Temporal Analysis of COVID - 19 Pandemic in India and R_0 prediction

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**

[+] equal first authors

Abstract

India now ranks 3rd in the list of countries affected due to the COVID-19 pandemic. With more than 0.7 million confirmed COVID–19 cases and a gradual withdrawal of nation-wide lockdown, it is expected India may see a further surge in the number of cases. So, predicting the expected number of cases is the need of the hour. Most of the existing methodologies work well for the short term but perform poorly when it comes to predicting the long term. Hence, in this study, we propose a novel strategy for the prediction of COVID–19 cases by employing Multiple Aggregation Prediction Algorithm (MAPA) with two different ways of predictions called the Principal and the Exponential Predictions. Exponential Prediction has been performed using MAPA on the number of cases, derived from predicted R_0. From the Principal and Exponential prediction, a third prediction called Mean prediction was derived by averaging both. We have validated this strategy using data of the different states of India and show that Principal, Mean, and the Exponential Predictions together capture a range in which the actual number of cases lies. We have used this strategy to forecast the range of COVID-19 cases for 45 days.

Keywords

COVID – 19; Machine Learning and Statistical modelling; MAPA; India; Time series; Basic Reproduction number

1. Introduction

The COVID pandemic has recorded over 11 million positive cases worldwide and India is one of the top countries that is severely affected. As India enters the Unlock 2.0 phase of the country wide lockdown, imposed to control the COVID - 19 pandemic, the tally of infected individuals crosses half a million. The disease has taken a toll on the lives of people suffering from other ailments or having a weak immune system. In this current scenario, researchers across the globe are striving hard to devise a treatment strategy to cure the critically ill patients as well as prevent the spread of the disease any further. Studies indicate that with the current rate of spread of the disease, the number of infected cases in India will rise exorbitantly by October, 2020, which may exceed the infrastructural capacity for providing the necessary treatment to all infected individuals. On the other hand, the huge economic crash suffered by the country and the precarious situation of the stranded migratory workers have compelled the dismissal of lockdowns any further. Under such a scenario, a validated model for long term prediction of the expected COVID – 19 cases has become an essential prerequisite to be able to form strategic plans for effective clinical management.

Although various models have already been proposed for India for short term predictions using mathematical modelling as well as data driven approaches (Chakraborty and Ghosh 2020; Poonia and Azad 2020), very few models perform well for long term predictions. Mathematical compartmental models developed for the study of epidemiology such as the SEIR (Susceptible-Exposed-Infected-Recovered) models are able to capture the disease dynamics over periods of time. The key advantage of SEIR models is that it can accommodate additional compartments (such as Asymptomatic Class, Mildly Symptomatic, Hospitalized, Quarantined, etc.) as well as various factors that influence the transmission of the disease (Kumar et al.; Das 2020; Dhanwant and Ramanathan 2020; Kumar 2020; Mandal et al. 2020; Pandey et al. 2020; Sardar et al. 2020; Sarkar and Khajanchi 2020; Ranjan 2020). Researchers have exploited this modelling strategy extensively for study of the spread the COVID – 19 pandemic where such factors such as lockdown implementations, migration, transport networks, influx of migrant workers have been considered for a near-accurate estimate of the when and how the pandemic can be controlled. However the major challenge for developing these models for the COVID – 19 predictions is that their accuracy is dependent on a host of unknown parameters that are often difficult to estimate given the scanty data that is available for the novel Coronavirus and also the due to the fact that the parameters regulating the implementation of the lockdowns, population density, and migration pattern is hugely different for the different states in India. This is another huge challenge for the development of parameterized models, especially now, when different states are gradually coming out of the nation-wide lockdown and following different rules customized for each state. In order to solve that, district level studies, using statistical modelling techniques such as hierarchical Bayesian, have been performed that predicts the expected number of cases that can be expected with the removal of lockdowns (Dutta et al. 2020). However, such statistical models have only been developed for a few states and hotspots that are severely affected by the pandemic (Dutta et al. 2020; Gupta and Shankar 2020).

On the other hand, most of the data driven models used to forecast the COVID – 19 pandemic relies less on the unknown parameters and more on the available time series data. On the other hand, most of the data driven models used to forecast the COVID – 19 pandemic relies less on the unknown parameters and more on the available time series data. Sujath *et. al.* have compared Linear Regression (LR), MultiLayer Perceptron (MLP), and Vector AutoRegression (VAR) models to predict the number of cases, deaths and recoveries for 69 days and have shown the MLP model to be the best amongst the three (Sujath et al. 2020). The AutoRegressive Integrated Moving Average (ARIMA) models have been the most commonly used strategy for predicting the number of COVID - 19 cases (Deb and Majumdar 2020; Gupta and Pal 2020; Marbaniang 2020; Sathish et al. 2020). For the study of the COVID – 19 cases in India, Khimya *et. al.* have employed a strategic implementation of ARIMA by combining multiple models (Tinani et al. 2020). Rafiq *et. al.* have employed predictive error minimization (PEM) based system identification technique to identify a discrete-time, single-input, single-output (SISO) model and have identified different models for different states based on the data collected (Rafiq et al. 2020). However, there are two main shortcomings of this family of models. The first is model selection. Even a theoretically sound selection of an ARIMA model may not result in selection of the optimal model order. The second is that the models are not suited over long term forecasts because the AutoRegressive part of the model converges to the mean of the time series. Yet, another challenge here is the parameterisation of the model.

In order to address these issues, we have developed a validated workflow for the forecast of COVID – 19 cases that gives a range of expected COVID – 19 cases that may be expected in the next 45 days in each state in India. The strategy developed uses the Multiple Aggregation Prediction Algorithm (MAPA) that creates multiple time series from the original data provided using temporal aggregation at multiple levels simultaneously (Kourentzes et al. 2014). This is one of the key advantages of using MAPA where the problem of model selection and parameterization are mitigated by repeating the process at each temporal aggregation level and combining the resulting models (Kourentzes et al. 2014). This ensures the utilization of the information of each aggregated level as well as the original time series data. Here, since no particular level is selected, multiple forecasts are projected for the same point. Thereafter, the resulting information is combined in a single prediction which is the output of MAPA. Hence the three main steps of this algorithm consist of temporal aggregation of original time series, forecasting at every aggregation level, and finally the combination of forecasts. The MAPA models the information of the past to predict values for the future. The data to be modeled for this study is COVID-19 incidence data. The incidence may be dependent on different factors for different regions and the nature of the data may vary as a complex function of these factors over time, especially for long term forecasts. We have performed MAPA on the data in two different ways to account for underlying complex functions governing the nature of data. The first method is to apply the general MAPA strategy directly onto the original time series data which gives us the Principal Prediction (PP) which forecasts the expected number of COVID – 19 cases based on the trend of the original time series data. In the second way, we have implemented MAPA in a novel strategy where we have utilized the prediction of the Reproductive number (R_0) of the disease. This strategy gives another forecast which accounts for sudden unprecedented exponential growth after the model training period which cannot be captured from the trend of the original time series data alone. We call this the Exponential Prediction (EP). The range in between the PP band and EP band, called the transition zone is the range which we expect the number of cases to lie within. The mean of the PP and EP is called the Mean Prediction (MP) which is calculated to make interpretation of the range easier and also provides better predictions in some states. In this study we have also shown that the actual number of cases lies in the range spanned by PP-MP-EP. Further we have shown that one of the PP, EP or MP predicts the expected number of COVID – 19 cases with better accuracy as compared to ARIMA for long term predictions. Thereafter, we have employed this validated novel integrative approach to forecast the expected range of the number of COVID – 19 cases for each state in India for 45 days till the 8th August, 2020 and suggest region wise selection of appropriate bands of the range, based on the verification of throw predictions with real data a few days into the 45 day long forecast horizon. Till the last date of this verification, the range spanned by EP, MP and PP is of reasonable width. This will help in gaining knowledge, a priori, of the expected number of cases in each state in India which will help in the development of better clinical management and policies to control the pandemic.

2. Materials & Methods

2.1 Data collation & processing

In this study, we require the incidence data of COVID - 19 in India and its states. The data is collected from publicly available repository COVID - 19 India (https://www.covid19india.org/) which sources data from the Ministry of Health and Family Welfare, Govt. of India (https://www.mohfw.gov.in/), Indian Council of Medical Research (https://covid.icmr.org.in/index.php) and other state government's COVID - 19 information portals.

The data is a time series of the daily incidence state wise. The cumulative series is derived from the data. First day in the dataset starts from 14th March, 2020, i.e. 45 days since COVID-19 arrival in India. (**Supplementary Table S1**)

2.2 Calculation of Basic Reproductive Number (R_0)

Basic reproduction number (R_0) is an epidemiological metric defined as the expected number of secondary infections that can develop from a single infectious source in a completely susceptible population (Delamater et al. 2019).

To estimate R_0 of COVID - 19 we consider the following equations (Lipsitch et al. 2003),

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

$$r_0 = \frac{ln\,(Y(t))}{t} \qquad \textbf{(Eq. 2)}$$

Where, r_0 is the exponential growth rate, Y(t) is the number of total confirmed cases on the t th day after the first reported case and hence the total number of cases has been derived. Taking the first difference of Y(t) will return the daily incidence series. In Eq. (1), we assume the serial interval (v) to be 5.1 days and assume the ratio of the latent period and mean serial interval (f) as 0.99 (Zhang et al. 2020). To get the number of cases from R_0 we make use of r_0 in Eq. (1) after estimating it from Eq. (2). As we are computing the value of exponential growth rate $r_{0h}(t)$ from Eq. (3) using the evaluated basic reproductive number from Eq. (1). Transforming the exponential growth rate gives an estimated total number of cases ($Y_h(t)$) from Eq. (4).

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

$$Y_h(t) = exp\ (tr_{0h}(t)) \qquad \textbf{(Eq. 4)}$$

2.3 Proposed Workflow using MAPA

Multiple Aggregation Prediction Algorithm (MAPA) creates multiple time series from the original series using temporal aggregation (Kourentzes and Petropoulos 2014). The workflow proposed in this study utilises MAPA in two ways. The first is to apply MAPA on the total number of cases data and the second is to apply MAPA on the R_0 series.

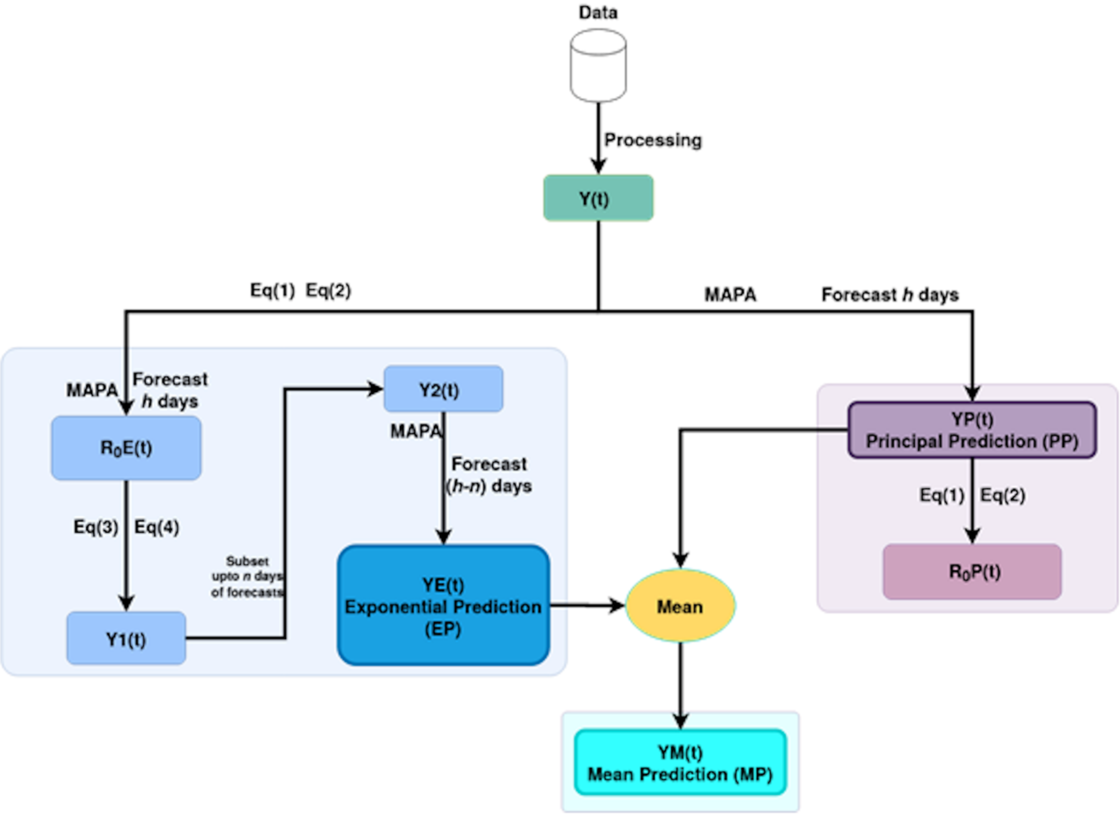

Fig. 1: Proposed Workflow: The data is processed in the required format and is treated in two ways (Blue and Purple Blocks). The Blue block results in the EP of the total number of cases (Y(t)) denoted by YE(t) and the Purple block results in the PP denoted by YP(t). $R_0E(t)$ and $R_0P(t)$ are the Exponential Prediction (EP) and Principal Prediction (PP) of $R_0(t)$ respectively. The mean of the YP(t) and YE(t) is the Mean Prediction (MP) denoted by YM(t) (cyan block).

In the first strategy (right purple panel, **Fig. 1**), MAPA has been directly implemented on the original time series data Y(t). To perform MAPA on our data we used the MAPA: Multiple Aggregation Prediction Algorithm 2.0.1 package in R (Kourentzes and Petropoulos 2014). The forecasts for *h* days after training on Y(t) is called the Principal Prediction (PP), denoted by YP(t). $R_0P(t)$, the Principal Prediction of $R_0(t)$ is derived by applying **Eq. (1) and Eq. (2)** on YP(t).

In the second way (right blue panel, **Fig. 1**), **Eq. (1) and Eq. (2)** are applied on Y(t) to obtain $R_0(t)$. MAPA is implemented on this $R_0(t)$. The forecast for *h* day is obtained after training on $R_0(t)$. This is called the Exponential Prediction (EP) of $R_0(t)$ and is denoted by $R_0E(t)$. **Eq. (3)** and **Eq. (4)** have been implemented on $R_0E(t)$ to obtain the corresponding total number of cases denoted by Y1(t). Since **Eq. (4)** is exponential in nature, so is Y1(t). Y1(t) is cut off after first *n* forecasts, where *n* is suitably chosen. This subset of Y1(t) is denoted by Y2(t). MAPA is performed on Y2(t) and forecasts are made for the remaining number of (*h* - *n*) days. These forecasts are the EP of Y(t). The second way is performed so as 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.

To allow for easier interpretation, the mean of the PP and EP is also calculated. This mean is called the Mean Prediction (MP) denoted by YM(t) = (YP(t)+YE(t))/2 (cyan block, **Fig. 1**). If the MP is a better predictor of the actual number of cases, it will justify the need for the EP, PP and the MP itself.

2.4 Error Metrics

Estimation of model performance is essential to determine the accuracy and precision of both short- and long-term predictions. To estimate the performance of our strategy, Root mean squared (RMSE) and Mean Absolute Percentage Error (MAPE) were calculated to quantify the deviation of estimates to the actual value. Equations are defined in **Eq. 5-7**.

a. RMSE:

$$RMSE = \sqrt{\sum_{i=1}^{n} \frac{|\, observed(i) - predicted(i)\, |^2}{n}} \qquad \textbf{(Eq. 5)}$$

We use the normalised RMSE (nRMSE), which is derived by dividing the RMSE by the mean of the actual data.

b. Percentage Error

$$PE(t) = \left(\frac{(\, observed(t) - predicted(t)\,)}{observed(t)}\right) \qquad \textbf{(Eq. 6)}$$

c. Mean Absolute Percentage Error:

It is the mean of the absolute values of PE for a set of n time points.

$$MAPE = \left(\frac{1}{n}\right)\left(\, \sum_{t=1}^{n} |PE\ (t)|\, \right) \qquad \textbf{(Eq. 7)}$$

2.5 Model Validation, Comparison and Use Case

In order to validate our strategy, the proposed workflow is performed on the state wise data from the first day of incidence or 14th March (whichever is later) in a region till 25th May, 2020 and predictions are made for the next 30 days, i.e. from 26th May, 2020 to 24th June, 2020 (**Supplementary Table S1**). Here *h = 30, n = 7*. It is to be mentioned here that no real data has been considered for modelling purposes after 25th May.

A comparison between ARIMA and MAPA was performed for India and the top 6 regions (Delhi, Gujarat, Madhya Pradesh, Maharashtra, Rajasthan and Tamil Nadu) with respect to the number of reported cases. Training was done till 30 April and forecasts were made for 30 days, from May 1 to May 30. Here *h = 30, n = 7*.

For the use case, the proposed workflow is performed on the state wise data from the first day of incidence in a region or 14th March (whichever is later) till 24th June, 2020 and predictions are made for the next 45 days, i.e. from 25th June, 2020 to 08th August, 2020 (**Supplementary Table S1**). Here *h = 45, n = 7*. No real data is taken for modelling purposes after 24th June**.** Here, ARIMA was 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.

Lakshadweep, Nagaland and Sikkim have not been considered in this study due to lack of data. COVID - 19 has not been reported in Lakshadweep yet and was first reported in Nagaland and Sikkim on 25th May and 23rd May respectively.

3. Results

3.1 Model development & validation

COVID - 19 dataset of all states was considered for training and testing. The proposed workflow provides two main prediction bands, one for the Principal Prediction (PP) (purple dots surrounded by yellow confidence bands of 99%, 90% and 80%) and one for Exponential Prediction (EP) (solid blue line surrounded by red confidence bands).The Mean Prediction (MP) (solid cyan line) is also depicted. The grey area in between these two bands is the transition zone (**Fig. 2-6**).

Based on the nature of the actual data from 26th May to 15th June, the 34 regions considered in this study (India, states and UTs) are categorised into 3 groups: A) Exponential, B) Rising and C) Plateauing. Some states were classified under an exception category, as the workflow was unable to capture the trends due to low order of magnitude of the data and/or lack of data. Principal prediction as such captures the underlying trend of data, whereas exponential prediction derived from R_0 value accounts for possible spurious cases (**Fig. 2-6**). The range between the PP bands and the EP bands is called the transition zone. The MP lies right in the middle of the range .In this transition zone, a state 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, the number of cases may eventually stride into the transition zone then to the principal prediction zone. As mentioned before, the data coming from different regions may be dependent on different factors and hence the generalisation of a single band of EP, PP or MP for all regions may not be optimal. This has been demonstrated during the testing of the model performance. Box plots of the EP, MP and PP were compared with the actual data that depicts the cumulative number of reported cases; and percentage errors that are observed temporally for the validation period are depicted in **Supplementary Fig. S1-S4**. Out of the 34 regions, 10 regions were categorised as exceptions (**Fig 6 - A, C, E, G, I, K, M, O, Q, S**). For the remaining 24 regions, with respect to MAPE, for 9 regions the PP (MAPE in the range 1.14-22.7%) performed better than EP and MP, for 9 regions the MP (MAPE in the range 2.23-19.38%) performed better than PP and EP and for 6 regions the EP (MAPE in the range 5.09-13.84%) performed better than PP and MP (**Supplementary Table S2**).

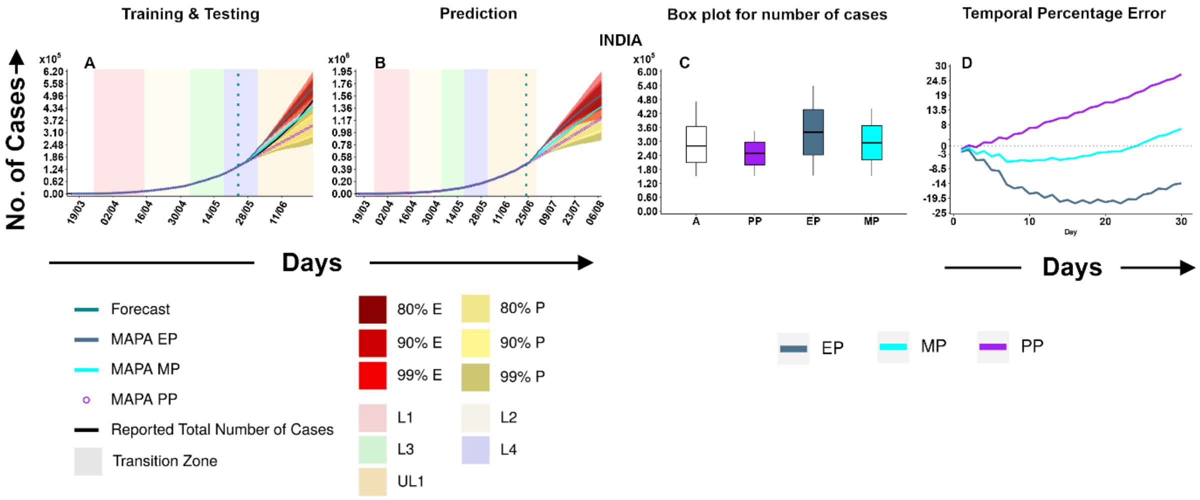

Fig. 2: Model Training and Validation of COVID-19 cases along with Use Case Prediction for India: Figure depicts (A) Model Training (up to dotted line) and Validation for 30 days from 26th May to 24th June, 2020 by comparison of Actual number of cases (black line) with the Principal Prediction (PP, purple circles in yellow confidence bands), Mean Prediction (MP, cyan line) and Exponential Prediction (EP, blue line in red confidence bands) for India; (B) depicts use case prediction for the subsequent 45 days for the duration 25th June to 08th August, 2020 for the same states respectively. The shaded background represents the 4 lockdowns (L1 to L4) and the first phase of opening (UL1). (C) The boxplot compares the structure of the actual data, PP, MP and EP. (D) Temporal percentage error for PP, MP and EP.

India is shown in **Fig. 2.** The number of cases is on the edge of the principal bands and the transition zone. Exponentially rising states are represented in **Fig. 3**. States included in this category did not show sudden spurt in COVID - 19 cases till training. Strategy used captures this sudden rise. In the case of Delhi, Jharkhand, Tripura, West Bengal, Bihar and Assam (**Fig. 3 - G, K, S, U, C, E**) this rise in number of cases can be attributed to the movement of stranded urban migrant labourers from the metropolitan cities such as Delhi (Economic Times 2020a), Bengaluru (Times of India 2020), Chennai (The Hindu 2020a), Mumbai (Times Now News 2020) and Thiruvananthapuram (The Hindu 2020b) who were severely affected due to lockdown and were transported back to their respective States via train. It is entirely possible that the migratory population moving away from the urban centres could have individuals who had contracted COVID - 19 and remained asymptomatic during the journey and spread it to the co-passengers and some even carrying the virus to their respective villages (Hindustan Times 2020a). It has to be noted that the states screened the arriving passengers and mandated them to institutional and home quarantine. On the other hand, Delhi being the centre of health infrastructure for adjacent districts of Uttar Pradesh and Haryana, back and forth movement of people could have carried the virus to and fro. This evidence can be validated by the closure of borders of districts adjoining Haryana and Uttar Pradesh government (ThePrint.in 2020). In Karnataka, COVID - 19 cases were rising steadily and Kerala showed signs of plateauing and the exponential rise in cases can be linked to internal migration and overseas Indian citizen expatriation. This could be noted as a probable second wave of COVID - 19 in the state of Kerala. Karnataka's case is curious. A sizable population from North Karnataka reside in various cities of Maharashtra namely, Mumbai, Pune, Solapur and Kolhapur, reverse migration resulted in formation of COVID - 19 clusters in Kalburgi, Vijayapura and Belagavi (Hindustan Times 2020b; The Hindu 2020c; The New Indian Express 2020a).

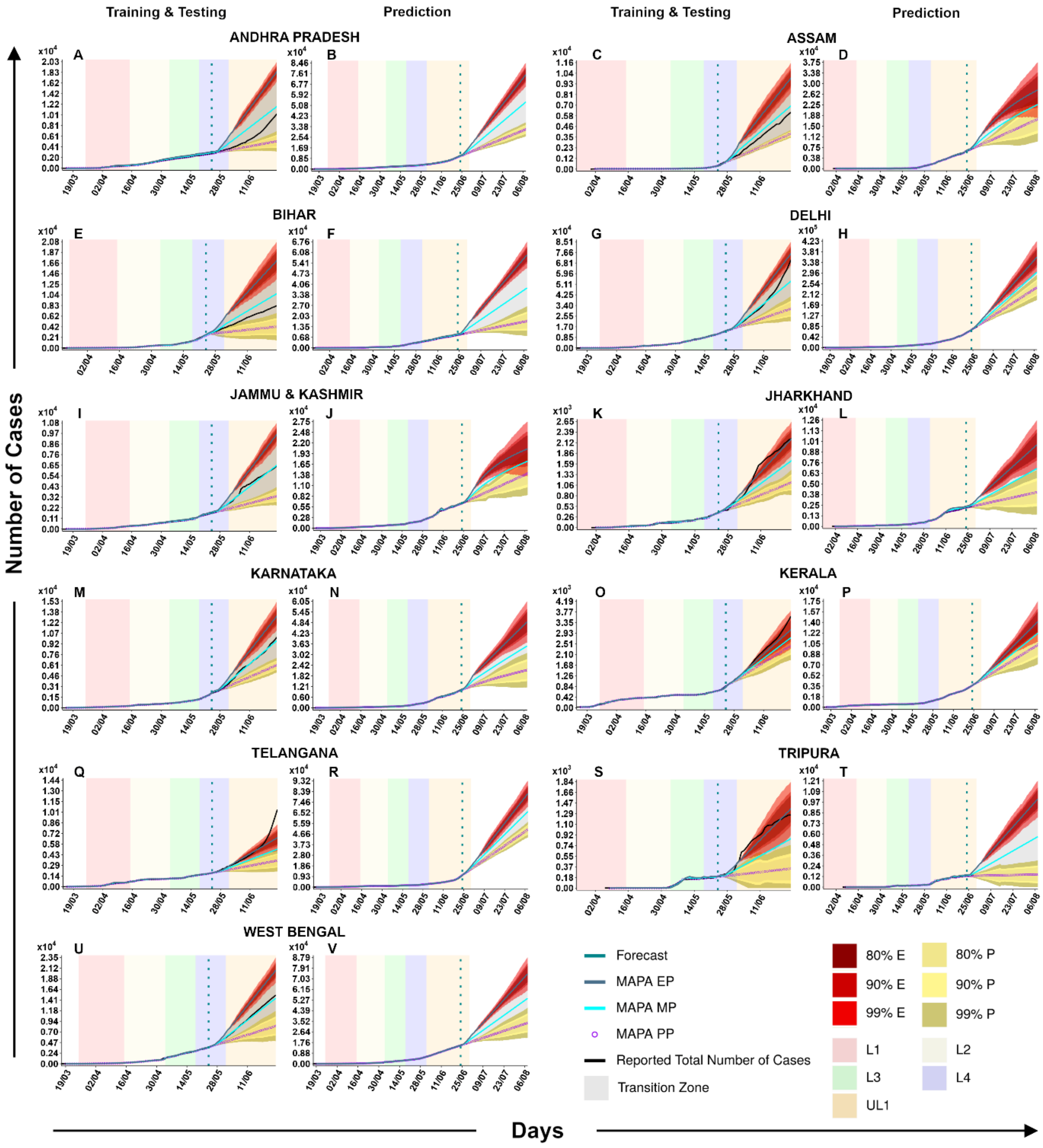

Fig. 3: Model Training and Validation of COVID-19 cases along with Use Case Prediction for Group A States: Figure depicts Model Training (up to dotted line) and Validation for 30 days from 26th May to 24th June, 2020 by comparison of Actual number of cases (black line) with the Principal Prediction (purple circles in yellow confidence bands), Mean Prediction (MP, cyan line) and Exponential Prediction (blue line in red confidence bands) for (A) Andhra Pradesh, (C) Assam, (E) Bihar, (G) Delhi, (I) Jammu & Kashmir, (K) Jharkhand, (M) Karnataka, (O) Kerala, (Q) Telangana, (S) Tripura and (U) West Bengal; (B, D, F, H, J, L, N, P, R, T, V) depicts use case prediction for the subsequent 45 days for the duration 25th June to 08th August, 2020 for the same states respectively. The shaded background represents the 4 lockdowns (L1 to L4) and the first phase of opening (UL1).

Fig. 4 represents the Rising states which showed sustained increase in COVID - 19 cases. This category includes states such as Maharashtra, Tamilnadu, Rajasthan, Punjab, Madhya Pradesh and Gujarat (**Fig. 4 - I, Q, O, M, G, C**) which saw a sudden increase in reporting of COVID - 19 during the early phases of Lockdown 1.0 and Lockdown 2.0. Mode of spread of COVID - 19 in these states are different. Maharashtra, Gujarat and Tamilnadu being industrialised states, movement of goods and people from the urban working centers such as Chennai, Mumbai City, Surat, Ahmedabad are the primary reason for the spread. Increased testing, tracing led to identification of 700 super-spreaders in Ahmedabad (Economic Times 2020b), whereas in Chennai's largest wholesale market Koyambedu, testing led to identification of more than 1700 COVID - 19 cases (Deccan Herald 2020). This market receives goods from several districts within Tamilnadu and neighbouring states, it is still possible that individuals exposed and infected in this market may have carried infection back to their districts. COVID - 19 cases in Madhya Pradesh and Punjab infection started with religious congregations. To compound that in Punjab it was reported that a single super - spreader was responsible for quarantining around 26000 people from 20 villages (The New Indian Express 2020b). Rajasthan was among the initial states that reported COVID - 19 infection, individuals returning from abroad were identified as a super - spreader which led to imposing of curfew in the town of Ramganj. Subsequently, the state saw spurt in COVID - 19 cases in Jaipur, Bhilwara, Jodhpur and Bharatpur (Hindustan Times 2020c).

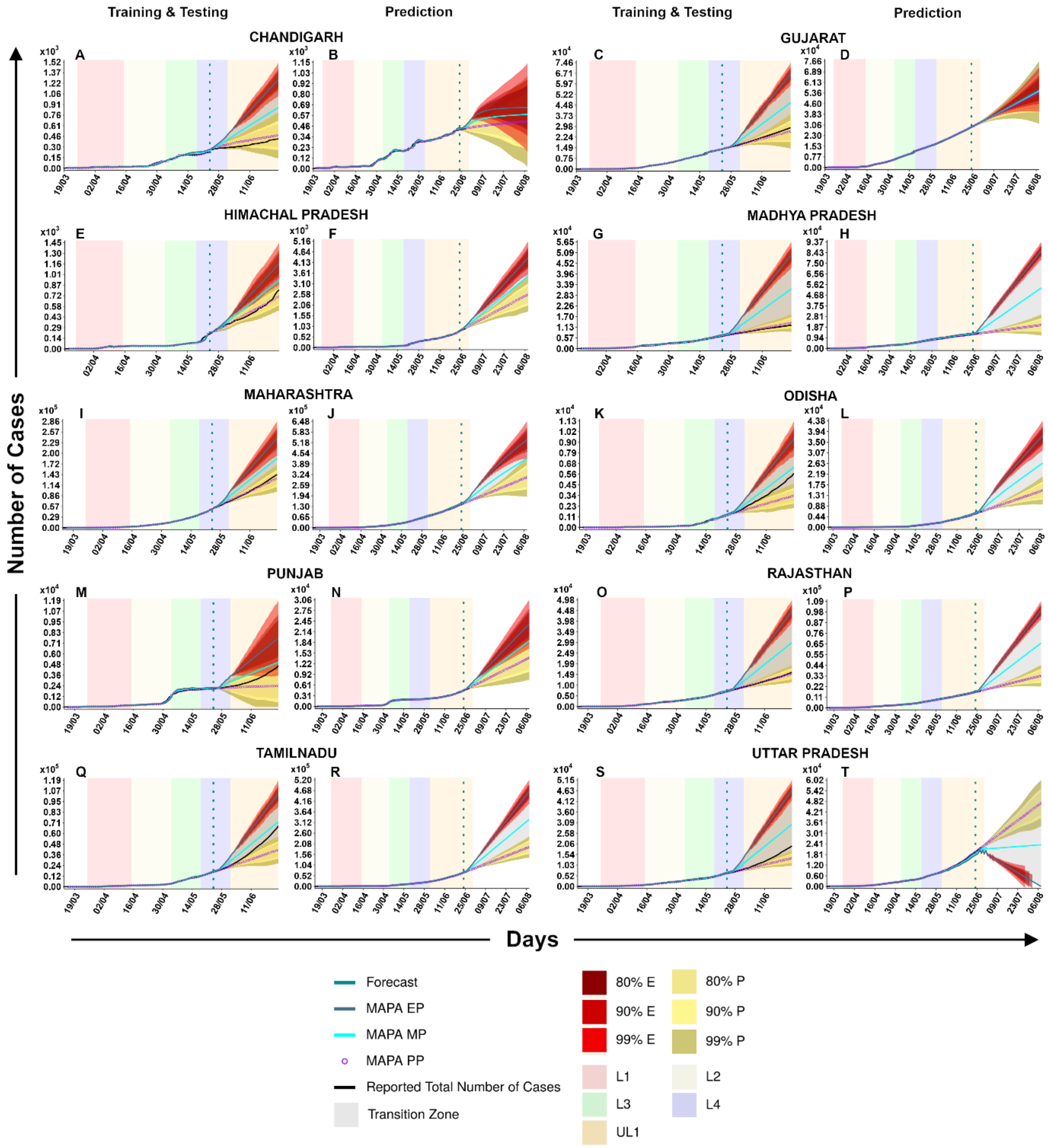

Fig. 4: Model Training and Validation of COVID-19 cases along with Use Case Prediction for Group B States: Figure depicts Model Training (up to dotted line) and Validation for 30 days from 26th May to 24th June, 2020 by comparison of Actual number of cases (black line) with the Principal Prediction (purple circles in yellow confidence bands), Mean Prediction (MP, cyan line) and Exponential Prediction (blue line in red confidence bands) for (A) Chandigarh, (C) Gujarat, (E) Himachal Pradesh, (G) Madhya Pradesh, (I) Maharashtra, (K) Odisha, (M) Punjab, (O) Rajasthan, (Q) Tamil Nadu and (S) Uttar Pradesh; (B, D, F, H, J, L, N, P, R, T) depicts use case prediction for the subsequent 45 days for the duration 25th June to 08th August, 2020 for the same states respectively. The shaded background represents the 4 lockdowns (L1 to L4) and the first phase of opening (UL1).

The Plateauing states are plotted in **Fig. 5**, which include Andaman and Nicobar Islands and Meghalaya (**Fig. 5 - A, C**). At the time of training these states showed signs of control over spread of COVID - 19. In addition to these three groups, states that didn’t fall under any of these categories were classified under Exception (**Fig.6**). This group consists of Goa, Haryana, Chhattisgarh, Uttarakhand, Union Territories and North - Eastern states (**Fig. 6 - G, I, C, S, E, K, Q, A, M, O**). Steep rise in COVID - 19 cases neither occur till training nor the number of cases followed Exponential bands. This sudden change in the slope of the curve could be attributed to relaxation of lockdown and reverse migration.

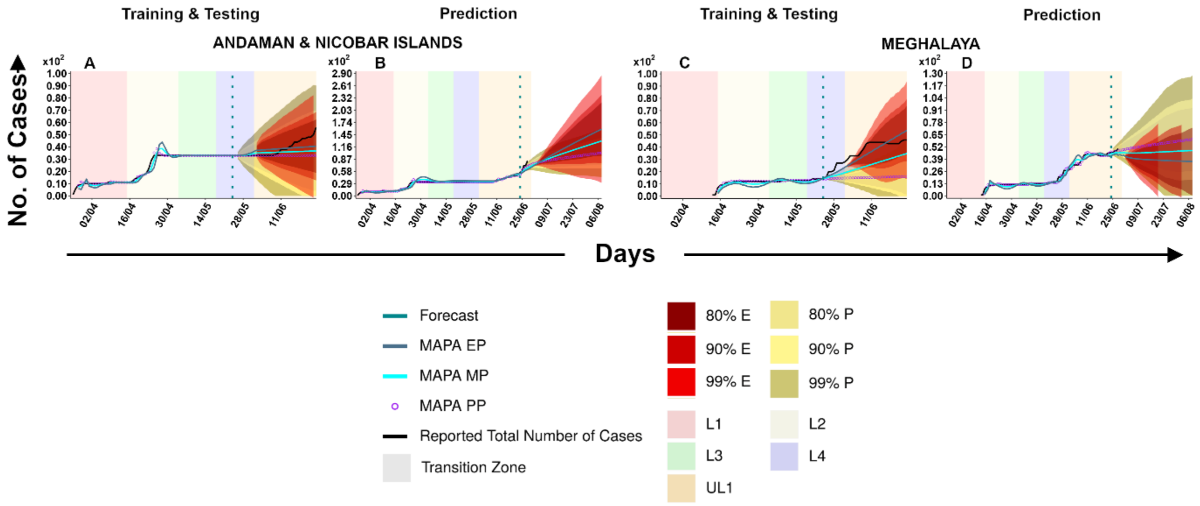

Fig. 5: Model Training and Validation of COVID-19 cases along with Use Case Prediction for Group C States: Figure depicts Model Training (up to dotted line) and Validation for 30 days from 26th May to 24th June, 2020 by comparison of Actual number of cases (black line) with the Principal Prediction (purple circles in yellow confidence bands), Mean Prediction (MP, cyan line) and Exponential Prediction (blue line in red confidence bands) for (A) Andaman & Nicobar Islands and (C) Meghalaya; (B, D) depicts use case prediction for the subsequent 45 days for the duration 25th June to 08th August, 2020 for the same states respectively. The shaded background represents the 4 lockdowns (L1 to L4) and the first phase of opening (UL1).

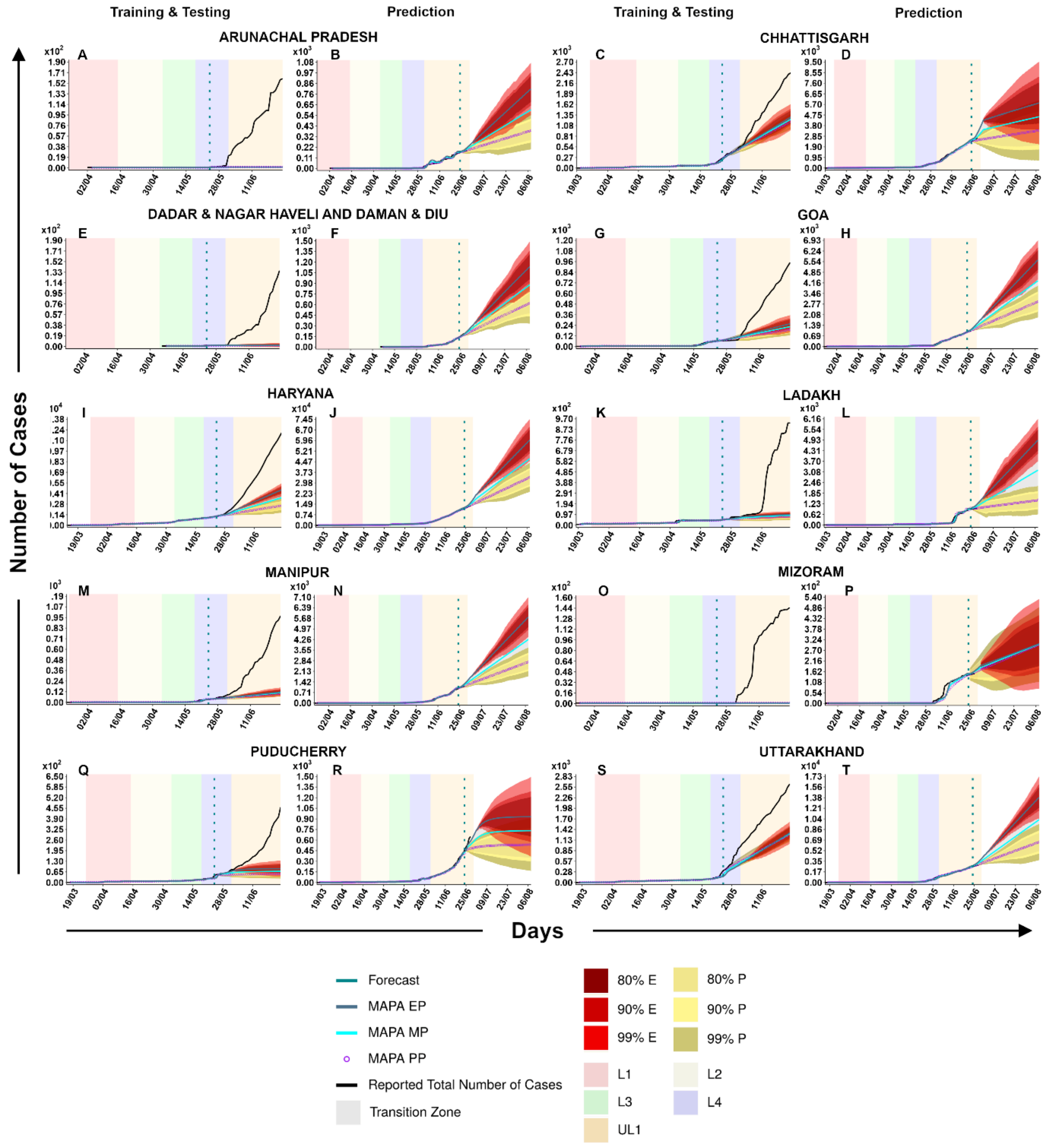

Fig. 6: Model Training and Validation of COVID-19 cases along with Use Case Prediction for Exception States: Figure depicts Model Training (up to dotted line) and Validation for 30 days from 26th May to 24th June, 2020 by comparison of Actual number of cases (black line) with the Principal Prediction (purple circles in yellow confidence bands), Mean Prediction (MP, cyan line) and Exponential Prediction (blue line in red confidence bands) for (A) Arunachal Pradesh, (C) Chhattisgarh, (E) Dadra & Nagar Haveli and Daman & Diu, (G) Goa, (I) Haryana, (K) Ladakh, (M) Manipur, (O) Mizoram, (Q) Puducherry, (S) and Uttarakhand; (B, D, F, H, J, L, N, P, R, T) depicts use case prediction for the subsequent 45 days for the duration 25th June to 08th August, 2020 for the same states respectively. The shaded background represents the 4 lockdowns (L1 to L4) and the first phase of opening (UL1).

The predictions of $R_0(t)$ for the first and last date of validation are compared with the values obtained from the real data (**Supplementary Table S3**). At the start of validation, as many as nine states showed R_0 above 1.57, highest among those were Gujarat and Maharashtra having values of 1.714 and 1.709 respectively. MAPA prediction closely matched R_0, closest among the three MAPA predictions was Principal prediction, closely matching the real R_0 value for six states and UTs. From the Actual data it is seen that fifteen states and UTs had R_0 value between 1.38 - 1.57, seven states and UTs between 1.19 - 1.38 and three between 1 - 1.19. It has to be noted from **Supplementary Table S3**, at the start of validation, value of R_0 was closer to Principal prediction barring few such as Assam, Tripura, Haryana, Telangana (Exponential prediction), Puducherry, Bihar, Odisha and Gujarat (Mean prediction). Mizoram is an exception as it had reported very few cases on 26th May, 2020, thereby R_0 remained 1 throughout the validation period for MAPA model.

Subsequently, high R_0 value decreased at the end of validation R0 for different states. Maharashtra reported the highest R_0 value of 1.563, Principal prediction predicted R_0 to be 1.559 which is closer to the actual value compared to Exponential Prediction (1.676) and Mean Prediction (1.618). Minimum was R_0 reported by Andaman and Nicobar Island. Overall, out of 34 states and UTs, Principal Prediction performed better for 13 states compared to 12 for Exponential and 9 for Mean prediction when compared with Actual R_0.

Many states and UTs see a drop in R_0 at the end of validation. Taking the difference of R_0 at the start and end of validation, we see that Gujarat and Madhya Pradesh report the biggest drop in R_0 value of 0.177 and 0.170 respectively. States classified in Exceptional category (except Chhattisgarh, Haryana and Uttarakhand) which did not see any rise in COVID -19 cases at the start of training, subsequently due to relaxation saw uptick in number of cases, thus leading to increase in R_0, Dadar and Nagar Haveli & Daman and Diu, Mizoram and Arunachal Pradesh see a rise in difference of R_0 to be 0.332, 0.276 and 0.244 respectively.

3.2 Model Comparison between ARIMA and MAPA

COVID - 19 dataset of all states was considered for training and testing. A comparison between ARIMA and MAPA was performed for India and the top 6 regions with respect to the number of reported cases. The ARIMA model is the green line (**Fig. 7**). For Gujarat, Madhya Pradesh and Rajasthan (**Fig. 7, C-D, E-F, I-J**) ARIMA and the PP are similar in capturing the nature of the actual data with MAPE values in the range of 10-15%. For Delhi, Tamil Nadu and India (**Fig. 7, A-B, K-L, M-N**) the EP is better than ARIMA and PP with MAPE values of EP in the range of 6-16%. For Maharashtra (**Fig. 7, G-H**) the number of cases is in the transition zone making the MP the best predictor of the actual data as compared to PP and EP. This comparison demonstrates the need of the range spanned by PP-MP-EP to capture the nature of the data over a long duration and in all cases.

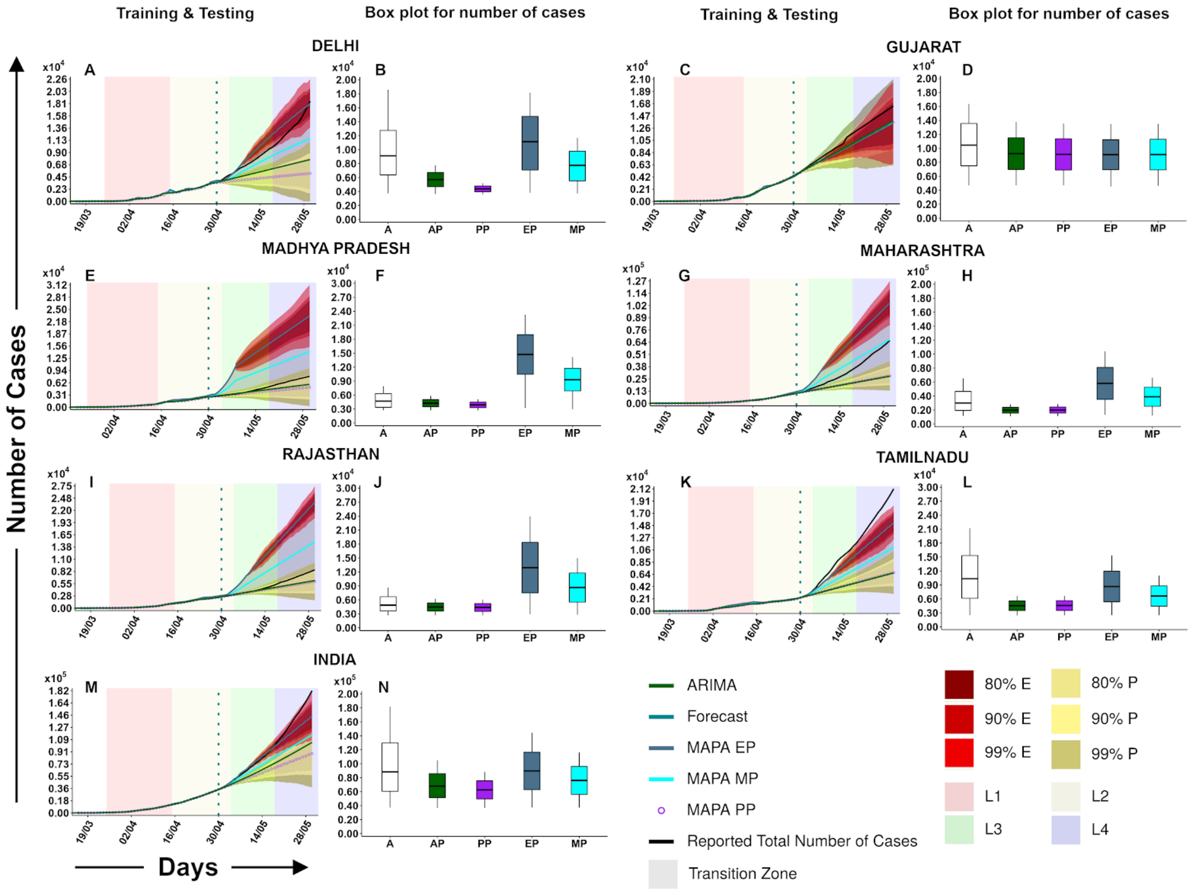

Fig. 7: Comparison of ARIMA and MAPA for six emerging hotspots: Figure depicts Model Training (up to dotted line) and Validation for 30 days from 1st May to 30th May, 2020 by comparison of Actual number of cases (black line) with ARIMA (green line), Principal Prediction (PP)(purple circles in yellow confidence bands), Exponential Prediction (EP)(blue line in red confidence bands) and Mean Prediction (MP)(cyan line) for (A) Delhi, (C) Gujarat, (E) Madhya Pradesh, (G) Maharashtra, (I) Rajasthan, (K) Tamilnadu, (M) India; (B, D, F, H, J, L, N) depicts box plots comparison of distribution of Actual number of cases (A, white) to ARIMA (AP) (green), PP (purple), EP (blue) and MP (cyan) for the same states respectively. The shaded background represents the 4 lockdowns (L1 to L4).

3.3 Use case prediction

After validation of our model for long term prediction, forecasting of cumulative number of COVID - 19 cases for 45 days (25th June to 08th August, 2020) was performed. States that were classified during validation were retained for prediction. Questions may arise regarding selection of Exponential or Principal band for Actual data during prediction. But it has to be noted that Actual curve will vary due to rampant spreading of the virus and the states will move from one band to another during the forecasting period.

Based on the nature of the actual data from 25th June to 01st July, the states of India are re-categorised into 3 groups: A) Exponential, B) Rising and C) Plateauing to accommodate for change in nature of the Exception group. Chhattisgarh, Manipur, Mizoram and Puducherry have been classified under Group A as these states see steep rise in reported cases and lower number base. During training and validation, data points for the above-mentioned states were minimum; therefore our pipeline was not able to provide correct predictions, hence classified under exceptional category. Likewise, Arunachal Pradesh, Goa, Haryana, Ladakh and Uttarakhand have been classified under group B, these states see their COVID -19 numbers rising. Only Dadra and Nagar Haveli & Daman and Diu is the only exception hence retained in the exceptional group.

Error metrics for the validated forecast is presented in **Supplementary Table S4.** The predictions of $R_0(t)$ for the first and last date of use case are compared between the PP,MP and EP (**Supplementary Table S5**). At the end of the 45 days, i.e. on 08th August, the PP,MP and EP for all states and UTs R_0 is lesser than that on the first day, but is greater than 1. The PP and MP of R_0 for all states and UTs is less than 1.5. Maharashtra, Tamil Nadu, Delhi and Gujarat have the highest R_0P between 1.39 and 1.42 values indicating a drop in values in the range 0.099-0.143 over the course of the forecast while Andaman Nicobar Islands, Meghalaya, MIzoram and Chandigarh have lowest R_0P values between 1.17 and 1.22 indicating a drop in the range 0.05-0.09 over the course of the forecast.

According to our analysis and predictions, by 08th August India can expect 1.21-1.57 million cases, Maharashtra 0.31-0.53 million cases, Tamil Nadu 0.19-0.46 million cases, Delhi 0.24-0.36 million cases, Gujarat 0.054-0.056 million cases, Uttar Pradesh 0.047-0.06 million cases and West Bengal 0.034-0.074 million cases. (**Supplementary Table S7, Fig. 2-6**).

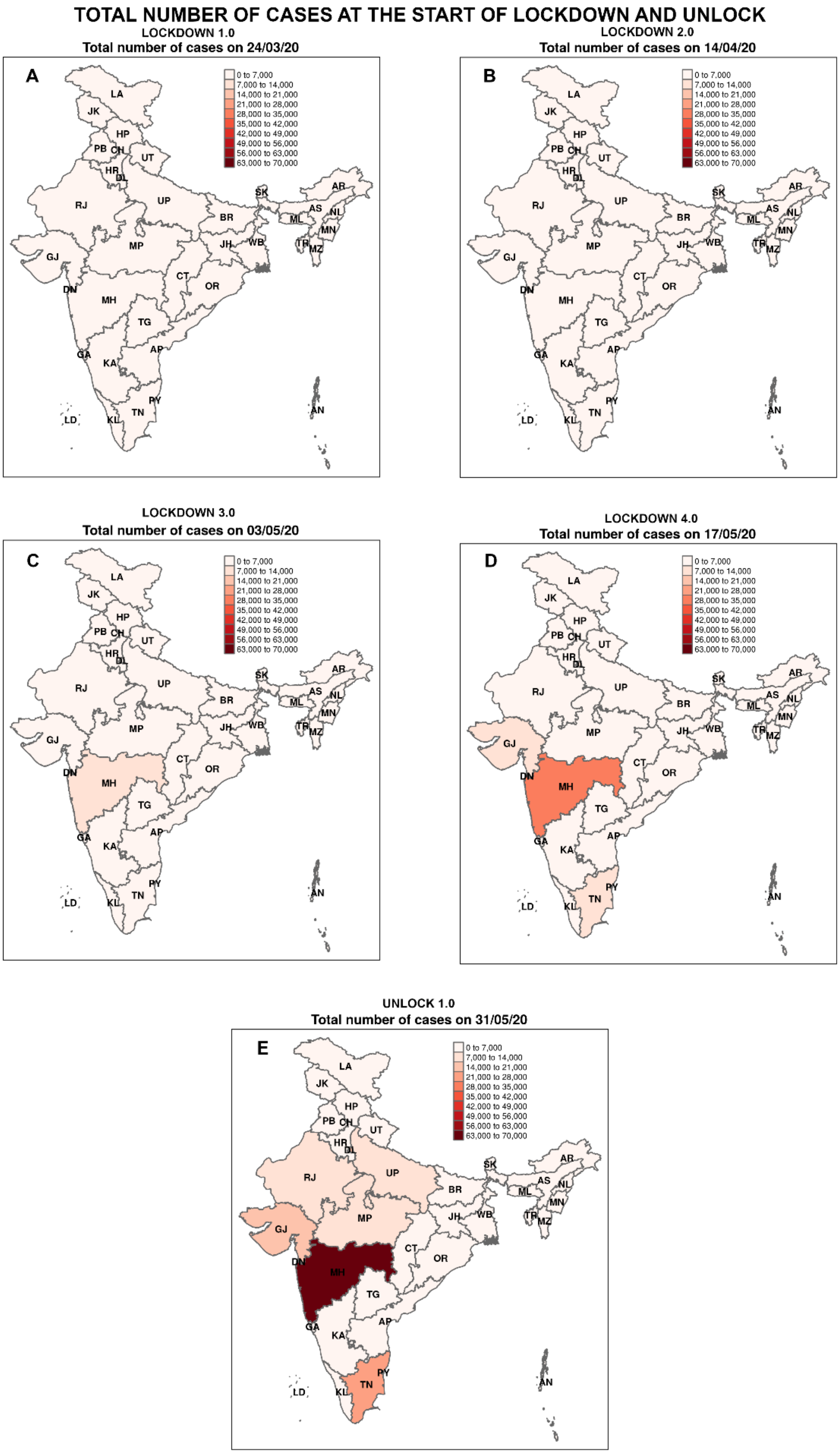

Fig. 8: Total Number of reported COVID - 19 Cases in India one day before each Lockdown phase: Figure depicts the distribution of the total number of reported COVID - 19 cases grouped state wise in India on (A) 24th March, 2020; (B) 14th April, 2020; (C) 03rd May,2020; (D) 17th May, 2020; and (E) 31st May, 2020

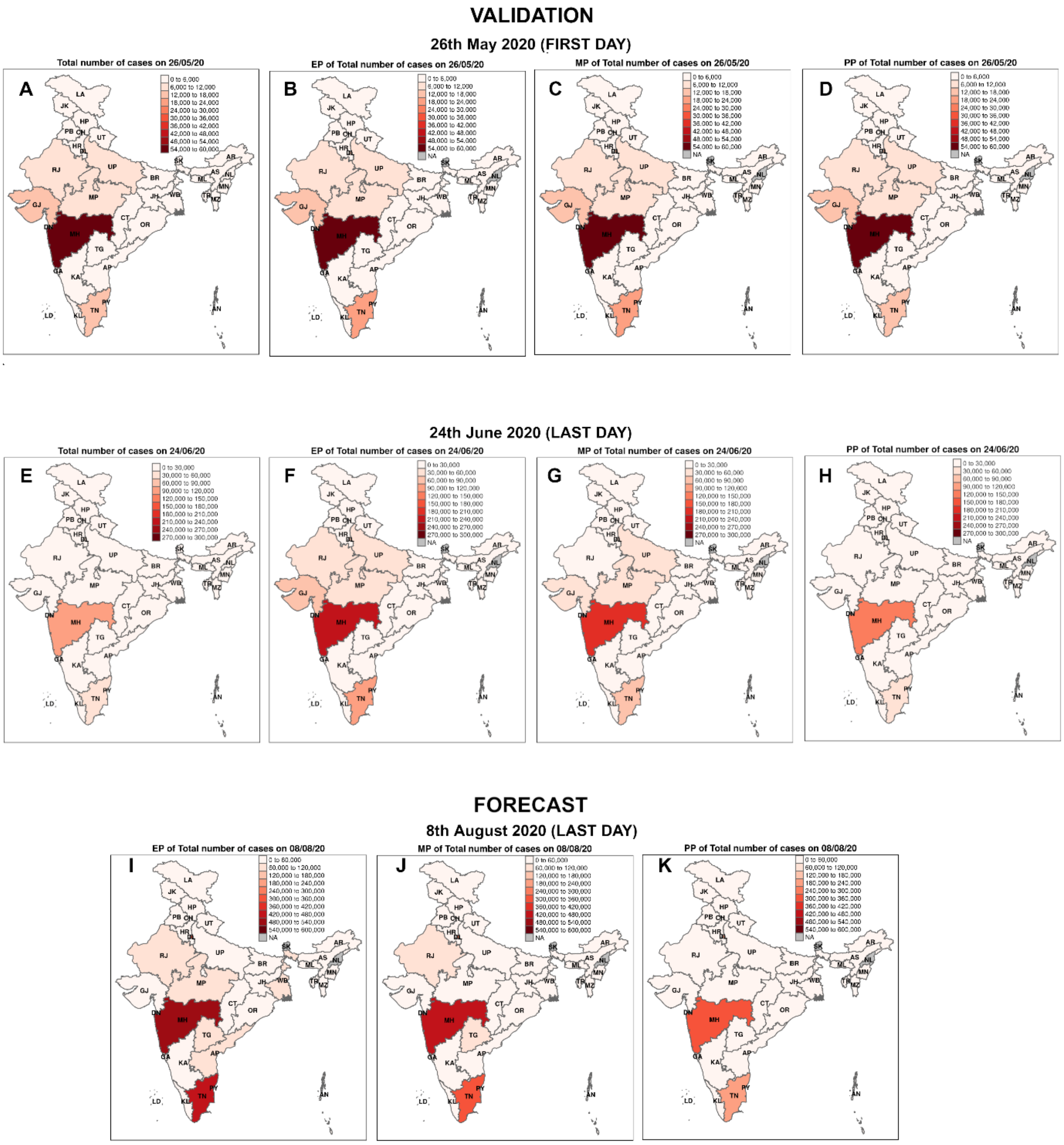

Fig. 9: Comparison of Total Number of reported COVID - 19 Cases in Indian states with the first and last date of Validation along with Forecast for 08th August, 2020: Figure depicts the comparison of the total number of reported COVID - 19 cases grouped state wise in India on 26th May, 2020 with the (A) Total number of cases, (B) Exponential Prediction (EP), (C) Mean Prediction (MP) and(D) Principal Prediction (PP); on 24th June,2020 with (E,F,G,H) for Total number of cases, EP, MP and PP respectively; Forecast for 08th August, 2020 with (I,J,K) for EP, MP and PP respectively.

4. Discussions

In this study, we propose a strategy to predict the expected number of positive COVID-19 cases for 45 days 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.

The model has been validated for 30 days prediction for the period 26th May, 2020 to 24th June, 2020 and then forecasts have been made for the period 25th June, 2020 to 08th August, 2020. In addition to our validation, we further report a verification of our forecasts with the recent available data from 25th June to 06th July, 2020. Here we observe that, out of the 34 regions (India, states & UT) in the study, for 4 states, the EP is performing better than MP and PP with MAPE values in the range 0.82-9.46%. For 8 states the MP is performing better with MAPE in the range 1.46-7.78%. For the remaining 22 regions, the PP is performing better than EP and MP with MAPE values in the range 0.99-10% for 20 of these regions (**Supplementary Table S4, Supplementary Table S6**). Based on this verification the PP, EP or MP can be chosen for the respective states. For R_0, a similar verification is done. Here it is noted that for all states and UTs, the MAPE values are in the range of 0.027-1.14%. Only for 3 regions (Gujarat, Karnataka and Puducherry) the EP is performing better than MP and PP. For 9 regions the MP is best (India, Andhra Pradesh, Arunachal Pradesh, Assam, Dadra and Nagar Haveli, Goa, Maharashtra, Tamil Nadu and Telangana). For the remaining 22 regions the PP is performing the best. For all 34 regions, the R_0 has been decreasing, but is still above 1.

Fig. 8 shows the heat map of the total number of cases state wise one day before each phase of the lockdown and the first unlock. Here we observe that the total number of cases have been increasing uniformly throughout each phase of lockdown. Nevertheless, through each phase of lockdown, the Government of India has intended to prevent the spread of COVID - 19 in the population. **Fig. 9 (A, B, C, D, E, F, G, H)** shows the heat map comparison between the first and last days of validation. Given the sheer population of India, it would not have been possible to mitigate the spread of the virus especially if it is carried from urban centres to the rural setting where medical infrastructure is scanty. The time thus sought using lockdown has been utilised to create awareness among the citizens regarding onset and symptoms of COVID - 19, benefit of masks and social distancing to control the spread. In addition, medical infrastructures such as hospital beds, ventilators, testing equipment were ramped to meet the rising cases (Hindustan Times 2020d; LiveMint 2020; The Hindu 2020d). Meanwhile, under the aegis of Ministry of Textile and Ministry of Commerce production of Personal Protective Equipments were increased from producing 47000 kits annually to 450000 per week (India News Briefing 2020) by involving Khadi and Village Industries Commission, Micro, Small and Medium Enterprises. As of 28th June, 2020 India stands 4th globally in number of COVID - 19 cases, given the costs the lockdown has come with (Economic, Human), nonetheless it has helped the Government of India to establish various infrastructure in a short period of time to mitigate the long-term damages that could have been caused if COVID - 19 spread was unchecked. Delhi, Gujarat, Maharashtra and Tamil Nadu have been in the top 5 states with highest number of reported cases throughout each phase of lockdown (**Fig. 8, A-D**) and naturally during the course of unlock (**Fig. 8.E**) as well. Although Kerala witnessed one of the first cases in the country

(**Fig. 8.A**) the state was able to control the number of cases throughout the course of the lockdown (**Fig. 8 - B, C, D, E**). Uttar Pradesh is the most densely populated state in India, but has not reported many cases (**Fig. 8 - A, B, C, D, E**). Recently it has seen a surge. This may be attributed to low testing rates earlier during infrastructure ramp-up.

Additionally, in this study, we have also tried to compare the long-term prediction results using our strategy against ARIMA and error minimization technique (Tinani et al. 2020; Rafiq et al. 2020). Here we are able to show that the range provided by our strategy is able to capture long term movement of the number of cases. Further, the EP, MP or PP predicts the actual number of cases with better accuracy (**Supplementary Table S8**). Selection of PP, MP or EP can be made by verification with real data a few days into the long forecast period.

However, one major shortcoming of this strategy is that the workflow can sometimes provide a large range of values in some cases depending on the nature of the EP. Also, in some exceptional cases, sudden changes in the number of new cases may sometimes not be detected even with the EP. In the future, an adaptive extension to the workflow may be included that will help to update the model every few days. A further improvement to the EP can be made by updating the *v* and *f* values state wise, if available.

Nevertheless, this strategy with the help of the R_0 values effectively captures the range spanned by PP-MP-EP within which the total number of cases will lie for the next 45 days. This will help with better insights and serve overall as a tool for decision makers that will aid them with planning more than a month in advance.

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.

Authors contributions

Ram Rup Sarkar: Conceptualization, Methodology, Project administration, Funding acquisition. **Kshitij Patil:** Methodology, Software, Validation, Formal analysis, Writing - Original Draft, Visualization. **Anirudh Murali:** Methodology, Software, Validation, Formal analysis, Writing - Original Draft, Visualization. **Piyali Ganguli:** Formal analysis, Writing - Original Draft, Visualization results. **Sutanu Nandi:** Formal analysis, Writing - Review & Editing, Visualization.

Conflict of interest statements

Authors declare no competing interests.

Data availability

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

References

Chakraborty T, Ghosh I (2020) Real-time forecasts and risk assessment of novel coronavirus (COVID-19) cases: A data-driven analysis. Chaos, Solitons and Fractals 135:. https://doi.org/10.1016/j.chaos.2020.109850

Das S (2020) Prediction of COVID-19 Disease Progression in India : Under the Effect of National Lockdown. 1–11. https://doi.org/arXiv:2004.03147

Deb S, Majumdar M (2020) A time series method to analyze incidence pattern and estimate reproduction number of COVID-19. 1–14. https://doi.org/arXiv:2003.10655

Deccan Herald (2020) Coronavirus super spreader Koyambedu market stings Tamilnadu, was the cluster avoidable? https://www.deccanherald.com/national/south/coronavirus-super-spreader-koyambedu-market-stings-tamil-nadu-was-the-cluster-avoidable-834911.html

Delamater PL, Street EJ, Leslie TF, et al (2019) Complexity of the Basic Reproduction Number (R(0)). Emerg Infect Dis 25:1–4. https://doi.org/10.3201/eid2501.171901

Dhanwant JN, Ramanathan V (2020) Forecasting COVID 19 growth in India using Susceptible-Infected-Recovered (S.I.R) model. https://doi.org/arXiv:2004.00696

Dutta S, Das K, Chatterjee K (2020) What if Lockdown is Removed ? District Level Predictions for Maharashtra and Gujarat. 18:209–221

Economic Times (2020a) Migrant labourers gather at Delhi - Uttar Pradesh border amid lockdown. https://economictimes.indiatimes.com/news/politics-and-nation/migrant-labourers-gather-at-delhi-uttar-pradesh-border-amid-lockdown/articleshow/75783884.cms?from=mdr

Economic Times (2020b) Ahmedabad 700 super spreaders found coronavirus positive in a week. https://economictimes.indiatimes.com/news/politics-and-nation/ahmedabad-700-super-spreaders-found-coronavirus-positive-in-a-week/articleshow/75780194.cms

Gupta R, Pal SK (2020) Trend Analysis and Forecasting of COVID-19 Outbreak in India. medRxiv 2020.03.26.20044511. https://doi.org/10.1101/2020.04.01.20049825

Gupta S, Shankar R (2020) Estimating the number of COVID-19 infections in Indian hot-spots using fatality data. 2:1–6. https://doi.org/arXiv:2004.04025

Hindustan Times (2020a) COVID - 19 latest over 300 rise in Jharkhand cases since May 1 all districts invaded. https://www.hindustantimes.com/india-news/covid-19-latest-over-300-rise-in-jharkhand-cases-since-may-1-all-districts-invaded/story-6wmefydBn86kdV4GCrxm7

Hindustan Times (2020b) Maharashtra returnees account for 84 of 135 new COVID - 19 cases in Karnataka. https://www.hindustantimes.com/india-news/maharashtra-returnees-account-for-84-of-135-new-covid-19-cases-in-karnataka/story-jW02YsU5oZV3xs0vq65nCN.html

Hindustan Times (2020c) Ramganj emerges as new Raj hot spot. https://www.hindustantimes.com/india-news/ramganj-emerges-as-new-raj-hot-spot/story-6KRiEiyAai6spqWxvY54gM_amp.html

Hindustan Times (2020d) From stadium to isolation centre to hospital a jumbo fight against COVID - 19 at Mumbai’s NSCI dome. https://www.hindustantimes.com/mumbai-news/from-stadium-to-isolation-centre-to-hospital-a-jumbo-fight-against-covid-19-at-mumbai-s-nsci-d

India News Briefing (2020) India’s PPE industry COVID - 19 demonstrated untapped local production capacity. https://www.india-briefing.com/news/indias-ppe-industry-covid-19-demonstrated-untapped-local-production-capacity-20486.html/

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

Kourentzes N, Petropoulos F, Trapero JR (2014) Improving forecasting by estimating time series structural components across multiple frequencies. Int J Forecast 30:291–302. https://doi.org/10.1016/j.ijforecast.2013.09.006

Kumar A, Yerra R, Rao S Spread of COVID-19 in Odisha (India) due to Influx of Migrants and Stability Analysis using Mathematical Modelling. 1–14. https://doi.org/10.21203/rs.3.rs-34007/v1

Kumar S (2020) Predication of Pandemic COVID-19 situation in Maharashtra, India. https://doi.org/10.1101/2020.04.10.20056697

Lipsitch M, Cohen T, Cooper B, et al (2003) Transmission dynamics and control of severe acute respiratory syndrome. Science 300:1966–1970. https://doi.org/10.1126/science.1086616

LiveMint (2020) Mumbai converts BKC to Nehru planetarium into quarantine centres as COVID -19 cases rise. https://www.livemint.com/news/india/mumbai-converts-bkc-to-nehru-planetarium-into-quarantine-centres-as-covid-19-cases-rise-11589872488746.html

Mandal M, Jana S, Nandi SK, et al (2020) A model based study on the dynamics of COVID-19: Prediction and control. Chaos, Solitons and Fractals 136:109889. https://doi.org/10.1016/j.chaos.2020.109889

Marbaniang SP (2020) Forecasting the Prevalence of COVID-19 in Maharashtra , Delhi , Kerala , and India using an ARIMA model. 1–16. https://doi.org/10.21203/rs.3.rs-34555/v1

Pandey G, Chaudhary P, Gupta R, Pal S (2020) SEIR and Regression Model based COVID-19 outbreak predictions in India. 1–10. https://doi.org/10.1101/2020.04.01.20049825

Poonia N, Azad S (2020) Short-term forecasts of COVID-19 spread across Indian states until 1 May 2020. https://doi.org/arXiv:2004.13538

Rafiq D, Suhail SA, Bazaz MA (2020) Evaluation and prediction of COVID-19 in India: A case study of worst hit states. Chaos, Solitons & Fractals 139:110014. https://doi.org/10.1016/j.chaos.2020.110014

Ranjan R (2020) Predictions for COVID-19 outbreak in India using Epidemiological models. medRxiv 2020.04.02.20051466. https://doi.org/10.1101/2020.04.02.20051466

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

Sarkar K, Khajanchi S (2020) Modeling and forecasting of the COVID-19 pandemic in India. https://doi.org/arXiv:2005.07071

Sathish T, Ray A, Gopal NN (2020) Predictions of COVID-19 patients raise , recovery and death rate in India by ARIMA model. 15:5–10. https://doi.org/10.9790/3008-1503030510

Sujath R, Chatterjee JM, Hassanien AE (2020) A machine learning forecasting model for COVID-19 pandemic in India. Stoch Environ Res Risk Assess 34:959–972. https://doi.org/10.1007/s00477-020-01827-8

The Hindu (2020a) Coronavirus migrant labourers take to the streets demand they be returned home. https://www.thehindu.com/news/cities/chennai/coronavirus-migrant-labourers-take-to-the-streets-demand-they-be-returned-home/article31490196.ece

The Hindu (2020b) COVID - 19 Thiruvananthapuram eateries hit by workers exodus.

https://www.thehindu.com/news/national/kerala/covid-19-thiruvananthapuram-eateries-hit-by-workers-exodus/article31094071.ece

The Hindu (2020c) Huge influx of travellers increases backlog of samples awaiting tests. https://www.thehindu.com/news/national/karnataka/huge-influx-of-travellers-increases-backlog-of-samples-awaiting-tests/article31705555.ece

The Hindu (2020d) ITBP to operate 10000 bed COVID facility in Chhatarpur. https://www.thehindu.com/news/cities/Delhi/itbp-to-operate-10000-bed-covid-facility-in-chhatarpur/article31901414.ece

The New Indian Express (2020a) Karnataka helpless as Maharashtra cases continue to stream in. https://www.newindianexpress.com/states/karnataka/2020/jun/05/karnataka-helpless-as-maharashtra-cases-continue-to-stream-in-2152505.html

The New Indian Express (2020b) India’s super spreader home quarantines 26000 people in 24 Punjab villages. https://www.newindianexpress.com/nation/2020/mar/28/indias-super-spreader-home-quarantines-26000-people-in-24-punjab-villages-2122724.html

ThePrint.in (2020) 45 COVID cases in Noida Ghaziabad tracked to Delhi, can’t open borders UP tells SC. In: https://theprint.in/judiciary/45-covid-cases-in-noida-ghaziabad-tracked-to-delhi-cant-open-borders-up-tells-sc/440318/

Times Now News (2020) Mumbai migrants CST from Uttar Pradesh basti flout social distancing norms in home bound trains. https://www.timesnownews.com/india/maharashtra-news/article/mumbai-migrants-cst-from-uttar-pradesh-basti-flout-social-distancing-norms-in-home

Times of India (2020) 2000 km and counting 15 men walk hitch ride to UP from Bengaluru. https://timesofindia.indiatimes.com/city/bengaluru/2000-km-and-counting-15-men-walk-hitch-ride-to-up-from-bengaluru/articleshow/75666285.cms

Tinani K, Muralidharan K, Deshmukh A, Patil B (2020) Analysis and Forecasting of COVID-19 Cases Across Hotspot States of India. 18:223–238

Zhang J, Litvinova M, Wang W, et al (2020) Evolving epidemiology and transmission dynamics of coronavirus disease 2019 outside Hubei province, China: a descriptive and modelling study. Lancet Infect Dis 3099:1–10. https://doi.org/10.1016/S1473-3099(20)30230-9