ARIMA based Time Series Analysis to Forecast Coronavirus (COVID-19) Disease

S. Sundarabalan[1] and S. Raguraman[2]

[1]Department of Statistics, DG Vaishnav College, Chennai – 600106, India.
[2]Department of Statistics, Manonmaniam Sundaranar University, Tirunelveli – 627012, India.

***Abstract*:** COVID-19, a novel coronavirus, is presently a major worldwide threat. It has infected more than a million people globally important to hundred-thousands of deaths. In such serious conditions, it is very important to predict the future infected cases to support the prevention of the disease and support in the healthcare service preparation. We have developed a model and then working it for forecasting future COVID-19 cases in Chennai, Tamil Nadu. This study indicates an ascending trend for the cases in the upcoming days. A time-series analysis similarly presents an exponential increase in the number of cases. It is supposed that the present prediction models will support the government and medical personnel to be prepared for the upcoming conditions.

Keywords: COVID-19, ARIMA models, exponential smoothing models, forecasting.

1. INTRODUCTION

Coronavirus disease 2019 (COVID-19) is an infectious disease affected by a new virus that has never been identified in humans before. This virus causes a respiratory illness with symptoms like cough, fever and in the most severe cases, pneumonia. The new COVID-19 is mostly spread through contact with an infected person after they cough or sneeze or through droplets of saliva or nasal secretions. The virus appeared for the first time on December 2019 in Wuhan, China[9,11].

COVID-19, generally known as Coronavirus, is a novel highly transmissible virus belonging to *the Coronaviridae* family that has been suspected to be transmitted to humans from animals. This virus causes mild to severe respiratory illness and death[6]. This pandemic has engulfed 185 countries/regions in merely four months infecting 1,949,210 people and taking the death toll to 123,348[3,8]. However, the early cases show the infection is less severe as associated with other coronaviruses such as SARS-CoV (Severe Acute Respiratory Syndrome Corona Virus) and MERS-CoV (Middle East Respiratory Syndrome Corona Virus), the cases of rapid human to human transmission signify that COVID-19 is highly infectious than others[4]. Although a local seafood market in Wuhan is believed to be the source of exposure[1], the scope of incidence of this disease is not clear since its occurrence at present is so dynamic[6]. An apparent variation is present in epidemiological examinations and detection abilities performed by different countries for detecting infected cases[5]. Presently, the highest cases of COVID-19 infections have been reported in the US, however, the cases are abruptly rising in Spain, Italy, France and Germany daily[3]. China, the place of origin of the disease, is now receiving very few cases[3]. The first case of coronavirus infection in India was reported on 30 January 2020 in Kerela, which was an imported case from Wuhan city of China[7]. In the early stage, the spread was extremely slow and only 3 people were positive for more than a month. However, the numbers started rising exponentially after one month and continue to do so. The numbers in India have reached up to 10,453 for confirmed COVID-19 infected cases with 358 deaths and 1181 recoveries as reported on 13 April 2020[3]. In the present day, there is neither a treatment nor an injection for the COVID-19 infection. Presently, it is a major health disaster around the world and it would not be wrong to say that it is 'an enemy to humanity'. In this situation, the only decision is preventing the occurrence of infection and preparing our healthcare system for the possible up-comings.

In that situation, it is particularly critical to construct models that are computationally competent in addition to realistic so that they can help policymakers, medical personals and also the general public. Modelling the disease and providing a future forecast of the possible number of daily cases can assist the medical system in getting prepared for the new patients. The statistical prediction models are suitable in forecasting in addition to controlling the global epidemic threat.

We have a working Auto-Regressive Integrated Moving Average (ARIMA) model for predicting the incidence of COVID-19 disease. As compared to other prediction models, for occurrence support vector machine (SVM), wavelet neural network (WNN) and ARIMA model is more capable in the prediction of natural adversities[10]. For our study, we have recognized the best ARIMA model and then forecast the number the cases for the next 20 days. The main objective of this study is to find the best predictive model and apply it to forecast the upcoming incidence of COVID-19 cases in Chennai, Tamil Nadu.

2. MATERIAL

Confirmed, recovered and death cases of COVID-19 infection are collected for Chennai, Tamil Nadu, as per World Health Organization region classification, from the official website of www.stopcorono.tn.gov.in from 22 January 2021 to 23 April 2021. This data is used to build predictive models.

3. METHODS

For forecasting a time series, ARIMA modelling is one of the best modelling procedures. ARIMA models are always represented with the support of particular parameters and the model is expressed as ARIMA (p, d, q). Here, p stands for the order of auto-regression, d indicates the degree of trend difference though q is the order of moving average. We must apply an ARIMA model to the time series data of confirmed COVID-19 cases in Chennai, Tamil Nadu. The autocorrelation

function (ACF) graph and partial autocorrelation (PACF) graph is used to find the early number of ARIMA models. These ARIMA models are then verified for variance in normality and stationary. Next, they are checked for accuracy by detecting their MAPE, MAD and MSD values to control the finest model to forecast. In addition, the best fit ARIMA model is associated with Linear Trend, Quadratic Trend, S- Curve Trend, Moving Average, Single Exponential as well as Double Exponential models using an output of measure of accuracy, MAPE, MAD, MSD, to select the finest model to forecast. The best model is the one that takes the lowest value for all the measures. Subsequently fitting the model, its parameters are predictably followed by verification of the model. The constructed model is active to forecast confirmed COVID-19 cases for the next 20 days, *i.e.* 24 April 2021 to 13 May 2021. The model for predicting upcoming confirmed COVID-19 cases is represented as,

$$ARIMA(p,d,f): X_t=\alpha_1 X_{t-1}+\alpha_2 X_{t-2}+\beta_1 Z_{t-1}+\beta_2 Z_{t-2}+Z_t \quad (1)$$

where

$$Z_t=X_t-X_{t-1} \quad (2)$$

Here, X_t is the predicted number of confirmed COVID-19 cases at t^{th} day, α_1, α_2, β_1 and β_2 are parameters whereas Z_t is the residual term for t^{th} day. The trend of upcoming incidences can be estimated after the earlier cases and a time series analysis is performed for this purpose. Time series forecasting refers to the employment of a model to forecast future data based on previously observed data[2]. In the current study, time series analysis is used to identify the trends in confirmed COVID-19 cases in Chennai, Tamil Nadu, from 22 January 2021 to 23 April 2021 and to predict future cases from 24 April 2021 till 13 May 2021. The level of statistical significance is set at 0.05. A graph is plotted for actual confirmed cases and forecast confirmed cases concerning time to confirm the efficiency of the model. To get an idea of the recovery and death trends in Chennai, Tamil Nadu, a graph is plotted concerning time.

A comparative study is also performed to examine the status of confirmed COVID-19 cases of Chennai, Tamil Nadu, concerning those of highly infected countries. All the model developments, computations and comparisons have been performed using SPSS and R software's.

4. RESULT AND DISCUSSION

The present work includes the development of a model to forecast COVID-19 incidences in the coming days. The results for the measure of model accuracy for ARIMA, Linear Trend, Quadratic Linear, S-Curve Trend, Moving Average, Single Exponential well as Double Exponential model are showed in Table 1. An expression at the MAPE, MAD and MSD values proposes that ARIMA (2, 2, 2) model is the most accurate of all for forecasting future incidences as it possesses the least value for all the measures.

Table 1: Measures of Model Accuracy

Models	MAPE	MAD	MSD
ARIMA (2, 2, 2)	4.3	57.9	25320.3
Single exponential method	9.8	98.5	58989.2
Double exponential method	8.9	53.0	158915.4
Moving average (MA)	11	144	103724
S-Curve Trend Model	67	1098	9798528
Quadratic Trend Model	13243	848	1085089
Linear Trend Model	14854	1315	3007514

Thus, parameters are estimated for the ARIMA (2, 2, 2) model which is shown in Table 2. It is observed that AR (2) and MA (2) parameters have a *p*-value of 0.000, 0.159, 0.000 and 0.000 respectively, thus implying that the parameters are significant in the model.

Table 2: Parameters Estimates of the ARIMA Model

Type	Coefficient	SE Coefficient	t Value	p-Value
AR (1)	0.5358	0.1331	4.03	0.000
AR (2)	-0.2052	0.1459	-1.39	0.159
MA (1)	1.5662	0.0567	27.88	0.000
MA (2)	-0.9491	0.0458	-19.12	0.000

Figure 1 showed the residual plots for confirmed COVID-19 cases in Chennai, Tamil Nadu from 30 January 2021 to 23 April 2021. A minor deviation of residuals from the straight line can be observed from the plot. This indicates that the errors are rather near to normal with a few outliers. Therefore, the normality assumption is followed. The residual histogram backs up this assumption. The graph between residuals and the fitted values shows a slight dispersion. This suggests that the assumption of constant variance is also satisfied by the model. The non-correlation of residuals is clear from the plot of residuals against the order of the data.

The Ljung Box statistics further corroborate this fact showed in Table 3. It is transparent that the *p*-value of all the lags is larger than the significance level (0.05) which means there is no violation of the independent assumption. The suitability of the ARIMA (2, 2, 2) model is showed by the non-significance of *p*-value and other statistics.

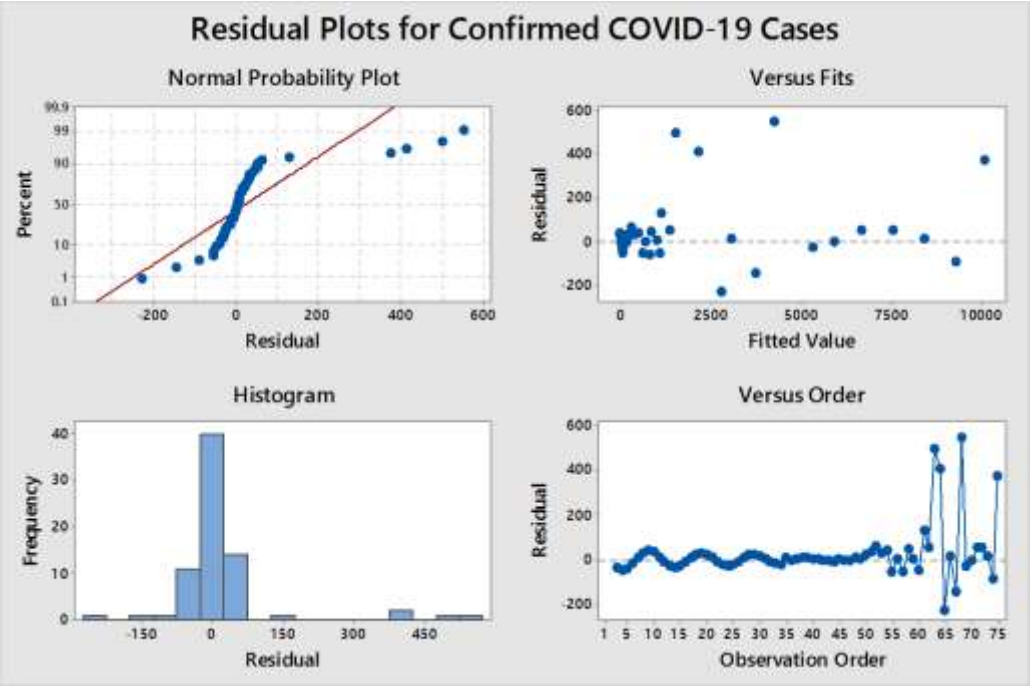

Figure 1: Residual plots for confirmed COVID-19 cases in Chennai, Tamil Nadu from 30 January 2021 to 23 April 2021.

Table 3: Ljung-Box χ^2 statistic

Lag	χ^2	Degrees of Freedom	p-Value
12	20.6	8	0.009
24	21.3	20	0.390
36	21.5	32	0.917
48	23.4	44	0.994

Hence, the model obtained after the substitution of estimated parameters is represented as,

$$X_t = 0.5358X_{t-1} - 0.2052X_{t-2} + 1.5662Z_{t-1} - 0.9491Z_{t-2} + Z_t \quad (3)$$

ARIMA (2, 2, 2) model Equation (3) is used to forecast confirmed COVID-19 cases in Chennai, Tamil Nadu for the next 20 days, 24 April 2021 to 13 May 2021. The forecast for cases is shown in Table 4 with a 95% confidence interval. According to the forecast, the number of confirmed COVID-19 cases is expected to increase considerably in the coming 20 days. This increase is highly suspected to be associated with the people involved in a large social gathering which took place just before the lockdown was imposed. Many of them have been tested positive while a number of them are still untraceable. Thus, these people may cause transmissions and lead to a higher number of infected figures. Another reason may be the carelessness on the part of a few people who didn't follow the suggested 14-day isolation after returning from abroad. Further, there is still a possibility that the transmission might be occurring from asymptomatic cases with/without a travel history. It is also supposed that many asymptomatic cases are still not tested. To some extent, social media is also contributing to some cases owing to the fake information being spread through the platform. It is very important to control such communications as they result in people moving out of their places due to wrong information's. All these situations can end up in transmissions. Apart from that, until now, it hasn't been confirmed whether a recovered person can act as a carrier of the virus or not. Further, if it is possible, then for how long.

It has been noted that some people have ignored the condition and warnings which resulted in a quick increase in the number of infected cases. Hence, it is extremely crucial that people are made aware of the situation and the lockdown is strictly imposed in the whole country to prevent further transmission of the infection. If severe measures are occupied, it is believed that the number of newly infected cases must begin decreasing in approximately 20 days. Looking at the prevention approach employed by China, that is, severe control and quarantine, it can be expected that Chennai, Tamil Nadu will also recover soon because of its similar preventive measures.

Table 4: Forecasted confirmed COVID-19 cases and their lower and upper limits for 20 days (24 April 2021 to 13 May 2021) with 95% CI.

Date	Forecast	LCL	UCL
24-Apr-21	11298.4	11072.1	11548.3
25-Apr-21	12212.2	11878.3	12546.2
26-Apr-21	13219.5	12787.8	13678.9
27-Apr-21	14269.6	13696.8	14859.2
28-Apr-21	15328.1	14541.7	16112.7
29-Apr-21	16369.2	15346.5	17439.1
30-Apr-21	17422.3	16110.7	18735.3

01-May-21	18468.6	16858.7	20089.7
02-May-21	19509.3	17569.7	21489.0
03-May-21	20549.4	19289.1	22852.8
04-May-21	21589.6	20999.8	24225.4
05-May-21	22638.8	20589.9	25678.8
06-May-21	23683.1	22227.1	27136.9
07-May-21	24727.3	22845.4	28689.9
08-May-21	25765.3	23457.4	30097.2
09-May-21	26817.5	24047.8	31593.1
10-May-21	27855.6	25617.1	33112.2
11-May-21	28989.8	26176.8	34634.8
12-May-21	29958.0	25718.3	36183.6
13-May-21	30931.1	29262.0	37736.2

Time series analysis presents the meaningful statistics for confirmed COVID-19 data. Figure 2 showed the time-series graph of the active infected COVID-19 cases from 30 January 2021 to 13 May 2021. It is clear from the plot that the time series is not stationary. An increasing trend is displayed by the time series suggesting a high rise in COVID-19 cases.

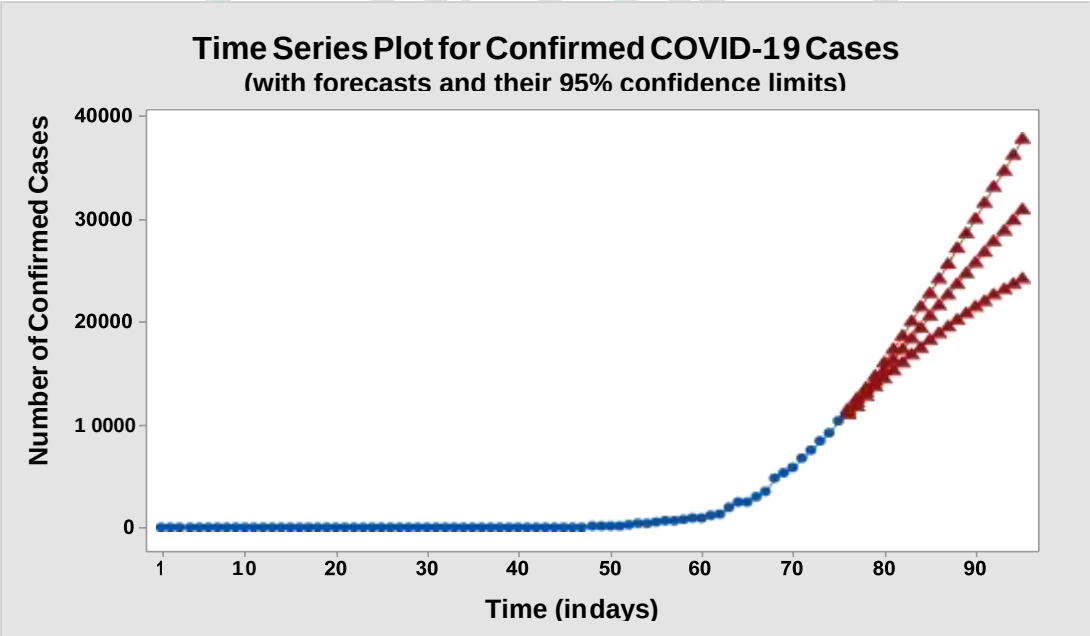

Figure 2: Times series plot for confirmed COVID-19 infections in Chennai, Tamil Nadu from 30 January 2021 to 13 May 2021 (Blue line represents actual confirmed cases and red lines represent case forecasts).

For comparing the actual and forecasted confirmed COVID-19 cases, a time series graph is plotted to start from 30 January 2021 till 23 April 2021. The plot is shown in Figure 3. The similarity of forecasted data with actual data is clear from these plots. This comparison reveals the precision of the model in forecasting.

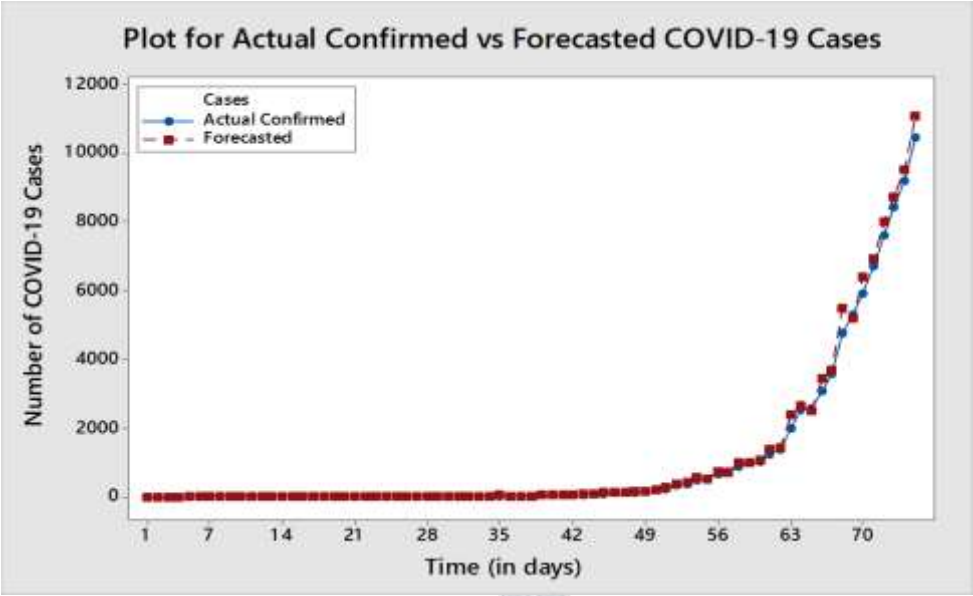

Figure 3: A comparative time series plot for actual conformed and forecasted COVID-19 cases from 30 January 2021 to 23 April 2021.

The trend for the number of recovery and death cases concerning time due to COVID-19 infections in Chennai, Tamil Nadu showed in Figure 4. It is observed that the number of recoveries as well as deaths increase with time, however, the rate of recovery is higher than the death rate. Thus, a low mortality rate could be expected from the disease.

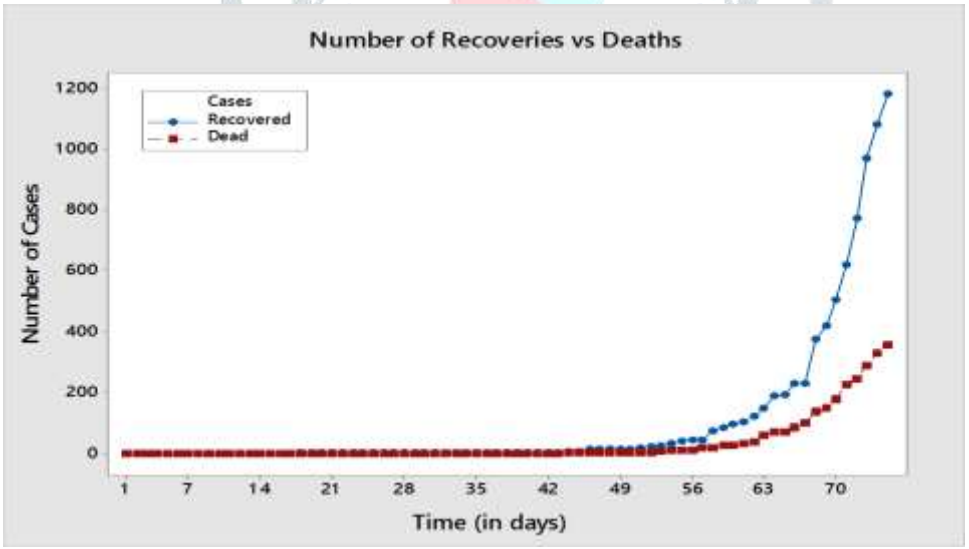

Figure 4: A comparative trend for the number of recoveries and deaths due to COVID-19 infections in Chennai, Tamil Nadu from 30 January 2021 to 23 April 2021.

Measures like quarantine and sanitization can decrease human exposure and control this pandemic. Thus, these measures should be stringently imposed in Chennai, Tamil Nadu and strict actions must be taken against those people who violate the rules and don't consider the severity of the situation. Although a large amount of data helps in providing a more exhaustive prediction and explanation, in the present circumstance, these models could be valuable in anticipating future cases of infection if the pattern of virus spread didn't change abnormally. It is obvious that this virus is new and can be transmitted intensely. Hence, it may influence the predictions, however as per our knowledge, in the present situation this model is the finest.

5. CONCLUSION

The novel coronavirus disease (COVID-19) has been declared as a pandemic by WHO and is currently a major global threat. To support the prevention of the disease and aid in the healthcare service preparation, we have conducted this study to examine the finest model for the prediction of confirmed COVID-19 infection cases and to employ that model for forecasting future COVID-19 infection cases in Chennai, Tamil Nadu. As per the model forecast, the confirmed cases are expected to greatly rise in the coming days. The time-series analysis shows an exponential enhancement in the infected cases. However, it is also anticipated that the efforts such as lockdown may affect this prediction and cases may start to decline after a month

approximately.

REFERENCES

1. Huang, C., Wang, Y., Li, X., Ren, L., Zhao, J., Hu, Y., Zhang, L., Fan, G., Xu, J., Gu, X., Cheng, Z., Yu, T., Xia, J., Wei, Y., Wu, W., Xie, X., Yin, W., Li, H., Liu, M., Xiao, Y., Gao, H., Guo, L., Xie, J., Wang, G., Jiang, R., Gao, Z., Jin, Q., Wang, J., and Cao, B. (2020). Clinical features of patients infected with 2019 novel coronavirus in Wuhan, China, *The Lancet*, 395, 497.
2. Imdadullah, M. *Time series analysis. Basic statistics and data analysis* 2014.
3. Johns Hopkins University Center for Systems Science and Engineering, Coronavirus (COVID-19) Cases.
4. N. C. P. E. R. E. Team (2020). The epidemiological characteristics of an outbreak of 2019 novel coronavirus diseases (COVID-19) in China, *China CDC Weekly*, 41, 145.
5. Niehus, R., De Salazar, P. M., Taylor, A., and Lipsitch, M. (2020). Quantifying bias of COVID-19 prevalence and severity estimates in Wuhan, China that depends on reported cases in international travellers, *medRxiv*.
6. Paules, C. I., Marston, H. D., and Fauci, A. S. (2020). Coronavirus infections more than just the common cold, *JAMA*, 323, 707.
7. Unnithan, P. S. G. (2020). Kerela confirmed the first novel coronavirus case in India, *India Today*. https://www.indiatoday.in/india/story/kerala-reports-first-confirmed-novel-coronavirus-case-in-india- 1641593-2020-01-30.
8. Wikipedia, 2019-20 coronavirus outbreak. https://en.wikipedia.org/wiki/2019-20_coronavirus_outbreak.
9. World Health Organization, Coronavirus disease (COVID-19).
10. Zhang, Y., Yang, H., Cui, H., and Chen, Q. (2019). Comparison of the Ability of ARIMA, WNN and SVM Models for Drought Forecasting in the Sanjiang Plain, China. *Nat. Resour. Res.*, 29, 1447.
11. Zhu, N., Zhang, D., Wang, W., Li, X., Yang, B., Song, J., Zhao, X., Huang, B., Shi, W., Lu, R., Niu, P., Zhan, F., Ma, X., Wang, D., Xu, W., Wu, G., Gao, G. F., and Tan, W. (2020). A novel coronavirus from patients with pneumonia in China, 2019. *N. Engl. J. Med.*, 382, 727.