medRxiv preprint doi: https://doi.org/10.1101/2020.10.01.20205021; this version posted October 2, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
All rights reserved. No reuse allowed without permission.

Theta autoregressive neural network model for COVID-19 outbreak predictions

Tanujit Chakraborty[a+], Arinjita Bhattacharyya[b+] and Monalisha Pattnaik[c+]

[a]Statistical Quality Control & Operations Research Unit, Indian Statistical Institute, Kolkata, India
[b] Department of Bioinformatics & Biostatistics, University of Louisville, KY, USA
[c] Department of Statistics, School of Mathematical Sciences, Sambalpur University, Sambalpur, India

ARTICLE HISTORY
Compiled September 30, 2020

ABSTRACT
An unprecedented outbreak of the novel coronavirus (COVID-19) in the form of peculiar pneumonia has spread globally since its first case at Wuhan, China, in December 2019, increasing infected cases and mortality at a pandemic speed. Thus, forecasting the COVID-19 pandemic became a key research interest for both the epidemiologists and statisticians. These future predictions are useful for the effective allocation of health care resources, stockpiling, and help in strategic planning for clinicians, government authorities, and public-health policymakers after understanding the extent of the effect. The main objective of this paper is to develop the most suitable forecasting model that can generate real-time short-term (ten days) and long-term (fifty days) out-of-sample forecasts of COVID-19 outbreaks for eight profoundly affected countries, namely the United States of America, Brazil, India, Russia, South Africa, Mexico, Spain, and Iran. A novel hybrid approach based on the Theta model and Autoregressive neural network (ARNN) model, named Theta-ARNN (TARNN) model, is proposed. The proposed method outperforms previously available single and hybrid forecasting models for COVID-19 predictions in most data sets. In addition, the ergodicity and asymptotic stationarity of the proposed TARNN model are established which is of particular interest in nonlinear time series literature. An R-Shiny application is created for implementation of TARNN model and is publicly available[1].

KEYWORDS
COVID-19 Forecasting; Theta model; Autoregressive Neural Networks; hybrid model; Asymptotic stationarity.

Abbreviations: ANN - Artificial Neural Network; ARIMA - Auto Regressive Integrated Moving Average; ARNN - Auto Regressive Neural Network; CFR- Case-Fatality Rate; IR- Incidence Rate; JB - Jarque–Bera test; SARS-Cov-2 - Severe Acute Respiratory Syndrome Coronavirus-2; TARNN - Theta-ARNN; USA- United States of America; WBF - Wavelet Based Forecasting.

Corresponding author: Tanujit Chakraborty. Email: tanujit_r@isical.ac.in
+All authors contributed equally.
[1]https://github.com/arinjita9/COVID-19-Forecasting-by-TARNN-

NOTE: This preprint reports new research that has not been certified by peer review and should not be used to guide clinical practice.

medRxiv preprint doi: https://doi.org/10.1101/2020.10.01.20205021; this version posted October 2, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
All rights reserved. No reuse allowed without permission.

1. Introduction

In December 2019, clusters of pneumonia cases caused by the novel severe acute respiratory syndrome coronavirus 2 (SARS-Cov-2) were identified at the Wuhan, Hubei province in China [21,25] after almost hundred years of the 1918 Spanish flu [59]. Soon after the emergence of the novel beta coronavirus, World Health Organization (WHO) characterized this contagious disease as a pandemic in March due to its rapid spread within and outside the highly mobile population of China supported by densely populated location of sea-food market and time of advent [51] with an exponential increase in the incidence rate (IR) and case-fatality rate (CFR) [41]. As of September 29, 2020, a total of 33,732,181 confirmed cases and 1,009,512 deaths have been reported worldwide [5].

Researchers are facing unprecedented challenges during this global pandemic to forecast future real-time cases with traditional mathematical, statistical and machine learning-based forecasting tools [17,31,33,61,65]. Studies in March with simple yet powerful forecasting methods like exponential smoothing model predicted cases ten days ahead, with large confidence intervals, that despite the positive bias, had reasonable forecast error [46]. Previously used linear and exponential model forecasts for better preparation regarding hospital beds, ICU admission estimation, resource allocation, emergency funding,and proposing strong containment measures were conducted [19] that projected about 869 ICU and 14542 ICU admissions in Italy for March 20, 2020. ICU admissions and mechanical ventilation use for critically ill patients reached its peak shattering the health system of Lombardy, Italy, by March-end [20]. Health-care workers had to go through the immense mental stress left with a formidable choice of prioritizing young and healthy adults over elderly for allocation of life support, especially unwanted ignoring of those who are extremely unlikely to survive [16,52]. Real estimates of mortality with 14-day delay demonstrated underestimating the COVID-19 outbreak and indicated a grave future with a global CFR of 5.7% in March [4]. The contact tracing, quarantine, and isolation efforts have a differential effect on the mortality due to COVID-19 among countries. Even though it seems that the CFR of COVID-19 is less compared to SARS (10%) and MERS (36%), there are concerns about it being eventually returning as the seasonal flu, causing a second wave or future pandemic [45,48]. Thus, real-time nowcasting and forecasting with foretelling predictions are required to reach a statistically validated conjecture in this current health crisis. Some of the impacting leading-edge research concerning real-time projections of COVID-19 confirmed cases, recovered cases, and mortality using statistical, epidemiological and machine learning models are given in Table 1. However, forecasting COVID-19 pandemics is harder and this is primarily due to the following major factors [29]:

- Very less amount of data is available;
- Less understanding of the factors that contribute to it;
- Model accuracy is constrained by our knowledge of the virus. With an emerging disease such as COVID-19, many biologic features of transmission are hard to measure and remain unknown.
- Another source of uncertainty, affecting all models, is that we don't know how many people are, or have been, infected.
- We are certainly missing a substantial number of cases due to virologic testing, so models fitted to confirmed cases are likely to be highly uncertain [23].

medRxiv preprint doi: https://doi.org/10.1101/2020.10.01.20205021; this version posted October 2, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
All rights reserved. No reuse allowed without permission.

Table 1. Related works on forecasting of the COVID-19 pandemic

Research Topic	Date	Countries	Model	Results	Main Conclusion
Forecasting confirmed & recovered cases [35]	Jan 22 - April 30, 2020	World data	TP–SMN–AR time series (Autoregressive series based on two-piece scale mixture normal distributions)	MAPE = 0.22% for confirmed cases; MAPE = 1.6% for recovered	Provided reasonable forecasts in terms of error & model selection
Short-term forecasting of cumulative confirmed cases [50]	Inception - April 18-19, 2020	Brazil	ARIMA, Cubist regression, Random Forest, Ridge, Support Vector Regression (SVR), Stacking-Ensemble Learning (SEL)	Forecast errors lower than 6.9%	SVR & SEL models are suitable tools for forecasting COVID-19.
Modelling & forecasting daily cases [1]	Jan 11 - Feb 10, 2020	China	Susceptible-Infectious-Recovered-Dead model	Estimated average reproduction number (R_0) $\sim 2.6; CI : (45,000 - 180,000); CFR \sim 0.15\%$	Simulations predicted a decline of the outbreak at the end of February
Forecasting & risk assessment [11]	Jan 30-31 - April 4, 2020	Canada, France, India, South Korea, UK	ARIMA,WBF, Hybrid ARIMA-WBF, Regression Tree	MAE & RMSE least for Hybrid Wavelet ARIMA	Hybrid ARIMA-WBF have better performance;country-wise total cases,% of elderly people, total country population, doctors/ 1000 people,lockdown period, time-period of cases, and hospital beds / 1000 people affect the CFR significantly.
Forecast [46]	Jan 22 - March 11, 2020	Global data	Exponential smoothing models	Ten-days-ahead forecasts have actual cases within 90% CI	Forecasts reflect the significant increase in the trend of global cases with growing uncertainty.
AI-based Forecast [24]	Jan 11 - Feb 27, 2020	China	Data driven AI-based methods	Using the multiple-step forecasting till April 19, 2020 for 34 provinces/cities.	The accuracy of the AI-based methods for forecasting the trajectory of COVID-19 was high.
Long-term trajectories of COVID-19 [10]	Inception - June 17, 2020	Spain & Italy	Integrated deterministic (SIRCX) - stochastic (AR) approach	Basic reproduction number and estimated future cases are computed.	Integrated model shows significant improvement in the long-term forecasting of COVID-19 cases for Italy and Spain.
Real-time forecast [51]	Feb 5 - February 24, 2020	China	Generalized logistic growth model, Richards model, sub-epidemic wave model	Mean case estimates and 95% prediction intervals emulsifies the global picture (5,10,& 15) days ahead.	All methods perform similarly & increase in data inclusion decreases the width of prediction intervals.
Predictions and role of interventions [49]	Live forecast	India	Extended state-space SIR epidemiological models	Live forecasts with broad confidence intervals	Lock down has a high chance of reducing the number of COVID-19 cases, to buy time for preparation.
Forecasting and now casting domestic & international spread [61]	Dec 31 - Jan 28, 2020	China	Susceptible-Exposed-Infectious-Recovered model	$R_0 = 2.68(95\% CI 2.47, 2.86)$; Epidemic doubling time 6.4 days (95% CI 5.8, 7.1)	COVID-19 is no longer contained withing China, & human-human transmission became evident.
Forecast [17]	Jan 22- March 15, 2020	China, Italy & France	Susceptible, Infected, Recovered, Dead model	Recovery rate is similar for Italy & China, while infection and death rate appear to be different.	There is a certain universality in the time to evolution of COVID-19.
Machine learning-based forecasts of COVID-19 [56]	Jan 22 - April 10, 2020	India	Multi layer perceptron (MLP)	Forecast for 69 days	MLP method gives good predictions.

Time series forecasting models work by taking a series of historical observations and extrapolating the patterns into the future. These are great when the data are accurate, the future is similar to the past. There are essentially two general approaches to forecasting a time series: (a) generating forecasts from a single model; and (b) combining

medRxiv preprint doi: https://doi.org/10.1101/2020.10.01.20205021; this version posted October 2, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
All rights reserved. No reuse allowed without permission.

forecasts from many models (Hybrid experts). In classical time series forecasting, the autoregressive integrated moving average (ARIMA) model is used predominantly for forecasting linear time series [6], which has a significant strong assumption of linearity in the system and homoscedastic error distribution; typically without sudden jumps and bursts. Individual models such as ARIMA, Wavelet ARIMA (WBF)[14], generalized autoregressive conditional heteroskedasticity (GARCH) [15] and Theta method [2,28] are inadequate to model such situations. Usage of non-linear time series techniques in infectious disease modeling have successfully demonstrated the success of artificial neural networks (ANN) and autoregressive neural networks (ARNN) [8,34,38,63]. There are also a vast literature available on the hybrid models motivated by the seminal work of Bates & Granger [3] and followed by a plethora of empirical applications showing that combination forecasts are often superior to their individual counterparts. The idea of hybridizing time series models are by no means new – see for example [3,9,11,13,30,63]. Most recently, some promising hybrid models combining linear and nonlinear time series models are proposed for COVID-19 [10,11] and dengue [9] forecasting and performed well for predicting these epidemics.

Motivated by this, this study considers the time series data sets of coronavirus confirmed cases which show non-linearity, non-stationary and non-Gaussian patterns, making decisions based on a discrete model critical and unreliable. Another difficulty in COVID-19 data from the modeling aspect is the unavailability of sufficient data points, which generates biased predictions and estimates, which can be well maneuvered by neural nets [42]. Most of the relevant studies focused on the outbreak's short-term and long-term forecasts of reported confirmed cases have a broad range of fluctuations, wide confidence intervals, poorly reported data/model specifics, and predictive performance being too optimistic the models are becoming unreliable [62]. Hybridization of two or more models is the most common solution [40] for optimizing forecasting performance, and efficacious with unknown complete data characteristics [32]. The importance of hybrid methodology with a fusion of linear and non-linear forecasting models, becomes evident in tackling such dynamic/non-linear time series and its inbuilt time-changing variance, complex autocorrelation structure [11,44]. With growing expectations of advanced parsimonious hybrid forecasting methods accompanied by precise accuracy and more accurate forecasts, the main objectives of this study are:

- To propose a simple and computationally efficient hybrid forecasting model which generates short-term and long-term out-of-sample forecasts for eight profoundly affected countries (United States of America (USA), India, Brazil, Russia, South Africa, Mexico, Spain, and Iran).
- To compare the suggested method with traditional discrete and hybrid forecasting models with finer accuracy.
- To prove the model's stationarity and ergodicity properties from statistician's point of view.
- To recommend policy-making decisions, resource allocation based on these forecasts.
- To discuss the merits and future challenges that has to be addressed while working with the proposal for epidemiological forecasting.

Therefore, this study proposes a novel hybrid Theta autoregressive neural network model (TARNN) model combining Theta and ARNN models that can capture complex COVID-19 data structures. The linearity is controlled by the Theta method in the initial phase, while the non-linear trend of the COVID-19 data sets is adjusted

medRxiv preprint doi: https://doi.org/10.1101/2020.10.01.20205021; this version posted October 2, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
All rights reserved. No reuse allowed without permission.

by the ARNN model using residual values obtained from the base Theta model. The proposed model has easy interpretability, robust predictability and can adapt seasonality indices. Desirable statistical properties like asymptotic stationarity and ergodicity of the proposed TARNN model are explored. Through experimental evaluation, we have shown the excellent performance of the proposed hybrid model for the COVID-19 pandemics forecasting for eight different countries' data sets.

The rest of the paper is organized as follows: Section 2 discusses the detailed formulation of the proposed hybrid TARNN model. The ergodicity and stationarity of the proposed hybrid model are discussed in Section 3. In Section 4, we discuss the country-wise COVID-19 confirmed case data sets, preliminary of data analysis, performance evaluation metrics, and the experimental results. The discussions about the results and practical implications are given in Section 5. Finally, we conclude the paper with a direction for future research in Section 6.

2. Methodology

This study proposes a novel hybrid model based on Theta and ARNN models to forecast the confirmed cases of COVID-19 for eight profoundly affected countries. We start by discussing the single forecasting models to be used in the hybridization followed by the detailed formulation of the proposed hybrid Theta-ARNN (TARNN) model.

2.1. Theta Method

The 'Theta method' or 'Theta model' is a univariate time series forecasting technique that performed particularly well in M3 forecasting competition and of interest to forecasters [2]. The method decomposes the original data into two or more lines, called theta lines, and extrapolates them using forecasting models. Finally, the predictions are combined to obtain the final forecasts. The theta lines can be estimated by simply modifying the 'curvatures' of the original time series. This change is obtained from a coefficient, called θ coefficient, which is directly applied to the second differences of the time series:

$$Y^{"}_{new}(\theta) = \theta Y^{"}_{data}, \tag{1}$$

where $Y^{"}_{data} = Y_t - 2Y_{t-1} + Y_{t-2}$ at time t for $t = 3, 4, \cdots, n$ and $\{Y_1, Y_2, \cdots, Y_n\}$ denote the observed univariate time series. In practice, coefficient θ can be considered as a transformation parameter which creates a series of the same mean and slope with that of the original data but having different variances. Now, Eqn. (1) is a second-order difference equation and has solution of the following form [28]:

$$Y_{new}(\theta) = a_\theta + b_\theta(t-1) + \theta Y_t, \tag{2}$$

where a_θ and b_θ are constants and $t = 1, 2, \cdots, n$. Thus, $Y_{new}(\theta)$ is equivalent to a linear function of Y_t with a linear trend added. The values of a_θ and b_θ are computed

medRxiv preprint doi: https://doi.org/10.1101/2020.10.01.20205021; this version posted October 2, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
All rights reserved. No reuse allowed without permission.

by minimizing the sum of squared differences:

$$\sum_{i=1}^{t} [Y_t - Y_{new}(\theta)]^2 = \sum_{i=1}^{t} [(1-\theta)Y_t - a_\theta - b_\theta(t-1)]^2 . \tag{3}$$

Forecasts from the Theta model are obtained by a weighted average of forecasts of $Y_{new}(\theta)$ for different values of θ. Also, the prediction intervals and likelihood-based estimation of the parameters can be obtained based on a state space model which is demonstrated in [28]. The generalized version of the Theta method is suitable for for automatic forecasting of time series [55].

2.2. ARNN Model

Artificial Neural Network-based forecasting methods received increasing interest in various applied domains in late 1990s. A wide variety of neural nets are popularly used for supervised classification, prediction and nonlinear time series forecasting [64]. The architecture of a simple feedforward neural network can be described as a network of neurons arranged in input layer, hidden layer, and output layer in a prescribed order. Each layer passes the information to the next layer using weights that are obtained using a learning algorithm [18]. ARNN model is a modification to the simple ANN model especially designed for prediction problems of time series data sets [18]. ARNN model uses a pre-specified number of lagged values of the time series as inputs and number of hidden neurons in its architecture is also fixed [26]. ARNN(p, k) model uses p lagged inputs of the time series data in a one hidden layered feedforward neural network with k hidden units in the hidden layer. Let $\underline{x}$ denotes a p-lagged inputs and f is a neural network of the following architecture:

$$f(\underline{x}) = c_0 + \sum_{j=1}^{k} w_j \phi\left(a_j + b_j' \underline{x}\right) ; \tag{4}$$

where c_0, a_j, w_j are connecting weights, b_j are p-dimensional weight vector and ϕ is a bounded nonlinear sigmoidal function (e.g., logistic squasher function or tangent hyperbolic activation function). These Weights are trained using a gradient descent backpropagation [53]. Standard ANN faces the dilemma to choose the number of hidden neurons in the hidden layer and optimal choice is unknown. But for ARNN model, we adopt the formula $k = [(p+1)/2]$ for non-seasonal time series data where p is the number of lagged inputs in an autoregressive model [26].

2.3. Proposed TARNN Model

In this section, we describe the propose hybrid model based on Theta method and ARNN model and we name it TARNN model. The proposed TARNN model is based on an error re-modeling approach and there are broadly two types of error calculations popular in the literature which are given below [39].

Definition 2.1. In the additive error model, the forecaster treats the expert's estimate

medRxiv preprint doi: https://doi.org/10.1101/2020.10.01.20205021; this version posted October 2, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
All rights reserved. No reuse allowed without permission.

as a variable, $\hat{Y}_t$, and thinks of it as the sum of two terms:

$$\hat{Y}_t = Y_t + e_t,$$

where Y_t is the true value and e_t be the additive error term.

Definition 2.2. In the multiplicative error model, the forecaster treats the expert's estimate $\hat{Y}_t$ as the product of two terms:

$$\hat{Y}_t = Y_t \times e_t,$$

where Y_t is the true value and e_t be the multiplicative error term.

Now, even if the relationship is of product type, in the log-log scale it becomes additive. Hence, without loss of generality, we may assume the relationship to be additive and expect errors (additive) of a forecasting model to be random shocks. But, this is violated when there are complex correlation structures in the time series data and less amount of knowledge is available about the data generating process. A simple example is the daily confirmed cases of the COVID-19 cases for various countries where very little is known about the structural properties of the current pandemic. Thus, we need two-stage modeling approach to formulate this complex time series problem. The proposed TARNN model is a hybrid model based on additive error re-modeling approach. The hybrid TARNN approach consists of three basic steps:

- In first step of the TARNN model, the Theta method is applied to the time series data to model the linear components of given time series data set.
- Theta model generates in-sample forecasts and the error series is calculated.
- In the next phase, the residuals (additive errors) generated by the Theta method are re-modeled using a nonlinear ARNN model. Finally, both the forecasts obtained from the Theta and ARNN models are combined together to get the final forecasts for the given time series.

The mathematical formulation of the proposed hybrid TARNN model (Z_t) is as follows:

$$Z_t = L_t + N_t,$$

where L_t is the linear part and N_t is the nonlinear part of the hybrid model. We can estimate both L_t and N_t from the available time series data. Let $\hat{L}_t$ be the forecast value of the Theta model at time t and ϵ_t represent the error residuals at time t, obtained from the Theta model. Then, we write

$$\epsilon_t = Z_t - \hat{L}_t.$$

These left-out residuals are further modeled by ARNN model and can be represented as follows:

$$\epsilon_t = f(\epsilon_{t-1}, \epsilon_{t-2}, ..., \epsilon_{t-p}) + \varepsilon_t,$$

where f is a nonlinear function and the modeling is done by the ARNN model as defined in Eqn. (4) and ε_t is supposed to be the random shocks. Therefore, the combined

medRxiv preprint doi: https://doi.org/10.1101/2020.10.01.20205021; this version posted October 2, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
All rights reserved. No reuse allowed without permission.

forecast can be obtained as follows:

$$\hat{Z}_t = \hat{L}_t + \hat{N}_t,$$

where $\hat{N}_t$ is the forecasted value of the ARNN model. An overall flow diagram of the proposed TARNN model is given in Figure 1.

In the proposed TARNN model, ARNN is applied to re-model the left-over autocorrelations in the residuals which Theta model individually could not model. Thus, the proposed TARNN model can be considered as an error re-modeling approach. This is important because due to model mis-specification and disturbances in the pandemic rate time series, the linear Theta model may fail to generate white noise behavior for the forecast residuals. The TARNN approach eventually can improve the predictions for the epidemiological forecasting problem as shown in Section 4.

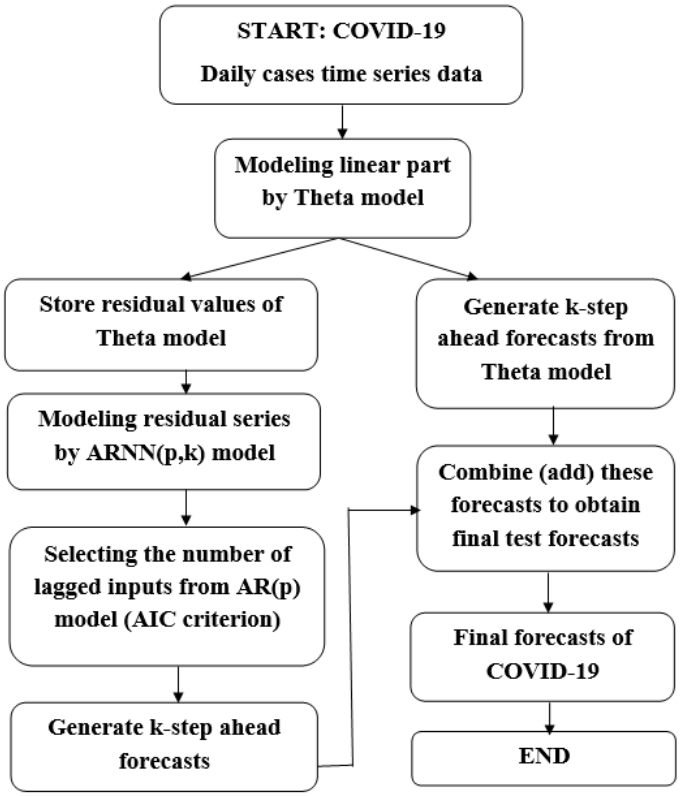


Figure 1. Flow diagram of the proposed TARNN model

Remark 1. The idea of the additive error modelling is useful for modeling complex time series for which achieving random shocks based on individual forecasting models is difficult. More precisely, the TARNN approach is developed for forecasting the COVID-19 confirmed cases for which the data generating process and the various characteristics of the epidemic are still unknown. The proposed TARNN model only assumes that the linear and nonlinear components of the epidemic time series can be separated individually.

medRxiv preprint doi: https://doi.org/10.1101/2020.10.01.20205021; this version posted October 2, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
All rights reserved. No reuse allowed without permission.

3. Ergodicity and Stationarity of the proposed TARNN model

In this section, we derive the results for ergodicity and asymptotic stationarity of the proposed TARNN model. The ergodicity and stationarity is of particular importance from a statistician's point of view in time series analysis since for such processes a single realization displays the whole probability of the data generating process. We use several previous results on nonlinear time series and Markov chains to find sufficient conditions for which the overall process is ergodic and stationary [7,12,58].
To start, we write the underlying stochastic model of the Theta method using the state space approach. We initialize the model by setting $Y_1 = l_1$ and then for $t = 2, 3, \cdots$; and drift term b, let

$$Y_t = l_{t-1} + b + \epsilon_t \quad \text{and} \quad l_t = l_{t-1} + b + \alpha\epsilon_t \ ;$$

where $\{\epsilon_t\}$ is assumed to follow Gaussian white noise with mean zero and variance σ^2 and α is the smoothing parameter for the simple exponential smoothing (SES) model. Now, Y_t follows a state space model which gives forecasts equivalent to SES with drift [28]. Also, Y_t can be written in the following form:

$$Y_t = Y_{t-1} + b + (\alpha - 1)\epsilon_{t-1} + \epsilon_t. \tag{5}$$

The above is an ARIMA(0,1,1) process with drift term [28]. The left-out residuals of Theta model is further modeled by ARNN process. We consider the ARNN process generated by the additive noise of the ARIMA(0,1,1) process with drift. Let ϵ_t be the process defined by the stochastic difference equation of the following form:

$$\epsilon_t = f(\epsilon_{t-1}, \epsilon_{t-2}, \cdots, \epsilon_{t-p}, \Theta) + \varepsilon_t, \tag{6}$$

where ε_t is an i.i.d. noise process and $f(\cdot, \Theta)$ is a feedforward neural network with weight (parameter) vector Θ and inputs $\epsilon_{t-1}, \epsilon_{t-2}, \cdots, \epsilon_{t-p}$. The definition of f is given in Eqn. (4).

3.1. Time Series as Markov chains

We start by defining the following notations:

$$\begin{aligned} z_{t-1} &= \Big(\epsilon_{t-1}, ..., \epsilon_{t-p}\Big)' \\ F(z_{t-1}) &= \Big(f(z_{t-1}), \epsilon_{t-1}, ..., \epsilon_{t-p+1}\Big)' \\ e_t &= \Big(\varepsilon_t, 0, ..., 0\Big)' \end{aligned}$$

Then Eqn. (6) can be written as follows [12]:

$$z_t = F(z_{t-1}) + e_t \tag{7}$$

with $z_t, e_t \in \mathbb{R}^p$. In this section, we show the (strict) stationarity of the state space form, as defined in Eqn. (7). The problem of showing $\{z_t\}$ to be stationary is closely related to the ergodicity of the process [60]. A Markov chain $\{z_t\}$ is called geometrically

medRxiv preprint doi: https://doi.org/10.1101/2020.10.01.20205021; this version posted October 2, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
All rights reserved. No reuse allowed without permission.

ergodic if there exists a probability measure π on the state space $(\mathbb{R}^p, \mathbb{B})$ and a constant $\rho > 1$ such that

$$\lim_{n\to\infty} \rho^n \|P^n(z, A) - \pi(A)\| \quad = \quad 0 \tag{8}$$

for each $z \in \mathbb{R}^p$, $\mathbb{B}$ is the Borel σ-algebra on $\mathbb{R}^p$ and $\|\cdot\|$ denotes the total variation norm. If Eqn. (8) holds for $\rho = 1$, then $\{z_t\}$ is called ergodic.

The definition for $P^n(z, A)$ can be given as the probability that $\{z_n\}$ moves from z to the set $A \in \mathbb{B}$ in n steps: $P^n(z, A) = P\left(z_{t+n} \in A | z_t = z\right)$. Also, expression for $\pi(A)$ is as follows:

$$\pi(A) = \int P^1(z, A)\pi(dz) \;\text{ for all } A \in \mathbb{B}.$$

Thus, we call π as the stationary measure and the distribution of z_t converges to π if $\{z_t\}$ is ergodic. Then, we say $\{z_t\}$ is asymptotically stationary [36]. To establish the ergodicity of TARNN processes, we need the concept of irreducibility and aperiodicity. A Markov chain $\{z_t\}$ is called irreducible if

$$\sum_{n=1}^{\infty} P^n(z, A) > 0 \;\text{ for all }\; z \in \mathbb{R}^p,$$

whenever $\lambda(A) > 0$ and λ denotes the Lebesgue measure on $(\mathbb{R}^p, \mathbb{B})$. Thus, for an irreducible Markov chain, all parts of the state space can be reached by the Markov chain irrespective of the starting point. Now, an irreducible Markov chain is aperiodic if there exists an $A \in \mathbb{B}$ with $\lambda(A) > 0$ and for all $C \in \mathbb{B}$, $C \subseteq A$ with $\lambda(C) > 0$, there exists a positive integer n such that

$$P^n(z, C) > 0 \;\text{ and }\; P^{n+1}(z, C) > 0 \;\text{ for all } z \in \mathbb{R}^p.$$

Hence, it is possible that the Markov chain returns to given sets only at specific time points for an aperiodic Markov chain. For most general time series models, irreducibility and aperiodicity cannot be assumed automatically. But for a TARNN process, these conditions can be checked. In general, it is sufficient to assume the distribution of the noise process to be an absolutely continuous component with respect to Lebesgue measure and the support of the probability density function (PDF) is sufficiently large.

3.2. Main Results

It is clear from the above discussion that if the Markov chain is geometrically ergodic then its distribution will converge to π and the corresponding time series will be called asymptotically stationary, see also [58]. Lemma 3.1 states that the state space of the Markov chain cannot be reduced depending on the starting point.

Lemma 3.1. *Let $\{z_t\}$ is defined by (7), and let $E|\varepsilon_t| < \infty$ and the PDF of ε_t is positive everywhere in $\mathbb{R}$. Then if f is defined by (4), the Markov chain $\{z_t\}$ is φ-irreducible and aperiodic.*

medRxiv preprint doi: https://doi.org/10.1101/2020.10.01.20205021; this version posted October 2, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
All rights reserved. No reuse allowed without permission.

Proof. Since the support of the PDF of ε_t is the whole real line, that is, the PDF is positive everywhere in $\mathbb{R}$, then we can say that $\{z_t\}$ is φ-irreducible by using [12]. In our case, every non-null p-dimensional hypercube can be reached in p steps with positive probability (and hence every non-null Borel set A). A necessary and sufficient condition for $\{z_t\}$ to be aperiodic is to have a set A and positive integer n such that $P^n(z, A) > 0$ and $P^{n+1}(z, A) > 0$ for all $z \in A$ [58]. In this case, this is true for all n due to consideration of the unbounded additive noise. □

The theorem below states the necessary condition for geometric ergodicity of a Markov chain. This can be obtained using the decomposition technique and ergodicity of stochastic difference equations [12].

Theorem 3.2. *Suppose $\{z_t\}$ is defined as in (6) and (7), F be a compact set that can be decomposed as $F = F_h + F_d$, and the following conditions hold:*
(i) $F_h(.)$ is continuous and homogeneous and $F_d(.)$ is bounded;
(ii) $E|\varepsilon_t| < \infty$ and probability distribution function of ε_t is positive everywhere in $\mathbb{R}$; then $\{z_t\}$ is geometrically ergodic.

Proof. $\{z_t\}$ satisfies the following equation:

$$z_t \quad = \quad F(z_{t-1}) + e_t \; ; \;\; t \geq 1, \text{ and } F : \mathbb{R}^p \to \mathbb{R}^p. \tag{9}$$

Let F_h be continuous and homogeneous, viz., $F_h(cz) = cF_h(z)$ for all $c > 0$, $z_t \in \mathbb{R}^p$, and F_d is bounded. Let the origin, O, be a fixed point of F_h. It is important to note that ε_t satisfies the condition (ii) in Theorem 3.2. We are going to show that the existence of a continuous Lyapunov function, V, in a neighbourhood of the origin which will ensure the geometric ergodicity of (7).
To start with we let $W \subseteq \mathbb{R}^p$, the closure of W by $\overline{W}$ and its boundary by $\widehat{W}$. We also let V be defined over the closure of the unit ball. We let $p_0 = \inf_{\|z\|=1} V(z)$, where $\|\cdot\|$ denote the Euclidean norm. Also, let G be the maximal connected component of $\{z : V(z) < \frac{p_0}{2}\}$ that contains the origin. Then we have $F(\overline{G}) \subset G$.
Let $g(z) = \inf\{r \geq 0, z \in rG\}$, $z \in \mathbb{R}^p$, where $rG = \{rz, z \in G\}$. Then $g(z)$ is well defined and it can easily be checked that g has the following properties:

(1) $g(cz) = cg(z)$, for all $c > 0$.
(2) There exists $0 < c < \mathbb{C} < \infty$ such that $c\|z\| \leq g(z) \leq \mathbb{C}\|z\|$.
(3) $\big(z/g(z)\big) \in \widehat{G}$.
(4) There exists $\epsilon > 0$, $0 < \theta < 1$ such that for all $z \in \widehat{G}$, $y \in \mathbb{R}^p$, we have $\|y - F(z)\| < \epsilon \Rightarrow y \in G$ and $g(y) < \theta$.

Now, for Eqn. (7), ε_t satisfies $E|\varepsilon_t| < \infty$. Let $A \in \mathbb{B}$ and $z \in \mathbb{R}^p$, we define $P(z, A)$ be the transition probability function as:

$$P(z, A) = \int_{A-F(z)} f(t)\mathbb{P}(dt).$$

medRxiv preprint doi: https://doi.org/10.1101/2020.10.01.20205021; this version posted October 2, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
All rights reserved. No reuse allowed without permission.

Thus, it holds that

$$\begin{aligned}
&\int g(y)P(z,dy) \\
&= \int g\big(F_h(z)+t\big)f(t)\mathbb{P}(dt) \\
&= \int_{\|t/g(z)\|<\epsilon} g(z)g\left(F_h\left(\frac{z}{g(z)}\right)+\frac{t}{g(z)}\right)f(t)\mathbb{P}(dt) \\
&\quad + \int_{\|t/g(z)\|\geq\epsilon} g(z)g\left(F_h\left(\frac{z}{g(z)}\right)+\frac{t}{g(z)}\right)f(t)\mathbb{P}(dt) \\
&< \big[\theta g(z)\big] + \big[M + g(z)\beta(z)\big] \text{ for some } M>0.
\end{aligned}$$

Here $|\beta(z)| < B < \infty$ for all z and $\beta(z) \to 0$ as $\|z\| \to \infty$. We let $h(z) = g(z)+1$, and r be such that

$$|\beta(z)| < \frac{(1-\theta)}{4} \quad \text{for} \quad ||z|| > r.$$

Then for $r_0 = \max\left(r, \frac{4(C+1)}{(1-\theta)p}\right)$, there exists B' such that

(1) $\int f(y)P(z,dy) < B' < \infty \quad \text{when} \quad ||z|| \leq r_0.$
(2) $\int f(y)P(z,dy) < \frac{1}{2}(1+\theta)h(z) \quad \text{when} \quad ||z|| > r_0.$

Using Theorem 4 of [60], we can conclude that $\{z_t\}$ is geometric ergodic. □

The next theorem gives the main result for asymptotic stationary of the TARNN model.

Theorem 3.3. *Let $E|\varepsilon_t| < \infty$ and the PDF of ε_t is positive everywhere in $\mathbb{R}$, and $\{\varepsilon_t\}$ and $\{z_t\}$ are defined as in (6) and (7), respectively. Then if f is a nonlinear neural network as defined in (4), then $\{z_t\}$ is geometrically ergodic and $\{\varepsilon_t\}$ is asymptotically stationary.*

Proof. The noise process ε_t satisfies $E|\varepsilon_t| < \infty$ by assumption (e.g., Gaussian noise). It is also important to note that neural network activation functions, more precisely logistics or tan-hyperbolic activation functions, are continuous compact functions and have bounded range. Thus $\{z_t\}$ satisfies all the criteria to be geometrically ergodic and using Theorem (3.2), one can write that for the ARNN process with $F_h \equiv 0$ and $F_d \equiv F$. Thus, the series $\{\varepsilon_t\}$ is asymptotically stationary. □

Remark 2. Some interpretations and practical implications of the theoretical results are given below:

- The geometric rate of convergence in Theorem 3.3 implies that the memory of TARNN process vanishes exponentially fast. This implies that the simplest version of the proposed model converges to a Wiener process.
- This is important for predictions over larger intervals of time, for example, one might train the network on an available sample and then use the trained network to generate new data with similar properties like the training sample. The results

medRxiv preprint doi: https://doi.org/10.1101/2020.10.01.20205021; this version posted October 2, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
All rights reserved. No reuse allowed without permission.

for the asymptotic stationarity can guarantee that the proposed hybrid model can not have growing variance over time.

- From practitioners point of view, when the data is generated by the irreducible TARNN process, the estimated weights are not too far from the true weights. Then, one can draw an indirect conclusions on the statistical nature of the observed data based on the estimated weights.

4. Experimental Analysis

4.1. COVID-19 Data Sets

Data is collected from the "Our World in Data" public repository (Link: `https://ourworldindata.org/coronavirus`) on eight countries: USA, Brazil, India, Russia, South Africa, Mexico, Spain and Iran with USA, India, Brazil and Russia leading in the number of confirmed cases of COVID-19. Eight univariate time-series of confirmed cases are analyzed for generating future outbreak predictions. The data is also available in this GitHub repository: `https://github.com/owid/COVID-19-data/tree/master/public/data`. The long term memory property of a time series is measured using the Hurst exponent (HE) [37]. The value of HE lying between 0.5 and 1 proves that the series has sufficiently long memory. To calculate the HE for the given data sets, we have used 'pracma' package in R statistical software. All the data sets have long-term memory and the nonlinearity is confirmed using Terasvirta's neural network test [57] applied on all the data sets.

4.2. Preliminary Data Analysis

A summary of the COVID-19 data sets of confirmed cases is shown in Table 2. USA has the maximum number of observations of 192 and highest mean,variability, followed by India (182), Russia (181) , Spain (180), Iran (162), Brazil (156), and Mexico (153). The Skewness values obtained from the data sets away from symmetric range of (-0.5,0.5) of and kurtosis ≥ 1 indicates towards skewed data. The values of the Jarque–Bera (JB) test statistic are away from zero with desired p-values, establishes the data sets are non-normally distributed, except for Iran. Also, it highlights that even though the inception of COVID-19 outbreak varies across borders, the epidemic curves hardly have any decline in new confirmed cases, though flattening the curve. Nonlinear and nonseasonal models are included in the analysis and compared with traditional ARIMA, GARCH, Theta, ARNN, hybrid WBF-ARIMA [11], hybrid WBF-GARCH [54], hybrid ARIMA-ARNN [11]. A pictorial view of the training data sets along with auto-correlation function (ACF) and partial ACF (PACF) plots are given in Figure 2.

Table 2. Descriptions of COVID-19 data sets of confirmed cases for USA, Brazil, India, Russia, South Africa, Mexico, Spain and Iran.

Country	Date	No. of Days	Minimum value	Max value	Mean	Standard deviation	Skewness	Kurtosis	JB test	p-value
USA	01/21 - 07/30/2020	192	0	78427	23057.20	20910.43	0.7290	2.8163	17.2741	0.0001
Brazil	02/26 - 07/30/2020	156	0	69074	16360.67	17172.34	0.8983	2.8981	21.0469	0.0000
India	01/30 - 07/30/2020	182	0	52123	8702.154	12921.09	1.7909	5.4357	142.2829	0.0000
Russia	02/01 - 07/30/2020	181	0	12640	4580.055	3956.639	0.0290	1.4616	17.8740	0.0000
South Africa	03/06 - 07/30/2020	147	0	13944	3204.918	4325.744	1.2776	3.2182	40.2798	0.0000
Mexico	02/29 - 07/30/2020	153	0	8438	2669.601	2452.852	0.4894	1.9025	13.7855	0.0010
Spain	02/01 - 07/30/2020	180	0	9181	1582.283	2201.895	1.7508	5.2113	128.6291	0.0000
Iran	02/20 - 07/30/2020	162	0	5275	1845.117	875.5673	-0.1613	3.5602	2.8206	0.2444

medRxiv preprint doi: https://doi.org/10.1101/2020.10.01.20205021; this version posted October 2, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
All rights reserved. No reuse allowed without permission.

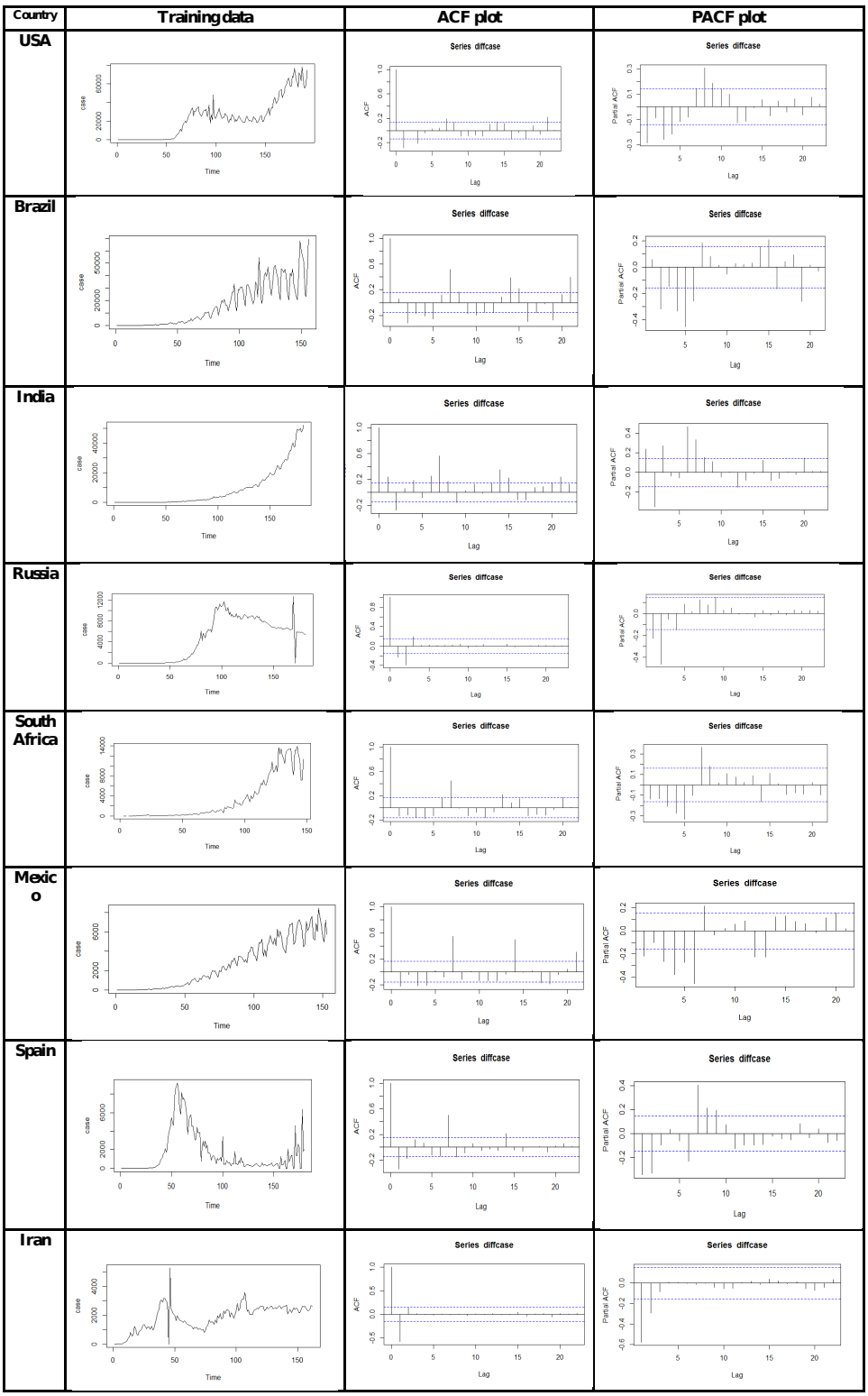


Figure 2. Training COVID-19 data sets of confirmed cases and corresponding ACF, PACF plots for USA, Brazil, India, Russia, South Africa, Mexico, Spain and Iran

medRxiv preprint doi: https://doi.org/10.1101/2020.10.01.20205021; this version posted October 2, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
All rights reserved. No reuse allowed without permission.

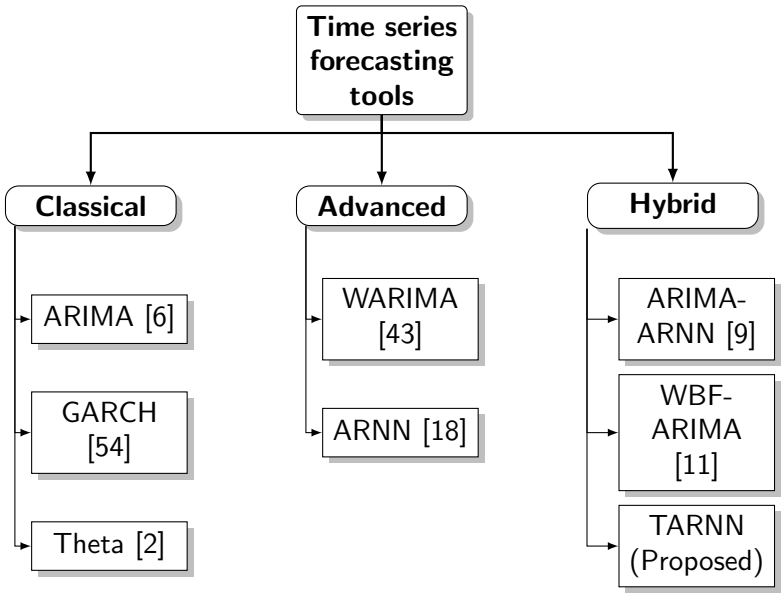


Figure 3. Time series forecasting tools (available and proposed) used in this study.

4.3. Performance Evaluation Metrics

The performance of different forecasting models are evaluated based on root mean square error (RMSE) and mean absolute error (MAE) metrics for eight COVID-19 data sets [26]:

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}; \quad MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|,$$

where, y_i is the actual value, $\hat{y}_i$ is the predicted value, and n denotes the number of data points. By definition, the lower the value of these performance metrics, the better is the performance of the concerned forecasting model.

4.4. Results

A schematic diagram is presented in Figure 3 to give an outline of the models to be used in this section. We start the experimental evaluation for all the data sets with the classical ARIMA(p, d, q) using '*forecast*' [27] statistical package in the R statistical software [47]. The proposed hybrid Theta-ARNN (TARNN) model with ten and fifty days ahead of prediction can indicate the extent of the pandemic. The nonlinear, non-stationary, and non-Gaussian structure of the data sets were confirmed by statistical tests in Table 2. The performances of traditional single models ARIMA, GARCH, Theta, ARNN, and hybrid models like hybrid WBF-ARIMA model, hybrid WBF-GARCH model, hybrid ARIMA-ARNN model were compared with the proposed novel hybrid Theta-ARNN (TARNN) model for all these eight COVID-19 data sets in Table 3.

In the proposed TARNN model, linear modelling is done with Theta model using

medRxiv preprint doi: https://doi.org/10.1101/2020.10.01.20205021; this version posted October 2, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
All rights reserved. No reuse allowed without permission.

'thetaf' function under the "forecast" package in R statistical software. Nonlinear modelling with ARNN approach is done with "caret" package using 'nnetar' function in R statistical software. After fitting the Theta model, we generate prediction for ten and fifty day time steps to compute the residual values (plots are given in Figure 4). Further, Theta residuals are modelled with ARNN(p, k) model having a pre-defined Box-Cox transformation set $\lambda = 0$ to ensure the forecast values to stay positive. The value of p and k are obtained by training the network and this is indeed a data reliant approach. Further, we add both the linear and non-linear forecasts to obtain the final forecast results. Theta model was fitted to eight data sets namely USA, Brazil, India, Russia, South Africa, Mexico, Spain and Iran. Further, model residuals of these eight countries were trained using ARNN(12,6) ARNN(20,10), ARNN(9,5), ARNN(7,4), ARNN(7,4), ARNN(14,8), ARNN(11,6) and ARNN(3,2) models with an average of 20 networks for all eight datasets, each of which is a 12-6-1, 20-10-1, 9-5-1, 7-4-1, 7-4-1, 14-8-1, 11-6-1 and 3-2-1 networks with 85, 221, 56, 37, 37, 129, 79 and 11 weights and with 492095, 33990, 40429, 196628, 59296, 646.9, 18603 and 102346 estimated σ^2, respectively. Finally, the predicted results of both Theta and ARNN models are added together to obtain the estimated forecasts of the proposed TARNN model. In a similar way, we applied hybrid WBF-ARIMA model, hybrid WBF-GARCH model, hybrid ARIMA-ARNN models over eight COVID-19 confirmed case datasets for comparison purposes.

Root mean square error (RMSE) and mean absolute error (MAE) were utilized to evaluate the predictive performance of the models [11]. Availability of data points is limited, thus implementation of the advanced deep learning techniques will result in over-fitting and biased estimates [22]. Actual vs. predicted values of well-performed model are plotted in Figure 5. As the Theta model is fitted on the residual time series, predictions are generated for the next ten (July 31 2020 to August 09 2020) and fifty (July 31 2020 to September 18 2020) time steps respectively. The real-time short-term forecasts using ARIMA, Theta, ARNN and hybrid ARIMA-ARNN and the proposed TARNN model are shown in Table 6 and the real-time long-term forecasts are illustrated in Table 7.

We compared our proposed TARNN model with traditional single models (ARIMA, GARCH, Theta, ARNN) along with hybrid WBF-ARIMA model, hybrid WBF-GARCH model and the experimental results are reported in Table 3. The performance of the proposed hybrid Theta-ARNN (TARNN) model is superior as compared to all the traditional individual and hybrid models on average. In comparison to other hybrid models, six out of eight data sets of COVID-19 confirmed cases, our proposed TARNN model outperformed all the hybrid and traditional individual models in the significant edge. The theoretically proven asymptotic stationarity of the proposed hybrid model also suggests that the model cannot have a growing variance over time. The consistency and adequacy in experimental results empirically approves the same. Thus, the efficacy of the proposed methodology of the proposed hybrid model is experimentally validated.

All the results can be effortlessly updated with the help of R-shiny application as new data becomes available. This publicly availble repository link `https://github.com/arinjita9/COVID-19-Forecasting-by-TARNN-` contains the current data files and R scripts for the TARNN model which ensures the repeatability and reproducibility of the results presented in this study.

medRxiv preprint doi: https://doi.org/10.1101/2020.10.01.20205021; this version posted October 2, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
All rights reserved. No reuse allowed without permission.

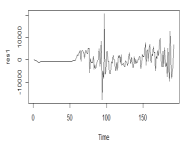
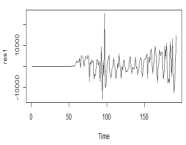
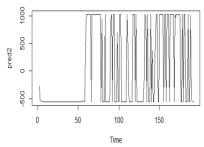
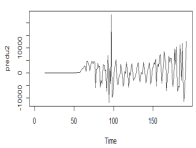
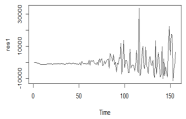
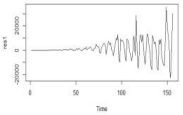
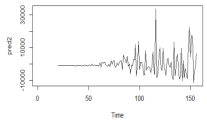
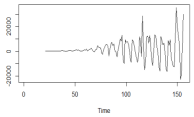
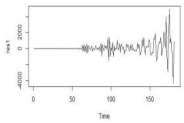
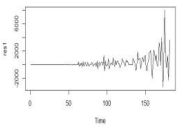
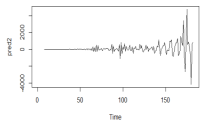
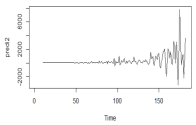
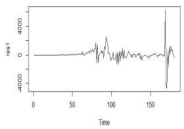
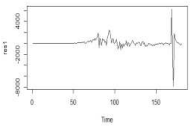
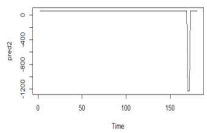
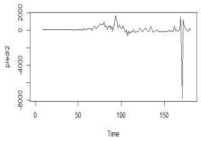
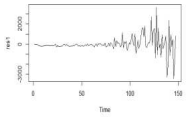
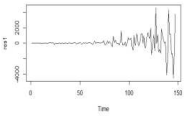
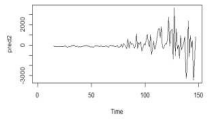
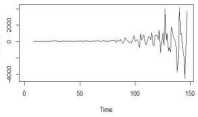
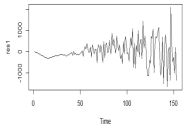
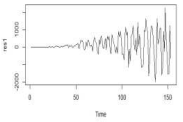
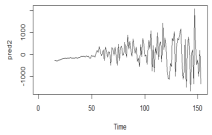
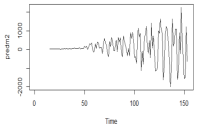
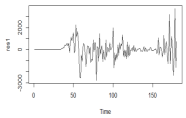
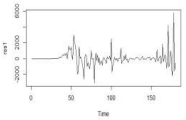
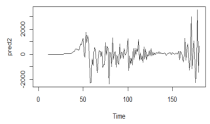
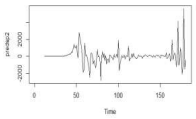
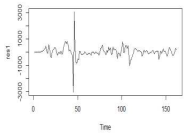
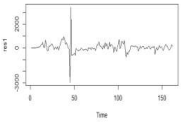
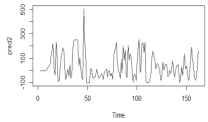
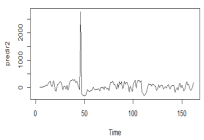
Country	ARIMA	Theta	Hybrid ARIMA-ARNN	Hybrid Theta-ARNN(TARNN)
USA				
Brazil				
India				
Russia				
South Africa				
Mexico				
Spain				
Iran				

Figure 4. Comparison of Residual Plots among different forecasting models with COVID-19 confirmed cases of USA, Brazil, India, Russia, South Africa, Mexico, Spain and Iran

medRxiv preprint doi: https://doi.org/10.1101/2020.10.01.20205021; this version posted October 2, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
All rights reserved. No reuse allowed without permission.

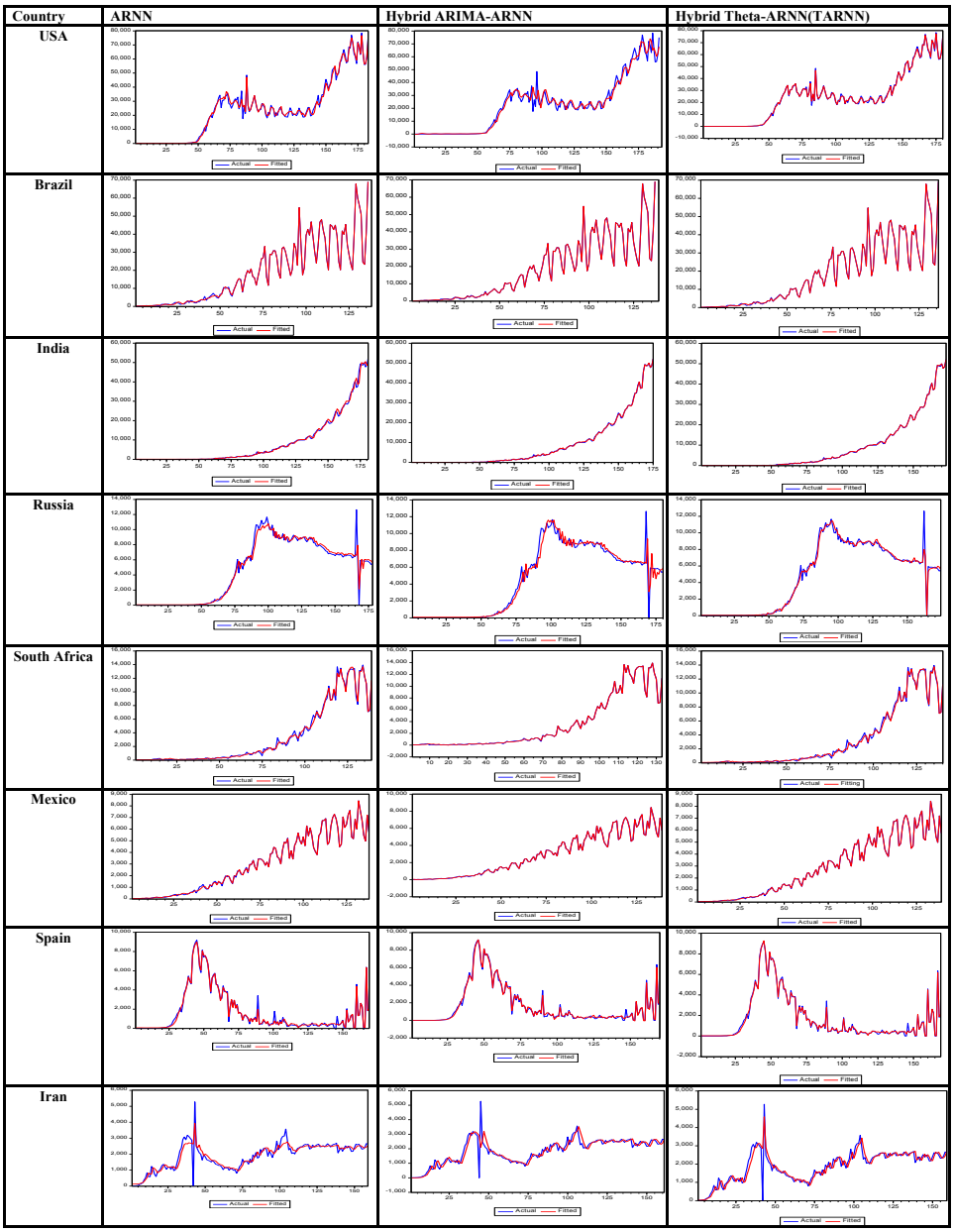


Figure 5. Actual and Predicted values of COVID-19 confirmed cases for different forecasting models for USA, Brazil, India, Russia, South Africa, Mexico, Spain and Iran

medRxiv preprint doi: https://doi.org/10.1101/2020.10.01.20205021; this version posted October 2, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
All rights reserved. No reuse allowed without permission.

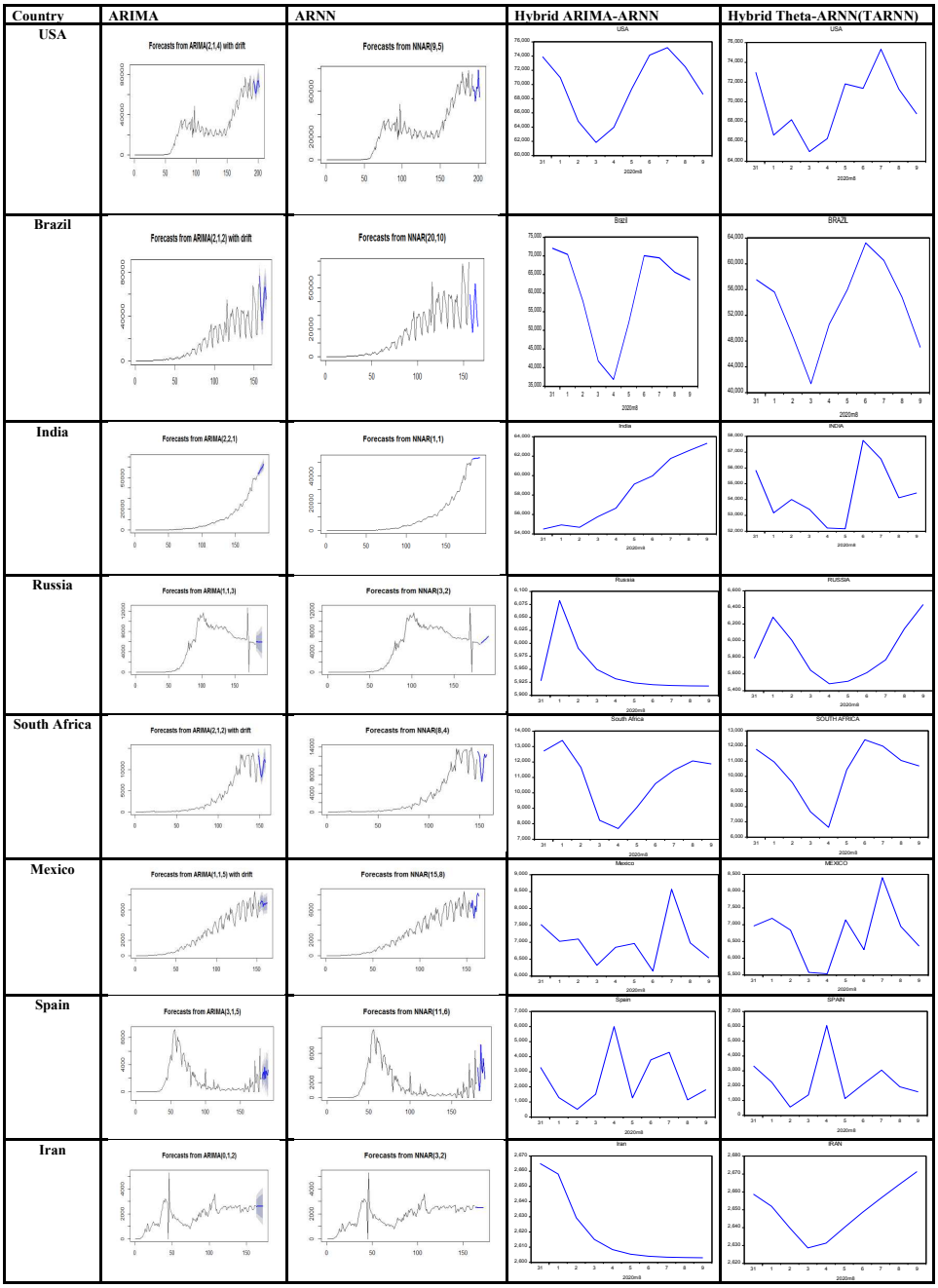


Figure 6. Real-time out of sample forecasts (10 days ahead) of COVID-19 confirmed cases for different forecasting models for USA, Brazil, India, Russia, South Africa, Mexico, Spain and Iran

medRxiv preprint doi: https://doi.org/10.1101/2020.10.01.20205021; this version posted October 2, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
All rights reserved. No reuse allowed without permission.

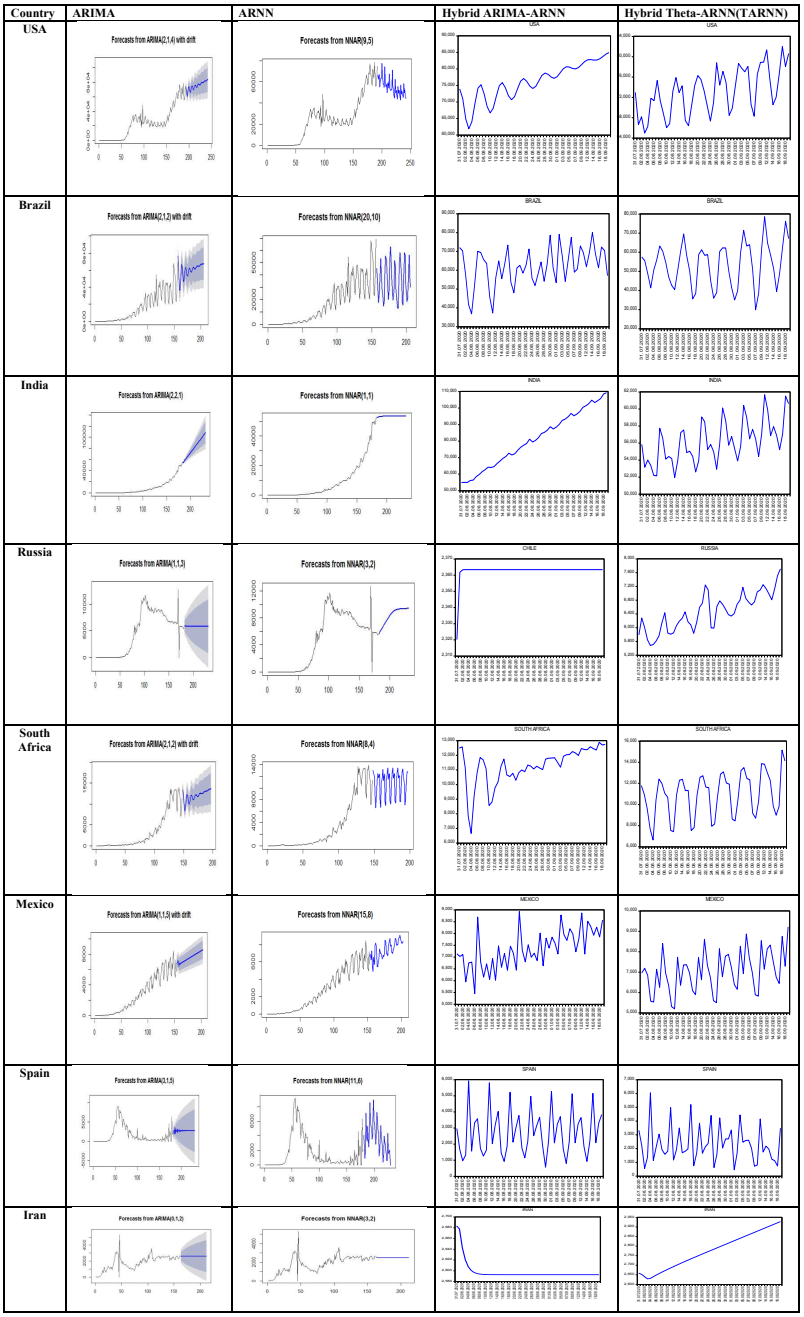


Figure 7. Real-time out of sample forecasts (50 days ahead) of COVID-19 confirmed cases for different forecasting models for USA, Brazil, India, Russia, South Africa, Mexico, Spain and Iran

medRxiv preprint doi: https://doi.org/10.1101/2020.10.01.20205021; this version posted October 2, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
All rights reserved. No reuse allowed without permission.

Table 3. Quantitative measures of performance for different forecasting models on eight time series (training data sets only) of COVID-19 confirmed cases for USA, Brazil, India, Russia, South Africa, Mexico, Spain and Iran.

	Performance							
Measures	ARIMA	GARCH	Theta	ARNN	WBF-ARIMA	WBF-GARCH	ARIMA-ARNN	Theta-ARNN(TARNN)
	RMSE							
USA	3714.3930	4827.5000	4524.3999	1787.7300	3961.0590	4594.2280	3639.0044	**721.5658**
Brazil	5626.7170	7687.0980	7761.1186	374.6499	6418.9380	7038.8090	**144.8500**	178.0458
India	819.4440	1040.8900	267.5700	908.2841	875.9043	712.7757	264.8608	**201.0696**
Russia	834.1711	1003.6880	920.3669	645.0402	887.1931	949.5511	776.6565	**443.4280**
South Africa	791.7877	1001.3170	980.9843	266.8777	770.2982	898.1525	**49.1428**	243.5067
Mexico	531.6361	679.1138	646.5100	57.1257	530.8732	681.1689	24.8734	**24.4335**
Spain	798.6674	1021.8120	914.8947	204.9399	909.3630	880.0793	171.5842	**136.3910**
Iran	440.9657	565.7666	453.3910	329.3327	486.8946	489.1197	418.5036	**319.9160**
	MAE							
USA	2468.3730	2794.1200	2837.2370	1153.2170	2489.6140	2913.2490	2220.4390	**468.6335**
Brazil	3344.2160	4169.0690	4678.6480	199.1100	3589.1310	4632.4600	**72.7800**	90.2053
India	425.9808	490.2410	487.5586	459.6912	465.4272	444.2688	176.4697	**128.7718**
Russia	370.7757	367.7794	361.7216	315.1817	373.6445	370.7152	383.5546	**202.6083**
South Africa	430.3521	472.0016	471.0693	175.7308	423.5596	497.7752	**30.7241**	160.3598
Mexico	377.0140	408.4892	428.7757	34.3695	361.7373	419.8514	15.1884	**15.1298**
Spain	482.3321	498.3005	479.3767	119.0491	533.3984	517.9961	109.8681	**87.5449**
Iran	237.4884	228.9164	239.7712	197.2631	286.3805	270.5897	209.4984	**182.8744**

5. Discussions

In this study, we proposed a novel hybrid Theta-ARNN (TARNN) model using residual modelling approach that performs considerably well for confirmed cases of COVID-19 forecasting for the countries that includes the ones with the highest number of cases USA, followed by India, Brazil, Russia. The proposed TARNN model filters linearity using the Theta model and can better explain the linear, nonlinear and non-stationary tendencies present in the selected COVID-19 data sets as compared to the traditional single and hybrid models. It also yields better forecast accuracy than various traditional single and hybrid models like ARIMA, GARCH, Theta, ARNN, WBF-ARIMA, WBF-GARCH and ARIMA-ARNN for six out of eight countries which are USA, India, Russia, Spain and Iran. The proposal will be useful in decision and policy makings for government officials and policymakers to allocate adequate health care resources for the coming days in responding to the crisis. Time series of epidemics can oscillate heavily due to various epidemiological factors and these fluctuations are challenging to be captured adequately for precise forecasting. This newly-developed model can still predict with better accuracy provided the conditions of asymptotic stationarity of hybrid model are satisfied. This method can be used to update real-time forecasts as more data becomes available. The study covering multiple countries can be utilized without geographical borders, and reflect the impact of of social distancing, wearing masks, lock down, shutdown, quarantine and sanitizing properly measures implemented by authorities. Both the short-term and long-term out of sample forecasts show oscillatory behaviour with upward trend and don't show any stiff decay sooner except Iran. All the seven different countries except Spain are going to face unlike uplifts in the number of new confirmed cases of COVID-19 pandemic. Followed by the both short-term and long-term out of sample forecasts reported in this paper, the lockdown and shutdown periods can be adjusted accordingly to handle the uncertain and vulnerable situations of COVID-19 pandemic. The newly-developed Theta-ARNN (TARNN) can efficiently predict COVID-19 cases, compared to traditional single and hybrid models.

medRxiv preprint doi: https://doi.org/10.1101/2020.10.01.20205021; this version posted October 2, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
All rights reserved. No reuse allowed without permission.

Prevalent techniques in literature were unable to completely capture the nonlinear behavior of stochastic time series containing inherent random shock component. This new method have significant theoretical (established ergodicity and stationarity of the proposed TARNN process) as well as practical implications. Authorities and health care can modify their planning in stockpile and hospital-beds depending on these forecasts of the COVID-19 pandemic.

6. Conclusions and Future Challenges

This work developed a novel hybrid Theta-ARNN (TARNN) to predict the subsequent COVID-19 outbreaks accurately and respond to pandemics more efficiently. Our proposed model is useful for nowcasting and forecasting of COVID-19 and the model can be further improved for multivariate time series set up when one get data sets on exogenous variables that impact on COVID-19 daily cases. Many parameters associated with COVID-19 transmission are still poorly understood. The resulting model uncertainty is not always calculated or reported in a standardized way. Once we can incorporate these variables, we can improve our estimates and update the TARNN model accordingly.

Since purely statistical approaches don't account for how transmission occurs, they are generally not well suited for long-term predictions about epidemiological dynamics (such as when the peak will occur and whether resurgence will happen) or for inference about intervention efficacy. Most forecasting models therefore limit their projections to one week or a few weeks ahead. Also, the problem of using confirmed cases to fit models is further complicated by the fact that the fraction of cases that are confirmed is spatially heterogeneous and time-varying. Amid enormous uncertainty about the future of the COVID-19 pandemic, the proposed TARNN model yields quantitative projections that policymakers may need in the short term to allocate resources or plan interventions. To conclude this model can further be extended for similar non-linear and non-Gaussian forecasting problems arising in other applied domains.

References

[1] C. Anastassopoulou, L. Russo, A. Tsakris, and C. Siettos, *Data-based analysis, modelling and forecasting of the covid-19 outbreak*, PloS one 15 (2020), p. e0230405.

[2] V. Assimakopoulos and K. Nikolopoulos, *The theta model: a decomposition approach to forecasting*, International journal of forecasting 16 (2000), pp. 521–530.

[3] J.M. Bates and C.W. Granger, *The combination of forecasts*, Journal of the Operational Research Society 20 (1969), pp. 451–468.

[4] D. Baud, X. Qi, K. Nielsen-Saines, D. Musso, L. Pomar, and G. Favre, *Real estimates of mortality following covid-19 infection*, The Lancet infectious diseases (2020).

[5] M.N.K. Boulos and E.M. Geraghty, *Geographical tracking and mapping of coronavirus disease covid-19/severe acute respiratory syndrome coronavirus 2 (sars-cov-2) epidemic and associated events around the world: how 21st century gis technologies are supporting the global fight against outbreaks and epidemics* (2020).

[6] G.E. Box, G.M. Jenkins, G.C. Reinsel, and G.M. Ljung, *Time series analysis: forecasting and control*, John Wiley & Sons, 2015.

[7] P.J. Brockwell and A. Lindner, *Strictly stationary solutions of autoregressive moving average equations*, Biometrika 97 (2010), pp. 765–772.

medRxiv preprint doi: https://doi.org/10.1101/2020.10.01.20205021; this version posted October 2, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
All rights reserved. No reuse allowed without permission.

[8] E. Cadenas and W. Rivera, *Wind speed forecasting in three different regions of mexico, using a hybrid arima–ann model*, Renewable Energy 35 (2010), pp. 2732–2738.

[9] T. Chakraborty, S. Chattopadhyay, and I. Ghosh, *Forecasting dengue epidemics using a hybrid methodology*, Physica A: Statistical Mechanics and its Applications (2019), p. 121266.

[10] T. Chakraborty and I. Ghosh, *An integrated deterministic-stochastic approach for predicting the long-term trajectories of covid-19*, medRxiv (2020).

[11] T. Chakraborty and I. Ghosh, *Real-time forecasts and risk assessment of novel coronavirus (covid-19) cases: A data-driven analysis*, Chaos, Solitons & Fractals (2020), p. 109850.

[12] K.S. Chan and H. Tong, *On the use of the deterministic lyapunov function for the ergodicity of stochastic difference equations*, Advances in applied probability 17 (1985), pp. 666–678.

[13] R.T. Clemen, *Combining forecasts: A review and annotated bibliography*, International journal of forecasting 5 (1989), pp. 559–583.

[14] A.J. Conejo, M.A. Plazas, R. Espinola, and A.B. Molina, *Day-ahead electricity price forecasting using the wavelet transform and arima models*, IEEE transactions on power systems 20 (2005), pp. 1035–1042.

[15] J.C. Duan, *The garch option pricing model*, Mathematical finance 5 (1995), pp. 13–32.

[16] E.J. Emanuel, G. Persad, R. Upshur, B. Thome, M. Parker, A. Glickman, C. Zhang, C. Boyle, M. Smith, and J.P. Phillips, *Fair allocation of scarce medical resources in the time of covid-19* (2020).

[17] D. Fanelli and F. Piazza, *Analysis and forecast of covid-19 spreading in china, italy and france*, Chaos, Solitons & Fractals 134 (2020), p. 109761.

[18] J. Faraway and C. Chatfield, *Time series forecasting with neural networks: a comparative study using the air line data*, Journal of the Royal Statistical Society: Series C (Applied Statistics) 47 (1998), pp. 231–250.

[19] G. Grasselli, A. Pesenti, and M. Cecconi, *Critical care utilization for the covid-19 outbreak in lombardy, italy: early experience and forecast during an emergency response*, Jama 323 (2020), pp. 1545–1546.

[20] G. Grasselli, A. Zangrillo, A. Zanella, M. Antonelli, L. Cabrini, A. Castelli, D. Cereda, A. Coluccello, G. Foti, R. Fumagalli, *et al.*, *Baseline characteristics and outcomes of 1591 patients infected with sars-cov-2 admitted to icus of the lombardy region, italy*, Jama 323 (2020), pp. 1574–1581.

[21] W.j. Guan, Z.y. Ni, Y. Hu, W.h. Liang, C.q. Ou, J.x. He, L. Liu, H. Shan, C.l. Lei, D.S. Hui, *et al.*, *Clinical characteristics of coronavirus disease 2019 in china*, New England journal of medicine 382 (2020), pp. 1708–1720.

[22] T. Hastie, R. Tibshirani, and J. Friedman, *The elements of statistical learning: data mining, inference, and prediction*, Springer Science & Business Media, 2009.

[23] I. Holmdahl and C. Buckee, *Wrong but useful—what covid-19 epidemiologic models can and cannot tell us*, New England Journal of Medicine (2020).

[24] Z. Hu, Q. Ge, L. Jin, and M. Xiong, *Artificial intelligence forecasting of covid-19 in china*, arXiv preprint arXiv:2002.07112 (2020).

[25] C. Huang, Y. Wang, X. Li, L. Ren, J. Zhao, Y. Hu, L. Zhang, G. Fan, J. Xu, X. Gu, *et al.*, *Clinical features of patients infected with 2019 novel coronavirus in wuhan, china*, The lancet 395 (2020), pp. 497–506.

[26] R.J. Hyndman and G. Athanasopoulos, *Forecasting: principles and practice*, OTexts, 2018.

[27] R.J. Hyndman, G. Athanasopoulos, C. Bergmeir, G. Caceres, L. Chhay, M. O'Hara-Wild, F. Petropoulos, S. Razbash, and E. Wang, *Package 'forecast'*, Online] https://cran. r-project. org/web/packages/forecast/forecast. pdf (2020).

[28] R.J. Hyndman and B. Billah, *Unmasking the theta method*, International Journal of Forecasting 19 (2003), pp. 287–290.

[29] J.P. Ioannidis, S. Cripps, and M.A. Tanner, *Forecasting for covid-19 has failed*, International journal of forecasting (2020).

[30] M. Khashei and M. Bijari, *An artificial neural network (p, d, q) model for timeseries*

medRxiv preprint doi: https://doi.org/10.1101/2020.10.01.20205021; this version posted October 2, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
All rights reserved. No reuse allowed without permission.

forecasting, Expert Systems with applications 37 (2010), pp. 479–489.

[31] A.J. Kucharski, T.W. Russell, C. Diamond, Y. Liu, J. Edmunds, S. Funk, R.M. Eggo, F. Sun, M. Jit, J.D. Munday, *et al.*, *Early dynamics of transmission and control of covid-19: a mathematical modelling study*, The lancet infectious diseases (2020).

[32] L.I. Kuncheva, *Combining pattern classifiers: methods and algorithms*, John Wiley & Sons, 2004.

[33] Q. Li, W. Feng, and Y.H. Quan, *Trend and forecasting of the covid-19 outbreak in china*, Journal of Infection 80 (2020), pp. 469–496.

[34] A. Maleki, S. Nasseri, M.S. Aminabad, and M. Hadi, *Comparison of arima and nnar models for forecasting water treatment plant's influent characteristics*, KSCE J. CIV. ENG. 22 (2018), pp. 3233–3245.

[35] M. Maleki, M.R. Mahmoudi, D. Wraith, and K.H. Pho, *Time series modelling to forecast the confirmed and recovered cases of covid-19*, Travel Medicine and Infectious Disease (2020), p. 101742.

[36] S.P. Meyn and R.L. Tweedie, *Markov chains and stochastic stability*, Springer Science & Business Media, 1993.

[37] J. Mielniczuk and P. Wojdyłło, *Estimation of hurst exponent revisited*, Computational statistics & data analysis 51 (2007), pp. 4510–4525.

[38] L. Milačić, S. Jović, T. Vujović, and J. Miljković, *Application of artificial neural network with extreme learning machine for economic growth estimation*, Physica A 465 (2017), pp. 285–288.

[39] A. Mosleh and G. Apostolakis, *The assessment of probability distributions from expert opinions with an application to seismic fragility curves*, Risk analysis 6 (1986), pp. 447–461.

[40] M.R. Oliveira and L. Torgo, *Ensembles for time series forecasting*, J. Mach. Learn. Res. 39 (2014), pp. 360–370.

[41] W.H. Organization, *et al.*, *Who director-general's opening remarks at the media briefing on covid-19-11 march 2020* (2020).

[42] A. Pasini, *Artificial neural networks for small dataset analysis*, Journal of thoracic disease 7 (2015), p. 953.

[43] D.B. Percival and A.T. Walden, *Wavelet methods for time series analysis*, Vol. 4, Cambridge university press, 2000.

[44] D.B. Percival and A.T. Walden, *Spectral Analysis for Univariate Time Series*, Vol. 51, Cambridge University Press, 2020.

[45] E. Petersen, M. Koopmans, U. Go, D.H. Hamer, N. Petrosillo, F. Castelli, M. Storgaard, S. Al Khalili, and L. Simonsen, *Comparing sars-cov-2 with sars-cov and influenza pandemics*, The Lancet infectious diseases (2020).

[46] F. Petropoulos and S. Makridakis, *Forecasting the novel coronavirus covid-19*, PloS one 15 (2020), p. e0231236.

[47] R Core Team, *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria (2017). Available at `https://www.R-project.org/`.

[48] D.D. Rajgor, M.H. Lee, S. Archuleta, N. Bagdasarian, and S.C. Quek, *The many estimates of the covid-19 case fatality rate*, The Lancet Infectious Diseases 20 (2020), pp. 776–777.

[49] D. Ray, M. Salvatore, R. Bhattacharyya, L. Wang, J. Du, S. Mohammed, S. Purkayastha, A. Halder, A. Rix, D. Barker, *et al.*, *Predictions, role of interventions and effects of a historic national lockdown in india's response to the covid-19 pandemic: data science call to arms*, Harvard data science review 2020 (2020).

[50] M.H.D.M. Ribeiro, R.G. da Silva, V.C. Mariani, and L. dos Santos Coelho, *Short-term forecasting covid-19 cumulative confirmed cases: Perspectives for brazil*, Chaos, Solitons & Fractals (2020), p. 109853.

[51] K. Roosa, Y. Lee, R. Luo, A. Kirpich, R. Rothenberg, J. Hyman, P. Yan, and G. Chowell, *Real-time forecasts of the covid-19 epidemic in china from february 5th to february 24th, 2020*, Infectious Disease Modelling 5 (2020), pp. 256–263.

medRxiv preprint doi: https://doi.org/10.1101/2020.10.01.20205021; this version posted October 2, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
All rights reserved. No reuse allowed without permission.

[52] L. Rosenbaum, *Facing covid-19 in italy—ethics, logistics, and therapeutics on the epidemic's front line*, New England Journal of Medicine 382 (2020), pp. 1873–1875.

[53] D.E. Rumelhart, G.E. Hinton, and R.J. Williams, *Learning internal representations by error propagation*, Tech. Rep., California Univ San Diego La Jolla Inst for Cognitive Science, 1985.

[54] M. Sabiruzzaman, M.M. Huq, R.A. Beg, and S. Anwar, *Modeling and forecasting trading volume index: Garch versus tgarch approach*, The Quarterly Review of Economics and Finance 50 (2010), pp. 141–145.

[55] E. Spiliotis, V. Assimakopoulos, and S. Makridakis, *Generalizing the theta method for automatic forecasting*, European Journal of Operational Research 284 (2020), pp. 550–558.

[56] R. Sujath, J.M. Chatterjee, and A.E. Hassanien, *A machine learning forecasting model for covid-19 pandemic in india*, Stochastic Environmental Research and Risk Assessment (2020), p. 1.

[57] T. Teräsvirta, C.F. Lin, and C.W. Granger, *Power of the neural network linearity test*, Journal of time series analysis 14 (1993), pp. 209–220.

[58] A. Trapletti, F. Leisch, and K. Hornik, *Stationary and integrated autoregressive neural network processes*, Neural Computation 12 (2000), pp. 2427–2450.

[59] A. Trilla, G. Trilla, and C. Daer, *The 1918 "spanish flu" in spain*, Clinical infectious diseases 47 (2008), pp. 668–673.

[60] R. Tweedie, *The existence of moments for stationary markov chains*, Journal of Applied Probability 20 (1983), pp. 191–196.

[61] J.T. Wu, K. Leung, and G.M. Leung, *Nowcasting and forecasting the potential domestic and international spread of the 2019-ncov outbreak originating in wuhan, china: a modelling study*, The Lancet 395 (2020), pp. 689–697.

[62] L. Wynants, B. Van Calster, M.M. Bonten, G.S. Collins, T.P. Debray, M. De Vos, M.C. Haller, G. Heinze, K.G. Moons, R.D. Riley, *et al.*, *Prediction models for diagnosis and prognosis of covid-19 infection: systematic review and critical appraisal*, bmj 369 (2020).

[63] G.P. Zhang, *Time series forecasting using a hybrid arima and neural network model*, Neurocomputing 50 (2003), pp. 159–175.

[64] G. Zhang, B.E. Patuwo, and M.Y. Hu, *Forecasting with artificial neural networks:: The state of the art*, International journal of forecasting 14 (1998), pp. 35–62.

[65] Z. Zhuang, P. Cao, S. Zhao, Y. Lou, W. Wang, S. Yang, L. Yang, and D. He, *Estimation of local novel coronavirus (covid-19) cases in wuhan, china from off-site reported cases and population flow data from different sources*, medRxiv (2020).