On the Temporal Analysis of COVID - 19 Pandemic and Prediction of R_0

Kshitij Patil[+,1], Anirudh Murali[+,1,2], Piyali Ganguli[1,2], Sutanu Nandi[1,2], Ram Rup Sarkar*[,1,2]

[1]CSIR National Chemical Laboratory, Dr Homi Bhabha Road, Pune 411008

[2]Academy of Scientific and Innovative Research (AcSIR), Ghaziabad- 201002, India

*corresponding author; Email: rr.sarkar@ncl.res.in

[+] these authors contributed equally to this work

ABSTRACT

The COVID-19 pandemic has affected millions of people and claimed numerous lives already. As of now, there are no available treatments and it has become both imperative and challenging to forecast the COVID-19 cases, which will help to design effective clinical management and policy to fight the pandemic. With the objective to forecast the COVID-19 cases and Basic Reproductive Number (R_0) country-wise, for more than a month ahead, we have adopted a data driven approach that employs Multiple Aggregation Prediction Algorithm (MAPA) for temporal predictions. Our strategy applies MAPA in two ways. The first is the direct application on the number of cases and second is by calculation of R_0, followed by MAPA. This novel workflow generates a Principal and an Exponential Prediction that provides a range of values within which the total number of cases is expected to lie. The strategy and workflow have been validated for long term predictions with 51 countries in different growth phases. Thereafter, we have made predictions of the possible number of COVID-19 cases for the next 45 days of these 51 countries, the world as a whole and the other 160 countries combined, that are affected by the pandemic.

Introduction

The spread of Coronavirus Disease (COVID-19) cases across 216 countries has brought the World under serious threat. Not only has the pandemic affected 8.8 million people and claimed over 0.46 million lives, but it has also hugely impacted the global economy as well [1]. As people patiently wait through the lockdowns imposed across different countries for the development of an effective treatment strategy and the development of a vaccine, it becomes imperative to be able to forecast the COVID-19 cases for designing effective strategies and policy making to fight the pandemic and heal the scars that it is leaving behind in lives of the people as well as on the global economy [1,2].

Although several researchers are striving hard to be able to predict when and how the spread of COVID-19 cases can be controlled, the world is witnessing a second wave of cases in few countries like China, South Korea, Iran, etc., where it was assumed to have been contained. This calls for an urgent need of effective tools and methodologies that can accurately predict the positive COVID-19 cases that can be expected at least more than a month ahead for advanced planning, resource allocation and strategy development in order to reduce the further spread of the disease and curb the pandemic. This is especially the need of the hour for the under-developed and developing countries, where lockdown is gradually being lifted now due to huge economic crashes and the question arises, what is the expected number of COVID-19 cases that is likely to be witnessed in the next one month? This is a very crucial question in these countries where the medical facilities and resources are currently limited and hence require a proper strategy design, a priori, for clinical management to combat the growing pandemic.

In order to predict the number of COVID-19 cases in advance, various modelling strategies have already been employed by several researchers across the globe, which includes compartmentalized models as well as data-driven models [3–20]. Among these, the compartmentalized models, that mostly consists of the

Ordinary Differential Equation based SEIR (Susceptible-Exposed-Infected-Recovered) models have been modified and improvised with the addition of several variables that represent the Asymptomatic Class, Pre-symptomatic Class, Mildly Symptomatic Class, Hospitalized Class, etc. [21,22]. Additionally, various factors regulating disease transmission have also been considered such as the effect of lockdowns and movement or migration patterns between cities by using the transport networks that contribute heavily towards the spread of the disease [23]. Over the years, although these kinds of SEIR models have been invaluable in the study of epidemiology and are intrinsically capable of predicting disease dynamics with precision, their accuracy, however, is heavily dependent on the correct estimations of the several unknown parameters that regulate the model outcome.

Here, in this case of novel COVID-19 pandemic, the difficulty of developing such parameterized models lies in the fact that most of the parameter values are still unknown and eventually will take more time to get a proper estimation to be used in such models. Hence, predicting the future dynamics of the epidemic becomes difficult due to the lack of knowledge of several of these parameter values that are required for an accurate prediction using SEIR modelling strategy. On the other hand, data-driven models, which make predictions based on the real-time data and observations, without any dependence on assumptions or unknown parameters may provide a much more realistic time series prediction for the study of the novel COVID-19 epidemiology. Data driven models, such as ARIMA, have been extensively used for predicting COVID-19 cases by several research groups [24–27]. Although it has shown high accuracy for short term predictions, ARIMA generally fails to perform properly for long term predictions. Additionally, yet another limitation of using ARIMA modelling is that the selection of the best model may not always be optimal even though they may seem to be good theoretically. This selection of the most appropriate time series model to achieve a good forecasting accuracy is tricky. Several Machine Learning approaches such as CUBIST regression, Random Forest, RIDGE regression, Support Vector Regression (SVR), and Stacking-Ensemble learning have also been tested for prediction of COVID-19 cases for 10 states in Brazil [28]. Although they perform reasonably well for predictions up to 6 days ahead, they may or may not perform adequately for long term predictions due to the accumulation of errors at each step.

Hence, in order to circumvent these shortcomings of model development, and with the objective to make at least 45 days long predictions for the total number of COVID-19 positive cases for different countries across the globe, for the first time, we have implemented the Multiple Aggregation Prediction Algorithm (MAPA) on the available datasets for 51 countries and the world as a whole. MAPA takes advantage of non-overlapping temporal aggregation of the time series, which can make many components of the time series distinct, and has direct effects on model identification and estimation [29]. Multiple time series are derived from the original one using temporal aggregation. Here, exponential smoothing is used on each derived series because it splits the series into level, trend and seasonal parts during modelling. One of the key advantages of this algorithm is that it reduces the importance of model selection with the help of forecast combination of each derived series. This algorithm has been shown to work better especially for long term forecasts [29].

Our strategy of using MAPA for long term predictions involves a Principal Prediction (PP) that is derived from the model training using past occurrences of the reported positive cases and an Exponential Prediction (EP) that is derived from the calculated the Basic Reproductive Number (R_0) of the COVID-19 for the month ahead. This Exponential Prediction accounts for the sudden exponential growth, after the model training period, that has been observed for a few countries such as India, Afghanistan, Peru among others. These Principal and Exponential Prediction bands proposed from our study essentially cover a range of values that is likely to be observed of COVID-19 cases for the next 45 days. In this study, we have validated this strategy for month long predictions. Additionally, we have also compared the performance of ARIMA and MAPA models for both short term and long-term predictions for several countries and we propose a novel workflow that can be used to predict the temporal dynamics of COVID-19 cases for 1.5 months ahead with better accuracy. This strategy being completely data-driven does not rely on unknown parameter values, which enhances its comprehensibility and reproducibility for future predictions.

Results

Model Development and Validation

COVID - 19 time-series datasets of 51 countries and the world that were considered for training and testing of the proposed model are non-linear, non-gaussian and non-stationary in nature. Based on the actual data for 10 days after the model training period, the countries have been categorized into three different groups namely, a) Exponential (Group A), b) Rising (Group B) and c) Plateauing (Group C) (**Supplementary Table S1**). Among these Group A countries, it can be observed that the exponential jump in the number of cases was noted either during the first or second week of May**.**

In order to validate the performance of our model for the month-long predictions, our strategy has been implemented on all the Groups A, B and C countries for the whole month of May 2020 (**Fig. 1, Fig. 2, Fig. 3** representing 6 countries from each group). Here we observe that the prediction for each country has two main bands, one for the Principal Prediction (PP) (blue dots surrounded by red confidence bands of 99%, 90% and 80%) and one for Exponential Prediction (EP) (solid yellow line surrounded by blue confidence bands). Principal prediction as such captures the underlying trend of data, whereas exponential prediction derived from the R_0 value acts as a good predictor for unforeseen and spurious cases.

The grey area in between these two bands has been termed as transition zone **(Fig. 1)**. In this transition zone, a country reporting a spike in the number of cases can tread into exponential prediction zones. Noting the fact that a country continuously can never be in the exponential prediction zone, number of cases will eventually stride into the transition zone then to the principal prediction zone. On a long-term basis, we expect that both principal and exponential prediction will effectively capture the persistent or spurious rise in cases.

For testing the model performance, box plots of the principal and exponential predictions were compared with the actual data that depicts the cumulative number of reported cases; and calculated the Percentage errors that are observed temporally for the month of May (**Supplementary Fig. S1-S8**). Here it was observed that the nature of the actual curve may vary but it mostly lies within our range of Principal and Exponential Predictions. During May, Brazil, Pakistan, Qatar and Russia (**Fig. 1**, **Supplementary Fig. S9**), moved from the principal band into the transition zone. Egypt, Mexico and Peru moved from the principal band through the transition zone to the exponential band. Sri Lanka had a similar movement towards the end of May, in the sense that it made one oscillation about the principal prediction in the principal band.

Among all countries, Chile, Nepal (**Fig. 1**) are outliers as our model under-predicted the sudden unforeseeable jump in the number of cases, as the underlying trend did not exist in the training data. Here it was observed that the Actual data lies even beyond our Exponential Prediction band.

While the predictions of most of the Group A countries lie in the exponential region, for some countries, nRMSE and MAPE point proximity of the Actual data towards Principal prediction. However, it has to be noted that these countries are gradually moving from the principal prediction band towards the transition zone. If the spread of the virus is not contained it is entirely possible that in the next 5 - 14 days the concerned country might experience a spurt in the number of cases being reported and lie in the exponential band.

Fig. 1 depicts the model training and validation of 6 countries from Group A (for other Group A countries refer **Supplementary Fig. S9**). Here we observe that India, which is the second most populated country in the world, falls under the Group A category and has witnessed an exponential rise in the number of COVID-19 cases across May. For India, our model had predicted the number of cases through the exponential band. The cases in India have inched into the Exponential Prediction Band. By the end of May the number of cases shoots out of the exponential band. Afghanistan has a similar trend. Brazil and Russia have been observed to lie in the transition zone.

The model training and validation of six countries of Group B have been depicted in **Fig. 2** (for other Group B countries refer **Supplementary Fig. S10**). This group consists of the countries where the numbers of COVID -19 cases are on steady rise. Here, the slope of these countries does not change sharply and hence the Principal prediction is better suited to predict the trends of the total number of cases. However, in the case of Ecuador (**Supplementary Fig. S10**), we observed an overlap of Exponential prediction and Principal prediction bands due to fluctuations and irregularity in the training data. Egypt observes complete movement from the principal prediction band to the exponential prediction band highlighting the importance of both predictions. For most of the countries in this group, the actual data is predicted better by the Principal prediction than the Exponential prediction even though the Actual data in the plot is closer to the 99% CI of Principal prediction in some cases. The slope of the above-mentioned countries does not change sharply in the principal zone and based on the predictions we can predict that the number of cases being reported by these countries will lie in the Principal prediction.

The Group C, consists of the countries where the sequential daily rise in the number of COVID -19 cases have reduced significantly and slope of the Actual data is almost constant (**Fig. 3 and Supplementary Fig. S11 – S12**). Plateauing of cases was possible due to restrictions and awareness measures taken up by the Government and citizens. Our predictions suggest that the total number of cases for the above-mentioned countries will not see any major rise and the predicted value will lie in the Principal prediction zone and the rise in case will move along the Mean of the Principal prediction. **Fig. 4** shows the validation for the world as a whole.

The nRMSE and MAPE values for the MAPA Principal Prediction and MAPA Exponential Prediction, calculated against actual data collected from May 1 to June 02 are tabulated in **Supplementary Table S1.** To gain different perspectives about how the suggested models capture the movement of the actual data, the boxplots of the actual and predicted values and the temporal plot of the percentage errors are shown in (**Supplementary Fig. S1-S8**). **Supplementary Table S2** provides comparison between the actual and predicted $R_0(t)$ values.

Model Comparison between ARIMA and MAPA

Short Term Comparison

In order to estimate how close predictions are to the actual number of cases predicted for 10 days during 1^{st} -10^{th} May 2020, we have compared our results using MAPA of both our Principal as well as the Exponential Predictions with the real data. **Supplementary Fig. S13** depicts the time series prediction alongside the real data for this duration on six countries representing the Group A, B and C categories. Here we note that the countries which have exponential growth in the first week of May (Group A), lie well within the ranges of the principal prediction (represented with blue open circles) and the exponential MAPA prediction (represented with yellow line) and are often seen moving between the two predicted bands.

In case of countries like India **(Supplementary Fig. S13.a, b**) and Peru (**Supplementary Fig. S13.c, d**), our model simulates the actual data well and it can be found that the actual data does not lie within the bands of principal prediction and moves towards the exponential prediction bands. MAPE points towards the closeness of the actual data towards the exponential band. (**Supplementary Table S3**)

Countries like Sweden (**Supplementary Fig. S13.e, f**) and Romania (**Supplementary Fig. S13.g, h**) were chosen to represent the group of countries where the number of COVID 19 cases are still rising (group B). The principal prediction models the actual data for Romania compared to ARIMA, while for Sweden, ARIMA is slightly better than MAPA. (**Supplementary Table S3**)

Austria (**Supplementary Fig. S13.i, j**) and Portugal (**Supplementary Fig. S13.k, l**) were categorized under plateauing (Group C) number of COVID-19 cases and the predicted values were closely lying in the principal prediction band as compared to ARIMA. (**Supplementary Table S3**)

Overall, from **Supplementary Table S3** and **Supplementary Fig. S13**, we note that MAPA outperforms ARIMA for all six countries except Sweden. From the plots it can be observed that the rate of change of new cases started to taper over time. Geographically Austria is bordered by Germany, Italy (worst affected Northern Italy) and Portugal by Spain. Germany, Italy and Spain are the worst affected countries in terms of COVID - 19 infections and deaths. In the case of Austria and Portugal, early closure of borders, closure of flight routes from China coupled with self-isolation and quarantine led to tackle COVID-19 better and achieve a shorter peak in the fastest time.

Long Term Comparison

In the long-term predictions obtained by using ARIMA and MAPA (**Supplementary Fig. S14)**, we observe that in the case of China **(Supplementary Fig. S14.a, b)** and South Korea **(Supplementary Fig. S14.c, d)**, the MAPA principal prediction matched the actual data well, ARIMA severely underperformed and MAPA Exponential prediction over performed for long term predictions. Whereas in the case of the USA **(Supplementary Fig. S14.e, f),** MAPA Principal prediction matched the Actual data, while both ARIMA and MAPA Exponential prediction over performed. For all 3 countries, we observe from the box plots that our MAPA principal prediction simulates the actual data values the better in comparison to ARIMA. (**Supplementary Table S4**)

Model Prediction for the Use Case: 45 days forecast (3rd June to 17th July)

After successful validation of our proposed workflow on 51 countries for long term predictions, we aim to forecast the expected number of cases in the subsequent 45 days, for the duration 3rd June to 17th July. For the sake of simplicity, the countries have been retained in the same group that they were classified into during validation. However, one important question that comes up is which band, i.e. the Principal or the Exponential, is more likely to contain the actual data through the 45 days of forecasts. As discussed in an earlier section, the nature of the actual curve may vary and a country may move from one band to the other over the course of the forecast period.

Forecast for 51 countries

According to our analysis, for the duration 3rd June to 17th July, India, which is in Group A, should brace itself as the number of cases is on the rise, and given that easing of Lockdown has started, social distancing will be harder to follow day by day. In May, while the lockdown was still in effect, the actual

number of cases lied in our exponential prediction band. If the cases continue along this exponential prediction band, India may expect to reach up to 1.18 million cases by mid-July. However, the principal prediction suggests about 0.58 million cases by mid-July. USA, which is in group B, should expect the Total Number of Cases to move into the Transition Zone and further into the Exponential bands based on the extent to which social distancing norms are practised in the current situation. Should the cases continue to be mapped by the principal prediction, USA can expect about 2.75 million total cases by mid-July.

For the countries in group C, it is unlikely that exponential growth will occur to such an extent that the Total Number of Cases will even be in the transition zone, and hence it is expected that the Principal Prediction for these countries is the far more likely scenario.

Forecast for the Rest of the World

The World as a whole will show a steady rise in the number of cases (behaviour similar to Group B countries). The 51 countries that have been considered in our study, have contributed in total at least to 91% of the total number of reported cases in the world as of June 02. They also show a gradual rise in the expected number of COVID-19 cases (behaviour similar to Group B countries). The remaining countries that contribute to less than 10% of the total number of cases also show a similar behaviour (**Supplementary Table S1, Fig. 4**). This implies that overall, the total number of cases is rising at the global scale.

Of the 51 selected countries, 25 countries were beginning to plateau or were already plateauing in May, while the remaining 26 countries showed signs of considerable growth. If things do not get worse, then the world can expect to report 11.2 -14.3 million cases by mid-July based on our Principal Prediction. What this essentially means is that, based on the Principal Prediction it is very likely that in the next 45 days, the virus will still be around in at least 25 countries.

$R_0P(t)$ and $R_0E(t)$ forecast values on 03rd June and 17th July are tabulated in **Supplementary Table S5**. The forecasted values $R_0P(t)$ of all the countries are monotonically decreasing. Values of $R_0E(t)$ is constant for most of the countries except Nepal (increasing), Israel (decreasing), Pakistan (decreasing), Poland (decreasing) and Turkey (decreasing).

Discussion

In this study, we propose a strategy to predict the expected number of positive COVID-19 cases for at least a month ahead using a data-driven approach. Here, using the MAPA, we provide the range of values from our principal prediction to the exponential prediction that provides an idea of the exact range where the number of cases may lie assuming that no drastic changes in nature (rate) of the calculated $R_0(t)$ occur during this period. This exception has been observed only in the case of Chile and Nepal. However, even though we provide this wide transition range, we observed that in most of the cases, the actual number of cases lie either within the principal prediction band or the exponential prediction band.

Through this study we demonstrate a comparison of ARIMA and MAPA strategy for long term predictions, using the dataset of China, USA and South Korea till 10th April 2020. Here it has been observed that for short term predictions although the two methodologies are almost comparable, with MAPA performing slightly better than ARIMA in most of the cases, the long-term predictions clearly highlight the advantages

of using MAPA over ARIMA. Here it is observed that MAPA shows much lower RMSE even for month long predictions.

During May, the number of cases for India and Peru lies within the exponential band even though they were under lockdown. In the case of India, the government underwent a strict phase of lockdown from 24th March to 14th April 2020. Subsequently, during phase 2, the lockdown was conditionally relaxed after 20th April to allow farm, dairy and allied activities. COVID-19 has an incubation period of 5-14 days, it is still possible that the infection could have occurred before we started training the data. On 29th April, the Indian government allowed inter-state travel for stranded people, interstate labourers and students subject to conditions like quarantining and periodic testing, it may be noted the spike in the number of cases may be due to travel, poor social distancing norms and better screening and testing. Reduction in the R_0 over the given period can be attributed to the precautions taken up by the countries and its people to mitigate the infection risk. Reduction in the R_0 of India and Sweden are comparable. But as mentioned previously Sweden hasn't implemented lockdown till date, whereas India has implemented three graded lockdowns. Even though it could be argued that lack of proper implementation of precautions led to the minimum change in R_0, it should also be noted that testing, contact tracing and ramping up of medical capacities led to better screening of COVID-19 infected. So, although China has the lowest value of R_0, Peru has a higher reduction in the daily number of newly infected people on average. Thus, from our results and analysis, we observe that a country with lower R_0 than another need not have a lower rate of growth, as we see with our comparison of Peru and China.

Based on the validation of 33 days of actual data with our predictions for May, we observe that our exponential prediction for India correctly predicts the date when India crosses 100,000 cases which is 18th May 2020. The world crossed 4 million total number of reported cases on 11th May 2020 while our principal prediction predicted 4 million total number of reported cases on 12th May 2020. The United States reported 1.5 million total cases on 19th May 2020 which was also captured by our principal prediction. We observe that the United States has the highest number of cases followed by Russia and Brazil in that order. Both Russia and Brazil are in our transition zone, and Brazil's exponential band is above Russia's. By the end of May, Brazil crossed Russia's total number of cases. This could have been inferred at the start of May by the exponential prediction of the countries. On the other hand, Nepal has just seen the start of the outbreak and can expect the cases to exponentially rise with a steeper slope, so the EP may be unable to capture the movement, as hinted in the validation of Chile. This can be accounted for by changing the value of *n* from 7 to a higher number.

Verification of the Use Case results was performed from 3rd June to 20th June, 2020 (till the date of submission). The errors are listed in **Supplementary Table S6**. So far, except for Nepal, the Principal Prediction is performing better than the Exponential Prediction. Out of the 51 countries and the World, MAPE values of 48 regions are less than 10%; 41 regions are less than 5%; and 33 regions are less than 2%. The maximum MAPE value is 13.86% for Bulgaria which is moving into the transition zone (Brazil, Pakistan, Peru and Sweden show similar movement). The comparison of the Actual Values with the Principal Prediction and Exponential Prediction of the countries (03 June to 20 June) depicted in **Fig.1-3** are tabulated in **Supplementary Table S7-S9.** The number of cases forecasted by the Principal Prediction and the Exponential Prediction for all countries on the last day of the forecast period, i.e. 17th July is tabulated in **Supplementary Table S10**.

New Zealand and Slovenia have declared themselves free of Covid19. The principal predictions of both countries have captured this fact (**Fig. 3; Supplementary Fig. S12**). Countries with principal predictions of similar nature can expect the same. (**Fig. 3, Supplementary Fig.S11-S12**)

Nonetheless, this strategy also has some limitations. Here the exponential prediction is a function of the modelled $R_0(t)$. Hence the exponential prediction is dependent on the nature of $R_0(t)$. We use a constant value of the serial interval for all countries. Our model cannot account from a drastic (unstable) change in the new number of cases or equivalently the rate of the total number of cases may not be detected even with the account for exponential growth. Also, for some countries the transition range of predictions is wide. The EP may not be required for countries that are plateauing or have lesser chances of such growth (which can be concluded with the help of additional region-specific information).

Further modifications can be made to specific country analysis by accounting for exponential growth at a certain time after training, not just the first *n* days after and selecting more region appropriate values of v (mean serial interval) and f (ratio of latent period to mean serial interval). A dynamic & adaptive extension can be made to the workflow to make *h* day forecasts, then updating the forecasts every *m* days. This will be helpful as it will highlight potential likely scenarios in the next *h* since the training. The first *m* of these *h* days can be used to strategize and make necessary arrangements for the required plan of action. On the m^{th} day, validation of the data can be done, and the training data can be updated on which the model will be applied. Now the $(h\text{-}m)^{th}$ day forecast will be the h^{th} day forecast and will surely provide a better idea of the true scenario. This will help in the formation of a structured plan to accommodate a large number of cases *h* days in advance.

The month-long predictions that we have made can be used by policymakers to make decisions by considering the entire movement range, or using the detailed information with them about their region and choosing either of the bands. Our predictions can also be used as data for analysis by our fellow scientists as this provides insights into the daily new number of cases which is an important indicator of the growth of the COVID-19 cases in a region.

Materials and Methods

Data Collation and Processing

Out of the 216 countries affected by the COVID-19 pandemic, 51 countries were chosen based on where they seem to be in their COVID-19 epidemic phase. This includes countries varying from rising cases, flattening cases and countries that have been declared COVID-19 free. We also perform our analysis on the World as a whole. These 51 countries selected for our analysis contribute towards 91% of the total number of reported COVID-19 cases till 2nd May.

The temporal data of the total number of confirmed COVID-19 cases reported daily over a span of approximately 99 - 103 days (depending on the country) were used for our analysis. These data were collated from sources like European CDC, WHO and Indian Council of Medical Research (ICMR), which have been made available publicly through various repositories such as Our World in Data (https://ourworldindata.org/) and The Humanitarian Data Exchange (https://data.humdata.org/).

The data of 51 countries were then processed to call the non-zero values of cases of each country to determine the first day of positive COVID-19 case for that country. In **Supplementary Table S7**, the 51 countries have been listed that have been used in our study along with the first reported day of the virus as per the collated data. The training dates for the long term and short-term comparison purposes with ARIMA, validation and use cases have been tabulated in **Supplementary Table S8**.

Calculation of Basic Reproductive Number (R_0)

The basic reproductive number of infections, R_0, is defined as the expected number of secondary infectious cases generated by an average infectious case in an entirely susceptible population. In order to estimate the basic reproductive number $R_0$of COVID-19 infection, we consider the formula as denoted in **Eq. (1)** [30].

$$R_0 = 1 + r_0 v + f(1-f)(r_0 v)^2 \quad \text{(Eq. 1)}$$

$$R_0 = \frac{\left(ln(Y(t))\right)}{t} \quad \text{(Eq. 2)}$$

where,r_0 is the exponential growth rate, $Y(t)$ being the number of total confirmed cases on the t^{th} day after the first reported case. In **Eq. (1)**, we assume the serial interval (v) to be 5.1 days [31] and assume f as 0.99, which represents the ratio of the latent period and mean serial interval. To get the number of cases from R_0 we make r_0 the subject in **Eq. (1)**. Since we will be using estimated values to recapture the total number of cases, we denote r_{0h} the estimated exponential growth rate, R_{0h} as the estimated basic reproductive number and $Y(t)_h$ as the estimated total number of cases. In terms of time series, we can think of **Eq. (1)** as a transformation of the data.

$$r_{0\,h}(t) = \frac{\left(-1 + \sqrt{1 - 4f(1-f)\left(1 - R_{0\,h}(t)\right)}\right)}{\left(2vf(1-f)\right)} \quad \text{(Eq. 3)}$$

$$Y_h(t) = e^{t r_{0h}(t)} \quad \text{(Eq. 4)}$$

Time Series Modelling using MAPA

Multiple Aggregation Prediction Algorithm (MAPA) creates multiple time series from the original series using temporal aggregation [29]. MAPA considers a time series under multiple temporal aggregation levels simultaneously, with the aim of capitalising on the advantages of each aggregated and original time series. Since no single level is selected, different forecasts are produced for the same point. The resulting information can be combined in a single prediction, the output of MAPA. Thus, the algorithm has three main steps. The first step consists of the temporal aggregation of original time series, then comes the forecasting at all aggregation levels which is followed by the combination of forecasts.

In this study, the input data for the MAPA algorithm consists of the total number of reported COVID-19 cases of each of the 51 countries, starting from the first day of report till the required date. **Fig. 5** illustrates the workflow followed for the analysis. Here, the total number of reported cases data has been

denoted as Y(t), which is our time series of interest. On this data we apply two different strategies to arrive at our time series prediction (denoted by blue and green panels in **Fig. 5**).

In the first strategy (left blue panel, **Fig. 5**), MAPA has been directly implemented on Y(t). To perform MAPA on our data we used the MAPA: Multiple Aggregation Prediction Algorithm 2.0.1 package in R [32]. Here, we make predictions for *h* days after training to obtain YP(t). This gives us the Principal Prediction (PP) (green box in the left blue panel, **Fig. 5**) from the data which is achieved by using the general MAPA strategy (**Fig. 1**). Then, $R_0P(t)$ is derived by applying **Eq. (1) and Eq. (2)** on this predicted YP(t).

In the second strategy (right green panel, **Fig. 5**), a novel modified MAPA strategy is then adapted to account for sudden changes in the number of cases being reported which may be due to preventive measures or the lack thereof or even a realization of under-reporting due to late recognition of infected individuals. In this strategy, **Eq. (1) and Eq. (2)** are applied on Y(t) to obtain $R_0(t)$. MAPA is implemented on this $R_0(t)$ to make predictions for *h* days which is denoted by $R_0E(t)$. We then use **Eq. (3) and Eq. (4)** on $R_0E(t)$ to get the corresponding total number of cases denoted by Y1(t). **Eq. (4)** being exponential in nature results in Y1(t) being exponential even if the predicted part $R_0P(t)$ is almost constant. Y1(t) grows exponentially throughout. We limit this growth to *n* number of days after training, where *n* is suitably chosen. We then create a subset Y1(t) from the starting to the first *n* days of prediction and call this subset Y2(t). We then perform MAPA on Y2(t) and make predictions for the remaining number of (*h* - *n*) days. This gives us the Exponential Prediction (EP) of Y(t) (green box in the right green panel, **Fig. 5**). Here, by exponential growth, we mean to account for the case where $R_0(t)$ may be almost constant or be increasing for a few days after the training is complete and is, therefore, relative to it. It is to be noted that each of these predictions includes both modelled and predicted values.

Error metrics

In order to estimate the model performance on testing datasets for both long term and short-term predictions, error metrics such as Root Mean Squared Error (RMSE) and Mean Absolute Percentage Error (MAPE) have been calculated to quantify how far the estimates have deviated from the actual values on an average. They have been defined in the **Eq. S1 - Eq. S3.**

Model Comparison for Long Term and Short-term Predictions: ARIMA versus MAPA

To show comparison of ARIMA and MAPA for long term predictions, we implement our strategy on two different data sets. In the first dataset we consider 6 countries, namely, India, Peru, Sweden, Romania, Austria and Portugal, where we have considered training data till 30th April and make predictions from 1st May to 10th May, i.e. for 10 days (**Fig. 5)**. This has been used to compare the accuracy short term predictions between MAPA and ARIMA. On the other hand, for the long-term prediction and comparison, our workflow has been implemented on the second dataset which consists of data from China, South Korea and USA, which are the first few countries to see the arrival of the virus. Here, we consider training data from the starting date till 10 April, 2020 and make predictions from 11th April to 10th May 2020 using both MAPA and ARIMA, i.e. for 30 days.

In this study, ARIMA has been performed using the *auto.arima* function from the *forecast package* in R after suitable data transformations and MAPA has been performed using the *mapa* function from the *MAPA package* in R [32,33]. The performances of the two strategies have been estimated using the RMSE and MAPE values calculated for the MAPA and ARIMA predictions.

Model Validation for the period 1st May to 2nd June 2020

In order to validate the model for long term predictions, our workflow have been implemented on all countries and the world data as a whole, starting from each of their respective first dates up to 30th April 2020. Predictions were then made from 1st May 2020 to 2nd June 2020 for the principal prediction. Thereafter, considering n=7, i.e. we assume exponential growth for 7 days after training (30th April), from 1st May to 7th May for all countries. This is performed only to account for possible exponential growth in the first week of May. Then, Y2(t) subset was created from the Y1(t) data from the starting date to May 7th. MAPA is then performed on Y2(t) and predictions are made for 8th May to 2nd June (33 - 7 = 26 days). We started this exercise on 30th April and waited till 2nd June to validate our results. Then using actual data from 1st May 2020 to 2nd June 2020 we compare our predictions and report both the nRMSE and MAPE values.

Use case: Forecasts for the period 3rd June 2020 to 17th July 2020

In order to predict the possible number of COVID-19 cases for all the 51 countries and also the rest of the world in the next 45 days, our strategy has been implemented on all countries and the world starting from each of their respective first dates up to 2nd June 2020. We then make forecasts for h = 45 days, i.e. from 3rd June 2020 to 17th July 2020 for our principal prediction. Thereafter, assuming exponential growth for 7 days (i.e., n=7) after 2nd June, from 3rd June to 9th June for all countries, we have tried to account for possible exponential growth. Following the similar workflow, a data subset Y2(t) has been created from Y1(t), i.e. from the starting date to June 9. MAPA is then performed on Y2(t) and predictions are made for 10th June to 17th July (45 - 7 = 38 days).

Acknowledgement

RRS acknowledges funding from Department of Science and Technology, Govt. of India (Sanction Order No. DST/ICPS/EDA/2018/General, dated 30.04.2019), PG acknowledges CSIR for Senior Research Fellowship. SN acknowledges DST INSPIRE Fellowship.

Conflict of interest statements

Authors declare no competing interests.

Data availability

All data needed to evaluate the conclusions in the paper are present in the paper or the Supplementary.

References

1. McKee, M. & Stuckler, D. If the world fails to protect the economy, COVID-19 will damage health not just now but also in the future. *Nat. Med.* **26**, 640–642 (2020).

2. Nicola, M. *et al.* The socio-economic implications of the coronavirus pandemic (COVID-19): A review. *Int. J. Surg.* **78**, 185–193 (2020).

3. Arino, J. & Portet, S. A simple model for COVID-19. *Infect. Dis. Model.* **5**, 309–315 (2020).

4. Cuevas, E. An agent-based model to evaluate the COVID-19 transmission risks in facilities. *Comput. Biol. Med.* **121**, 103827 (2020).

5. Mollalo, A., Vahedi, B. & Rivera, K. M. GIS-based spatial modeling of COVID-19 incidence rate in the continental United States. *Sci. Total Environ.* **728**, 138884 (2020).

6. Marimuthu, Y., Nagappa, B., Sharma, N., Basu, S. & Chopra, K. K. COVID-19 and Tuberculosis: A mathematical model based forecasting in Delhi, India. *Indian J. Tuberc.* (2020).

7. Ivorra, B., Ferrández, M. R., Vela-Pérez, M. & Ramos, A. M. Mathematical modeling of the spread of the coronavirus disease 2019 (COVID-19) taking into account the undetected infections. The case of China. *Commun. Nonlinear Sci. Numer. Simul.* **88**, 105303 (2020).

8. Shen, C. Y. A logistic growth model for COVID-19 proliferation: experiences from China and international implications in infectious diseases. *Int. J. Infect. Dis.* (2020).

9. Vega, D. I. Lockdown, one, two, none, or smart. Modeling containing covid-19 infection. A conceptual model. *Sci. Total Environ.* 138917 (2020).

10. Saba, A. I. & Elsheikh, A. H. Forecasting the prevalence of COVID-19 outbreak in Egypt using nonlinear autoregressive artificial neural networks. *Process Saf. Environ. Prot.* **141**, 1–8 (2020).

11. Chatterjee, K., Chatterjee, K., Kumar, A. & Shankar, S. Healthcare impact of COVID-19 epidemic in India: A stochastic mathematical model. *Med. J. Armed Forces India* **76**, 147–155 (2020).

12. Reis, R. F. *et al.* Characterization of the COVID-19 pandemic and the impact of uncertainties, mitigation strategies, and underreporting of cases in South Korea, Italy, and Brazil. *Chaos, Solitons and Fract.* **136**, (2020).

13. Ravinder, R. *et al.* An Adaptive, Interacting, Cluster-Based Model Accurately Predicts the Transmission Dynamics of COVID-19. *medRxiv* 2020.04.21.20074211 (2020).

14. Qi, H. *et al.* COVID-19 transmission in Mainland China is associated with temperature and humidity: A time-series analysis. *Sci. Total Environ.* **728**, 138778 (2020).

15. Iwata, K. & Miyakoshi, C. A Simulation on Potential Secondary Spread of Novel Coronavirus in an Exported Country Using a Stochastic Epidemic SEIR Model. *J. Clin. Med.* **9**, (2020).

16. Zhu, X. *et al.* Spatially Explicit Modeling of 2019-nCoV Epidemic Trend based on Mobile Phone Data in Mainland China. *medRxiv* 2020.02.09.20021360 (2020).

17. Li, H., Liu, S. M., Yu, X. H., Tang, S. L. & Tang, C. K. Coronavirus disease 2019 (COVID-19):

current status and future perspectives. *Int. J. Antimicrob. Agents* **55**, 105951 (2020).

18. Kissler, S. M., Tedijanto, C., Goldstein, E., Grad, Y. H. & Lipsitch, M. Projecting the transmission dynamics of SARS-CoV-2 through the postpandemic period. *Science.* **368**, 860–868 (2020).

19. Gatto, M. *et al.* Spread and dynamics of the COVID-19 epidemic in Italy: Effects of emergency containment measures. *Proc. Natl. Acad. Sci.* **117**, 10484 LP – 10491 (2020).

20. Wells, C. R. *et al.* Impact of international travel and border control measures on the global spread of the novel 2019 coronavirus outbreak. *Proc. Natl. Acad. Sci.* **117**, 7504 LP – 7509 (2020).

21. Sardar, T., Nadim, S. S. & Chattopadhyay, J. Assessment of 21 Days Lockdown Effect in Some States and Overall India: A Predictive Mathematical Study on COVID-19 Outbreak. *arXiv:2004.03487v1* 1–36 (2020).

22. Wood, F. *et al.* Planning as Inference in Epidemiological Models. *arXiv:2003.13221v2* 1–26 (2020).

23. Pujari, B. S. & Shekatkar, S. M. Multi-city modeling of epidemics using spatial networks: Application to 2019-nCov (COVID-19) coronavirus in India. *medRxiv* 2020.03.13.20035386 (2020).

24. Chakraborty, T. & Ghosh, I. Real-time forecasts and risk assessment of novel coronavirus (COVID-19) cases: A data-driven analysis. *medRxiv* **135**, 2020.04.09.20059311 (2020).

25. Tandon, H., Ranjan, P., Chakraborty, T. & Suhag, V. Coronavirus (COVID-19): ARIMA based time-series analysis to forecast near future. *arXiv:2004.07859v1* 1–11

26. Benvenuto, D., Giovanetti, M., Vassallo, L., Angeletti, S. & Ciccozzi, M. Application of the ARIMA model on the COVID-2019 epidemic dataset. *Data Br.* **29**, 105340 (2020).

27. Bayyurt, L. & Bayyurt, B. Forecasting of COVID-19 Cases and Deaths Using ARIMA Models. *medRxiv* 2020.04.17.20069237 (2020).

28. Ribeiro, M. H. D. M., da Silva, R. G., Mariani, V. C. & Coelho, L. dos S. Short-term forecasting COVID-19 cumulative confirmed cases: Perspectives for Brazil. *Chaos, Solitons & Fract.* **135**, 109853 (2020).

29. Kourentzes, N., Petropoulos, F. & Trapero, J. R. Improving forecasting by estimating time series structural components across multiple frequencies. *Int. J. Forecast.* **30**, 291–302 (2014).

30. Lipsitch, M. *et al.* Transmission dynamics and control of severe acute respiratory syndrome. *Science* **300**, 1966–1970 (2003).

31. Zhang, J. *et al.* Evolving epidemiology and transmission dynamics of coronavirus disease 2019 outside Hubei province, China: a descriptive and modelling study. *Lancet Infect. Dis.* **3099**, 1–10 (2020).

32. Kourentzes, N. & Petropoulos, F. MAPA: Multiple Aggregation Prediction Algorithm, R package version 2.0.4. https://github.com/trnnick/mapa/ (2014).

33. Hyndman, R. J. & Khandakar, Y. Automatic Time Series Forecasting: The forecast Package for R. *J. Stat. Software; Vol 1, Issue 3* (2008).

Plots and Figures:

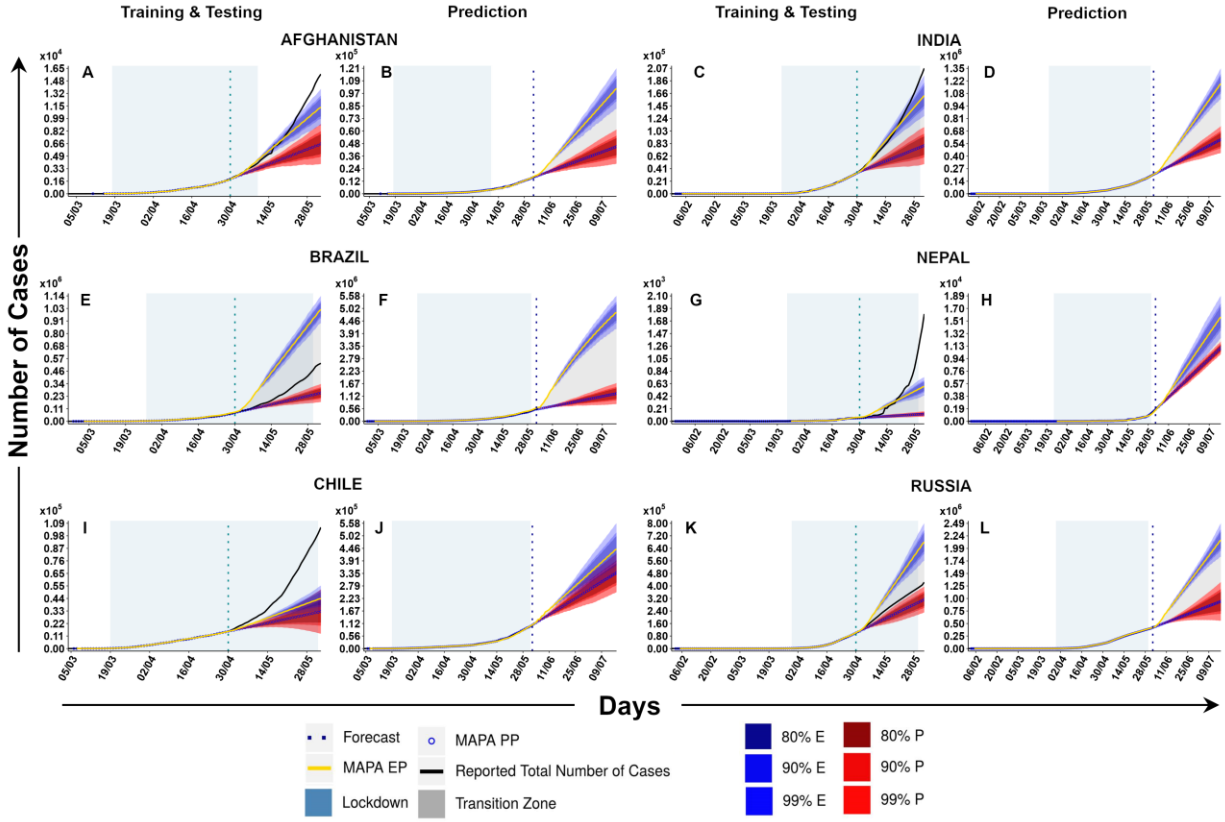

Fig. 1: Model Training and Validation of COVID-19 cases along with Use Case Prediction for Group A countries: Figure depicts Model Training (up to dotted line) and Validation for 33 days by comparison of Actual number of cases (black line) with the Principal Prediction (blue circles in red confidence band) and Exponential Prediction (yellow line in blue confidence band) for (A) Afghanistan, (C) India, (E) Brazil, (G) Nepal, (I) Chile, (K) Russia**;** (B, D, F, H, J, L) depicts use case prediction for the subsequent 45 days for the duration 3rd June to 17th July, 2020 for the same countries respectively; Blue shaded region depicts the lockdown for each country

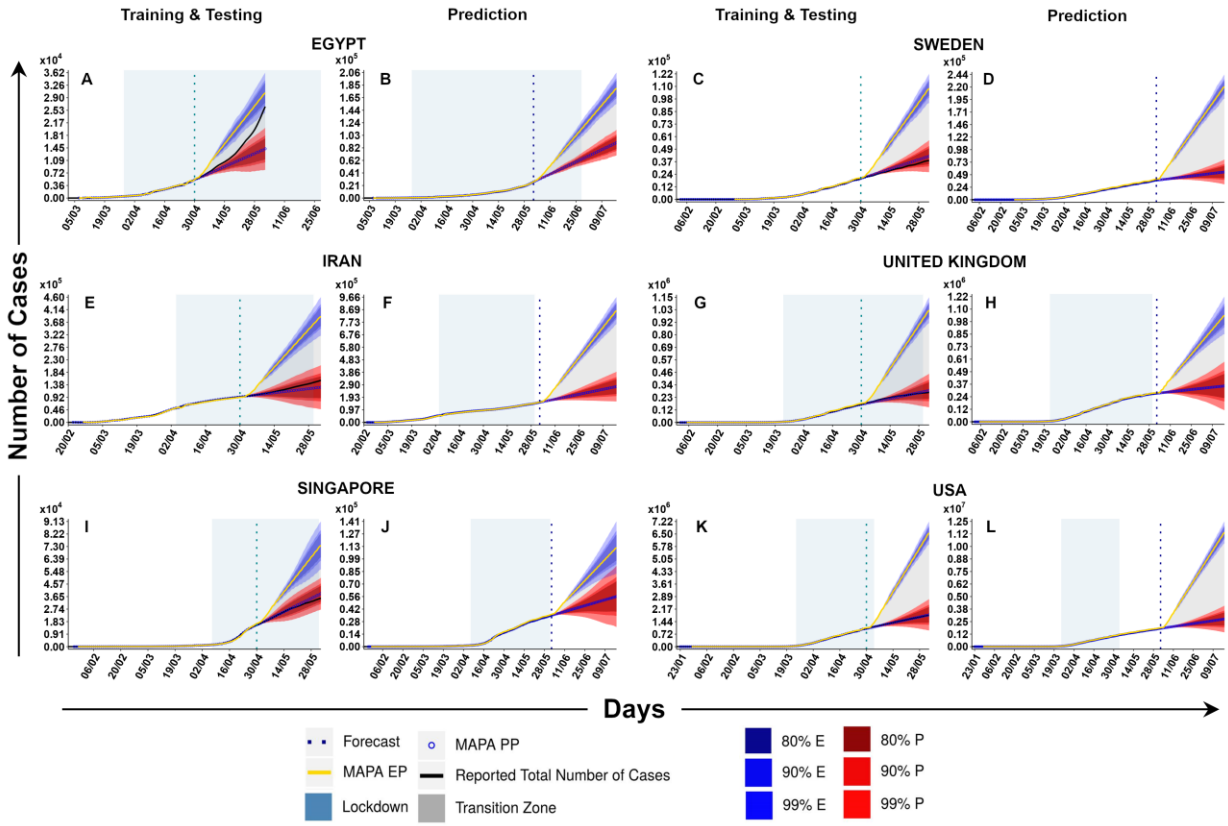

Fig. 2: Model Training and Validation of COVID-19 cases along with Use Case Prediction for Group B countries: Figure depicts Model Training (up to dotted line) and Validation for 33 days by comparison of Actual number of cases (black line) with the Principal Prediction (blue circles in red confidence band) and Exponential Prediction (yellow line in blue confidence band) for (A) Egypt, (C) Sweden, (E) Iran, (G) United Kingdom, (I) Singapore, (K) USA; (B, D, F, H, J, L) depicts use case prediction for the subsequent 45 days for the duration 3rd June to 17th July, 2020 for the same countries respectively; Blue shaded region depicts the lockdown for each country

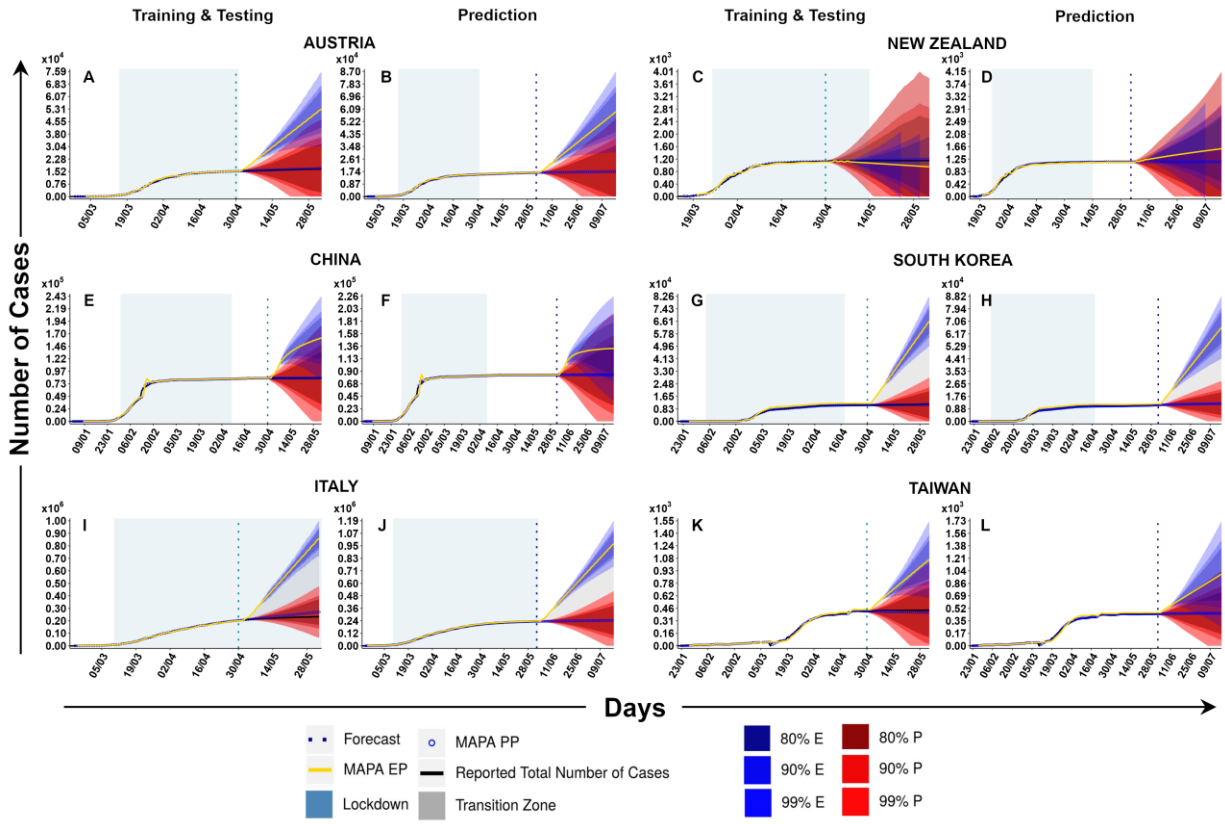

Fig. 3: Model Training and Validation of COVID-19 cases along with Use Case Prediction for Group C countries: Figure depicts Model Training (up to dotted line) and Validation for 33 days by comparison of Actual number of cases (black line) with the Principal Prediction (blue circles in red confidence band) and Exponential Prediction (yellow line in blue confidence band) for (A) Austria, (C) New Zealand, (E) China, (G) South Korea, (I) Italy, (K) Taiwan; (B, D, F, H, J, L) depicts use case prediction for the subsequent 45 days for the duration 3rd June to 17th July, 2020 for the same countries respectively; Blue shaded region depicts the lockdown for each country

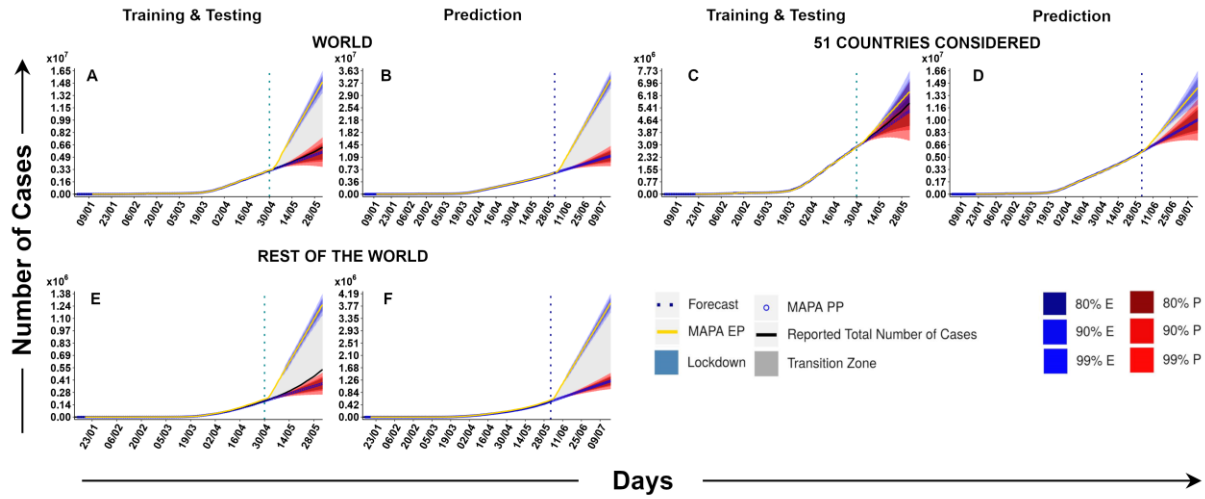

Fig. 4: Model Training and Validation of COVID-19 cases along with Use Case Prediction for the World as a whole, the 51 Countries considered in this study combined, and the Rest of the World consisting of 160 other countries affected by the COVID-19 pandemic: Figure depicts Model Training (up to dotted line) and Validation for 33 days by comparison of Actual number of cases (black line) with the Principal Prediction (blue circles in red confidence band) and Exponential Prediction (yellow line in blue confidence band) for (A) World, (C) 51 Countries combined, (E) Rest of the World**;** (B, D, F) depicts use case prediction for the subsequent 45 days for the duration 3rd June to 17th July, 2020

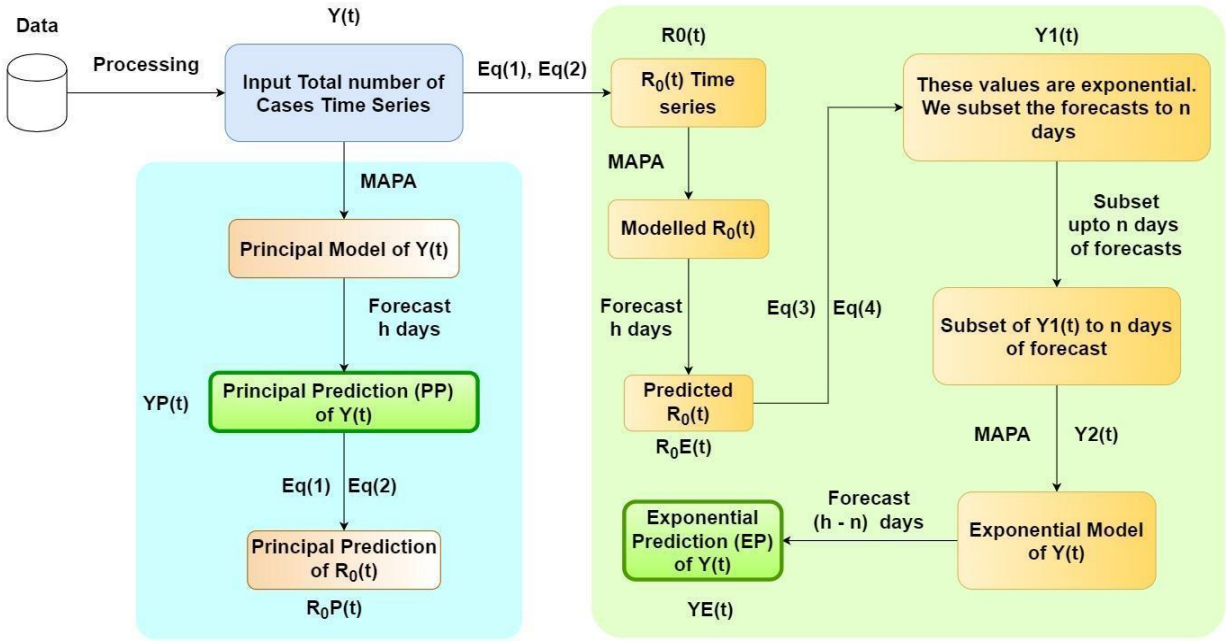

Fig 5: Workflow developed for the Time Series Analysis and Prediction of COVID-19 cases: The workflow comprises of Data processing followed by the Principal Prediction (left, blue panel) using MAPA, and Exponential Prediction (right, green panel) using calculated R_0 followed by MAPA.