistan, Peru, Russia, Saudi Arabia, Spain, UK, USA, and Bangladesh will be ≈950,000, ≈8,500,000, ≈625,000, ≈405,000, ≈6,250,000, ≈410,000, ≈220,000, ≈800,000, ≈356,000, ≈440,000, ≈1,500,000, ≈395,000, ≈300,000, ≈375,000, ≈7,600,000, ≈400,000 respectively. To avoid the surge in the number of new confirmed cases, local businesses, schools etc. should follow the guidelines for organizing events and gatherings as published by the Center for Disease Control and Prevention (CDC) [38].

5.2. Cumulative recovered cases

The ARIMA based forecasted trends of cumulative recovered cases for all the top-16 countries are given in Fig. 8. It is clear

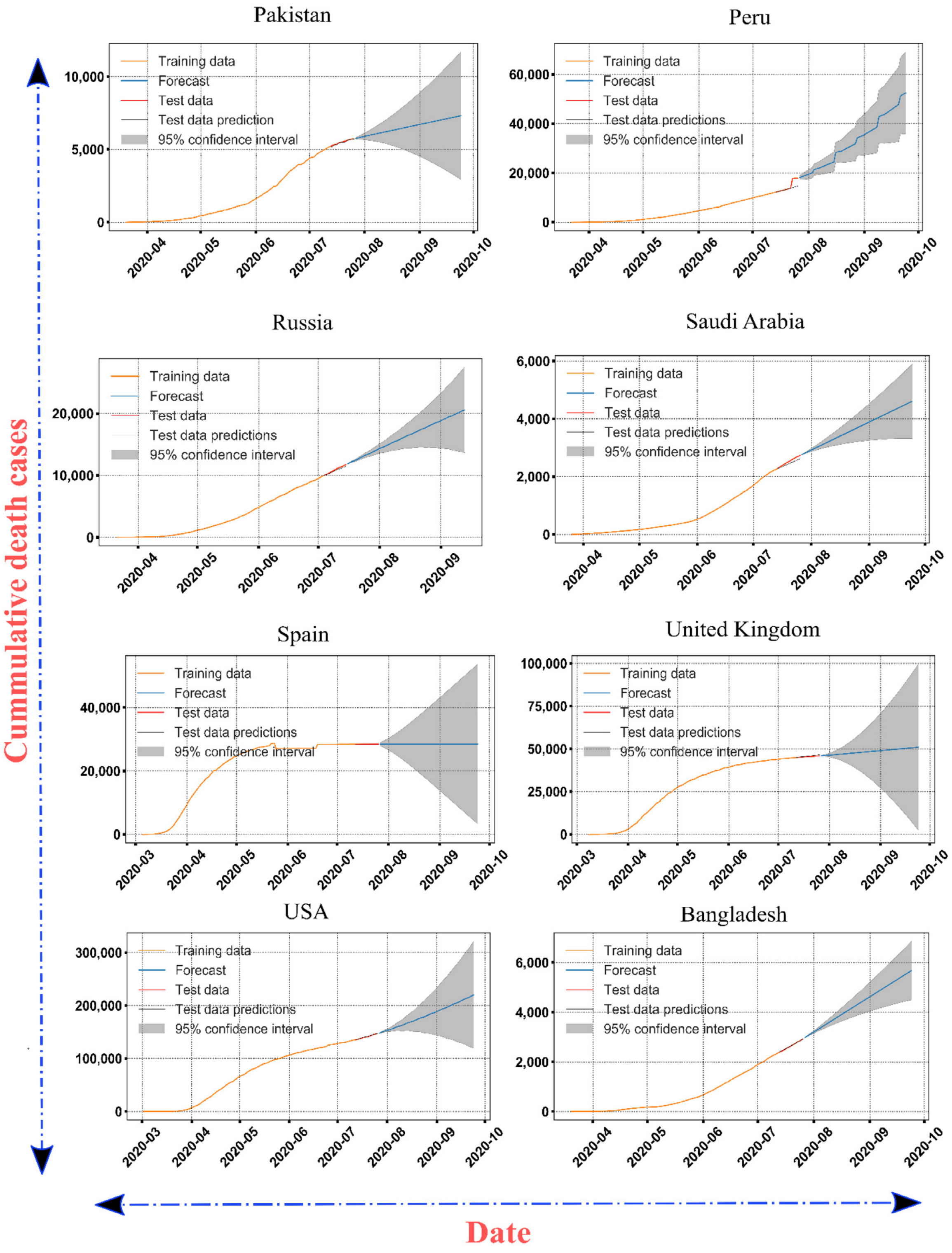

Fig. 11(B). 60-day ahead forecast of the cumulative death cases in the top-16 countries (9–16) generated on July 26th of 2020 based on the best SARIMA models selected for each country.

from Fig. 8, that the recovery rate has shown three different trends such as exponential rise, steep linear increase and gradual linear increase were observed. After reviewing Fig. 8, the selected ARIMA models projected that the number of recovered cases in South Africa, Brazil, Colombia, Chile, India, Iran, Italy, Mexico, Pakistan, Peru, Russia, Saudi Arabia, Spain, UK, USA, and Bangladesh will reach ≈1,600,000, ≈4,500,000, ≈425,000, ≈475,000, ≈4,000,000, ≈410,000, ≈210,000, ≈650,000, ≈610,000, ≈585,000, ≈1,150,000, ≈410,000, ≈150,000, ≈1,850, ≈2,510,000, ≈325,000 by the end of the September 2nd week, respectively. Fig. 9 reports the forecasted recovered cases of 16 countries based on the SARIMA model, it is evident that the recovered cases in South Africa, Brazil, Colombia, India, Mexico, Pakistan, the USA, and Bangladesh are increasing exponentially.

Table 5
Selected ARIMA models for forecasting cumulative death cases.

Country	ARIMA (p,d,q)	AIC	BIC	MSE	MAE	RMSE	MAPE
South Africa	(0,2,1)	9.27E+02	9.36E+02	1.30E+03	2.65E+01	3.60E+01	4.62E−01
Bangladesh	(2,2,2)	7.18E+02	7.30E+02	1.65E+03	3.70E+01	4.00E+01	1.37E+00
Brazil	(6,2,3)	1.45E+03	1.48E+03	2.04E+05	3.55E+02	4.52E+02	4.26E−01
Chile	(0,2,1)	1.27E+03	1.28E+03	2.04E+03	3.60E+01	4.52E+01	4.20E−01
Columbia	(1,2,1)	1.22E+03	1.23E+03	4.01E+02	1.40E+01	2.90E+01	2.10E−01
India	(0,2,1)	1.75E+03	1.75E+03	1.47E+04	8.40E+01	1.21E+02	2.95E−01
Iran	(2,2,1)	1.16E+03	1.17E+03	4.64E+04	1.80E+02	2.16E+02	1.22E+00
Italy	(3,1,6)	1.55E+03	1.55E+03	3.61E+03	4.91E+01	6.01E+01	1.40E−01
Mexico	(3,2,2)	1.44E+03	1.46E+03	2.43E+04	1.23E+02	1.56E+02	3.03E−01
Pakistan	(3,2,3)	1.08E+03	1.10E+03	2.84E+03	4.27E+01	5.33E+01	7.77E−01
Peru	(0,2,1)	1.81E+03	1.82E+03	5.95E+03	1.08E+03	1.87E+03	6.16E+00
Russia	(5,2,1)	1.07E+03	1.09E+03	5.29E+03	5.27E+01	7.27E+01	4.68E−01
Saudi Arabia	(1,2,0)	6.89E+02	6.97E+02	2.94E+02	1.38E+01	1.72E+01	5.46E−01
Spain	(0,2,1)	1.80E+03	1.81E+03	1.27E+01	3.18E+00	3.56E+00	1.00E−02
UK	(6,2,2)	1.57E+03	1.60E+03	2.39E+05	4.14E+02	4.88E+02	6.17E−01
USA	(4,2,4)	1.88E+03	1.91E+03	1.33E+05	3.16E+02	3.65E+02	2.80E−01

Table 6
Selected SARIMA models for forecasting cumulative death cases.

Country	SARIMA (p,d,q)(P,D,Q,m)	AIC	BIC	MSE	MAE	RMSE	MAPE
South Africa	(2,2,2)(1,0,1,7)	8.56E+02	8.74E+02	1.94E+05	3.14E+02	4.41E+02	5.10E+00
Bangladesh	(0,2,1)(0,0,1,7)	6.69E+02	6.77E+02	1.28E+02	9.20E+00	1.13E+01	3.50E−01
Brazil	(5,2,2)(3,2,3,7)	9.96E+02	1.03E+03	2.33E+04	1.36E+02	1.53E+02	2.54E−01
Chile	(1,2,1)(2,0,1,7)	1.35E+03	1.37E+03	6.25E+05	6.99E+02	7.90E+02	8.05E+00
Columbia	(0,2,1)(2,0,2,3)	1.12E+03	1.13E+03	7.83E+04	2.19E+02	2.80E+02	2.90E+00
India	(1,2,1)(1,0,1,12)	1.43E+03	1.44E+03	2.24E+06	1.20E+03	1.50E+03	4.02E+00
Iran	(1,2,2)(1,0,1,12)	1.17E+03	1.18E+03	1.04E+05	2.60E+02	3.21E+02	1.74E+00
Italy	(6,2,2)(1,0,1,7)	1.41E+03	1.44E+03	3.12E+04	1.47E+02	1.77E+02	4.20E−01
Mexico	(0,2,1)(1,0,1,7)	1.34E+03	1.35E+03	1.36E+05	3.06E+02	3.68E+02	7.50E−01
Pakistan	(2,2,3)(2,0,2,7)	1.03E+03	1.05E+03	3.65E+04	1.58E+03	1.91E+02	3.10E+00
Peru	(1,2,1)(2,0,2,12)	9.01E+02	9.18E+02	3.71E+06	1.09E+03	1.93E+03	6.10E+00
Russia	(0,2,2)(1,0,2,7)	9.09E+02	9.24E+02	2.81E+04	1.50E+02	1.67E+02	1.34E+00
Saudi Arabia	(1,2,0)(1,0,0,3)	6.67E+02	6.75E+02	6.47E+03	7.20E+01	8.00E+01	2.80E+00
Spain	(1,2,2)(0,0,1,3)	1.90E+03	1.92E+03	3.10E+00	1.48E+00	1.77E+00	5.00E−03
UK	(6,2,4)(2,0,2,3)	1.42E+03	1.47E+03	1.63E+05	3.61E+02	4.03E+02	8.00E−01
USA	(5,2,1)(1,0,1,7)	1.89E+03	1.92E+03	1.96E+05	4.07E+02	4.42E+02	2.00E−01

Recovered cases in countries like Peru, Russia, Iran, and Saudi Arabia are increasing at higher linear rate as compared to that in the Chile, Italy, Spain, and UK. This observation is very similar to the observation made by ARIMA (Fig. 8). The predicted number of cumulative recovered cases according to SARIMA models (Table 4) is lesser than the prediction of ARIMA models of the respective countries. The number of recovered cases in countries such as South Africa, Brazil, Colombia, Chile, India, Iran, Italy, Mexico, Pakistan, Peru, Russia, Saudi Arabia, Spain, UK, USA, and Bangladesh will reach ≈2,560,000, ≈4,100,000, ≈790,000, ≈490,000, ≈6,100,000, ≈400,000, ≈210,000, ≈585,000, ≈605,000, ≈475,000, ≈1,000,000, ≈401,000, ≈150,000, ≈1,500, ≈2,500,000, ≈250,000 respectively (Fig. 9).

Interestingly in Spain, the number of cumulative recovered cases remained constant. It can be explained due to the fact that the constant number of the recovered cases in the forecast was influenced by the constant recovered data reported by Spain, over a recent couple of months. If we observe the cumulative recovered cases between, May and July the number is constant. The forecast of Spain remained same with ARIMA(2,2,2) and SARIMA(1,1,3)(2,0,1,7) but the AIC and BIC values were decreased when we used the SARIMA model as shown in Tables 3 and 4. From Fig. 8(A), it is evident that the number of recovered COVID-19 patients is ≈4,000,000 in India. If we compare the number of recovered cases of India the with number of cumulative confirmed cases in India (Fig. 6(A)), the percentage of recovered COVID-19 patients will be more than 65% by the end of September. Whereas in the USA, the percentage of recovered COVID-19 patients will be ≈35% by the end of September. In the ARIMA(1,2,6) forecast of Iran (Fig. 8(A)), there will be 410,000 COVID-19 patients recovered by the end of the second week of September. Whereas the SARIMA(4,2,4)(3,1,2,3) of Iran (Fig. 9(A)), predicted that the number of COVID-19 patients recovered will be equal to 400,000. When SARIMA models were used the number of predicted cumulative recovered cases was less than that of the ARIMA models for countries — Brazil, Iran, Italy, Mexico, Pakistan, Peru, Russia Saudi Arabia, USA, and Bangladesh (Figs. 8 and 9). However, the cumulative recovered cases of countries — South Africa, Colombia, Chile, India increased after using SARIMA models for forecasting the 60 days. The SARIMA(1,2,1)(1,0,1,7) model of Chile predicted that a total of ≈490,000 will be recovered from the COVID-19 disease which is 10,000 greater than the ARIMA(0,2,1) (Table 3) predictions. In the case of USA, the cumulative recovered cases were 900 less when SARIMA(2,2,2)(1,0,1,7) was used.

5.3. Cumulative death cases

As of today (09/08/2020), we are 33 weeks into the COVID-19 pandemic. According to the CDC's weekly summary released on 14th august the current percentage of deaths attributed to COVID-19 is 8.1% which is higher than the epidemic threshold. The percentage of deaths are expected to increase in the coming weeks as more the death certificates are being handled [39] which is also being supported by our results. The forecasted trends of the confirmed cumulative deaths were very similar to the trends observed for the confirmed cases and recovered cases. Though the USA is leading the world with a high number of confirmed deaths, it is found that the USA has a steep increase in the number of deaths along with Pakistan and Saudi Arabia. Countries such as Spain, UK, Italy have shown a gradual linear increase in the number of deaths whereas, South Africa, Brazil, Colombia, Chile, India, Iran, Mexico, Peru, Russia, and Bangladesh have shown an exponential increase in the number of deaths (Figs. 10 and 11). The selected ARIMA models projected the number of deaths for South Africa, Brazil, Colombia, Chile, India, Iran, Italy, Mexico, Pakistan, Peru, Russia, Saudi Arabia, Spain, UK, USA, and Bangladesh will be ≈23,000, ≈152,000, ≈23,000, ≈24,000, ≈80,000, ≈30,000, ≈38,000, ≈95,000, ≈7,900, ≈41,000, ≈24,500, ≈,4,900, ≈29,000, ≈55,000, ≈210,000, ≈6,500 respectively (Figs. 10(A) and 10(B)). Similarly, selected SARIMA models projected the number of deaths for South Africa, Brazil, Colombia, Chile, India, Iran, Italy, Mexico, Pakistan, Peru, Russia, Saudi Arabia, Spain, UK, USA, and Bangladesh will be ≈58,000, 150,000, ≈20,000, ≈16,000, ≈76,000, ≈28,000, ≈37,000, ≈85,000, ≈7,600, ≈52,000, ≈21,000, ≈5,600, ≈30,000, ≈50,000, ≈225,000, ≈5,650 respectively (Figs. 11(A) and 11(B)).

For example, the cumulative death cases in South Africa might increase exponentially to ≈40,000 (Fig. 10(B)) in the second week of September according to the SARIMA(2,2,2)(1,0,1,7) model listed in Table 6. The SARIMA(5,2,2)(2,1,2,7) for Bangladesh, has lower AIC and BIC values of 699 and 677 as shown in Table 6, when compared to ARIMA(2,2,2) model of Bangladesh which has AIC and BIC values are 718 and 730 as shown in Table 5. Similarly, other countries' models are described in Tables 5 and 6. The seasonality has a key role in determining the number of cumulative death cases, which is evident from Figs. 11(A) & 11(B). When the SARIMA models were used, the forecasts of the countries showed the number of cumulative death cases was less than that of the ARIMA models' forecasts. For instance, in case of Brazil, Chile, India, Iran, Mexico, Pakistan, Russia, Saudi Arabia, Spain, and Bangladesh the forecasted cases on 3rd week of September are ≈10,000, ≈8000, ≈6000, ≈1000, ≈8,000 ≈200, ≈3,900 ≈4000 and ≈350 lesser than their respective ARIMA models' predictions. However, in the case of the USA, South Africa, Colombia, Italy, Peru, the SARIMA models predicted the cumulative death cases were ≈20,000 ≈33,100, ≈6000, ≈10,000, ≈10,000 greater than the predictions of their respective ARIMA models.

The 95% CI of the forecast of UK and Spain remain broad for both ARIMA and SARIMA based forecasts as shown in Figs. 10(B) & 11(B). Moreover, the lower limit of 95% CI of the UK's forecast as shown in Fig. 11(B) declined to near zero deaths. This scenario is a deviation from the current dynamics of the COVID-19 pandemic. However, the upper limit of 95% CI of the UK is a more realistic projection as shown in Fig. 10(B). The deaths might increase to ≈99,000 by the end of September. A similar trend was observed with Spain (Fig. 11(B)) with ≈45,000 deaths by the end of September. In the case of the USA, ARIMA(4,2,4) model suggests the number of deaths will increase to 200,000 in the next few weeks, the upper limit of 95% CI implies that the number of cumulative death cases might even cross 250,000. Whereas the lower limit indicates the number of cumulative deaths might remain at 150,000. But by further inspection of SARIMA(5,2,1)(1,1,1,7) forecast of the USA (Fig. 11(B)), we can see the upper limit of 95% CI of the forecast reveals that the cumulative death cases might increase to ≈310,000. On the contrary, the lower limit of 95% CI of forecast displays a sharp decline in the number of deaths to less than 100,000. This decline in the trend can be achieved by enforcing strict social and physical distancing measures and implementing lockdowns at the federal level. By implementing lockdown at the country level, India was able to control the pandemic for a while [4,40].

6. Conclusions

In this study, we have forecasted COVID-19 cases (confirmed, recovered and deaths) for 60 days, until 21st September 2020, using ARIMA and SARIMA statistical models. Our forecast indicates that the COVID-19 trends in top-16 countries can be classified into three classes as exponential, steep linear increase, gradual linear increase. The reasons for this observation can be the population density, infection rate, lifestyle etc. The exponential rise of the COVID-19 forecast has a very narrow width of the shared region of the 95% CI, whereas the width of the shared region increases for both linear increment cases.

Countries such as the United States, Brazil, South Africa, Colombia, Bangladesh, India, Mexico and Pakistan have shown exponential growth in confirmed cases and recovered cases for the upcoming 60 days. In the case of deaths, countries such as Brazil, South Africa, Chile, Colombia, Bangladesh, India, Mexico, Iran, Peru, and Russia have shown an exponential increase in trends. Spain, UK, Italy the projections are stable with not much increase in COVID-19 cases. It is found that the COVID-19 forecasted value of the 60th day from the ARIMA and SARIMA models are more or less the same but to capture the seasonality or trends of the data SARIMA models outperform the ARIMA models. For most of the countries including the USA and India have a 7-day seasonal pattern, as selecting 7 in the SARIMA model generated the lowest AIC and BIC values. When we considered seasonality the SARIMA models predicted a number of COVID-19 cases was less than that of the ARIMA models' predictions. The SARIMA forecasts are more realistic numbers because they considered the variations that occurred in the past few weeks (June–July 2020) of the COVID-19 time-series and projected into the future.

Based on our predictions and forecasts, health care strategy administrators should take proper decision on the right time in supplying equipment to hospitals and other healthcare aids to the public. To keep the COVID-19 pandemic under control all countries must be prepared with their health care workers and hospital facilities. These results shed light on the approaching surge in cases thereby emphasizing the importance of social distancing and implementation of preventive measures of COVID-19.

CRediT authorship contribution statement

K.E. ArunKumar: Conceptualization, Methodology, Software, Validation, Formal analysis, Writing - original draft, Visualization, Data curation, Resources. **Dinesh V. Kalaga:** Writing - review & editing, Supervision, Data curation, Project administration. **Ch. Mohan Sai Kumar:** Data curation and preprocessing. **Govinda Chilkoor:** Data curation and preprocessing. **Masahiro Kawaji:** Data curation, Project administration. **Timothy M. Brenza:** Data curation, Project administration.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Appendix A. Snapshots of the Python script for the forecasting the time-series data

# STEP 1: Data collection,Import required libraries and modules, Read data into dataframe, convert data into date–time format.

```
raw_excel_data = pd.read_excel("US_Data.xlsx")
df_complex=raw_excel_data.copy()
df_complex.date = pd.to_datetime(df_complex.Date, dayfirst = True)
df_complex.set_index("Date", inplace=True)
df= df_complex.asfreq(’d’)
df = df.rename(columns=’USA’: ’Actual_data’})
```

# STEP 2: Identify the data trends and seasonality, do differencing to introduce stationarity (identify ‘d’ in ARIMA parameters (p,d,q))

```
seasonality = seasonal_decompose(df, model=’multiplicative’)
df["d1"] = diff(df["Actual_data"],k_diff = 1)
df[’d2’] = diff(df["Actual_data"],k_diff = 2)
```

# Step 3: Check the ACF and PACF plots of the actual data and compare them with differenced data

```
plot_acf(df["Actual_data"], lags = 40, label = "90");
plot_pacf(df["Actual_data"], lags = 40, label = "90");
plot_acf(df[’d1’], lags = 40);
plot_pacf(df[’d1’], lags = 40);
```

# Step 4: Split data into test and training data and build the basic model using auto_Arima

```
size = int(len(df)*0.9)
train_data= df_comp.iloc[:size]
test_data =df_comp.iloc[size:]
len(test_data)
step_fit = auto_arima(df[’Actual_data’], start_p=0, start_q=0,
                max_p=6, max_q=6,
                seasonal=False,# for SARIMA models seasonality is set to True
d=None, trace=True,enforce_stationarity =False,enforce_invertibility = False,
                error_action=’ignore’,
                suppress_warnings=True,maxiter = 50,
                stepwise=True)
step_fit.summary()
get_ipython().run_cell_magic(’time’, ’’, ’model_base = ARIMA(train_data["USA"].astype(float),
order =(0,2,0))\nresults_base = model_base.fit()\nresults_base.summary()’)
```

# Step 5: Predict the test data using the basic model parameters and plot actual data vs predictions based on the basic model

```
start=len(train_data)
end=len(train_data)+len(test_data)-1
predictions_base = results_base.predict(start=start, end=end, dynamic=False, typ=’levels’).rename
(’BASE_model Predictions’)
for i in range(len(predictions_base)):
print(f"predicted={predictions_base[i]:<11.10}, expected={test_data[’USA’][i]}")
plt.rc(’axes’, axisbelow=True)
fig = plt.figure(figsize = (15,9))
Test_data, = plt.plot(test_data[’USA’],"o",color = "#ff7f0e", label = "Test data (USA)")
predicted, =plt.plot(predictions_base, color = ’#1f77b4’, label = ’Predictions(Basic model)’, linewidth =2)
```

# Step 6: Evaluation of the Basic model

```
error_1 = mean_squared_error(test_data[’USA’], predictions_base)
error_2 = mean_absolute_error(test_data[’USA’], predictions_base)
error_3 = rmse(test_data[’USA’], predictions_base)
```

# Step 7: Development of the basic model based on the ACF and PACF plots of the residuals

```
model = ARIMA(df[’Actual_data’].astype(float),order=(7,2,1))
results = model.fit(start_ar_lags = 8)
fcast=results.predict(len(df),len(df)+60,typ=’levels’).rename(’ARIMA(7,2,1) Forecast’)
fig, ax = plt.subplots(figsize=(10, 5),dpi=900)
plot_acf(results.resid, lags =20,ax=ax,color =’#1f77b4’,linewidth =0.1)
fig, ax = plt.subplots(figsize=(10, 5),dpi=900)
plot_pacf(results.resid, lags =20,ax=ax,color =’#1f77b4’,linewidth =0.1)
```

# Step 8: Plot predictions based on developed model and actual data.

```
start=len(train_data)
end=len(train_data)+len(test_data)-1
predictions = results.predict(start=(start-2), end=(end-2), dynamic=False, typ=’levels’).rename(’Selected
model Predictions’)
plt.rc(’axes’, axisbelow=True)
```

```
fig = plt.figure(figsize = (15,9))
Test_data, = plt.plot(test_data['USA'],"o",color = "#ff7f0e", label = "Test data (USA)")
predicted, =plt.plot(predictions, color = '#1f77b4', label = 'Predictions(ARIMA 7,2,1)', linewidth =2)
# # # Step 9: Evaluation metrics for developed model
MSE = mean_squared_error(test_data['USA'], predictions)
MAE = mean_absolute_error(test_data['USA'], predictions)
RMSE = rmse(test_data['USA'], predictions)
# # Step 10: Diagnosing the developed model with kde/q–q plots
fig, ax = plt.subplots(figsize=(10,5), dpi=900)
results.resid.plot(kind = "kde")
# # Step 11: Forecasting time-series data based on the selected model
figs, ax = plt.subplots(figsize=(10,5),dpi = 900)
model_fit = model.fit(disp = -1000)
figs = model_fit.plot_predict(10,235,dynamic = False,plot_insample = True, ax=ax)
# # # The procedure is same for ARIMA and SARIMA. The additional three parameters(p,d,q)(P,D,Q) of the SARIMA model were optimized.
```

Appendix B. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.asoc.2021.107161.

References

[1] N. Zhu, et al., A novel coronavirus from patients with pneumonia in China, 2019, New Engl. J. Med. (2020).

[2] D. Cucinotta, M. Vanelli, WHO declares COVID-19 a pandemic, Acta Bio-Med.: Atenei Parmensis 91 (1) (2020) 157–160.

[3] W.H. Organization, WHO announces COVID-19 outbreak a pandemic, 2020, Available from: https://www.euro.who.int/en/health-topics/health-emergencies/coronavirus-covid-19/news/news/2020/3/who-announces-covid-19-outbreak-a-pandemic.

[4] T. Lancet, India under COVID-19 lockdown, Lancet 395 (10233) (2020) 1315.

[5] J. Peto, et al., Universal weekly testing as the UK COVID-19 lockdown exit strategy, Lancet 395 (10234) (2020) 1420–1421.

[6] S.M. Parodi, V.X. Liu, From containment to mitigation of COVID-19 in the US, JAMA 323 (15) (2020) 1441–1442.

[7] M. Andersen, Early evidence on social distancing in response to COVID-19 in the United States. Available at SSRN: https://ssrn.com/abstract=3569368 or http://dx.doi.org/10.2139/ssrn.3569368.

[8] J.H.U.a.M.C.V.R. Center, Corona Virus Resource Center, 2020, Available from: https://coronavirus.jhu.edu.

[9] H.D. Meares, M.P. Jones, When a system breaks: a queuing theory model for the number of intensive care beds needed during the COVID-19 pandemic, Med. J. Aust. 212 (10) (2020) 1.

[10] J. Watkins, Preventing a covid-19 pandemic, BMJ 368 (2020) m810, British Medical Journal Publishing Group. https://doi.org/10.1136/bmj.m810.

[11] T. Tran, L. Pham, Q. Ngo, Forecasting epidemic spread of SARS-CoV-2 using ARIMA model (case study: Iran), Glob. J. Environ. Sci. Manag. 6 (2020) 1–10, (Special Issue (Covid-19)).

[12] X. Zhang, et al., Comparative study of four time series methods in forecasting typhoid fever incidence in China, PLoS One 8 (5) (2013) e63116.

[13] Y. Chen, et al., Epidemiological features and time-series analysis of influenza incidence in urban and rural areas of Shenyang, China, 2010–2018, Epidemiol. Infect. (2020) 148.

[14] O.S. Olayemi, O.E. Oluwatosin, O.E. Segun, Time series analysis on reported cases of Tuberculosis in Minna Niger state Nigeria, Open J. Stat. 10 (3) (2020) 412–430.

[15] M.S.D.P. Nayak, K. Narayan, Forecasting dengue fever incidence using ARIMA analysis, Int. J. Collab. Res. Intern. Med. Publ. Health 11 (3) (2019) 924–932.

[16] W. Wu, et al., Time series analysis of human brucellosis in mainland China by using elman and Jordan recurrent neural networks, BMC Infect. Dis. 19 (1) (2019) 1–11.

[17] Z. Ceylan, Estimation of COVID-19 prevalence in Italy, Spain, and France, Sci. Total Environ. (2020) 138817.

[18] L. Moftakhar, M. Seif, The exponentially increasing rate of patients infected with COVID-19 in Iran, Arch. Iran. Med. 23 (4) (2020) 235–238.

[19] S.P. Marbaniang, Forecasting the prevalence of COVID-19 in Maharashtra, Delhi, Kerala, and India using an ARIMA model, 2020, http://dx.doi.org/10.21203/rs.3.rs-34555/v1.

[20] G. Perone, An ARIMA model to forecast the spread and the final size of COVID-2019 epidemic in Italy, 2020, medRxiv. https://doi.org/10.1101/2020.04.27.20081539.

[21] S. Ghosal, et al., Prediction of the number of deaths in India due to SARS-CoV-2 at 5–6 weeks., Diabetes Metab. Syndr.: Clin. Res. Rev. (2020).

[22] D. Parbat, M. Chakraborty, A python based support vector regression model for prediction of COVID19 cases in India, Chaos Solitons Fractals 138 (2020) 109942.

[23] M. Maleki, et al., Time series modelling to forecast the confirmed and recovered cases of COVID-19, Travel Med. Infect. Dis. (2020) 101742.

[24] M.H.D.M. Ribeiro, et al., Short-term forecasting COVID-19 cumulative confirmed cases: Perspectives for Brazil, Chaos Solitons Fractals (2020) 109853.

[25] R. Salgotra, M. Gandomi, A.H. Gandomi, Time series analysis and forecast of the COVID-19 pandemic in India using genetic programming, Chaos Solitons Fractals (2020) 109945.

[26] V.K.R. Chimmula, L. Zhang, Time series forecasting of COVID-19 transmission in Canada using LSTM networks, Chaos Solitons Fractals (2020) 109864.

[27] M. Azarafza, M. Azarafza, J. Tanha, COVID-19 infection forecasting based on deep learning in Iran, 2020, medRxiv.

[28] S.F. Ardabili, et al., Covid-19 outbreak prediction with machine learning, 2020, Available at SSRN 3580188.

[29] J.H.U.C.f.S.S.a. Engineering, Novel coronavirus (COVID-19) cases data, 2020, Available from: https://data.humdata.org/dataset/novel-coronavirus-2019-ncov-cases.

[30] R.A. Yaffee, M. McGee, An Introduction to Time Series Analysis and Forecasting: With Applications of SAS® and SPSS®, Elsevier, 2000.

[31] B. Fanoodi, B. Malmir, F.F. Jahantigh, Reducing demand uncertainty in the platelet supply chain through artificial neural networks and ARIMA models, Comput. Biol. Med. 113 (2019) 103415.

[32] G.E. Box, G.M. Jenkins, Time Series Analysis: Forecasting and Control, revised ed., San Francisco, 1976.

[33] A. Ewis, et al., ARIMA models for predicting the end of COVID-19 pandemic and the risk of a second rebound, Neural Comput. Appl. (2020) 1–20.

[34] P.C. Cooley, et al., Weekends as social distancing and their effect on the spread of influenza, Comput. Math. Organ. Theory 22 (1) (2016) 71–87.

[35] R. Kapoor, et al., God is in the rain: The impact of rainfall-induced early social distancing on COVID-19 outbreaks, 2020, https://ssrn.com/abstract=3605549 or http://dx.doi.org/10.2139/ssrn.3605549. Available at SSRN 3605549.

[36] T. Elhassan, A. Gaafar, Mathematical modeling of the COVID-19 prevalence in Saudi Arabia, 2020, medRxiv.

[37] K.E. ArunKumar, Dinesh V. Kalaga, Masahiro Kawaji, Timothy M. Brenza, Forecasting of COVID-19 using deep layer Recurrent Neural Networks (RNNs) with Gated Recurrent Units (GRUs) and Long Short-Term Memory (LSTM) cells, Chaos Solitons Fractals (2021) submitted for publication.

[38] CDC, Considerations for events and gatherings, 2020, Available from: https://www.cdc.gov/coronavirus/2019-ncov/community/large-events/considerations-for-events-gatherings.html.

[39] CDC, COVIDView weekly summary, 2020, Available from: https://www.cdc.gov/coronavirus/2019-ncov/covid-data/covidview/index.html.

[40] P. Pulla, Covid-19: India imposes lockdown for 21 days and cases rise, BMJ 368 (2020) m1251, British Medical Journal Publishing Group. https://doi.org/10.1136/bmj.m1251.